A Course in Statistical Theory
by David Olive
Preprint M-02-006
Copyright August 2005, May 2012

Course notes for statistical inference. This manuscript is unfinished. Revisions are ongoing. This .PDF version below is as of May 2012.
http://www.math.siu.edu/olive/infbook.htm

A Course in Statistical Theory

David J. Olive
Southern Illinois University
Department of Mathematics
Mailcode 4408
Carbondale, IL 62901-4408
dolive@math.siu.edu

May 13, 2012
# Contents

Preface vi

1 Probability and Expectations 1
   1.1 Probability .............................................. 1
   1.2 Counting .................................................. 3
   1.3 Conditional Probability and Independence .......... 4
   1.4 The Expected Value and Variance ...................... 7
   1.5 The Kernel Method ....................................... 13
   1.6 Mixture Distributions .................................... 16
   1.7 Summary .................................................... 19
   1.8 Complements ............................................... 22
   1.9 Problems ................................................... 22

2 Multivariate Distributions 31
   2.1 Joint, Marginal and Conditional Distributions ...... 31
   2.2 Expectation, Covariance and Independence ........... 35
   2.3 Conditional Expectation and Variance ................ 45
   2.4 Location–Scale Families .................................. 48
   2.5 Transformations ............................................ 49
   2.6 Sums of Random Variables ............................... 58
   2.7 Random Vectors ............................................. 64
   2.8 The Multinomial Distribution ............................ 67
   2.9 The Multivariate Normal Distribution .................. 69
   2.10 Elliptically Contoured Distributions .................. 72
   2.11 Summary .................................................... 76
   2.12 Complements ............................................... 78
   2.13 Problems ................................................... 79
3 Exponential Families
  3.1 Regular Exponential Families ........................................... 92
  3.2 Properties of \( (t_1(Y), ..., t_k(Y)) \) ........................................... 102
  3.3 Summary ........................................................................ 107
  3.4 Complements ...................................................................... 108
  3.5 Problems ........................................................................ 110

4 Sufficient Statistics ................................................................ 114
  4.1 Statistics and Sampling Distributions ....................................... 114
  4.2 Minimal Sufficient Statistics ..................................................... 121
  4.3 Summary ........................................................................ 135
  4.4 Complements ...................................................................... 138
  4.5 Problems ........................................................................ 139

5 Point Estimation ........................................................................ 147
  5.1 Maximum Likelihood Estimators .............................................. 147
  5.2 Method of Moments Estimators ................................................ 161
  5.3 Summary ........................................................................ 163
  5.4 Complements ...................................................................... 164
    5.4.1 Two “Proofs” of the Invariance Principle ......................... 166
  5.5 Problems ........................................................................ 169

6 UMVUEs and the FCRLB ............................................................... 178
  6.1 MSE and Bias .................................................................... 178
  6.2 Exponential Families, UMVUEs and the FCRLB ......................... 181
  6.3 Summary ........................................................................ 192
  6.4 Complements ...................................................................... 193
  6.5 Problems ........................................................................ 193

7 Testing Statistical Hypotheses ..................................................... 203
  7.1 Exponential Families, the Neyman Pearson Lemma, and UMP Tests ........................................... 204
  7.2 Likelihood Ratio Tests ............................................................ 212
  7.3 Summary ........................................................................ 215
  7.4 Complements ...................................................................... 219
  7.5 Problems ........................................................................ 221
### 8 Large Sample Theory

| 8.1 The CLT, Delta Method and an Exponential Family Limit Theorem | 228 |
| ASymptotically Efficient Estimators | 236 |
| Modes of Convergence and Consistency | 240 |
| Slutsky’s Theorem and Related Results | 245 |
| Order Relations and Convergence Rates | 249 |
| Multivariate Limit Theorems | 252 |
| More Multivariate Results | 257 |
| Summary | 259 |
| Complements | 260 |
| Problems | 262 |

### 9 Confidence Intervals

| 9.1 Introduction | 274 |
| Some Examples | 281 |
| Bootstrap and Randomization Tests | 300 |
| Complements | 307 |
| Problems | 307 |

### 10 Some Useful Distributions

| 10.1 The Beta Distribution | 314 |
| The Beta–Binomial Distribution | 317 |
| The Binomial and Bernoulli Distributions | 318 |
| The Birnbaum Saunders Distribution | 321 |
| The Burr Type III Distribution | 322 |
| The Burr Type X Distribution | 323 |
| The Burr Type XII Distribution | 324 |
| The Cauchy Distribution | 325 |
| The Chi Distribution | 326 |
| The Chi–square Distribution | 328 |
| The Discrete Uniform Distribution | 329 |
| The Double Exponential Distribution | 331 |
| The Exponential Distribution | 333 |
| The Two Parameter Exponential Distribution | 335 |
| The F Distribution | 337 |
| The Gamma Distribution | 337 |
| The Generalized Gamma Distribution | 340 |
10.18 The Generalized Negative Binomial Distribution . . . 341
10.19 The Geometric Distribution . . . . . . . . . . . . . . . . . . . . . . 342
10.20 The Gompertz Distribution . . . . . . . . . . . . . . . . . . . . . . . 343
10.21 The Half Cauchy Distribution . . . . . . . . . . . . . . . . . . . . . . 344
10.22 The Half Logistic Distribution . . . . . . . . . . . . . . . . . . . . . . 344
10.23 The Half Normal Distribution . . . . . . . . . . . . . . . . . . . . . . . 345
10.24 The Hypergeometric Distribution . . . . . . . . . . . . . . . . . . . . . 347
10.25 The Inverse Exponential Distribution . . . . . . . . . . . . . . . . . 347
10.26 The Inverse Gaussian Distribution . . . . . . . . . . . . . . . . . . . 348
10.27 The Inverse Weibull Distribution . . . . . . . . . . . . . . . . . . . . 351
10.28 The Inverted Gamma Distribution . . . . . . . . . . . . . . . . . . . . 352
10.29 The Largest Extreme Value Distribution . . . . . . . . . . . . . . . . 353
10.30 The Logarithmic Distribution . . . . . . . . . . . . . . . . . . . . . . . 354
10.31 The Logistic Distribution . . . . . . . . . . . . . . . . . . . . . . . . . 355
10.32 The Log-Cauchy Distribution . . . . . . . . . . . . . . . . . . . . . . . 356
10.33 The Log-Gamma Distribution . . . . . . . . . . . . . . . . . . . . . . . 356
10.34 The Log-Logistic Distribution . . . . . . . . . . . . . . . . . . . . . . . 356
10.35 The Lognormal Distribution . . . . . . . . . . . . . . . . . . . . . . . . 357
10.36 The Maxwell-Boltzmann Distribution . . . . . . . . . . . . . . . . . . 360
10.37 The Modified DeMoivre’s Law Distribution . . . . . . . . . . . . . . 361
10.38 The Negative Binomial Distribution . . . . . . . . . . . . . . . . . . . 362
10.39 The Normal Distribution . . . . . . . . . . . . . . . . . . . . . . . . . 364
10.40 The One Sided Stable Distribution . . . . . . . . . . . . . . . . . . . . 368
10.41 The Pareto Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 369
10.42 The Poisson Distribution . . . . . . . . . . . . . . . . . . . . . . . . . 372
10.43 The Power Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 373
10.44 The Two Parameter Power Distribution . . . . . . . . . . . . . . . . . 375
10.45 The Rayleigh Distribution . . . . . . . . . . . . . . . . . . . . . . . . 376
10.46 The Smallest Extreme Value Distribution . . . . . . . . . . . . . . . 378
10.47 The Student’s t Distribution . . . . . . . . . . . . . . . . . . . . . . . 379
10.48 The Topp-Leone Distribution . . . . . . . . . . . . . . . . . . . . . . . 379
10.49 The Truncated Extreme Value Distribution . . . . . . . . . . . . . . 381
10.50 The Uniform Distribution . . . . . . . . . . . . . . . . . . . . . . . . 382
10.51 The Weibull Distribution . . . . . . . . . . . . . . . . . . . . . . . . 383
10.52 The Zero Truncated Poisson Distribution . . . . . . . . . . . . . . . 385
10.53 The Zeta Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 385
10.54 The Zipf Distribution . . . . . . . . . . . . . . . . . . . . . . . . . . 386
10.55 Complements . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 387
Preface

Many math and some statistics departments offer a one semester graduate course in statistical inference using texts such as Casella and Berger (2002), Bickel and Doksum (2007) or Mukhopadhyay (2000, 2006). The course typically covers minimal and complete sufficient statistics, maximum likelihood estimators (MLEs), method of moments, bias and mean square error, uniform minimum variance estimators (UMVUEs) and the Fréchet-Cramér-Rao lower bound (FCRLB), an introduction to large sample theory, likelihood ratio tests, and uniformly most powerful (UMP) tests and the Neyman Pearson Lemma. A major goal of this text is to make these topics much more accessible to students by using the theory of exponential families.

This material is essential for Masters and PhD students in biostatistics and statistics, and the material is often very useful for graduate students in economics, psychology and electrical engineering (especially communications and control).

The material is also useful for actuaries. According to (www.casact.org), topics for the CAS Exam 3 (Statistics and Actuarial Methods) include the MLE, method of moments, consistency, unbiasedness, mean square error, testing hypotheses using the Neyman Pearson Lemma and likelihood ratio tests, and the distribution of the max. These topics make up about 20% of the exam.

One of the most important uses of exponential families is that the theory often provides two methods for doing inference. For example, minimal sufficient statistics can be found with either the Lehmann-Scheffé theorem or by finding $T$ from the exponential family parameterization. Similarly, if $Y_1, ..., Y_n$ are iid from a one parameter regular exponential family with complete sufficient statistic $T(Y)$, then one sided UMP tests can be found by using the Neyman Pearson lemma or by using exponential family theory.

The prerequisite for this text is a calculus based course in statistics at


A less satisfactory alternative prerequisite is a calculus based course in probability at the level of Hoel, Port and Stone (1971), Parzen (1960) or Ross (1984, 2009).

A course in Real Analysis at the level of Bartle (1964), Gaughan (2009), Rosenlicht (1985), Ross (1980) or Rudin (1964) would be useful for the large sample theory chapter.


Some other useful references include a good low level probability text Ash (1993) and a good introduction to probability and statistics Dekking, Kraaikamp, Lopuhaä and Meester (2005). Also see Ash (2011), Spiegel (1975), Romano and Siegel (1986) and see online lecture notes by Ash at (www.math.uiuc.edu/~r-ash/).

Many of the most important ideas in statistics are due to Fisher, Neyman, E.S. Pearson and K. Pearson. For example, David (2006-7) says that the following terms were due to Fisher: consistency, covariance, degrees of freedom, efficiency, information, information matrix, level of significance, likelihood, location, maximum likelihood, multinomial distribution, null hypothesis, pivotal quantity, probability integral transformation, sampling distribution, scale, statistic, Student’s t, studentization, sufficiency, sufficient statistic, test of significance, uniformly most powerful test and variance.

David (2006-7) says that terms due to Neyman and E.S. Pearson include alternative hypothesis, composite hypothesis, likelihood ratio, power, power
function, simple hypothesis, size of critical region, test criterion, test of hypotheses, type I and type II errors. Neyman also coined the term confidence interval.

David (2006-7) says that terms due to K. Pearson include binomial distribution, bivariate normal, method of moments, moment, random sampling, skewness, and standard deviation.


This book covers some of these ideas and begins by reviewing probability, counting, conditional probability, independence of events, the expected value and the variance. Chapter 1 also covers mixture distributions and shows how to use the kernel method to find $E(g(Y))$. Chapter 2 reviews joint, marginal, and conditional distributions; expectation; independence of random variables and covariance; conditional expectation and variance; location–scale families; univariate and multivariate transformations; sums of random variables; random vectors; the multinomial, multivariate normal and elliptically contoured distributions. Chapter 3 introduces exponential families while Chapter 4 covers sufficient statistics. Chapter 5 covers maximum likelihood estimators and method of moments estimators. Chapter 6 examines the mean square error and bias as well as uniformly minimum variance unbiased estimators, Fisher information and the Fréchet-Cramér-Rao lower bound. Chapter 7 covers uniformly most powerful and likelihood ratio tests. Chapter 8 gives an introduction to large sample theory while Chapter 9 covers confidence intervals. Chapter 10 gives some of the properties of 54 univariate distributions, many of which are exponential families. Chapter 10 also gives over 30 exponential family parameterizations, over 28 MLE examples and over 27 UMVUE examples. Chapter 11 gives some hints for the problems.

Some highlights of this text follow.

- Exponential families, indicator functions and the support of the distribution are used throughout the text to simplify the theory.

- Section 1.5 describes the kernel method, a technique for computing $E(g(Y))$, in detail rarely given in texts.

- Theorem 2.2 shows the essential relationship between the independence
of random variables $Y_1, \ldots, Y_n$ and the support in that the random variables are dependent if the support is not a cross product. If the support is a cross product and if the joint pdf or pmf factors on the support, then $Y_1, \ldots, Y_n$ are independent.

- Theorems 2.17 and 2.18 give the distribution of $\sum Y_i$ when $Y_1, \ldots, Y_n$ are iid for a wide variety of distributions.

- Chapter 3 presents exponential families. The theory of these distributions greatly simplifies many of the most important results in mathematical statistics.

- Corollary 4.6 presents a simple method for finding sufficient, minimal sufficient and complete statistics for $k$-parameter exponential families.

- Section 5.4.1 compares the “proofs” of the MLE invariance principle due to Zehna (1966) and Berk (1967). Although Zehna (1966) is cited by most texts, Berk (1967) gives a correct elementary proof.

- Theorem 6.5 compares the UMVUE and the estimator that minimizes the MSE for a large class of one parameter exponential families.

- Theorem 7.3 provides a simple method for finding uniformly most powerful tests for a large class of 1–parameter exponential families. Power is examined in Example 7.10.

- Theorem 8.4 gives a simple proof of the asymptotic efficiency of the complete sufficient statistic as an estimator of its expected value for 1–parameter regular exponential families.

- Theorem 8.21 provides a simple limit theorem for the complete sufficient statistic of a $k$–parameter regular exponential family.

- Chapter 10 gives information for 54 “brand name” distributions.

Much of the text material is on parametric frequentist methods, but the most used methods in statistics tend to be semiparametric. Many of the most used methods originally based on the univariate or multivariate normal distribution are also semiparametric methods. For example the t-interval works for a large class of distributions if $\sigma^2$ is finite and $n$ is large. Similarly, least squares regression is a semiparametric method. Multivariate analysis
procedures originally based on the multivariate normal distribution tend to also work well for a large class of elliptically contoured distributions.

In a semester, I cover Sections 1.1–1.6, 2.1–2.9, 3.1, 3.2, 4.1, 4.2, 5.1, 5.2, 6.1, 6.2, 7.1, 7.2 and 8.1-8.4.

**Warning:** For parametric methods that are not based on the normal distribution, often the methods work well if the parametric distribution is a good approximation to the data, but perform very poorly otherwise.

**Acknowledgements**

Teaching the material to Math 580 students at Southern Illinois University in 2001, 2004, 2008 (when the text was first used) and 2012 was very useful. Some of the Chapter 8 material came from a reading course in Large Sample Theory taught to SIU students on two occasions. Some of the SIU QUAL problems were written by Bhaskar Bhattacharya, Sakthivel Jeyaratnam, and Abdel Mugdadi, who also contributed several solutions. A small part of the research in Chapters 9 and 10 was supported by NSF grant DMS 0600933.
Chapter 1

Probability and Expectations

1.1 Probability

Definition 1.1. **Statistics** is the science of extracting useful information from data.

This chapter reviews some of the tools from probability that are useful for statistics, and the following terms from set theory should be familiar. A set consists of distinct elements enclosed by braces, eg \{1, 5, 7\}. The *universal set* \( S \) is the set of all elements under consideration while the *empty set* \( \emptyset \) is the set that contains no elements. The set \( A \) is a *subset* of \( B \), written \( A \subseteq B \), if every element in \( A \) is in \( B \). The *union* \( A \cup B \) of \( A \) with \( B \) is the set of all elements in \( A \) or \( B \) or in both. The *intersection* \( A \cap B \) of \( A \) with \( B \) is the set of all elements in \( A \) and \( B \). The *complement* of \( A \), written \( \overline{A} \) or \( A^c \), is the set of all elements in \( S \) but not in \( A \).

Theorem 1.1. DeMorgan’s Laws:

a) \( \overline{A \cup B} = \overline{A} \cap \overline{B} \).
b) \( \overline{A \cap B} = \overline{A} \cup \overline{B} \).

Sets are used in probability, but often different notation is used. For example, the universal set is called the sample space \( S \). In the definition of an event below, the special field of subsets \( \mathcal{B} \) of the sample space \( S \) forming the class of events will not be formally given. However, \( \mathcal{B} \) contains all “interesting” subsets of \( S \) and every subset that is easy to imagine. The point is that not necessarily all subsets of \( S \) are events, but every event \( A \) is a subset of \( S \).
Definition 1.2. The sample space $S$ is the set of all possible outcomes of an experiment.

Definition 1.3. Let $\mathcal{B}$ be a special field of subsets of the sample space $S$ forming the class of events. Then $A$ is an event if $A \in \mathcal{B}$.

Definition 1.4. If $A \cap B = \emptyset$, then $A$ and $B$ are mutually exclusive or disjoint events. Events $A_1, A_2, \ldots$ are pairwise disjoint or mutually exclusive if $A_i \cap A_j = \emptyset$ for $i \neq j$.

A simple event is a set that contains exactly one element $s_i$ of $S$, e.g. $A = \{s_3\}$. A sample point $s_i$ is a possible outcome.

Definition 1.5. A discrete sample space consists of a finite or countable number of outcomes.

Notation. Generally we will assume that all events under consideration belong to the same sample space $S$.

The relative frequency interpretation of probability says that the probability of an event $A$ is the proportion of times that event $A$ would occur if the experiment was repeated again and again infinitely often.

Definition 1.6: Kolmogorov’s Definition of a Probability Function. Let $\mathcal{B}$ be the class of events of the sample space $S$. A probability function $P : \mathcal{B} \to [0, 1]$ is a set function satisfying the following three properties:

P1) $P(A) \geq 0$ for all events $A$,

P2) $P(S) = 1$, and

P3) if $A_1, A_2, \ldots$ are pairwise disjoint events, then $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

Example 1.1. Flip a coin and observe the outcome. Then the sample space $S = \{H, T\}$. If $P(\{H\}) = 1/3$, then $P(\{T\}) = 2/3$. Often the notation $P(H) = 1/3$ will be used.

Theorem 1.2. Let $A$ and $B$ be any two events of $S$. Then

i) $0 \leq P(A) \leq 1$.

ii) $P(\emptyset) = 0$ where $\emptyset$ is the empty set.

iii) Complement Rule: $P(A) = 1 - P(\overline{A})$.

iv) General Addition Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

v) If $A \subseteq B$, then $P(A) \leq P(B)$.
vi) **Boole’s Inequality:** 
\[ P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) \]
for any events \( A_1, A_2, \ldots \).

vii) **Bonferroni’s Inequality:** 
\[ P(\bigcap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n - 1) \]
for any events \( A_1, A_2, \ldots, A_n \).

The general addition rule for two events is very useful. Given three of the 4 probabilities in iv), the 4th can be found. 
\[ P(A \cup B) \]
can be found given 
\[ P(A), P(B) \]
and that \( A \) and \( B \) are disjoint or independent. The addition rule can also be used to determine whether \( A \) and \( B \) are independent (see Section 1.3) or disjoint.

### 1.2 Counting

The **sample point method** for finding the probability for event \( A \) says that if \( S = \{s_1, \ldots, s_k\} \) then \( 0 \leq P(s_i) \leq 1 \), \( \sum_{i=1}^{k} P(s_i) = 1 \), and \( P(A) = \sum_{i: s_i \in A} P(s_i) \). That is, \( P(A) \) is the sum of the probabilities of the sample points in \( A \). If all of the outcomes \( s_i \) are *equally likely*, then \( P(s_i) = 1/k \) and \( P(A) = \text{(number of outcomes in } A)/k \) if \( S \) contains \( k \) outcomes.

Counting or combinatorics is useful for determining the number of elements in \( S \). The **multiplication rule** says that if there are \( n_1 \) ways to do a first task, \( n_2 \) ways to do a 2nd task, ..., and \( n_k \) ways to do a \( k \)th task, then the number of ways to perform the total act of performing the 1st task, then the 2nd task, ..., then the \( k \)th task is \( \prod_{i=1}^{k} n_i = n_1 \cdot n_2 \cdot n_3 \cdots n_k \).

#### Techniques for the multiplication principle:

a) use a slot for each task and write \( n_i \) above the \( i \)th task. There will be \( k \) slots, one for each task.

b) Use a tree diagram.

**Definition 1.7.** A **permutation** is an ordered arrangements using \( r \) of \( n \) distinct objects and the **number of permutations** = \( P_r^n \). A special case of permutation formula is

\[ P_r^n = n! = n \cdot (n - 1) \cdot (n - 2) \cdot (n - 3) \cdots 4 \cdot 3 \cdot 2 \cdot 1 = n \cdot (n-1)! = n \cdot (n-1) \cdot (n-2)! = n \cdot (n-1) \cdot (n-2) \cdot (n-3)! = \cdots . \]

Generally \( n \) is a positive integer, but define 0! = 1. An application of the multiplication rule can be used to show that \( P_r^n = n \cdot (n-1) \cdot (n-2) \cdots (n-r+1) = \frac{n!}{(n-r)!} \).
The quantity $n!$ is read “n factorial.” Typical problems using $n!$ include the number of ways to arrange $n$ books, to arrange the letters in the word CLIPS (5!), et cetera.

Recognizing when a story problem is asking for the permutation formula: The story problem has $r$ slots and order is important. No object is allowed to be repeated in the arrangement. Typical questions include how many ways are there to “to choose $r$ people from $n$ and arrange in a line,” “to make $r$ letter words with no letter repeated” or “to make 7 digit phone numbers with no digit repeated.” Key words include order, no repeated and different.

Notation. The symbol $\equiv$ below means the first three symbols are equivalent and equal, but the fourth term is the formula used to compute the symbol. This notation will often be used when there are several equivalent symbols that mean the same thing. The notation will also be used for functions with subscripts if the subscript is usually omitted, eg $g_X(x) \equiv g(x)$. The symbol $\binom{n}{r}$ is read “$n$ choose $r$,” and is called a binomial coefficient.

Definition 1.8. A combination is an unordered selection using $r$ of $n$ distinct objects. The number of combinations is

$$C(n,r) \equiv C^n_r \equiv \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$ 

Combinations are used in story problems where order is not important. Key words include committees, selecting (eg 4 people from 10), choose, random sample and unordered.

1.3 Conditional Probability and Independence

Definition 1.9. The conditional probability of $A$ given $B$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

if $P(B) > 0$.

It is often useful to think of this probability as an experiment with sample space $B$ instead of $S$. 
Definition 1.10. Two events $A$ and $B$ are independent, written $A \perp B$, if
\[ P(A \cap B) = P(A)P(B). \]
If $A$ and $B$ are not independent, then $A$ and $B$ are dependent.

Definition 1.11. A collection of events $A_1, ..., A_n$ are mutually independent if for any subcollection $A_{i_1}, ..., A_{i_k}$,
\[ P(\cap_{j=1}^k A_{i_j}) = \prod_{j=1}^k P(A_{i_j}). \]
Otherwise the $n$ events are dependent.

Theorem 1.3. Assume that $P(A) > 0$ and $P(B) > 0$. Then the two events $A$ and $B$ are independent if any of the following three conditions hold:
i) $P(A \cap B) = P(A)P(B)$,
ii) $P(A|B) = P(A)$, or
iii) $P(B|A) = P(B)$.
If any of these conditions fails to hold, then $A$ and $B$ are dependent.

The above theorem is useful because only one of the conditions needs to be checked, and often one of the conditions is easier to verify than the other two conditions.

Theorem 1.4. a) Multiplication rule: If $A_1, ..., A_k$ are events with $P(A_1 \cap A_2 \cap \cdots \cap A_{k-1}) > 0$, then
\[ P(\cap_{i=1}^k A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)\cdots P(A_k|A_1 \cap A_2 \cap \cdots \cap A_{k-1}). \]
In particular, $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$.

b) Multiplication rule for independent events: If $A_1, A_2, ..., A_k$ are independent, then $P(A_1 \cap A_2 \cap \cdots \cap A_k) = P(A_1) \cdots P(A_k)$. If $A$ and $B$ are independent ($k = 2$), then $P(A \cap B) = P(A)P(B)$.

c) Addition rule for disjoint events: If $A$ and $B$ are disjoint, then $P(A \cup B) = P(A) + P(B)$. If $A_1, ..., A_k$ are pairwise disjoint, then $P(\cup_{i=1}^k A_i) = P(A_1 \cup A_2 \cup \cdots \cup A_k) = P(A_1) + \cdots + P(A_k) = \sum_{i=1}^k P(A_i)$.

Example 1.2. The above rules can be used to find the probabilities of more complicated events. The following probabilities are closed related to Binomial experiments. Suppose that there are $n$ independent identical trials, that $Y$ counts the number of successes and that $\rho =$ probability of success
for any given trial. Let $D_i$ denote a success in the $i$th trial. Then
i) $P(\text{none of the n trials were successes}) = (1 - \rho)^n = P(\overline{D_1} \cap \overline{D_2} \cap \cdots \cap \overline{D_n})$.

ii) $P(\text{at least one of the trials was a success}) = 1 - (1 - \rho)^n = P(\overline{Y} = 0) = P(D_1 \cap D_2 \cap \cdots \cap D_n)$. 

iii) $P(\text{all n trials were successes}) = \rho^n = P(Y = n) = P(D_1 \cap D_2 \cap \cdots \cap D_n)$. 

iv) $P(\text{not all n trials were successes}) = 1 - \rho^n = P(Y < n) = 1 - P(Y = n) = 1 - P(\text{all})$. 

v) $P(Y \text{ was at least } k) = P(Y \geq k)$. 

vi) $P(Y \text{ was at most } k) = P(Y \leq k)$.

If $A_1, A_2, \ldots$ are pairwise disjoint and if $\bigcup_{i=1}^{\infty} A_i = S$, then the collection of sets $A_1, A_2, \ldots$ is a partition of $S$. By taking $A_j = \emptyset$ for $j > k$, the collection of pairwise disjoint sets $A_1, A_2, \ldots, A_k$ is a partition of $S$ if $\bigcup_{i=1}^{k} A_i = S$.

**Theorem 1.5: Law of Total Probability.** If $A_1, A_2, \ldots, A_k$ form a partition of $S$ such that $P(A_i) > 0$ for $i = 1, \ldots, k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

**Theorem 1.6: Bayes’ Theorem.** Let $A_1, A_2, \ldots, A_k$ be a partition of $S$ such that $P(A_i) > 0$ for $i = 1, \ldots, k$, and let $B$ be an event such that $P(B) > 0$. Then

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

**Proof.** Notice that $P(A_i|B) = P(A_i \cap B)/P(B)$ and $P(A_i \cap B) = P(B|A_i)P(A_i)$. Since $B = (B \cap A_1) \cup \cdots \cup (B \cap A_k)$ and the $A_i$ are disjoint, $P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j)$. QED

**Example 1.3.** There are many medical tests for rare diseases and a positive result means that the test suggests (perhaps incorrectly) that the person has the disease. Suppose that a test for disease is such that if the person has the disease, then a positive result occurs 99% of the time. Suppose that a person without the disease tests positive 2% of the time. Also assume that 1 in 1000 people screened have the disease. If a randomly selected person tests positive, what is the probability that the person has the disease?
CHAPTER 1. PROBABILITY AND EXPECTATIONS

Solution: Let $A_1$ denote the event that the randomly selected person has the disease and $A_2$ denote the event that the randomly selected person does not have the disease. If $B$ is the event that the test gives a positive result, then we want $P(A_1|B)$. By Bayes’ theorem,

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2)} = \frac{0.99(0.001)}{0.99(0.001) + 0.02(0.999)} \approx 0.047.$$  

Hence instead of telling the patient that she has the rare disease, the doctor should inform the patient that she is in a high risk group and needs further testing.

1.4 The Expected Value and Variance

Definition 1.12. A random variable (RV) $Y$ is a real valued function with a sample space as a domain: $Y : S \to \mathbb{R}$ where the set of real numbers $\mathbb{R} = (-\infty, \infty)$.

Definition 1.13. Let $S$ be the sample space and let $Y$ be a random variable. Then the (induced) probability function for $Y$ is $P_Y(Y = y_i) \equiv P(Y = y_i) = P_S(\{s \in S : Y(s) = y_i\})$. The sample space of $Y$ is $S_Y = \{y_i \in \mathbb{R} : \text{there exists an } s \in S \text{ with } Y(s) = y_i\}$.

Definition 1.14. The population is the entire group of objects from which we want information. The sample is the part of the population actually examined.

Example 1.4. Suppose that 5 year survival rates of 100 lung cancer patients are examined. Let a 1 denote the event that the $i$th patient died within 5 years of being diagnosed with lung cancer, and a 0 if the patient lived. Then outcomes in the sample space $S$ are 100-tuples (sequences of 100 digits) of the form $s = 1010111 \cdots 0111$. Let the random variable $X(s) =$ the number of 1’s in the 100-tuple = the sum of the 0’s and 1’s = the number of the 100 lung cancer patients who died within 5 years of being diagnosed with lung cancer. Notice that $X(s) = 82$ is easier to understand than a 100-tuple with 82 ones and 18 zeroes.

For the following definition, $F$ is a right continuous function if for every real number $x$, $\lim_{y \uparrow x} F(y) = F(x)$. Also, $F(\infty) = \lim_{y \to \infty} F(y)$ and $F(-\infty) = \lim_{y \to -\infty} F(y)$.
Definition 1.15. The cumulative distribution function (cdf) of any RV \( Y \) is \( F(y) = P(Y \leq y) \) for all \( y \in \mathbb{R} \). If \( F(y) \) is a cumulative distribution function, then \( F(-\infty) = 0, F(\infty) = 1 \), \( F \) is a nondecreasing function and \( F \) is right continuous.

Definition 1.16. A RV is discrete if it can assume only a finite or countable number of distinct values. The collection of these probabilities is the probability distribution of the discrete RV. The probability mass function (pmf) of a discrete RV \( Y \) is \( f(y) = P(Y = y) \) for all \( y \in \mathbb{R} \) where \( 0 \leq f(y) \leq 1 \) and \( \sum_{y_i: f(y_i) > 0} f(y_i) = 1 \).

Remark 1.1. The cdf \( F \) of a discrete RV is a step function.

Example 1.5: Common low level problem. The sample space of \( Y \) is \( S_Y = \{y_1, y_2, \ldots, y_k\} \) and a table of \( y_j \) and \( f(y_j) \) is given with one \( f(y_j) \) omitted. Find the omitted \( f(y_j) \) by using the fact that \( \sum_{i=1}^{k} f(y_i) = f(y_1) + f(y_2) + \cdots + f(y_k) = 1 \).

Definition 1.17. A RV \( Y \) is continuous if its distribution function \( F(y) \) is absolutely continuous.

The notation \( \forall y \) means “for all \( y \).”

Definition 1.18. If \( Y \) is a continuous RV, then the probability density function (pdf) \( f(y) \) of \( Y \) is a continuous function such that

\[
F(y) = \int_{-\infty}^{y} f(t)dt
\]

for all \( y \in \mathbb{R} \). If \( f(y) \) is a pdf, then \( f(y) \geq 0 \ \forall y \) and \( \int_{-\infty}^{\infty} f(t)dt = 1 \).

Theorem 1.7. If \( Y \) has pdf \( f(y) \), then \( f(y) = \frac{d}{dy} F(y) \equiv F'(y) \) wherever the derivative exists (in this text the derivative will exist everywhere except possibly for a finite number of points).

Theorem 1.8. i) \( P(a < Y \leq b) = F(b) - F(a) \).
ii) If \( Y \) has pdf \( f(y) \), then \( P(a < Y < b) = P(a < Y \leq b) = P(a \leq Y < b) = P(a \leq Y \leq b) = \int_{a}^{b} f(y)dy = F(b) - F(a) \).
iii) If \( Y \) has a probability mass function \( f(y) \), then \( Y \) is discrete and \( P(a < Y \leq b) = F(b) - F(a) \), but \( P(a \leq Y \leq b) \neq F(b) - F(a) \) if \( f(a) > 0 \).

Definition 1.19. Let \( Y \) be a discrete RV with probability mass function
CHAPTER 1. PROBABILITY AND EXPECTATIONS

$f(y)$. Then the **mean** or **expected value** of $Y$ is

$$EY \equiv E(Y) = \sum_{y:f(y)>0} y f(y)$$

if the sum exists when $y$ is replaced by $|y|$. If $g(Y)$ is a real valued function of $Y$, then $g(Y)$ is a random variable and

$$E[g(Y)] = \sum_{y:f(y)>0} g(y) f(y)$$

if the sum exists when $g(y)$ is replaced by $|g(y)|$. If the sums are not absolutely convergent, then $E(Y)$ and $E[g(Y)]$ do not exist.

**Definition 1.20.** If $Y$ has pdf $f(y)$, then the mean or expected value of $Y$ is

$$EY \equiv E(Y) = \int_{-\infty}^{\infty} y f(y)dy$$

and

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

provided the integrals exist when $y$ and $g(y)$ are replaced by $|y|$ and $|g(y)|$. If the modified integrals do not exist, then $E(Y)$ and $E[g(Y)]$ do not exist.

**Definition 1.21.** If $E(Y^2)$ exists, then the variance of a RV $Y$ is

$$\text{VAR}(Y) \equiv \text{Var}(Y) \equiv VY \equiv V(Y) = E[(Y - E(Y))^2]$$

and the **standard deviation** of $Y$ is $\text{SD}(Y) = \sqrt{\text{V}(Y)}$. If $E(Y^2)$ does not exist, then $V(Y)$ does not exist.

The following theorem is used over and over again, especially to find $E(Y^2) = V(Y) + (E(Y))^2$. The theorem is valid for all random variables that have a variance, including continuous and discrete RVs. If $Y$ is a Cauchy $(\mu, \sigma)$ RV (see Chapter 10), then neither $E(Y)$ nor $V(Y)$ exist.

**Theorem 1.9:** Short cut formula for variance.

$$V(Y) = E(Y^2) - (E(Y))^2.$$ (1.6)
CHAPTER 1. PROBABILITY AND EXPECTATIONS

If $Y$ is a discrete RV with sample space $\mathcal{S}_Y = \{y_1, y_2, \ldots, y_k\}$ then

$$E(Y) = \sum_{i=1}^{k} y_i f(y_i) = y_1 f(y_1) + y_2 f(y_2) + \cdots + y_k f(y_k)$$

and $E[g(Y)] = \sum_{i=1}^{k} g(y_i) f(y_i) = g(y_1) f(y_1) + g(y_2) f(y_2) + \cdots + g(y_k) f(y_k)$.

In particular,

$$E(Y^2) = y_1^2 f(y_1) + y_2^2 f(y_2) + \cdots + y_k^2 f(y_k).$$

Also

$$V(Y) = \sum_{i=1}^{k} (y_i - E(Y))^2 f(y_i) =$$

$$(y_1 - E(Y))^2 f(y_1) + (y_2 - E(Y))^2 f(y_2) + \cdots + (y_k - E(Y))^2 f(y_k).$$

For a continuous RV $Y$ with pdf $f(y)$, $V(Y) = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy$. Often using $V(Y) = E(Y^2) - (E(Y))^2$ is simpler.

**Example 1.6: Common low level problem.**

i) Given a table of $y$ and $f(y)$, find $E[g(Y)]$ and the standard deviation $\sigma = SD(Y)$.

ii) Find $f(y)$ from $F(y)$.

iii) Find $F(y)$ from $f(y)$.

iv) Given that $f(y) = c \, g(y)$, find $c$.

v) Given the pdf $f(y)$, find $P(a < Y < b)$, et cetera.

vi) Given the pmf or pdf $f(y)$ find $E[Y]$, $V(Y)$, $SD(Y)$, and $E[g(Y)]$. The functions $g(y) = y$, $g(y) = y^2$, and $g(y) = e^{ty}$ are especially common.

**Theorem 1.10.** Let $a$ and $b$ be any constants and assume all relevant expectations exist.

i) $E(a) = a$.

ii) $E(aY + b) = aE(Y) + b$.

iii) $E(aX + bY) = aE(X) + bE(Y)$.

iv) $V(aY + b) = a^2 V(Y)$.

**Definition 1.22.** The **moment generating function** (mgf) of a random variable $Y$ is

$$m(t) = E[e^{tY}]$$

(1.7)

if the expectation exists for $t$ in some neighborhood of 0. Otherwise, the mgf does not exist. If $Y$ is discrete, then $m(t) = \sum_{y} e^{ty} f(y)$, and if $Y$ is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$. 

Definition 1.23. The characteristic function (cf) of a random variable $Y$ is $c(t) = E[e^{itY}]$ where the complex number $i = \sqrt{-1}$.

This text does not require much knowledge of theory of complex variables, but know that $i^2 = -1$, $i^3 = -i$ and $i^4 = 1$. Hence $i^{4k-3} = i$, $i^{4k-2} = -1$, $i^{4k-1} = -i$ and $i^{4k} = 1$ for $k = 1, 2, 3, \ldots$. To compute the cf, the following result will be used. Moment generating functions do not necessarily exist in a neighborhood of zero, but a characteristic function always exists.

Proposition 1.11. Suppose that $Y$ is a RV with an mgf $m(t)$ that exists for $|t| < b$ for some constant $b > 0$. Then the cf of $Y$ is $c(t) = m(it)$.

Definition 1.24. Random variables $X$ and $Y$ are identically distributed, written $X \sim Y$ or $Y \sim F_X$, if $F_X(y) = F_Y(y)$ for all real $y$.

Proposition 1.12. Let $X$ and $Y$ be random variables. Then $X$ and $Y$ are identically distributed, $X \sim Y$, if any of the following conditions hold.

a) $F_X(y) = F_Y(y)$ for all $y$,

b) $f_X(y) = f_Y(y)$ for all $y$,

c) $c_X(t) = c_Y(t)$ for all $t$ or

d) $m_X(t) = m_Y(t)$ for all $t$ in a neighborhood of zero.

Definition 1.25. The $k$th moment of $Y$ is $E[Y^k]$ while the $k$th central moment is $E[(Y - E[Y])^k]$.

Theorem 1.13. Suppose that the mgf $m(t)$ exists for $|t| < b$ for some constant $b > 0$, and suppose that the $k$th derivative $m^{(k)}(t)$ exists for $|t| < b$. Then $E[Y^k] = m^{(k)}(0)$. In particular, $E[Y] = m'(0)$ and $E[Y^2] = m''(0)$.

Notation. The natural logarithm of $y$ is $\log(y) = \ln(y)$. If another base is wanted, it will be given, eg $\log_{10}(y)$.

Example 1.7: Common problem. Let $h(y)$, $g(y)$, $n(y)$ and $d(y)$ be functions. Review how to find the derivative $g'(y)$ of $g(y)$ and how to find the $k$th derivative

$$g^{(k)}(y) = \frac{d^k}{dy^k}g(y)$$

for $k \geq 2$. Recall that the product rule is

$$(h(y)g(y))' = h'(y)g(y) + h(y)g'(y).$$
The quotient rule is
\[
\left( \frac{n(y)}{d(y)} \right)' = \frac{d(y)n'(y) - n(y)d'(y)}{[d(y)]^2}.
\]

The chain rule is
\[
[h(g(y))]' = [h'(g(y))][g'(y)].
\]

Know the derivative of \( \log(y) \) and \( e^y \) and know the chain rule with these functions. Know the derivative of \( y^k \).

Then given the mgf \( m(t) \), find \( E[Y] = m'(0) \), \( E[Y^2] = m''(0) \) and \( V(Y) = E[Y^2] - (E[Y])^2 \).

Definition 1.26. Let \( f(y) \equiv f_Y(y|\theta) \) be the pdf or pmf of a random variable \( Y \). Then the set \( \mathcal{Y}_\theta = \{ y | f_Y(y|\theta) > 0 \} \) is called the sample space or support of \( Y \). Let the set \( \Theta \) be the set of parameter values \( \theta \) of interest. Then \( \Theta \) is the parameter space of \( Y \). Use the notation \( Y = \{ y | f(y|\theta) > 0 \} \) if the support does not depend on \( \theta \). So \( \mathcal{Y} \) is the support of \( Y \) if \( \mathcal{Y}_\theta \equiv \mathcal{Y} \forall \theta \in \Theta \).

Definition 1.27. The indicator function \( I_A(x) \equiv I(x \in A) = 1 \) if \( x \in A \) and \( I_A(x) = 0 \), otherwise. Sometimes an indicator function such as \( I_{(0,\infty)}(y) \) will be denoted by \( I(y > 0) \).

Example 1.8. Often equations for functions such as the pmf, pdf or cdf are given only on the support (or on the support plus points on the boundary of the support). For example, suppose
\[ f(y) = P(Y = y) = \binom{k}{y} \rho^y (1 - \rho)^{k-y} \]
for \( y = 0,1,\ldots,k \) where \( 0 < \rho < 1 \). Then the support of \( Y \) is \( \mathcal{Y} = \{0,1,\ldots,k\} \), the parameter space is \( \Theta = (0,1) \) and \( f(y) = 0 \) for \( y \) not \( \in \mathcal{Y} \). Similarly, if \( f(y) = 1 \) and \( F(y) = y \) for \( 0 \leq y \leq 1 \), then the support \( \mathcal{Y} = [0,1] \), \( f(y) = 0 \) for \( y < 0 \) and \( y > 1 \), \( F(y) = 0 \) for \( y < 0 \) and \( F(y) = 1 \) for \( y > 1 \).

Since the pmf and cdf are defined for all \( y \in \mathbb{R} = (-\infty,\infty) \) and the pdf is defined for all but finitely many \( y \), it may be better to use indicator functions when giving the formula for \( f(y) \). For example,
\[ f(y) = 1I(0 \leq y \leq 1) \]
is defined for all \( y \in \mathbb{R} \).
1.5 The Kernel Method

**Notation.** Notation such as \( E(Y|\theta) \equiv E_\theta(Y) \) or \( f_Y(y|\theta) \) is used to indicate that the formula for the expected value or pdf are for a family of distributions indexed by \( \theta \in \Theta \). A major goal of parametric inference is to collect data and estimate \( \theta \) from the data.

**Example 1.9.** If \( Y \sim N(\mu,\sigma^2) \), then \( Y \) is a member of the normal family of distributions with \( \theta = \{ (\mu,\sigma) | -\infty < \mu < \infty \text{ and } \sigma > 0 \} \). Then \( E[Y|(\mu,\sigma)] = \mu \) and \( V(Y|(\mu,\sigma)) = \sigma^2 \). This family has uncountably many members.

The kernel method is a widely used technique for finding \( E[g(Y)] \).

**Definition 1.28.** Let \( f_Y(y) \) be the pdf or pmf of a random variable \( Y \) and suppose that \( f_Y(y|\theta) = c(\theta)k(y|\theta) \). Then \( k(y|\theta) \geq 0 \) is the kernel of \( f_Y \) and \( c(\theta) > 0 \) is the constant term that makes \( f_Y \) sum or integrate to one. Thus \( \int_{-\infty}^{\infty} k(y|\theta) dy = 1/c(\theta) \) or \( \sum_{y \in Y} k(y|\theta) = 1/c(\theta) \).

Often \( E[g(Y)] \) is found using “tricks” tailored for a specific distribution. The word “kernel” means “essential part.” Notice that if \( f_Y(y) \) is a pdf, then
\[
E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y|\theta) dy = \int_{-\infty}^{\infty} g(y)f(y|\theta) dy.
\]
\[E[g(Y)] = a \left( \int_{-\infty}^{\infty} k(y|\tau) dy \right)
\]
for some constant \( a \) where \( \tau \in \Theta \) and \( \Theta \) is the parameter space. Then the kernel method says that
\[
E[g(Y)] = a \left( \int_{-\infty}^{\infty} \frac{c(\tau)}{c(\tau)} k(y|\tau) dy \right) = \frac{a \int_{-\infty}^{\infty} c(\tau)k(y|\tau) dy}{c(\tau)}.
\]

Similarly, if \( f_Y(y) \) is a pmf, then
\[
E[g(Y)] = \sum_{y : f_Y(y) > 0} g(y)f(y|\theta) = \sum_{y \in \mathcal{Y}} g(y)f(y|\theta)
\]
where \( \mathcal{Y} = \{ y : f_Y(y) > 0 \} \) is the support of \( Y \). Suppose that after algebra, it is found that
\[
E[g(Y)] = a \left( \sum_{y \in \mathcal{Y}} k(y|\tau) \right)
\]
for some constant $a$ where $\tau \in \Theta$. Then the kernel method says that

$$E[g(Y)] = a \frac{c(\theta)}{c(\tau)} \sum_{y \in Y} c(\tau) k(y|\tau) = \frac{a c(\theta)}{c(\tau)} \sum_{y \in Y} c(\tau) k(y|\tau) = \frac{a c(\theta)}{c(\tau)}.$$

The kernel method is often useful for finding $E[g(Y)]$, especially if $g(y) = y$, $g(y) = y^2$ or $g(y) = e^{ty}$. The kernel method is often easier than memorizing a trick specific to a distribution because the kernel method uses the same trick for every distribution: $\sum_{y \in Y} f(y) = 1$ and $\int_{y \in Y} f(y) dy = 1$. Of course sometimes tricks are needed to get the kernel $f(y|\tau)$ from $g(y|\theta)$. For example, complete the square for the normal (Gaussian) kernel.

**Example 1.10.** To use the kernel method to find the mgf of a gamma $(\nu, \lambda)$ distribution, refer to Chapter 10 and note that

$$m(t) = E(e^{tY}) = \int_0^\infty e^{ty} y^{\nu-1} e^{-y/\lambda} dy = \frac{1}{\lambda \nu \Gamma(\nu)} \int_0^\infty y^{\nu-1} \exp[-y(\frac{1}{\lambda} - t)] dy.$$

The integrand is the kernel of a gamma $(\nu, \eta)$ distribution with

$$\frac{1}{\eta} = \frac{1}{\lambda} - t = \frac{1 - \lambda t}{\lambda} \quad \text{so} \quad \eta = \frac{\lambda}{1 - \lambda t}.$$

Now

$$\int_0^\infty y^{\nu-1} e^{-y/\lambda} dy = \frac{1}{c(\nu, \lambda)} = \lambda^\nu \Gamma(\nu).$$

Hence

$$m(t) = \frac{1}{\lambda^\nu \Gamma(\nu)} \int_0^\infty y^{\nu-1} \exp[-y/\eta] dy = c(\nu, \lambda) \frac{1}{c(\nu, \eta)} = \frac{1}{\lambda^\nu \eta^\nu \Gamma(\nu)} \left( \frac{\eta}{\lambda} \right)^\nu = \left( \frac{1}{1 - \lambda t} \right)^\nu$$

for $t < 1/\lambda$.

**Example 1.11.** The zeta($\nu$) distribution has probability mass function

$$f(y) = P(Y = y) = \frac{1}{\zeta(\nu)y^\nu}$$
where \( \nu > 1 \) and \( y = 1, 2, 3, \ldots \). Here the zeta function

\[
\zeta(\nu) = \sum_{y=1}^{\infty} \frac{1}{y^\nu}
\]

for \( \nu > 1 \). Hence

\[
E(Y) = \sum_{y=1}^{\infty} \frac{1}{\zeta(\nu)} \frac{1}{y^\nu}
\]

\[
= \frac{1}{\zeta(\nu)} \zeta(\nu - 1) \sum_{y=1}^{\infty} \frac{1}{\zeta(\nu - 1)} \frac{1}{y^{\nu-1}} = \frac{\zeta(\nu - 1)}{\zeta(\nu)}
\]

if \( \nu > 2 \). Similarly

\[
E(Y^k) = \sum_{y=1}^{\infty} y^k \frac{1}{\zeta(\nu)} \frac{1}{y^\nu}
\]

\[
= \frac{1}{\zeta(\nu)} \zeta(\nu - k) \sum_{y=1}^{\infty} \frac{1}{\zeta(\nu - k)} \frac{1}{y^{\nu-k}} = \frac{\zeta(\nu - k)}{\zeta(\nu)}
\]

if \( \nu - k > 1 \) or \( \nu > k + 1 \). Thus if \( \nu > 3 \), then

\[
V(Y) = E(Y^2) - [E(Y)]^2 = \frac{\zeta(\nu - 2)}{\zeta(\nu)} - \left[ \frac{\zeta(\nu - 1)}{\zeta(\nu)} \right]^2.
\]

**Example 1.12.** The generalized gamma distribution has pdf

\[
f(y) = \frac{\phi y^{\nu-1}}{\lambda^\nu \Gamma(\nu)} \exp(-y^\phi/\lambda^\phi)
\]

where \( \nu, \lambda, \phi \) and \( y \) are positive, and

\[
E(Y^k) = \frac{\lambda^k \Gamma(\nu + k)}{\Gamma(\nu)} \text{ if } k > -\phi \nu.
\]

To prove this result using the kernel method, note that

\[
E(Y^k) = \int_{0}^{\infty} y^k \frac{\phi y^{\nu-1}}{\lambda^\nu \Gamma(\nu)} \exp(-y^\phi/\lambda^\phi) dy = \int_{0}^{\infty} \frac{\phi y^{\nu+k-1}}{\lambda^\nu \Gamma(\nu)} \exp(-y^\phi/\lambda^\phi) dy.
\]
This integrand looks much like a generalized gamma pdf with parameters $\nu_k$, $\lambda$ and $\phi$ where $\nu_k = \nu + (k/\phi)$ since

$$E(Y^k) = \int_0^\infty \frac{\phi y^{\nu+k/\phi-1}}{\lambda^\nu \Gamma(\nu)} \exp(-y^{\phi}/\lambda^\phi) dy.$$ 

Multiply the integrand by

$$1 = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\lambda^k \Gamma(\nu + \frac{k}{\phi})}$$

to get

$$E(Y^k) = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\Gamma(\nu)} \int_0^\infty \frac{\phi y^{\nu+k/\phi-1}}{\lambda^\phi(\nu+k/\phi) \Gamma(\nu + \frac{k}{\phi})} \exp(-y^{\phi}/\lambda^\phi) dy.$$ 

Then the result follows since the integral of a generalized gamma pdf with parameters $\nu_k$, $\lambda$ and $\phi$ over its support is 1. Notice that $\nu_k > 0$ implies $k > -\phi \nu$.

### 1.6 Mixture Distributions

Mixture distributions are often used as outlier models. The following two definitions and proposition are useful for finding the mean and variance of a mixture distribution. Parts a) and b) of Proposition 1.14 below show that the definition of expectation given in Definition 1.30 is the same as the usual definition for expectation if $Y$ is a discrete or continuous random variable.

**Definition 1.29.** The distribution of a random variable $Y$ is a mixture distribution if the cdf of $Y$ has the form

$$F_Y(y) = \sum_{i=1}^k \alpha_i F_{W_i}(y)$$

(1.8)

where $0 < \alpha_i < 1$, $\sum_{i=1}^k \alpha_i = 1$, $k \geq 2$, and $F_{W_i}(y)$ is the cdf of a continuous or discrete random variable $W_i$, $i = 1, \ldots, k$.

**Definition 1.30.** Let $Y$ be a random variable with cdf $F(y)$. Let $h$ be a function such that the expected value $E[h(Y)]$ exists. Then

$$E[h(Y)] = \int_{-\infty}^\infty h(y) dF(y).$$

(1.9)
Chapter 1. Probability and Expectations

Proposition 1.14. a) If $Y$ is a discrete random variable that has a pmf $f(y)$ with support $Y$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y) = \sum_{y \in Y} h(y) f(y).$$

b) If $Y$ is a continuous random variable that has a pdf $f(y)$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y) = \int_{-\infty}^{\infty} h(y) f(y) dy.$$

c) If $Y$ is a random variable that has a mixture distribution with cdf $F_Y(y) = \sum_{i=1}^{k} \alpha_i F_{W_i}(y)$, then

$$E[h(Y)] = \int_{-\infty}^{\infty} h(y) dF(y) = \sum_{i=1}^{k} \alpha_i E_{W_i}[h(W_i)]$$

where $E_{W_i}[h(W_i)] = \int_{-\infty}^{\infty} h(y) dF_{W_i}(y)$.

Example 1.13. Proposition 1.14c implies that the pmf or pdf of $W_i$ is used to compute $E_{W_i}[h(W_i)]$. As an example, suppose the cdf of $Y$ is $F(y) = (1 - \epsilon) \Phi(y) + \epsilon \Phi(y/k)$ where $0 < \epsilon < 1$ and $\Phi(y)$ is the cdf of $W_1 \sim N(0,1)$. Then $\Phi(x/k)$ is the cdf of $W_2 \sim N(0,k^2)$. To find $E[Y]$, use $h(y) = y$. Then

$$E[Y] = (1 - \epsilon)E[W_1] + \epsilon E[W_2] = (1 - \epsilon)0 + \epsilon 0 = 0.$$

To find $E[Y^2]$, use $h(y) = y^2$. Then

$$E[Y^2] = (1 - \epsilon)E[W_1^2] + \epsilon E[W_2^2] = (1 - \epsilon)1 + \epsilon k^2 = 1 - \epsilon + \epsilon k^2.$$

Thus $\text{VAR}(Y) = E[Y^2] - (E[Y])^2 = 1 - \epsilon + \epsilon k^2$. If $\epsilon = 0.1$ and $k = 10$, then $EY = 0$, and $\text{VAR}(Y) = 10.9$.

Remark 1.2. Warning: Mixture distributions and linear combinations of random variables are very different quantities. As an example, let

$$W = (1 - \epsilon)W_1 + \epsilon W_2$$

where $\epsilon$, $W_1$ and $W_2$ are as in the previous example and suppose that $W_1$ and $W_2$ are independent. Then $W$, a linear combination of $W_1$ and $W_2$, has a normal distribution with mean

$$E[W] = (1 - \epsilon)E[W_1] + \epsilon E[W_2] = 0$$
Figure 1.1: PDF $f$ of $(W_1 + W_2)/2$ and $f = 0.5f_1(y) + 0.5f_2(y)$
and variance
\[ \text{VAR}(W) = (1 - \epsilon)^2 \text{VAR}(W_1) + \epsilon^2 \text{VAR}(W_2) = (1 - \epsilon)^2 + \epsilon^2 k^2 < \text{VAR}(Y) \]

where \( Y \) is given in the example above. Moreover, \( W \) has a unimodal normal distribution while \( Y \) does not follow a normal distribution. In fact, if \( W_1 \sim N(0,1) \), \( W_2 \sim N(10,1) \), and \( W_1 \) and \( W_2 \) are independent, then \( (W_1 + W_2)/2 \sim N(5,0.5) \); however, if \( Y \) has a mixture distribution with cdf \( F_Y(y) = 0.5F_{W_1}(y) + 0.5F_{W_2}(y) = 0.5\Phi(y) + 0.5\Phi(y-10) \), then the pdf of \( Y \) is bimodal. See Figure 1.1.

1.7 Summary

Referring to Chapter 10, memorize the pmf or pdf \( f \), \( E(Y) \) and \( V(Y) \) for the following 10 RVs. You should recognize the mgf of the binomial, \( \chi^2 \), exponential, gamma, normal and Poisson distributions. You should recognize the cdf of the exponential, normal and uniform distributions.

1) beta(\( \delta, \nu \))
\[
f(y) = \frac{\Gamma(\delta + \nu)}{\Gamma(\delta)\Gamma(\nu)} y^{\delta-1}(1-y)^{\nu-1}
\]
where \( \delta > 0, \nu > 0 \) and \( 0 \leq y \leq 1 \).
\[ E(Y) = \frac{\delta}{\delta + \nu} \]
\[ \text{VAR}(Y) = \frac{\delta \nu}{(\delta + \nu)^2(\delta + \nu + 1)} \]

2) Bernoulli(\( \rho \)) = binomial(k = 1, \( \rho \)) \( f(y) = \rho^y (1-\rho)^{1-y} \) for \( y = 0, 1 \).
\[ E(Y) = \rho \]
\[ \text{VAR}(Y) = \rho(1-\rho) \]
\[ m(t) = [(1-\rho) + \rho e^t] \]

3) binomial(k, \( \rho \))
\[
f(y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}
\]
for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$.

$E(Y) = k\rho.$

$\text{VAR}(Y) = k\rho(1 - \rho).$

$m(t) = [(1 - \rho) + \rho e^t]^k.$

4) Cauchy($\mu, \sigma$)

$f(y) = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$

where $y$ and $\mu$ are real numbers and $\sigma > 0$.

$E(Y) = \infty = \text{VAR}(Y).$

5) chi-square($p$) = gamma($\nu = p/2, \lambda = 2$)

$f(y) = \frac{y^{\frac{p}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{p}{2}}\Gamma(\frac{p}{2})}$

$E(Y) = p.$

$\text{VAR}(Y) = 2p.$

$m(t) = \left(\frac{1}{1 - 2t}\right)^{p/2} = (1 - 2t)^{-p/2}$

for $t < 1/2.$

6) exponential($\lambda$) = gamma($\nu = 1, \lambda$)

$f(y) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right) I(y \geq 0)$

where $\lambda > 0$.

$E(Y) = \lambda.$

$\text{VAR}(Y) = \lambda^2.$

$m(t) = 1/(1 - \lambda t)$

for $t < 1/\lambda.$

$F(y) = 1 - \exp(-y/\lambda), \ y \geq 0.$

7) gamma($\nu, \lambda$)

$f(y) = \frac{y^{\nu-1}e^{-y/\lambda}}{\lambda^\nu\Gamma(\nu)}$
where $\nu, \lambda$, and $y$ are positive.

$E(Y) = \nu \lambda$.

$\text{VAR}(Y) = \nu \lambda^2$.

$$m(t) = \left(\frac{1}{1 - \lambda t}\right)^\nu$$

for $t < 1/\lambda$.

8) $N(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $\mu$ and $y$ are real.

$E(Y) = \mu$. $\text{VAR}(Y) = \sigma^2$.

$$m(t) = \exp(t\mu + t^2\sigma^2/2).$$

$$F(y) = \Phi\left(\frac{y - \mu}{\sigma}\right).$$

9) Poisson($\theta$)

$$f(y) = \frac{e^{-\theta} \theta^y}{y!}$$

for $y = 0, 1, \ldots$, where $\theta > 0$.

$E(Y) = \theta = \text{VAR}(Y)$.

$$m(t) = \exp(\theta(e^t - 1)).$$

10) uniform($\theta_1, \theta_2$)

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).$$

$F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \leq y \leq \theta_2$.

$E(Y) = (\theta_1 + \theta_2)/2$.

$\text{VAR}(Y) = (\theta_2 - \theta_1)^2/12$.

The terms sample space $S$, events, disjoint, partition, probability function, sampling with and without replacement, conditional probability, Bayes’ theorem, mutually independent events, random variable, cdf, continuous RV, discrete RV, identically distributed, pmf and pdf are important.
I) Be able to find $E[g(Y)]$, especially $E(Y) = m'(0)$, $E(Y^2) = m''(0)$, $V(Y) = E(Y^2) - [E(Y)]^2$ and the mgf $m(t) = m_Y(t) = E[e^{tY}]$.

II) Let $f_Y(y|\theta) = c(\theta)k(y|\theta)$ where $k(y|\theta)$ is the kernel of $f_Y$. Thus $\int_{-\infty}^{\infty} k(y|\theta)dy = 1/c(\theta)$. The kernel method is useful for finding $E[g(Y)]$ if $E[g(Y)] = a c(\theta) \int_{-\infty}^{\infty} k(y|\tau)dy = a c(\theta) \frac{1}{c(\tau)} \int_{-\infty}^{\infty} c(\tau)k(y|\tau)dy = a c(\theta) c(\tau)$ for some constant $a$. Replace the integral by a sum for a discrete distribution.

III) If the cdf of $X$ is $F_X(x) = (1-\epsilon)F_Z(x) + \epsilon F_W(x)$ where $0 \leq \epsilon \leq 1$ and $F_Z$ and $F_W$ are cdfs, then $E[g(X)] = (1-\epsilon)E[g(Z)] + \epsilon E[g(W)]$. In particular, $E(X^2) = (1-\epsilon)E[Z^2] + \epsilon E[W^2] = (1-\epsilon)[E(Z)+(E[Z])^2] + \epsilon[E(W)+(E[W])^2]$.

1.8 Complements

Kolmogorov’s definition of a probability function makes a probability function a normed measure. Hence many of the tools of measure theory can be used for probability theory. See, for example, Ash and Doleans-Dade (1999), Billingsley (1995), Dudley (2002), Durrett (1995), Feller (1971) and Resnick (1999). Feller (1957) and Tucker (1984) are good references for combinatorics.

1.9 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL. Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

1.1*. For the Binomial($k, \rho$) distribution,
- a) find $E Y$.
- b) Find $\text{Var} Y$.
- c) Find the mgf $m(t)$.

1.2*. For the Poisson($\theta$) distribution,
- a) find $E Y$. 


b) Find $\text{Var} \ Y$. (Hint: Use the kernel method to find $E \ Y(Y-1)$.)

c) Find the mgf $m(t)$.

1.3*. For the Gamma($\nu, \lambda$) distribution,

a) find $E \ Y$.

b) Find $\text{Var} \ Y$.

c) Find the mgf $m(t)$.

1.4*. For the Normal($\mu, \sigma^2$) (or Gaussian) distribution,

a) find the mgf $m(t)$. (Hint: complete the square to get a Gaussian kernel.)

b) Use the mgf to find $E \ Y$.

c) Use the mgf to find $\text{Var} \ Y$.

1.5*. For the Uniform($\theta_1, \theta_2$) distribution

a) find $E \ Y$.

b) Find $\text{Var} \ Y$.

c) Find the mgf $m(t)$.

1.6*. For the Beta($\delta, \nu$) distribution,

a) find $E \ Y$.

b) Find $\text{Var} \ Y$.

1.7*. See Mukhopadhyay (2000, p. 39). Recall integrals by u-substitution:

$$I = \int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(u)du = \int_c^d f(u)du =$$

$$F(u)|_c^d = F(d) - F(c) = F(u)|_{g(a)}^{g(b)} = F(g(x))|_a^b = F(g(b)) - F(g(a))$$

where $F'(x) = f(x), u = g(x), du = g'(x)dx, d = g(b), \text{and } c = g(a)$.

This problem uses the Gamma function and u-substitution to show that the normal density integrates to 1 (usually shown with polar coordinates). When you perform the u-substitution, make sure you say what $u = g(x), du = g'(x)dx, d = g(b), \text{and } c = g(a)$ are.

a) Let $f(x)$ be the pdf of a $N(\mu, \sigma^2)$ random variable. Perform u-substitution on

$$I = \int_{-\infty}^{\infty} f(x)dx$$

with $u = (x - \mu)/\sigma$. 
b) Break the result into two parts,
\[ I = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-u^2/2} du + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u^2/2} du. \]

Then perform u-substitution on the first integral with \( v = -u \).

c) Since the two integrals are now equal,
\[ I = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-v^2/2} dv = \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-v^2/2} \frac{1}{\sqrt{v}} dv. \]

Perform u-substitution with \( w = v^2/2 \).

d) Using the Gamma function, show that \( I = \Gamma(1/2)/\sqrt{\pi} = 1. \)

1.8. Let \( X \) be a \( N(0, 1) \) (standard normal) random variable. Use integration by parts to show that \( EX^2 = 1 \). Recall that integration by parts is used to evaluate \( \int f(x)g'(x)dx = \int udv = uv - \int vdu \) where \( u = f(x) \), \( dv = g'(x)dx \), \( du = f'(x)dx \) and \( v = g(x) \). When you do the integration, clearly state what these 4 terms are (eg \( u = x \)).

1.9. Verify the formula for the cdf \( F \) for the following distributions. That is, either show that \( F'(y) = f(y) \) or show that \( \int_{-\infty}^{y} f(t)dt = F(y) \forall y \in \mathbb{R} \).

a) Cauchy (\( \mu, \sigma \)).
b) Double exponential (\( \theta, \lambda \)).
c) Exponential (\( \lambda \)).
d) Logistic (\( \mu, \sigma \)).
e) Pareto (\( \sigma, \lambda \)).
f) Power (\( \lambda \)).
g) Uniform (\( \theta_1, \theta_2 \)).
h) Weibull \( W(\phi, \lambda) \).

1.10. Verify the formula for the expected value \( E(Y) \) for the following distributions.
a) Double exponential (\( \theta, \lambda \)).
b) Exponential ($\lambda$).

c) Logistic ($\mu, \sigma$). (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma W]$ so $E(Y) = \mu + \sigma E(W)$ where $W \sim L(0, 1)$. Hence

$$E(W) = \int_{-\infty}^{\infty} y \frac{e^y}{[1 + e^y]^2} dy.$$  

Use substitution with

$$u = \frac{e^y}{1 + e^y}.$$  

Then

$$E(W^k) = \int_{0}^{1} [\log(u) - \log(1 - u)]^k du.$$  

Also use the fact that

$$\lim_{v \to 0} v \log(v) = 0$$  

to show $E(W) = 0.)$

d) Lognormal ($\mu, \sigma^2$).

e) Pareto ($\sigma, \lambda$).

f) Weibull ($\phi, \lambda$).

1.11. Verify the formula for the variance $\text{VAR}(Y)$ for the following distributions.

a) Double exponential ($\theta, \lambda$).

b) Exponential ($\lambda$).

c) Logistic ($\mu, \sigma$). (Hint from deCani and Stine (1986): Let $Y = [\mu + \sigma X]$ so $V(Y) = \sigma^2 V(X) = \sigma^2 E(X^2)$ where $X \sim L(0, 1)$. Hence

$$E(X^2) = \int_{-\infty}^{\infty} y^2 \frac{e^y}{[1 + e^y]^2} dy.$$  

Use substitution with

$$v = \frac{e^y}{1 + e^y}.$$  

Then

$$E(X^2) = \int_{0}^{1} [\log(v) - \log(1 - v)]^2 dv.$$  

Let $w = \log(v) - \log(1 - v)$ and $du = [\log(v) - \log(1 - v)] dv$. Then

$$E(X^2) = \int_{0}^{1} wdu = uw|_{0}^{1} - \int_{0}^{1} udw.$$
CHAPTER 1. PROBABILITY AND EXPECTATIONS

Now
\[ uw|_0^1 = [v \log(v) + (1 - v) \log(1 - v)] w|_0^1 = 0 \]

since
\[ \lim_{v \to 0} v \log(v) = 0. \]

Now
\[ -\int_0^1 u \, dw = -\int_0^1 \frac{\log(v)}{1 - v} \, dv - \int_0^1 \frac{\log(1 - v)}{v} \, dv = \frac{2\pi^2}{6} = \frac{\pi^2}{3} \]

using
\[ \int_0^1 \frac{\log(v)}{1 - v} \, dv = \int_0^1 \frac{\log(1 - v)}{v} \, dv = -\frac{\pi^2}{6}. \]

d) Lognormal \((\mu, \sigma^2)\).
e) Pareto \((\sigma, \lambda)\).
f) Weibull \((\phi, \lambda)\).

Problems from old quizzes and exams.

1.12. Suppose the random variable \(X\) has cdf \(F_X(x) = 0.9 \Phi(x - 10) + 0.1 F_W(x)\) where \(\Phi(x - 10)\) is the cdf of a normal \(N(10, 1)\) random variable with mean 10 and variance 1 and \(F_W(x)\) is the cdf of the random variable \(W\) that satisfies \(P(W = 200) = 1\).
a) Find \(E W\).
b) Find \(E X\).

1.13. Suppose the random variable \(X\) has cdf \(F_X(x) = 0.9 F_Z(x) + 0.1 F_W(x)\) where \(F_Z\) is the cdf of a gamma\((\alpha = 10, \beta = 1)\) random variable with mean 10 and variance 10 and \(F_W(x)\) is the cdf of the random variable \(W\) that satisfies \(P(W = 400) = 1\).
a) Find \(E W\).
b) Find \(E X\).
1.14. Suppose the cdf \( F_X(x) = (1 - \epsilon)F_Z(x) + \epsilon F_W(x) \) where \( 0 \leq \epsilon \leq 1 \), \( F_Z \) is the cdf of a random variable \( Z \), and \( F_W \) is the cdf of a random variable \( W \). Then \( E g(X) = (1 - \epsilon)E g(Z) + \epsilon E g(W) \) where \( E g(Z) \) means that the expectation should be computed using the pmf or pdf of \( Z \).

Suppose the random variable \( X \) has cdf \( F_X(x) = 0.9 F_Z(x) + 0.1 F_W(x) \) where \( F_Z \) is the cdf of a gamma(\( \alpha = 10 \), \( \beta = 1 \)) random variable with mean 10 and variance 10 and \( F_W(x) \) is the cdf of the RV \( W \) that satisfies \( P(W = 400) = 1 \).

a) Find \( E W \).

b) Find \( E X \).

1.15. Let \( A \) and \( B \) be positive integers. A hypergeometric random variable \( X = W_1 + W_2 + \cdots + W_n \) where the random variables \( W_i \) are identically distributed random variables with \( P(W_i = 1) = A/(A + B) \) and \( P(W_i = 0) = B/(A + B) \).

a) Find \( E(W_i) \).

b) Find \( E(X) \).

1.16. Suppose \( P(X = x_o) = 1 \) for some constant \( x_o \).

a) Find \( E g(X) \) in terms of \( x_o \).

b) Find the moment generating function \( m(t) \) of \( X \).

c) Find \( m^{(n)}(t) = \frac{d^n}{dt^n} m(t) \). (Hint: find \( m^{(n)}(t) \) for \( n = 1, 2, \) and 3. Then the pattern should be apparent.)

1.17. Suppose \( P(X = 1) = 0.5 \) and \( P(X = -1) = 0.5 \). Find the moment generating function of \( X \).

1.18. Suppose that \( X \) is a discrete random variable with pmf \( f(x) = P(X = x) \) for \( x = 0, 1, \ldots, n \) so that the moment generating function of \( X \) is \( m(t) = \sum_{x=0}^{n} e^{tx} f(x) \).

a) Find \( \frac{d}{dt} m(t) = m'(t) \).

b) Find \( m'(0) \).

c) Find \( m''(t) = \frac{d^2}{dt^2} m(t) \).

d) Find \( m''(0) \).

e) Find \( m^{(k)}(t) = \frac{d^k}{dt^k} m(t) \). (Hint: you found \( m^{(k)}(t) \) for \( k = 1, 2, \) and the
pattern should be apparent.)

1.19. Suppose that the random variable \( W = e^X \) where \( X \sim N(\mu, \sigma^2) \). Find \( E(W^r) = E[(e^X)^r] \) by recognizing the relationship of \( E[(e^X)^r] \) with the moment generating function of a \( normal(\mu, \sigma^2) \) random variable.

1.20. Let \( X \sim N(\mu, \sigma^2) \) so that \( EX = \mu \) and \( Var \ X = \sigma^2 \).
   a) Find \( E(X^2) \).
   b) If \( k \geq 2 \) is an integer, then \( E(X^k) = (k-1)\sigma^2 E(X^{k-2}) + \mu E(X^{k-1}) \). Use this recursion relationship to find \( E(X^3) \).

1.21*. Let \( X \sim gamma(\nu, \lambda) \). Using the kernel method, find \( EX^r \) where \( r > -\nu \).

1.22. Find \( \int_{-\infty}^{\infty} \exp(-\frac{1}{2}y^2)dy \).
(Hint: the integrand is a Gaussian kernel.)

1.23. Let \( X \) have a Pareto \((\sigma, \lambda = 1/\theta)\) pdf
\[
f(x) = \frac{\theta\sigma^\theta}{x^{\theta+1}}
\]
where \( x > \sigma, \sigma > 0 \) and \( \theta > 0 \). Using the kernel method, find \( EX^r \) where \( \theta > r \).

1.24. Let \( Y \sim beta(\delta, \nu) \). Using the kernel method, find \( EY^r \) where \( r > -\delta \).

1.25. Use the kernel method to find the mgf of the logarithmic \((\theta)\) distribution.

1.26. Suppose that \( X \) has pdf
\[
f(x) = \frac{h(x)e^{\theta x}}{\lambda(\theta)}
\]
for \( x \in \mathcal{X} \) and for \(-\infty < \theta < \infty\) where \( \lambda(\theta) \) is some positive function of \( \theta \) and \( h(x) \) is some nonnegative function of \( x \). Find the moment generating function of \( X \) using the kernel method. Your final answer should be written in terms of \( \lambda, \theta \) and \( t \).

1.27. Use the kernel method to find \( E(Y^r) \) for the \( chi(p, \sigma) \) distribution.
1.28. Suppose the cdf $F_X(x) = (1 - \epsilon)F_Z(x) + \epsilon F_W(x)$ where $0 \leq \epsilon \leq 1$, $F_Z$ is the cdf of a random variable $Z$, and $F_W$ is the cdf of a random variable $W$. Then $E g(X) = (1 - \epsilon)E_Z g(Z) + \epsilon E_W g(W)$ where $E_Z g(Z)$ means that the expectation should be computed using the pmf or pdf of $Z$.

Suppose the random variable $X$ has cdf $F_X(x) = 0.9 F_Z(x) + 0.1 F_W(x)$ where $F_Z$ is the cdf of a gamma($\nu = 3$, $\lambda = 4$) random variable and $F_W(x)$ is the cdf of a Poisson(10) random variable.

a) Find $E X$.

b) Find $E X^2$.

1.29. If $Y$ has an exponential distribution truncated at 1, $Y \sim TEXP(\theta, 1)$, then the pdf of $Y$ is

$$f(y) = \frac{\theta}{1 - e^{-\theta}} e^{-\theta y}$$

for $0 < y < 1$ where $\theta > 0$. Find the mgf of $Y$ using the kernel method.

1.30. Following Morris (1982), let

$$f(y) = \frac{\cos(\theta)}{2 \cosh(\pi y/2)} \exp(\theta y)$$

where $y$ is real and $|\theta| < \pi/2$. Find the mgf of $Y$ using the kernel method.

1.31. If $Y$ has a log-gamma distribution, the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda^\nu \Gamma(\nu)} \exp(\nu y + (-1/\lambda)e^y)$$

where $y$ is real, $\nu > 0$ and $\lambda > 0$. Find the mgf of $Y$ using the kernel method.

1.32. If $Y$ has an inverted gamma distribution, $Y \sim INVG(\nu, \lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{y^{\nu+1} \Gamma(\nu)} I(y > 0) \frac{1}{\lambda^\nu} \exp\left(\frac{-1}{\lambda} \frac{1}{y}\right)$$

where $\lambda$, $\nu$ and $y$ are all positive. Using the kernel method, show

$$E(Y^r) = \frac{\Gamma(\nu - r)}{\lambda^\nu \Gamma(\nu)}$$
for \( \nu > r \).

1.33. If \( Y \) has a zero truncated Poisson distribution, \( Y \sim ZTP(\theta) \), then the pmf of \( Y \) is

\[
f(y) = \frac{e^{-\theta} \theta^y}{(1 - e^{-\theta}) y!}
\]

for \( y = 1, 2, 3, \ldots \) where \( \theta > 0 \). Find the mgf of \( Y \) using the kernel method.

1.34. If \( Y \) has a Zipf distribution, \( Y \sim Zipf(\nu) \), then the pmf of \( Y \) is

\[
f(y) = \frac{1}{y^\nu z(\nu)}
\]

where \( y \in \{0, 1, \ldots, m\} \) and \( m \) is known, \( \nu \) is real and

\[
z(\nu) = \sum_{y=1}^{m} \frac{1}{y^\nu}.
\]

Using the kernel method, show

\[
E(Y^r) = \frac{z(\nu - r)}{z(\nu)}
\]

for real \( r \).
Chapter 2
Multivariate Distributions
and Transformations

2.1 Joint, Marginal and Conditional Distributions

Often there are $n$ random variables $Y_1, \ldots, Y_n$ that are of interest. For example, age, blood pressure, weight, gender and cholesterol level might be some of the random variables of interest for patients suffering from heart disease.

**Notation.** Let $\mathbb{R}^n$ be the $n$-dimensional Euclidean space. Then the vector $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ if $y_i$ is an arbitrary real number for $i = 1, \ldots, n$.

**Definition 2.1.** If $Y_1, \ldots, Y_n$ are discrete random variables, then the joint pmf (probability mass function) of $Y_1, \ldots, Y_n$ is

$$f(y_1, \ldots, y_n) = P(Y_1 = y_1, \ldots, Y_n = y_n) \quad (2.1)$$

for any $(y_1, \ldots, y_n) \in \mathbb{R}^n$. A joint pmf $f$ satisfies $f(\mathbf{y}) \equiv f(y_1, \ldots, y_n) \geq 0 \ \forall \mathbf{y} \in \mathbb{R}^n$ and

$$\sum \cdots \sum_{\mathbf{y} : f(\mathbf{y}) > 0} f(y_1, \ldots, y_n) = 1.$$

For any event $A \in \mathbb{R}^n$,

$$P[(Y_1, \ldots, Y_n) \in A] = \sum \cdots \sum_{\mathbf{y} : \mathbf{y} \in A \text{ and } f(\mathbf{y}) > 0} f(y_1, \ldots, y_n).$$
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Definition 2.2. The joint cdf (cumulative distribution function) of \( Y_1, \ldots, Y_n \) is \( F(y_1, \ldots, y_n) = P(Y_1 \leq y_1, \ldots, Y_n \leq y_n) \) for any \((y_1, \ldots, y_n) \in \mathbb{R}^n\).

Definition 2.3. If \( Y_1, \ldots, Y_n \) are continuous random variables, then the joint pdf (probability density function) of \( Y_1, \ldots, Y_n \) is a function \( f(y_1, \ldots, y_n) \) that satisfies
\[
F(y_1, \ldots, y_n) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} f(t_1, \ldots, t_n) dt_1 \cdots dt_n
\]
where the \( y_i \) are any real numbers. A joint pdf \( f \) satisfies \( f(y) \equiv f(y_1, \ldots, y_n) \geq 0 \) \( \forall y \in \mathbb{R}^n \) and \( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, \ldots, t_n) dt_1 \cdots dt_n = 1 \). For any event \( A \in \mathbb{R}^n \),
\[
P[(Y_1, \ldots, Y_n) \in A] = \int_A f(t_1, \ldots, t_n) dt_1 \cdots dt_n
\]

Definition 2.4. If \( Y_1, \ldots, Y_n \) has a joint pdf or pmf \( f \), then the sample space or support of \( Y_1, \ldots, Y_n \) is
\[
\mathcal{Y} = \{(y_1, \ldots, y_n) \in \mathbb{R}^n : f(y_1, \ldots, y_n) > 0\}.
\]
If \( Y \) comes from a family of distributions \( f(y|\theta) \) for \( \theta \in \Theta \), then the support \( \mathcal{Y}_\theta = \{y : f(y|\theta) > 0\} \) may depend on \( \theta \).

Theorem 2.1. Let \( Y_1, \ldots, Y_n \) have joint cdf \( F(y_1, \ldots, y_n) \) and joint pdf \( f(y_1, \ldots, y_n) \). Then
\[
f(y_1, \ldots, y_n) = \frac{\partial^n}{\partial y_1 \cdots \partial y_n} F(y_1, \ldots, y_n)
\]
wherever the partial derivative exists.

Definition 2.5. The marginal pmf of any subset \( Y_{i_1}, \ldots, Y_{i_k} \) of the coordinates \( (Y_1, \ldots, Y_n) \) is found by summing the joint pmf over all possible values of the other coordinates where the values \( y_{i_1}, \ldots, y_{i_k} \) are held fixed. For example,
\[
f_{Y_{i_1}, \ldots, Y_{i_k}}(y_{i_1}, \ldots, y_{i_k}) = \sum_{y_{i_{k+1}}} \cdots \sum_{y_n} f(y_1, \ldots, y_n)
\]
where \( y_1, \ldots, y_k \) are held fixed. In particular, if \( Y_1 \) and \( Y_2 \) are discrete RVs with joint pmf \( f(y_1, y_2) \), then the marginal pmf for \( Y_1 \) is
\[
f_{Y_1}(y_1) = \sum_{y_2} f(y_1, y_2) \quad (2.2)
\]
where \( y_1 \) is held fixed. The marginal pmf for \( Y_2 \) is
\[
f_{Y_2}(y_2) = \sum_{y_1} f(y_1, y_2) \quad (2.3)
\]
where $y_2$ is held fixed.

**Example 2.1.** For $n = 2$, double integrals are used to find marginal pdfs (defined below) and to show that the joint pdf integrates to 1. If the region of integration $\Omega$ is bounded on top by the function $y_2 = \phi_T(y_1)$, on the bottom by the function $y_2 = \phi_B(y_1)$ and to the left and right by the lines $y_1 = a$ and $y_2 = b$ then

$$
\int_{a}^{b} \left[ \int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1, y_2) \, dy_2 \right] \, dy_1.
$$

Within the inner integral, treat $y_2$ as the variable, anything else, including $y_1$, is treated as a constant.

If the region of integration $\Omega$ is bounded on the left by the function $y_1 = \psi_L(y_2)$, on the right by the function $y_1 = \psi_R(y_2)$ and to the top and bottom by the lines $y_2 = c$ and $y_2 = d$ then

$$
\int_{c}^{d} \left[ \int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1, y_2) \, dy_1 \right] \, dy_2.
$$

Within the inner integral, treat $y_1$ as the variable, anything else, including $y_2$, is treated as a constant.

**Definition 2.6.** The **marginal pdf** of any subset $Y_{i1}, ..., Y_{ik}$ of the coordinates $(Y_1, ..., Y_n)$ is found by integrating the joint pdf over all possible values of the other coordinates where the values $y_{i1}, ..., y_{ik}$ are held fixed. For example, $f(y_1, ..., y_k) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(t_1, ..., t_n) \, dt_{k+1} \cdots dt_n$ where $y_1, ..., y_k$ are held fixed. In particular, if $Y_1$ and $Y_2$ are continuous RVs with joint pdf $f(y_1, y_2)$, then the marginal pdf for $Y_1$ is

$$
f_{Y_1}(y_1) = \int_{-\infty}^{\infty} f(y_1, y_2) \, dy_2 = \int_{\phi_B(y_1)}^{\phi_T(y_1)} f(y_1, y_2) \, dy_2 \quad (2.4)
$$

where $y_1$ is held fixed (to get the region of integration, draw a line parallel to the $y_2$ axis and use the functions $y_2 = \phi_B(y_1)$ and $y_2 = \phi_T(y_1)$ as the lower and upper limits of integration). The marginal pdf for $Y_2$ is

$$
f_{Y_2}(y_2) = \int_{-\infty}^{\infty} f(y_1, y_2) \, dy_1 = \int_{\psi_L(y_2)}^{\psi_R(y_2)} f(y_1, y_2) \, dy_1 \quad (2.5)
$$
where $y_2$ is held fixed (to get the region of integration, draw a line parallel to the $y_1$ axis and use the functions $y_1 = \psi_L(y_2)$ and $y_1 = \psi_R(y_2)$ as the lower and upper limits of integration).

**Definition 2.7.** The **conditional pmf** of any subset $Y_{i_1}, ..., Y_{i_k}$ of the coordinates $(Y_1, ..., Y_n)$ is found by dividing the joint pmf by the marginal pmf of the remaining coordinates assuming that the values of the remaining coordinates are fixed and that the denominator $> 0$. For example,

$$f(y_1, ..., y_k | y_{k+1}, ..., y_n) = \frac{f(y_1, ..., y_n)}{f(y_{k+1}, ..., y_n)}$$

if $f(y_{k+1}, ..., y_n) > 0$. In particular, the conditional pmf of $Y_1$ given $Y_2 = y_2$ is a function of $y_1$ and

$$f_{Y_1 | Y_2 = y_2}(y_1 | y_2) = \frac{f(y_1, y_2)}{f_{Y_2}(y_2)}$$

(2.6)

if $f_{Y_2}(y_2) > 0$, and the conditional pmf of $Y_2$ given $Y_1 = y_1$ is a function of $y_2$ and

$$f_{Y_2 | Y_1 = y_1}(y_2 | y_1) = \frac{f(y_1, y_2)}{f_{Y_1}(y_1)}$$

(2.7)

if $f_{Y_1}(y_1) > 0$.

**Definition 2.8.** The **conditional pdf** of any subset $Y_{i_1}, ..., Y_{i_k}$ of the coordinates $(Y_1, ..., Y_n)$ is found by dividing the joint pdf by the marginal pdf of the remaining coordinates assuming that the values of the remaining coordinates are fixed and that the denominator $> 0$. For example,

$$f(y_1, ..., y_k | y_{k+1}, ..., y_n) = \frac{f(y_1, ..., y_n)}{f(y_{k+1}, ..., y_n)}$$

if $f(y_{k+1}, ..., y_n) > 0$. In particular, the conditional pdf of $Y_1$ given $Y_2 = y_2$ is a function of $y_1$ and

$$f_{Y_1 | Y_2 = y_2}(y_1 | y_2) = \frac{f(y_1, y_2)}{f_{Y_2}(y_2)}$$

(2.8)

if $f_{Y_2}(y_2) > 0$, and the conditional pdf of $Y_2$ given $Y_1 = y_1$ is a function of $y_2$ and

$$f_{Y_2 | Y_1 = y_1}(y_2 | y_1) = \frac{f(y_1, y_2)}{f_{Y_1}(y_1)}$$

(2.9)
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

35

if \( f_{Y_1}(y_1) > 0 \).

**Example 2.2: Common Problem.** If the joint pmf \( f(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \) is given by a table, then the function \( f(y_1, y_2) \) is a joint pmf if \( f(y_1, y_2) \geq 0, \forall y_1, y_2 \) and if

\[
\sum_{(y_1, y_2): f(y_1, y_2) > 0} f(y_1, y_2) = 1.
\]

The marginal pmfs are found from the row sums and column sums using Definition 2.5, and the conditional pmfs are found with the formulas given in Definition 2.7.

**Example 2.3: Common Problem.** Given the joint pdf \( f(y_1, y_2) = kg(y_1, y_2) \) on its support, find \( k \), find the marginal pdfs \( f_{Y_1}(y_1) \) and \( f_{Y_2}(y_2) \) and find the conditional pdfs \( f_{Y_1|Y_2}(y_1|y_2) \) and \( f_{Y_2|Y_1}(y_2|y_1) \). Also,

\[
P(a_1 < Y_1 < b_1, a_2 < Y_2 < b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(y_1, y_2) dy_1 dy_2.
\]

Tips: Often using symmetry helps.

The support of the marginal pdf does not depend on the 2nd variable.

The support of the conditional pdf can depend on the 2nd variable. For example, the support of \( f_{Y_1|Y_2}(y_1|y_2) \) could have the form \( 0 \leq y_1 \leq y_2 \).

The support of continuous random variables \( Y_1 \) and \( Y_2 \) is the region where \( f(y_1, y_2) > 0 \). The support is generally given by one to three inequalities such as \( 0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1, \) and \( 0 \leq y_1 \leq y_2 \leq 1 \). For each variable, set the inequalities to equalities to get boundary lines. For example \( 0 \leq y_1 \leq y_2 \leq 1 \) yields 5 lines: \( y_1 = 0, y_1 = 1, y_2 = 0, y_2 = 1, \) and \( y_2 = y_1 \). Generally \( y_2 \) is on the vertical axis and \( y_1 \) is on the horizontal axis for pdfs.

To determine the limits of integration, examine the dummy variable used in the inner integral, say \( dy_1 \). Then within the region of integration, draw a line parallel to the same \((y_1)\) axis as the dummy variable. The limits of integration will be functions of the other variable \((y_2)\), never of the dummy variable \((dy_1)\).

2.2 Expectation, Covariance and Independence

For joint pmfs with \( n = 2 \) random variables \( Y_1 \) and \( Y_2 \), the marginal pmfs and conditional pmfs can provide important information about the data. For joint pdfs the integrals are usually too difficult for the joint, conditional
and marginal pdfs to be of practical use unless the random variables are independent. (Exceptions are the multivariate normal distribution and the elliptically contoured distributions. See Sections 2.9 and 2.10.)

For independent random variables, the joint cdf is the product of the marginal cdfs, the joint pmf is the product of the marginal pmfs, and the joint pdf is the product of the marginal pdfs. Recall that $\forall$ is read “for all.”

**Definition 2.9.** i) The random variables $Y_1, Y_2, ..., Y_n$ are independent if $F(y_1, y_2, ..., y_n) = F_{Y_1}(y_1)F_{Y_2}(y_2) \cdots F_{Y_n}(y_n) \forall y_1, y_2, ..., y_n$.

ii) If the random variables have a joint pdf or pmf $f$ then the random variables $Y_1, Y_2, ..., Y_n$ are independent if $f(y_1, y_2, ..., y_n) = f_{Y_1}(y_1)f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) \forall y_1, y_2, ..., y_n$.

If the random variables are not independent, then they are dependent.

In particular random variables $Y_1$ and $Y_2$ are independent, written $Y_1 \perp Y_2$, if either of the following conditions holds.

i) $F(y_1, y_2) = F_{Y_1}(y_1)F_{Y_2}(y_2) \forall y_1, y_2$.

ii) $f(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2) \forall y_1, y_2$.

Otherwise, $Y_1$ and $Y_2$ are dependent.

**Definition 2.10.** Recall that the support $\mathcal{Y}$ of $(Y_1, Y_2, ..., Y_n)$ is $\mathcal{Y} = \{y : f(y) > 0\}$. The support is a cross product or Cartesian product if $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 \times \cdots \times \mathcal{Y}_n = \{y : y_i \in \mathcal{Y}_i \text{ for } i = 1, ..., n\}$ where $\mathcal{Y}_i$ is the support of $Y_i$. If $f$ is a joint pdf then the support is rectangular if $\mathcal{Y}_i$ is an interval for each $i$. If $f$ is a joint pmf then the support is rectangular if the points in $\mathcal{Y}_i$ are equally spaced for each $i$.

**Example 2.4.** In applications the support is often rectangular. For $n = 2$ the support is a cross product if $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2 = \{(y_1, y_2) : y_1 \in \mathcal{Y}_1 \text{ and } y_2 \in \mathcal{Y}_2\}$ where $\mathcal{Y}_i$ is the support of $Y_i$. The support is rectangular if $\mathcal{Y}_1$ and $\mathcal{Y}_2$ are intervals. For example, if $\mathcal{Y} = \{(y_1, y_2) : a < y_1 < \infty \text{ and } c \leq y_2 \leq d\}$, then $\mathcal{Y}_1 = (a, \infty)$ and $\mathcal{Y}_2 = [c, d]$. For a joint pmf, the support is rectangular if the grid of points where $f(y_1, y_2) > 0$ is rectangular.
Cross Product of $(1,2,3,4,9)$ with $(1,3,4,5,9)$

Figure 2.1: Cross Product for a Joint PMF
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Figure 2.1 shows the cross product of $\mathcal{Y}_1 \times \mathcal{Y}_2$ where $\mathcal{Y}_1 = \{1, 2, 3, 4, 9\}$ and $\mathcal{Y}_2 = \{1, 3, 4, 5, 9\}$. Each dot occurs where $f(y_1, y_2) > 0$. Notice that each point in $\mathcal{Y}_1$ occurs with each point in $\mathcal{Y}_2$. This support would not be a cross product if any point was deleted, but would be a cross product if any row of dots or column of dots was deleted.

Theorem 2.2a below is useful because it is often immediate from the formula for the joint pdf or the table for the joint pmf that the support is not a cross product. Hence $Y_1$ and $Y_2$ are dependent. For example, if the support of $Y_1$ and $Y_2$ is a triangle, then $Y_1$ and $Y_2$ are dependent. A necessary condition for independence is that the support is a cross product. Theorem 2.2b is useful because factorizing the joint pdf on cross product support is easier than using integration to find the marginal pdfs. Many texts give Theorem 2.2c, but 2.2b is easier to use. Recall that $\prod_{i=1}^n a_i = a_1a_2\cdots a_n$. For example, let $n = 3$ and $a_i = i$ for $i = 1, 2, 3$. Then $\prod_{i=1}^n a_i = a_1a_2a_3 = (1)(2)(3) = 6$.

**Theorem 2.2.** a) Random variables $Y_1, \ldots, Y_n$ with joint pdf or pmf $f$ are dependent if their support $\mathcal{Y}$ is not a cross product. In particular, $Y_1$ and $Y_2$ are dependent if $\mathcal{Y}$ does not have the form $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$.

b) If random variables $Y_1, \ldots, Y_n$ with joint pdf or pmf $f$ have support $\mathcal{Y}$ that is a cross product, then $Y_1, \ldots, Y_n$ are independent iff $f(y_1, y_2, \ldots, y_n) = h_1(y_1)h_2(y_2)\cdots h_n(y_n)$ for all $y \in \mathcal{Y}$ where $h_i$ is a positive function of $y_i$ alone. In particular, if $\mathcal{Y} = \mathcal{Y}_1 \times \mathcal{Y}_2$, then $Y_1 \perp Y_2$ iff $f(y_1, y_2) = h_1(y_1)h_2(y_2)$ for all $(y_1, y_2) \in \mathcal{Y}$ where $h_i(y_i) > 0$ for $y_i \in \mathcal{Y}_i$ and $i = 1, 2$.

c) $Y_1, \ldots, Y_n$ are independent iff $f(y_1, y_2, \ldots, y_n) = g_1(y_1)g_2(y_2)\cdots g_n(y_n)$ for all $y$ where $g_i$ is a nonnegative function of $y_i$ alone.

d) If discrete $Y_1$ and $Y_2$ have cross product support given by a table, find the row and column sums. If $f(y_1, y_2) \neq f_{Y_1}(y_1)f_{Y_2}(y_2)$ for some entry $(y_1, y_2)$, then $Y_1$ and $Y_2$ are dependent. If $f(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$ for all table entries, then $Y_1$ and $Y_2$ are independent.

**Proof.** a) If the support is not a cross product, then there is a point $y$ such that $f(y) = 0$ but $f_{Y_i}(y_i) > 0$ for $i = 1, \ldots, n$. Hence $f(y) \neq \prod_{i=1}^n f_{Y_i}(y_i)$ at the point $y$ and $Y_1, \ldots, Y_n$ are dependent.

b) The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. If $Y_1, \ldots, Y_n$ are independent, take $h_i(y_i) = f_{Y_i}(y_i) > 0$ for $y_i \in \mathcal{Y}_i$ and $i = 1, \ldots, n$. 
If \( f(y) = h_1(y_1) \cdots h_n(y_n) \) for \( y \in \mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n \) then \( f(y) = 0 = f_{Y_1}(y_1) \cdots f_{Y_n}(y_n) \) if \( y \) is not in \( \mathcal{Y} \). Hence we need to show that \( f(y) = f_{Y_1}(y_1) \cdots f_{Y_n}(y_n) = h_1(y_1) \cdots h_{n}(y_n) \) if \( y \in \mathcal{Y} \). Since \( f \) is a joint pdf,

\[
1 = \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} f(y) \, dy = \prod_{i=1}^{n} \int_{\mathcal{Y}_i} h_i(y_i) \, dy_i = \prod_{i=1}^{n} a_i
\]

where \( a_i = \int_{\mathcal{Y}_i} h_i(y_i) \, dy_i > 0 \). For \( y_i \in \mathcal{Y}_i \), the marginal pdfs \( f_{Y_i}(y_i) = \)

\[
\int_{\mathcal{Y}_i} \cdots \int_{\mathcal{Y}_{i+1}} \cdots \int_{\mathcal{Y}_{i-1}} \cdots \int_{\mathcal{Y}_1} h_1(y_1) \cdots h_i(y_i) \cdots h_{n}(y_n) \, dy_1 \cdots dy_{i-1}dy_{i+1} \cdots dy_n = h_i(y_i) \prod_{j=1, j \neq i}^{n} \int_{\mathcal{Y}_j} h_j(y_j) \, dy_j = h_i(y_i) \prod_{j=1, j \neq i}^{n} a_j = h_i(y_i) \frac{1}{a_i}.
\]

Thus \( a_i f_{Y_i}(y_i) = h_i(y_i) \) for \( y_i \in \mathcal{Y}_i \). Since \( \prod_{i=1}^{n} a_i = 1 \),

\[
f(y) = \prod_{i=1}^{n} h_i(y_i) = \prod_{i=1}^{n} a_i f_{Y_i}(y_i) = \left( \prod_{i=1}^{n} a_i \right) \left( \prod_{i=1}^{n} f_{Y_i}(y_i) \right) = \prod_{i=1}^{n} f_{Y_i}(y_i)
\]

if \( y \in \mathcal{Y} \).

c) Take

\[
g_i(y_i) = \begin{cases} h_i(y_i), & \text{if } y_i \in \mathcal{Y}_i \\ 0, & \text{otherwise.} \end{cases}
\]

Then the result follows from b).

d) Since \( f(y_1, y_2) = 0 = f_{Y_1}(y_1)f_{Y_2}(y_2) \) if \( (y_1, y_2) \) is not in the support of \( Y_1 \) and \( Y_2 \), the result follows by the definition of independent random variables. QED

The following theorem shows that finding the marginal and conditional pdfs or pmfs is simple if \( Y_1, \ldots, Y_n \) are independent. Also subsets of independent random variables are independent: if \( Y_1, \ldots, Y_n \) are independent and if \( \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\} \) for \( k \geq 2 \), then \( Y_{i_1}, \ldots, Y_{i_k} \) are independent.

**Theorem 2.3.** Suppose that \( Y_1, \ldots, Y_n \) are independent random variables with joint pdf or pmf \( f(y_1, \ldots, y_n) \). Then
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

a) the marginal pdf or pmf of any subset $Y_{i_1},...,Y_{i_k}$ is $f(y_{i_1},...,y_{i_k}) = \prod_{j=1}^{k} f_{Y_{ij}}(y_{ij})$. Hence $Y_{i_1},...,Y_{i_k}$ are independent random variables for $k \geq 2$.

b) The conditional pdf or pmf of $Y_{i_1},...,Y_{i_k}$ given any subset of the remaining random variables $Y_{j_1} = y_{j_1},...,Y_{j_m} = y_{j_m}$ is equal to the marginal:

$$f(y_{i_1},...,y_{i_k}|y_{j_1},...,y_{j_m}) = f(y_{i_1},...,y_{i_k}) = \prod_{j=1}^{k} f_{Y_{ij}}(y_{ij})$$

if $f(y_{j_1},...,y_{j_m}) > 0$.

Proof. The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. a) The marginal

$$f(y_{i_1},...,y_{i_k}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left[ \prod_{j=1}^{n} f_{Y_{ij}}(y_{ij}) \right] dy_{i_{k+1}} \cdots dy_{i_n}$$

$$= \left[ \prod_{j=1}^{k} f_{Y_{ij}}(y_{ij}) \right] \left[ \prod_{j=k+1}^{n} \int_{-\infty}^{\infty} f_{Y_{ij}}(y_{ij}) dy_{ij} \right]$$

$$= \left[ \prod_{j=1}^{k} f_{Y_{ij}}(y_{ij}) \right] (1)^{n-k} = \prod_{j=1}^{k} f_{Y_{ij}}(y_{ij}).$$

b) follows from a) and the definition of a conditional pdf assuming that $f(y_{j_1},...,y_{j_m}) > 0$. QED

Definition 2.11. Suppose that random variables $\mathbf{Y} = (Y_1,\ldots,Y_n)$ have support $\mathcal{Y}$ and joint pdf or pmf $f$. Then the expected value of $h(\mathbf{Y}) = h(Y_1,\ldots,Y_n)$ is

$$E[h(\mathbf{Y})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\mathbf{y})f(\mathbf{y}) \, d\mathbf{y} = \int \cdots \int_{\mathcal{Y}} h(\mathbf{y})f(\mathbf{y}) \, d\mathbf{y} \quad (2.10)$$

if $f$ is a joint pdf and if

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} |h(\mathbf{y})|f(\mathbf{y}) \, d\mathbf{y}$$

exists. Otherwise the expectation does not exist. The expected value is

$$E[h(\mathbf{Y})] = \sum_{y_1} \cdots \sum_{y_n} h(y_1) \cdots h(y_n) = \sum_{\mathbf{y} \in \mathbb{R}^n} h(\mathbf{y})f(\mathbf{y}) = \sum_{\mathbf{y} \in \mathcal{Y}} h(\mathbf{y})f(\mathbf{y}) \quad (2.11)$$
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

if \( f \) is a joint pmf and if \( \sum_{\mathbf{y} \in \mathbb{R}^n} |h(\mathbf{y})|f(\mathbf{y}) \) exists. Otherwise the expectation does not exist.

The following theorem is useful since multiple integrals with smaller dimension are easier to compute than those with higher dimension.

**Theorem 2.4.** Suppose that \( Y_1, ..., Y_n \) are random variables with joint pdf or pmf \( f(y_1, ..., y_n) \). Let \( \{i_1, ..., i_k\} \subset \{1, ..., n\} \), and let \( f(y_{i_1}, ..., y_{i_k}) \) be the marginal pdf or pmf of \( Y_{i_1}, ..., Y_{i_k} \) with support \( \mathcal{Y}_{i_1}, ..., \mathcal{Y}_{i_k} \). Assume that \( E[h(Y_{i_1}, ..., Y_{i_k})] \) exists. Then

\[
E[h(Y_{i_1}, ..., Y_{i_k})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1}, ..., y_{i_k}) f(y_{i_1}, ..., y_{i_k}) \, dy_{i_1} \cdots dy_{i_k} = \\
\int \cdots \int_{\mathcal{Y}_{i_1}, ..., \mathcal{Y}_{i_k}} h(y_{i_1}, ..., y_{i_k}) f(y_{i_1}, ..., y_{i_k}) \, dy_{i_1} \cdots dy_{i_k}
\]

if \( f \) is a pdf, and

\[
E[h(Y_{i_1}, ..., Y_{i_k})] = \sum_{y_{i_1}} \cdots \sum_{y_{i_k}} h(y_{i_1}, ..., y_{i_k}) f(y_{i_1}, ..., y_{i_k})
\]

if \( f \) is a pmf.

**Proof.** The proof for a joint pdf is given below. For a joint pmf, replace the integrals by appropriate sums. Let \( g(Y_1, ..., Y_n) = h(Y_{i_1}, ..., Y_{i_k}) \). Then \( E[g(Y)] = \)

\[
\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1}, ..., y_{i_k}) f(y_{i_1}, ..., y_{i_k}) \, dy_{i_1} \cdots dy_{i_k} = \\
\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(y_{i_1}, ..., y_{i_k}) \left[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_{i_1}, ..., y_{i_k}) \, dy_{i_{k+1}} \cdots dy_{i_n} \right] \, dy_{i_1} \cdots dy_{i_k}
\]

since the term in the brackets gives the marginal. QED
Example 2.5. Typically $E(Y_i), E(Y_i^2)$ and $E(Y_i Y_j)$ for $i \neq j$ are of primary interest. Suppose that $(Y_1, Y_2)$ has joint pdf $f(y_1, y_2)$. Then $E[h(Y_1, Y_2)]$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) \, dy_2 \, dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) \, dy_1 \, dy_2$$

where $-\infty$ to $\infty$ could be replaced by the limits of integration for $dy_1$. In particular,

$$E(Y_1 Y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) \, dy_2 \, dy_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y_1 y_2 f(y_1, y_2) \, dy_1 \, dy_2.$$

Since finding the marginal pdf is usually easier than doing the double integral, if $h$ is a function of $Y_i$ but not of $Y_j$, find the marginal for $Y_i$:

$$E[h(Y_1)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(y_1, y_2) f(y_1, y_2) \, dy_2 \, dy_1 = \int_{-\infty}^{\infty} h(y_1) f_{Y_1}(y_1) \, dy_1.$$

Similarly, $E[h(Y_2)] = \int_{-\infty}^{\infty} h(y_2) f_{Y_2}(y_2) \, dy_2$.

In particular, $E(Y_1) = \int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1) \, dy_1$, and $E(Y_2) = \int_{-\infty}^{\infty} y_2 f_{Y_2}(y_2) \, dy_2$.

Suppose that $(Y_1, Y_2)$ have a joint pmf $f(y_1, y_2)$. Then the expectation $E[h(Y_1, Y_2)] = \sum_{y_2} \sum_{y_1} h(y_1, y_2) f(y_1, y_2) = \sum_{y_1} \sum_{y_2} h(y_1, y_2) f(y_1, y_2)$. In particular,

$$E[Y_1 Y_2] = \sum_{y_1} \sum_{y_2} y_1 y_2 f(y_1, y_2).$$

Since finding the marginal pmf is usually easier than doing the double summation, if $h$ is a function of $Y_i$ but not of $Y_j$, find the marginal for pmf for $Y_i$:

$$E[h(Y_1)] = \sum_{y_2} \sum_{y_1} h(y_1, y_2) f(y_1, y_2) = \sum_{y_1} h(y_1) f_{Y_1}(y_1).$$

Similarly, $E[h(Y_2)] = \sum_{y_2} h(y_2) f_{Y_2}(y_2)$. In particular, $E(Y_1) = \sum_{y_1} y_1 f_{Y_1}(y_1)$ and $E(Y_2) = \sum_{y_2} y_2 f_{Y_2}(y_2)$.

For pdfs it is sometimes possible to find $E[h(Y_i)]$ but for $k \geq 2$ these expected values tend to be too difficult to compute unless the problem is impractical. Independence makes finding some expected values simple.

Theorem 2.5. Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of $Y_i$ alone and if the relevant expected values exist, then

$$E[h_1(Y_1)h_2(Y_2) \cdots h_n(Y_n)] = E[h_1(Y_1)] \cdots E[h_n(Y_n)].$$

In particular, $E[Y_i Y_j] = E[Y_i] E[Y_j]$ for $i \neq j$. 
**CHAPTER 2. MULTIVARIATE DISTRIBUTIONS**

**Proof.** The result will be shown for the case where $Y = (Y_1, ..., Y_n)$ has a joint pdf $f$. For a joint pmf, replace the integrals by appropriate sums. By independence, the support of $Y$ is a cross product: $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_n$. Since $f(y) = \prod_{i=1}^{n} f_{Y_i}(y_i)$, the expectation $E[h_1(Y_1)h_2(Y_2)\cdots h_n(Y_n)] = \int_{\mathcal{Y}} \cdots \int_{\mathcal{Y}} h_1(y_1)h_2(y_2)\cdots h_n(y_n) f(y_1, ..., y_n) \, dy_1 \cdots dy_n$

$$= \int_{\mathcal{Y}_n} \cdots \int_{\mathcal{Y}_1} \left[ \prod_{i=1}^{n} h_i(y_i) f_{Y_i}(y_i) \right] \, dy_1 \cdots dy_n
= \prod_{i=1}^{n} \left[ \int_{\mathcal{Y}_i} h_i(y_i) f_{Y_i}(y_i) \, dy_i \right] = \prod_{i=1}^{n} E[h_i(Y_i)].$$

**QED**

**Corollary 2.6.** Let $Y_1, ..., Y_n$ be independent random variables. If $h_j(Y_{ij})$ is a function of $Y_{ij}$ alone and if the relevant expected values exist, then

$$E[h_1(Y_{i1}) \cdots h_k(Y_{ik})] = E[h_1(Y_{i1})] \cdots E[h_k(Y_{ik})].$$

**Proof.** Method 1: Take $X_j = Y_{ij}$ for $j = 1, ..., k$. Then $X_1, ..., X_k$ are independent and Theorem 2.5 applies.

Method 2: Take $h_j(Y_{ij}) \equiv 1$ for $j = k + 1, ..., n$ and apply Theorem 2.5. QED

**Theorem 2.7.** Let $Y_1, ..., Y_n$ be independent random variables. If $h_i(Y_i)$ is a function of $Y_i$ alone and $X_i = h_i(Y_i)$, then the random variables $X_1, ..., X_n$ are independent.

**Definition 2.12.** The **covariance** of $Y_1$ and $Y_2$ is

$$\text{Cov}(Y_1, Y_2) = E[(Y_1 - E(Y_1))(Y_2 - E(Y_2))]$$

provided the expectation exists. Otherwise the covariance does not exist.

**Theorem 2.8: Short cut formula.** If $\text{Cov}(Y_1, Y_2)$ exists then $\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2)$.

**Theorem 2.9.** Let $Y_1$ and $Y_2$ be independent random variables.

a) If $\text{Cov}(Y_1, Y_2)$ exists, then $\text{Cov}(Y_1, Y_2) = 0$. 
b) **The converse is false:** $\text{Cov}(Y_1, Y_2) = 0$ does not imply $Y_1 \perp Y_2$.

**Example 2.6.** When $f(y_1, y_2)$ is given by a table, a common problem is to determine whether $Y_1$ and $Y_2$ are independent or dependent, find the marginal pmfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ and find the conditional pmfs $f_{Y_1|Y_2=y_2}(y_1|y_2)$ and $f_{Y_2|Y_1=y_1}(y_2|y_1)$. Also find $E(Y_1)$, $E(Y_2)$, $V(Y_1)$, $V(Y_2)$, $E(Y_1Y_2)$ and $\text{Cov}(Y_1, Y_2)$.

**Example 2.7.** Given the joint pdf $f(y_1, y_2) = kg(y_1, y_2)$ on its support, a common problem is to find $k$, find the marginal pdfs $f_{Y_1}(y_1)$ and $f_{Y_2}(y_2)$ and find the conditional pdfs $f_{Y_1|Y_2=y_2}(y_1|y_2)$ and $f_{Y_2|Y_1=y_1}(y_2|y_1)$. Also determine whether $Y_1$ and $Y_2$ are independent or dependent, and find $E(Y_1)$, $E(Y_2)$, $V(Y_1)$, $V(Y_2)$, $E(Y_1Y_2)$ and $\text{Cov}(Y_1, Y_2)$.

**Example 2.8.** Suppose that the joint probability mass function of $Y_1$ and $Y_2$ is $f(y_1, y_2)$ is tabled as shown.

<table>
<thead>
<tr>
<th>$f(y_1, y_2)$</th>
<th>$y_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>$y_1$</td>
<td>1/9</td>
</tr>
<tr>
<td>2</td>
<td>1/9</td>
</tr>
</tbody>
</table>

a) Are $Y_1$ and $Y_2$ independent? Explain.
b) Find the marginal pmfs.
c) Find $E(Y_1)$.
d) Find $E(Y_2)$.
e) Find $\text{Cov}(Y_1, Y_2)$.

**Solution:** a) No, the support is not a cross product. Alternatively, $f(2, 2) = 0 < f_{Y_1}(2)f_{Y_2}(2)$.
b) Find $f_{Y_1}(y_1)$ by finding the row sums. Find $f_{Y_2}(y_2)$ by finding the column sums. In both cases, $f_{Y_1}(0) = f_{Y_1}(1) = 4/9$ and $f_{Y_2}(2) = 1/9$.
c) $E(Y_1) = \sum y_1 f_{Y_1}(y_1) = 0 + 4 \cdot 1/9 + 2 \cdot 1/9 = 6/9 \approx 0.6667$.
d) $E(Y_2) \approx 0.6667$ is found as in c) with $y_2$ replacing $y_1$.
e) $E(Y_1Y_2) = \sum \sum y_1 y_2 f(y_1, y_2) = 0 + 0 + 0 + 1 \cdot 1/9 + 2 \cdot 1/9 = 2/9$.

Hence $\text{Cov}(Y_1, Y_2) = E(Y_1Y_2) - E(Y_1)E(Y_2) = 2/9 - (6/9)(6/9) = -2/9 \approx -0.2222$.

**Example 2.9.** Suppose that the joint pdf of the random variables $Y_1$
and $Y_2$ is given by

$$f(y_1, y_2) = 10y_1y_2^2, \text{ if } 0 < y_1 < y_2 < 1$$

and $f(y_1, y_2) = 0$, otherwise. a) Find the marginal pdf of $Y_1$. Include the support. b) Is $Y_1 \perp Y_2$?

Solution: a) Notice that for a given value of $y_1$, the joint pdf is positive for $y_1 < y_2 < 1$. Thus

$$f_{Y_1}(y_1) = \int_{y_1}^{1} 10y_1y_2^2dy_2 = 10y_1\left[\frac{y_2^3}{3}\right]_{y_1}^{1} = 10y_1\left(1 - \frac{y_1^3}{3}\right), 0 < y_1 < 1.$$  

b) No, the support is not a cross product.

**Example 2.10.** Suppose that the joint pdf of the random variables $Y_1$ and $Y_2$ is given by

$$f(y_1, y_2) = 4y_1(1 - y_2), \text{ if } 0 \leq y_1 \leq 1, 0 \leq y_2 \leq 1$$

and $f(y_1, y_2) = 0$, otherwise.

a) Find the marginal pdf of $Y_1$. Include the support.

b) Find $E(Y_1)$.

c) Find $V(Y_1)$.

d) Are $Y_1$ and $Y_2$ independent? Explain.

Solution: a) $f_{Y_1}(y_1) = \int_{0}^{1} 4y_1(1 - y_2)dy_2 = 4y_1\left[1 - \frac{y_2^2}{2}\right]_{0}^{1} = 4y_1(1 - \frac{1}{2}) = 2y_1, 0 < y_1 < 1.$

b) $E(Y_1) = \int_{0}^{1} y_1f_{Y_1}(y_1)dy_1 = \int_{0}^{1} y_12y_1dy_1 = 2\int_{0}^{1} y_1^2dy_1 = 2\left[\frac{y_1^3}{3}\right]_{0}^{1} = 2/3.$

c) $E(Y_1^2) = \int_{0}^{1} y_1^2f_{Y_1}(y_1)dy_1 = \int_{0}^{1} y_1^22y_1dy_1 = 2\int_{0}^{1} y_1^3dy_1 = 2\left[\frac{y_1^4}{4}\right]_{0}^{1} = 1/2.$

So $V(Y_1) = E(Y_1^2) - [E(Y_1)]^2 = \frac{1}{2} - \frac{4}{9} = \frac{1}{18} \approx 0.0556$.

d) Yes, use Theorem 2.2b with $f(y_1, y_2) = (4y_1)(1 - y_2) = h_1(y_1)h_2(y_2)$ on cross product support.

### 2.3 Conditional Expectation and Variance

**Notation:** $Y|X = x$ is a single conditional distribution while $Y|X$ is a family of distributions. For example, if $Y|X = x \sim N(c + dx, \sigma^2)$, then
\( Y \mid X \sim N(c + dX, \sigma^2) \) is the family of normal distributions with variance \( \sigma^2 \) and mean \( \mu_{Y \mid X = x} = c + dx \).

**Definition 2.13.** Suppose that \( f(y \mid x) \) is the conditional pmf or pdf of \( Y \mid X = x \) and that \( h(Y) \) is a function of \( Y \). Then the *conditional expected value* \( E[h(Y) \mid X = x] \) of \( h(Y) \) given \( X = x \) is

\[
E[h(Y) \mid X = x] = \sum_y h(y)f(y \mid x) \tag{2.12}
\]

if \( f(y \mid x) \) is a pmf and if the sum exists when \( h(y) \) is replaced by \(|h(y)|\). In particular,

\[
E[Y \mid X = x] = \sum_y yf(y \mid x). \tag{2.13}
\]

Similarly,

\[
E[h(Y) \mid X = x] = \int_{-\infty}^{\infty} h(y)f(y \mid x)dy \tag{2.14}
\]

if \( f(y \mid x) \) is a pdf and if the integral exists when \( h(y) \) is replaced by \(|h(y)|\). In particular,

\[
E[Y \mid X = x] = \int_{-\infty}^{\infty} yf(y \mid x)dy. \tag{2.15}
\]

**Definition 2.14.** Suppose that \( f(y \mid x) \) is the conditional pmf or pdf of \( Y \mid X = x \). Then the *conditional variance*

\[
\text{VAR}(Y \mid X = x) = E(Y^2 \mid X = x) - [E(Y \mid X = x)]^2
\]

whenever \( E(Y^2 \mid X = x) \) exists.

Recall that \( f(y \mid x) \) is a function of \( y \) with \( x \) fixed, but \( E(Y \mid X = x) \equiv m(x) \) is a function of \( x \). In the definition below, both \( E(Y \mid X) \) and \( \text{VAR}(Y \mid X) \) are random variables since \( m(X) \) and \( v(X) \) are random variables.

**Definition 2.15.** If \( E(Y \mid X = x) = m(x) \), then \( E(Y \mid X) = m(X) \). Similarly if \( \text{VAR}(Y \mid X = x) = v(x) \), then \( \text{VAR}(Y \mid X) = v(X) = E(Y^2 \mid X) - [E(Y \mid X)]^2 \).

**Example 2.11.** Suppose that \( Y = \text{weight} \) and \( X = \text{height} \) of college students. Then \( E(Y \mid X = x) \) is a function of \( x \). For example, the weight of 5 feet tall students is less than the weight of 6 feet tall students, on average.
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Notation: When computing \(E(h(Y))\), the marginal pdf or pmf \(f(y)\) is used. When computing \(E[h(Y)|X = x]\), the conditional pdf or pmf \(f(y|x)\) is used. In a formula such as \(E[E(Y|X)]\) the inner expectation uses \(f(y|x)\) but the outer expectation uses \(f(x)\) since \(E(Y|X)\) is a function of \(X\). In the formula below, we could write \(E_Y(Y) = E_X[E_{Y|X}(Y|X)]\), but such notation is usually omitted.

Theorem 2.10: Iterated Expectations. Assume the relevant expected values exist. Then

\[
E(Y) = E[E(Y|X)].
\]

Proof: The result will be shown for the case where \((Y, X)\) has a joint pmf \(f\). For a joint pdf, replace the sums by appropriate integrals. Now

\[
E(Y) = \sum_x \sum_y yf(x, y) = \sum_x \sum_y yf_{Y|X}(y|x)f_X(x)
\]

\[
= \sum_x \left[ \sum_y yf_{Y|X}(y|x) \right] f_X(x) = \sum_x E(Y|X = x)f_X(x) = E[E(Y|X)]
\]

since the term in brackets is \(E(Y|X = x)\). QED

Theorem 2.11: Steiner’s Formula or the Conditional Variance Identity. Assume the relevant expectations exist. Then

\[
\text{VAR}(Y) = E[\text{VAR}(Y|X)] + \text{VAR}[E(Y|X)].
\]

Proof: Following Rice (1988, p. 132), since \(\text{VAR}(Y|X) = E(Y^2|X) - [E(Y|X)]^2\) is a random variable,

\[
E[\text{VAR}(Y|X)] = E[E(Y^2|X)] - E([E(Y|X)]^2).
\]

If \(W\) is a random variable then \(E(W) = E[E(W|X)]\) by Theorem 2.10 and \(\text{VAR}(W) = E(W^2) - [E(W)]^2\) by the short cut formula. Letting \(W = E(Y|X)\) gives

\[
\text{VAR}(E(Y|X)) = E([E(Y|X)]^2) - (E[E(Y|X)])^2.
\]

Since \(E(Y^2) = E[E(Y^2|X)]\) and since \(E(Y) = E[E(Y|X)]\),

\[
\text{VAR}(Y) = E(Y^2) - [E(Y)]^2 = E[E(Y^2|X)] - (E[E(Y|X)])^2.
\]
Adding 0 to \( \text{VAR}(Y) \) gives
\[
\text{VAR}(Y) = E[E(Y^2|X)] - E([E(Y|X)]^2) + E([E(Y|X)]^2) - (E[E(Y|X)])^2
\]
\[
= E[\text{VAR}(Y|X)] + \text{VAR}[E(Y|X)]. \quad \text{QED}
\]

A hierarchical model models a complicated process by a sequence of models placed in a hierarchy. Interest might be in the marginal expectation \( E(Y) \) and marginal variance \( \text{VAR}(Y) \). One could find the joint pmf from \( f(x,y) = f(y|x)f(x) \), then find the marginal distribution \( f_{Y}(y) \) and then find \( E(Y) \) and \( \text{VAR}(Y) \). Alternatively, use Theorems 2.10 and 2.11.

**Example 2.12.** Suppose \( Y|X \sim \text{BIN}(X, \rho) \) and \( X \sim \text{Poisson}(\lambda) \). Then \( E(Y|X) = X\rho, \text{VAR}(Y|X) = X\rho(1 - \rho) \) and \( E(X) = \text{VAR}(X) = \lambda \). Hence \( E(Y) = E[E(Y|X)] = E(X\rho) = \rho E(X) = \rho \lambda \) and \( \text{VAR}(Y) = E[\text{VAR}(Y|X)] + \text{VAR}[E(Y|X)] = E[X\rho(1 - \rho)] + \text{VAR}(X\rho) = \lambda \rho(1 - \rho) + \rho^2 \text{VAR}(X) = \lambda \rho(1 - \rho) + \rho^2 \lambda = \lambda \rho \).

### 2.4 Location–Scale Families

Many univariate distributions are location, scale or location–scale families. Assume that the random variable \( Y \) has a pdf \( f_Y(y) \).

**Definition 2.16.** Let \( f_Y(y) \) be the pdf of \( Y \). Then the family of pdfs
\[
f_W(w) = f_Y(w - \mu)
\]
indexed by the location parameter \( \mu, -\infty < \mu < \infty \), is the location family for the random variable \( W = \mu + Y \) with standard pdf \( f_Y(y) \).

**Definition 2.17.** Let \( f_Y(y) \) be the pdf of \( Y \). Then the family of pdfs
\[
f_W(w) = (1/\sigma)f_Y(w/\sigma)
\]
indexed by the scale parameter \( \sigma > 0 \), is the scale family for the random variable \( W = \sigma Y \) with standard pdf \( f_Y(y) \).

**Definition 2.18.** Let \( f_Y(y) \) be the pdf of \( Y \). Then the family of pdfs
\[
f_W(w) = (1/\sigma)f_Y((w - \mu)/\sigma)
\]
indexed by the location and scale parameters \( \mu, -\infty < \mu < \infty \), and \( \sigma > 0 \), is the location–scale family for the random variable \( W = \mu + \sigma Y \) with standard pdf \( f_Y(y) \).

The most important scale family is the exponential \( \text{EXP}(\lambda) \) distribution. Other scale families from Chapter 10 include the chi \( (p,\sigma) \) distribution if \( p \) is known, the Gamma \( G(\nu,\lambda) \) distribution if \( \nu \) is known, the lognormal...
(\mu, \sigma^2) distribution with scale parameter \tau = e^\mu if \sigma^2 is known, the one sided stable OSS(\sigma) distribution, the Pareto PAR(\sigma, \lambda) distribution if \lambda is known, and the Weibull W(\phi, \lambda) distribution with scale parameter \sigma = \lambda^{1/\phi} if \phi is known.

A location family can be obtained from a location–scale family by fixing the scale parameter while a scale family can be obtained by fixing the location parameter. The most important location–scale families are the Cauchy C(\mu, \sigma), double exponential DE(\theta, \lambda), logistic L(\mu, \sigma), normal N(\mu, \sigma^2) and uniform U(\theta_1, \theta_2) distributions. Other location–scale families from Chapter 10 include the two parameter exponential EXP(\theta, \lambda), half Cauchy HC(\mu, \sigma), half logistic HL(\mu, \sigma), half normal HN(\mu, \sigma), largest extreme value LEV(\theta, \sigma), Maxwell Boltzmann MB(\mu, \sigma), Rayleigh R(\mu, \sigma) and smallest extreme value SEV(\theta, \sigma) distributions.

2.5 Transformations

Transformations for univariate distributions are important because many “brand name” random variables are transformations of other brand name distributions. These transformations will also be useful for finding the distribution of the complete sufficient statistic for a 1 parameter exponential family. See Chapter 10.

Example 2.13: Common problem. Suppose that X is a discrete random variable with pmf \( f_X(x) \) given by a table. Let the transformation \( Y = t(X) \) for some function \( t \) and find the probability function \( f_Y(y) \).

Solution: Step 1) Find \( t(x) \) for each value of \( x \).
Step 2) Collect \( x : t(x) = y \), and sum the corresponding probabilities:
\[
   f_Y(y) = \sum_{x : t(x) = y} f_X(x),
\]
and table the resulting pmf \( f_Y(y) \) of \( Y \).

For example, if \( Y = X^2 \) and \( f_X(-1) = 1/3, f_X(0) = 1/3, \) and \( f_X(1) = 1/3 \), then \( f_Y(0) = 1/3 \) and \( f_Y(1) = 2/3 \).

Definition 2.19. Let \( h : D \to \mathbb{R} \) be a real valued function with domain \( D \). Then \( h \) is increasing if \( h(y_1) < h(y_2) \), nondecreasing if \( h(y_1) \leq h(y_2) \), decreasing if \( h(y_1) > h(y_2) \) and nonincreasing if \( h(y_1) \geq h(y_2) \) provided that \( y_1 \) and \( y_2 \) are any two numbers in \( D \) with \( y_1 < y_2 \). The function \( h \) is a monotone function if \( h \) is either increasing or decreasing.
Figure 2.2: Increasing and Decreasing $t(x)$
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Recall that if \( h \) is differentiable on an open interval \( D \) or continuous on a closed interval \( D \) and differentiable on the interior of \( D \), then \( h \) is increasing if \( h'(y) > 0 \) for all \( y \) in the interior of \( D \) and \( h \) is decreasing if \( h'(y) < 0 \) for all \( y \) in the interior of \( D \). Also if \( h \) is increasing then \(-h\) is decreasing. Similarly, if \( h \) is decreasing then \(-h\) is increasing.

Suppose that \( X \) is a continuous random variable with pdf \( f_X(x) \) on support \( \mathcal{X} \). Let the transformation \( Y = t(X) \) for some monotone function \( t \). Then there are two ways to find the support \( \mathcal{Y} \) of \( Y = t(x) \) if the support \( \mathcal{X} \) of \( X \) is an interval with endpoints \( a < b \) where \( a = -\infty \) and \( b = \infty \) are possible. Let \( t(a) \equiv \lim_{y \to a} t(y) \) and let \( t(b) \equiv \lim_{y \to b} t(y) \). A graph can help. If \( t \) is an increasing function, then \( \mathcal{Y} \) is an interval with endpoints \( t(a) < t(b) \). If \( t \) is a decreasing function, then \( \mathcal{Y} \) is an interval with endpoints \( t(b) < t(a) \). The second method is to find \( x = t^{-1}(y) \). Then if \( \mathcal{X} = [a, b] \), say, solve \( a \leq t^{-1}(y) \leq b \) in terms of \( y \).

If \( t(x) \) is increasing then \( P(\{Y \leq y\}) = P(\{X \leq t^{-1}(y)\}) \) while if \( t(x) \) is decreasing \( P(\{Y \leq y\}) = P(\{X \geq t^{-1}(y)\}) \). To see this, look at Figure 2.2. Suppose the support of \( Y \) is \([0, 9]\) and the support of \( X \) is \([0, 3]\). Now the height of the curve is \( y = t(x) \). Mentally draw a horizontal line from \( y \) to \( t(x) \) and then drop a vertical line to the \( x \)-axis. The value on the \( x \)-axis is \( t^{-1}(y) \) since \( t(t^{-1}(y)) = y \). Hence in Figure 2.2 a) \( t^{-1}(4) = 2 \) and in Figure 2.2 b) \( t^{-1}(8) = 1 \). If \( w < y \) then \( t^{-1}(w) < t^{-1}(y) \) if \( t(x) \) is increasing as in Figure 2.2 a), but \( t^{-1}(w) > t^{-1}(y) \) if \( t(x) \) is decreasing as in Figure 2.2 b). Hence \( P(Y \leq y) = P(t^{-1}(Y) \geq t^{-1}(y)) = P(X \geq t^{-1}(y)) \).

**Theorem 2.12: the CDF Method or Method of Distributions:**
Suppose that the continuous cdf \( F_X(x) \) is known and that \( Y = t(X) \). Find the support \( \mathcal{Y} \) of \( Y \).

i) If \( t \) is an increasing function then, \( F_Y(y) = P(Y \leq y) = P(t(X) \leq y) = P(X \leq t^{-1}(y)) = F_X(t^{-1}(y)) \).

ii) If \( t \) is a decreasing function then, \( F_Y(y) = P(Y \leq y) = P(t(X) \leq y) = P(X \geq t^{-1}(y)) = 1 - P(X < t^{-1}(y)) = 1 - P(X \leq t^{-1}(y)) = 1 - F_X(t^{-1}(y)) \).

iii) The special case \( Y = X^2 \) is important. If the support of \( X \) is positive, use i). If the support of \( X \) is negative, use ii). If the support of \( X \) is \((-\infty, a)\) (where \( a = \infty \) is allowed), then \( F_Y(y) = P(Y \leq y) = \)
Figure 2.3: Pdfs for N(0,1) and HN(0,1) Distributions

\[ P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \]
\[ \int_{-\sqrt{y}}^{\sqrt{y}} f_X(x) \, dx = F_X(\sqrt{y}) - F_X(-\sqrt{y}), \quad 0 \leq y < a^2. \]

After finding the cdf \( F_Y(y) \), the pdf of \( Y \) is \( f_Y(y) = \frac{d}{dy} F_Y(y) \) for \( y \in \mathcal{Y} \).

**Example 2.14.** Suppose \( X \) has a pdf with support on the real line and that the pdf is symmetric about \( \mu \) so \( f_X(\mu - w) = f_X(\mu + w) \) for all real \( w \). It can be shown that \( X \) has a symmetric distribution about \( \mu \) if \( Z = X - \mu \) and \( -Z = \mu - X \) have the same distribution. Several named right skewed distributions with support \( y \geq \mu \) are obtained by the transformation \( Y = \mu + |X - \mu| \). Similarly, let \( U \) be a U(0,1) random variable that is independent of \( Y \), then \( X \) can be obtained from \( Y \) by letting \( X = Y \) if \( U \leq 0.5 \) and \( X = 2\mu - Y \) if \( U > 0.5 \). Pairs of such distributions include the
exponential and double exponential, normal and half normal, Cauchy and half Cauchy, and logistic and half logistic distributions. Figure 2.3 shows the N(0,1) and HN(0,1) pdfs.

Notice that for \( y \geq \mu \),

\[
F_Y(y) = P(Y \leq y) = P(\mu + |X - \mu| \leq y) = P(|X - \mu| \leq y - \mu) = P(\mu - y \leq X - \mu \leq y - \mu) = P(2\mu - y \leq X \leq y) = F_X(y) - F_X(2\mu - y).
\]

Taking derivatives and using the symmetry of \( f_X \) gives

\[
f_Y(y) = f_X(y) + f_X(2\mu - y) = f_X(\mu + (y - \mu)) + f_X(\mu - (y - \mu)) = 2f_X(\mu + (y - \mu)) = 2f_X(y) \text{ for } y \geq \mu.
\]

Hence \( f_Y(y) = 2f_X(y)I(y \geq \mu) \).

Then \( X \) has pdf

\[
f_X(x) = \frac{1}{2} f_Y(\mu + |x - \mu|)
\]

for all real \( x \), since this pdf is symmetric about \( \mu \) and \( f_X(x) = 0.5f_Y(x) \) if \( x \geq \mu \).

**Example 2.15.** Often the rules of differentiation such as the multiplication, quotient and chain rules are needed. For example if the support of \( X \) is \([-a, a]\) and if \( Y = X^2 \), then

\[
f_Y(y) = \frac{1}{2\sqrt{y}} [f_X(\sqrt{y}) + f_X(-\sqrt{y})]
\]

for \( 0 \leq y \leq a^2 \).

**Theorem 2.13: the Transformation Method.** Assume that \( X \) has pdf \( f_X(x) \) and support \( \mathcal{X} \). Find the support \( \mathcal{Y} \) of \( Y = t(X) \). If \( t(x) \) is either increasing or decreasing on \( \mathcal{X} \) and if \( t^{-1}(y) \) has a continuous derivative on \( \mathcal{Y} \), then \( Y = t(X) \) has pdf

\[
f_Y(y) = f_X(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right| \quad (2.16)
\]

for \( y \in \mathcal{Y} \). As always, \( f_Y(y) = 0 \) for \( y \) not in \( \mathcal{Y} \).

**Proof:** Examining Theorem 2.12, if \( t \) is increasing then \( F_Y(y) = F_X(t^{-1}(y)) \) and

\[
f_Y(y) = \frac{d}{dy} F_Y(y)
\]
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

\[ \frac{d}{dy} F_X(t^{-1}(y)) = f_X(t^{-1}(y)) \frac{dt^{-1}(y)}{dy} t^{-1}(y) = f_X(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right| \]

for \( y \in \mathcal{Y} \) since the derivative of a differentiable increasing function is positive.

If \( t \) is a decreasing function then from Theorem 2.12, \( F_Y(y) = 1 - F_X(t^{-1}(x)) \). Hence

\[ f_Y(y) = \frac{d}{dy} [1 - F_X(t^{-1}(y))] = -f_X(t^{-1}(y)) \frac{dt^{-1}(y)}{dy} = f_X(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right| \]

for \( y \in \mathcal{Y} \) since the derivative of a differentiable decreasing function is negative.

Tips: To be useful, formula (2.16) should be simplified as much as possible.

a) The pdf of \( Y \) will often be that of a gamma random variable. In particular, the pdf of \( Y \) is often the pdf of an exponential(\( \lambda \)) random variable.

b) To find the inverse function \( x = t^{-1}(y) \), solve the equation \( y = t(x) \) for \( x \).

c) The log transformation is often used. Know how to sketch \( \log(x) \) and \( e^x \) for \( x > 0 \). Recall that in this text, \( \log(x) \) is the natural logarithm of \( x \).

d) If \( \mathcal{X} \) is an interval with endpoints \( a \) and \( b \), find

\[ \mathcal{Y} = (\min\{t(a), t(b)\}, \max\{t(a), t(b)\}) \]

as described two paragraphs above Theorem 2.12.

**Example 2.16.** Let \( X \) be a random variable with pdf

\[ f_X(x) = \frac{1}{x \sqrt{2\pi\sigma^2}} \exp \left( -\frac{(\log(x) - \mu)^2}{2\sigma^2} \right) \]

where \( x > 0 \), \( \mu \) is real and \( \sigma > 0 \). Let \( Y = \log(X) \) and find the distribution of \( Y \).

Solution: \( X = e^Y = t^{-1}(Y) \). So

\[ \left| \frac{dt^{-1}(y)}{dy} \right| = |e^y| = e^y, \]

and

\[ f_Y(y) = f_X(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right| = f_X(e^y) e^y = \]
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

\begin{equation}
\frac{1}{e^y\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(\log(e^y) - \mu)^2}{2\sigma^2} \right)e^y =
\end{equation}

\begin{equation}
\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left( -\frac{(y - \mu)^2}{2\sigma^2} \right)
\end{equation}

for \( y \in (-\infty, \infty) \) since \( x > 0 \) implies that \( y = \log(x) \in (-\infty, \infty) \). Notice that \( X \) is lognormal \( (\mu, \sigma^2) \) and \( Y \sim N(\mu, \sigma^2) \).

**Example 2.17.** If \( Y \) has a Topp–Leone distribution, then pdf of \( Y \) is

\begin{equation}
f(y) = \nu(2 - 2y)(2y - y^2)^{\nu-1}
\end{equation}

for \( \nu > 0 \) and \( 0 < y < 1 \). Notice that \( F(y) = (2y - y^2)^\nu \) for \( 0 < y < 1 \) since \( F'(y) = f(y) \). Then the distribution of \( W = -\log(2Y - Y^2) \) will be of interest for later chapters.

Let \( X = Y - 1 \). Then the support of \( X \) is \((-1, 0)\) and \( F_X(x) = P(X \leq x) = P(Y - 1 \leq x) = P(Y \leq x + 1) = F_Y(x + 1) \)

\begin{equation}
= (2(x+1) - (x+1)^2)\nu = ((x+1)(2-(x+1)))\nu = [(x+1)(1-x)]\nu = (1-x^2)\nu.
\end{equation}

So \( F_X(x) = (1-x^2)\nu \) for \(-1 < x < 0\). Now the support of \( W \) is \( w > 0 \) and \( F_W(w) = P(W \leq w) = P(-\log(2Y - Y^2) \leq w) = P(2Y - Y^2 \geq e^{-w}) = P(2Y - Y^2 - 1 \geq e^{-w} - 1) = P(-(Y-1)^2 \geq e^{-w} - 1) = P((Y-1)^2 \leq 1 - e^{-w}). \) So \( F_W(w) = P(X^2 \leq 1 - e^{-w}) = P(-\sqrt{a} \leq X \leq \sqrt{a}) \) where \( a = 1 - e^{-w} \in (0, 1) \). So \( F_W(w) = F_X(\sqrt{a}) - F_X(-\sqrt{a}) = 1 - F_X(-\sqrt{a}) = 1 - F_X(-\sqrt{1-e^{-w}}) \)

\begin{equation}
= 1 - [1 - (-\sqrt{1-e^{-w}})^2]^{\nu} = 1 - [1 - (1-e^{-w})]^{\nu} = 1 - e^{-w\nu}
\end{equation}

for \( w > 0 \). Thus \( W = -\log(2Y - Y^2) \sim EXP(1/\nu) \).

Transformations for vectors are often less useful in applications because the transformation formulas tend to be impractical to compute. For the theorem below, typically \( n = 2 \). If \( Y_1 = t_1(X_1, X_2) \) is of interest, choose \( Y_2 = t_2(X_1, X_2) \) such that the determinant \( J \) is easy to compute. For example, \( Y_2 = X_2 \) may work. Finding the support \( \mathcal{Y} \) can be difficult, but if the joint pdf of \( X_1, X_2 \) is \( g(x_1, x_2) = h(x_1, x_2) I[(x_1, x_2) \in \mathcal{X}] \), then the joint pdf of \( Y_1, Y_2 \) is

\begin{equation}
f(y_1, y_2) = h(t_1^{-1}(y), t_2^{-1}(y)) I[(t_1^{-1}(y), t_2^{-1}(y)) \in \mathcal{X}] |J|,
\end{equation}
and using \( I[(t_1^{-1}(y), t_2^{-1}(y)) \in X] \) can be useful for finding \( Y \). The fact that \( \prod_{j=1}^{k} I_{A_j}(y) = I_{\bigcap_{j=1}^{k} A_j}(y) \) can also be useful. See Problem 2.67. Also sketch \( X \) with \( x_1 \) on the horizontal axis and \( x_2 \) on the vertical axis, and sketch \( Y \) with \( y_1 \) on the horizontal axis and \( y_2 \) on the vertical axis.

**Theorem 2.14: the Multivariate Transformation Method.** Let \( X_1, \ldots, X_n \) be random variables with joint pdf \( g(x_1, \ldots, x_n) \) and support \( X \). Let \( Y_i = t_i(X_1, \ldots, X_n) \) for \( i = 1, \ldots, n \). Suppose that \( f(y_1, \ldots, y_n) \) is the joint pdf of \( Y_1, \ldots, Y_n \) and that the multivariate transformation is one to one. Hence the transformation is invertible and can be solved for the equations \( x_i = t_i^{-1}(y_1, \ldots, y_n) \) for \( i = 1, \ldots, n \). Then the Jacobian of this multivariate transformation is

\[
J = \det \begin{bmatrix}
\frac{\partial t_1^{-1}}{\partial y_1} & \cdots & \frac{\partial t_1^{-1}}{\partial y_n} \\
\vdots & \ddots & \vdots \\
\frac{\partial t_n^{-1}}{\partial y_1} & \cdots & \frac{\partial t_n^{-1}}{\partial y_n}
\end{bmatrix}
\]

Let \(|J|\) denote the absolute value of the determinant \( J \). Then the pdf of \( Y_1, \ldots, Y_n \) is

\[
f(y_1, \ldots, y_n) = g(t_1^{-1}(y), \ldots, t_n^{-1}(y)) \ |J|.
\]  

(2.17)

**Example 2.18.** Let \( X_1 \) and \( X_2 \) have joint pdf

\[
g(x_1, x_2) = 2e^{-(x_1+x_2)}
\]

for \( 0 < x_1 < x_2 < \infty \). Let \( Y_1 = X_1 \) and \( Y_2 = X_1 + X_2 \). An important step is finding the support \( \mathcal{Y} \) of \( (Y_1, Y_2) \) from the support of \( (X_1, X_2) \)

\[
= \mathcal{X} = \{(x_1, x_2) | 0 < x_1 < x_2 < \infty \}.
\]

Now \( x_1 = y_1 = t_1^{-1}(y_1, y_2) \) and \( x_2 = y_2 - y_1 = t_2^{-1}(y_1, y_2) \). Hence \( x_1 < x_2 \) implies \( y_1 < y_2 - y_1 \) or \( 2y_1 < y_2 \), and

\[
\mathcal{Y} = \{(y_1, y_2) | 0 < 2y_1 < y_2 \}.
\]

Now

\[
\frac{\partial t_1^{-1}}{\partial y_1} = 1, \quad \frac{\partial t_1^{-1}}{\partial y_2} = 0,
\]

\[
\frac{\partial t_2^{-1}}{\partial y_1} = -1, \quad \frac{\partial t_2^{-1}}{\partial y_2} = 1,
\]
and the Jacobian
\[ J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1. \]

Hence \(|J| = 1\). Using indicators,
\[ g_{X_1,X_2}(x_1, x_2) = 2e^{-(x_1+x_2)}I(0 < x_1 < x_2 < \infty), \]
and
\[ f_{Y_1,Y_2}(y_1, y_2) = g_{X_1,X_2}(y_1, y_2-y_1)|J| = 2e^{-(y_1+y_2-y_1)}I(0 < y_1 < y_2-y_1) = 2e^{-y_2}I(0 < 2y_1 < y_2). \]

Notice that \(Y_1\) and \(Y_2\) are not independent since the support \(Y\) is not a cross product. The marginals
\[ f_{Y_1}(y_1) = \int_{-\infty}^{\infty} 2e^{-y_2}I(0 < 2y_1 < y_2)dy_2 = \int_{2y_1}^{\infty} 2e^{-y_2}dy_2 = -2e^{-y_2} \bigg|_{y_2=2y_1}^{\infty} = 0 - 2e^{-2y_1} = 2e^{-2y_1} \]
for \(0 < y_1 < \infty\), and
\[ f_{Y_2}(y_2) = \int_{-\infty}^{\infty} 2e^{-y_2}I(0 < 2y_1 < y_2)dy_1 = \int_{0}^{y_2/2} 2e^{-y_2}dy_1 = 2e^{-y_2} \bigg|_{y_1=0}^{y_1=y_2/2} = y_2e^{-y_2} \]
for \(0 < y_2 < \infty\).

**Example 2.19.** Following Bickel and Doksum (2007, p. 489-490), let \(X_1\) and \(X_2\) be independent gamma \((\nu_i, \lambda)\) RVs for \(i = 1, 2\). Then \(X_1\) and \(X_2\) have joint pdf \(g(x_1, x_2) = g_1(x_1)g_2(x_2) =
\[
\frac{x_1^{\nu_1-1}e^{-x_1/\lambda} \cdot x_2^{\nu_2-1}e^{-x_2/\lambda}}{\lambda^{\nu_1+\nu_2}\Gamma(\nu_1)\Gamma(\nu_2)} = \frac{1}{\lambda^{\nu_1+\nu_2}\Gamma(\nu_1)\Gamma(\nu_2)} x_1^{\nu_1-1} x_2^{\nu_2-1} \exp[-(x_1 + x_2)/\lambda]
\]
for \(0 < x_1\) and \(0 < x_2\). Let \(Y_1 = X_1 + X_2\) and \(Y_2 = X_1/(X_1+X_2)\). An important step is finding the support \(Y\) of \((Y_1, Y_2)\) from the support of \((X_1, X_2)\)
\[ = \mathcal{X} = \{(x_1, x_2)|0 < x_1\ and \ 0 < x_2\}. \]
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Now \( y_2 = x_1/y_1 \), so \( x_1 = y_1 y_2 = t_1^{-1}(y_1, y_2) \) and \( x_2 = y_1 - x_1 = y_1 - y_1 y_2 = t_2^{-1}(y_1, y_2) \). Notice that \( 0 < y_1 \) and \( 0 < x_1 < x_1 + x_2 \). Thus \( 0 < y_2 < 1 \), and

\[
\mathcal{Y} = \{(y_1, y_2) | 0 < y_1 \text{ and } 0 < y_2 < 1\}.
\]

Now

\[
\frac{\partial t_1^{-1}}{\partial y_1} = y_2, \quad \frac{\partial t_1^{-1}}{\partial y_2} = y_1,
\]

\[
\frac{\partial t_2^{-1}}{\partial y_1} = 1 - y_2, \quad \frac{\partial t_2^{-1}}{\partial y_2} = -y_1,
\]

and the Jacobian

\[
J = \begin{vmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{vmatrix} = -y_1 y_2 - (y_1 - y_1 y_2) = -y_1,
\]

and \( |J| = y_1 \). So the joint pdf

\[
f(y_1, y_2) = g(t_1^{-1}(y), t_2^{-1}(y)) \ |J| = g(y_1 y_2, y_1 - y_1 y_2) y_1 =
\]

\[
\frac{1}{\lambda^{\nu_1+\nu_2} \Gamma(\nu_1) \Gamma(\nu_2)} y_1^{\nu_1-1} y_2^{\nu_2-1} (1 - y_2)^{\nu_1-1} \exp[-(y_1 y_2 + y_1 - y_1 y_2)/\lambda] y_1 =
\]

\[
\frac{1}{\lambda^{\nu_1+\nu_2} \Gamma(\nu_1) \Gamma(\nu_2)} y_1^{\nu_1+\nu_2-1} y_2^{\nu_2-1} (1 - y_2)^{\nu_1-1} e^{-y_1/\lambda} =
\]

\[
\frac{1}{\lambda^{\nu_1+\nu_2} \Gamma(\nu_1+\nu_2)} y_1^{\nu_1+\nu_2-1} e^{-y_1/\lambda} \frac{\Gamma(\nu_1+\nu_2)}{\Gamma(\nu_1) \Gamma(\nu_2)} y_2^{\nu_2-1} (1 - y_2)^{\nu_2-1}.
\]

Thus \( f(y_1, y_2) = f_1(y_1) f_2(y_2) \) on \( \mathcal{Y} \), and \( Y_1 \sim \text{gamma}(\nu_1 + \nu_2, \lambda) \) \( \| \) \( Y_2 \sim \text{beta}(\nu_1, \nu_2) \) by Theorem 2.2b.

### 2.6 Sums of Random Variables

An important multivariate transformation of the random variables \( \mathbf{Y} = (Y_1, \ldots, Y_n) \) is \( T(Y_1, \ldots, Y_n) = \sum_{i=1}^{n} Y_i \). Some properties of sums are given below.

**Theorem 2.15.** Assume that all relevant expectations exist. Let \( a, a_1, \ldots, a_n \) and \( b_1, \ldots, b_m \) be constants. Let \( Y_1, \ldots, Y_n \), and \( X_1, \ldots, X_m \) be random variables. Let \( g_1, \ldots, g_k \) be functions of \( Y_1, \ldots, Y_n \).

i) \( E(a) = a \).
ii) \( E[aY] = aE[Y] \)

iii) \( V(aY) = a^2V(Y) \).

iv) \( E[g_1(Y_1, ..., Y_n) + \cdots + g_k(Y_1, ..., Y_n)] = \sum_{i=1}^k E[g_i(Y_1, ..., Y_n)] \).

Let \( W_1 = \sum_{i=1}^n a_i Y_i \) and \( W_2 = \sum_{i=1}^m b_i X_i \).

v) \( E(W_1) = \sum_{i=1}^n a_i E(Y_i) \).

vi) \( V(W_1) = \text{Cov}(W_1, W_1) = \sum_{i=1}^n a_i^2 V(Y_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n a_i a_j \text{Cov}(Y_i, Y_j) \).

vii) \( \text{Cov}(W_1, W_2) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(Y_i, X_j) \).

viii) \( E(\sum_{i=1}^n Y_i) = \sum_{i=1}^n E(Y_i) \).

ix) If \( Y_1, ..., Y_n \) are independent, \( V(\sum_{i=1}^n Y_i) = \sum_{i=1}^n V(Y_i) \).

Let \( Y_1, ..., Y_n \) be iid RVs with \( E(Y_i) = \mu \) and \( V(Y_i) = \sigma^2 \), then the sample mean \( \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \). Then

x) \( E(\bar{Y}) = \mu \) and

xi) \( V(\bar{Y}) = \sigma^2/n \).

Definition 2.20. \( Y_1, ..., Y_n \) are a random sample or iid if \( Y_1, ..., Y_n \) are independent and identically distributed (all of the \( Y_i \) have the same distribution).

Example 2.20: Common problem. Let \( Y_1, ..., Y_n \) be independent random variables with \( E(Y_i) = \mu_i \) and \( V(Y_i) = \sigma_i^2 \). Let \( W = \sum_{i=1}^n Y_i \). Then

a) \( E(W) = E(\sum_{i=1}^n Y_i) = \sum_{i=1}^n E(Y_i) = \sum_{i=1}^n \mu_i \), and

b) \( V(W) = V(\sum_{i=1}^n Y_i) = \sum_{i=1}^n V(Y_i) = \sum_{i=1}^n \sigma_i^2 \).

A statistic is a function of the random sample and known constants. A statistic is a random variable and the sampling distribution of a statistic is the distribution of the statistic. Important statistics are \( \sum_{i=1}^n Y_i, \ \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \) and \( \sum_{i=1}^n a_i Y_i \) where \( a_1, ..., a_n \) are constants. The following theorem shows how to find the mgf and characteristic function of such statistics.

Theorem 2.16. a) The characteristic function uniquely determines the
b) If the moment generating function exists, then it uniquely determines the distribution.

c) Assume that $Y_1, \ldots, Y_n$ are independent with characteristic functions $\phi_{Y_i}(t)$. Then the characteristic function of $W = \sum_{i=1}^{n} Y_i$ is

$$\phi_W(t) = \prod_{i=1}^{n} \phi_{Y_i}(t).$$  \hspace{1cm} (2.18)


d) Assume that $Y_1, \ldots, Y_n$ are iid with characteristic functions $\phi_{Y}(t)$. Then the characteristic function of $W = \sum_{i=1}^{n} Y_i$ is

$$\phi_W(t) = [\phi_{Y}(t)]^n.$$  \hspace{1cm} (2.19)

e) Assume that $Y_1, \ldots, Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^{n} Y_i$ is

$$m_W(t) = \prod_{i=1}^{n} m_{Y_i}(t).$$  \hspace{1cm} (2.20)

f) Assume that $Y_1, \ldots, Y_n$ are iid with mgf $m_{Y}(t)$. Then the mgf of $W = \sum_{i=1}^{n} Y_i$ is

$$m_W(t) = [m_{Y}(t)]^n.$$  \hspace{1cm} (2.21)

g) Assume that $Y_1, \ldots, Y_n$ are independent with characteristic functions $\phi_{Y_i}(t)$. Then the characteristic function of $W = \sum_{j=1}^{n} (a_j + b_j Y_j)$ is

$$\phi_W(t) = \exp(it \sum_{j=1}^{n} a_j) \prod_{j=1}^{n} \phi_{Y_j}(b_j t).$$  \hspace{1cm} (2.22)

h) Assume that $Y_1, \ldots, Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^{n} (a_i + b_i Y_i)$ is

$$m_W(t) = \exp(t \sum_{i=1}^{n} a_i) \prod_{i=1}^{n} m_{Y_i}(b_i t).$$  \hspace{1cm} (2.23)

Proof of g): Recall that $\exp(w) = e^w$ and $\exp(\sum_{j=1}^{n} d_j) = \prod_{j=1}^{n} \exp(d_j)$. It can be shown that for the purposes of this proof, that the complex constant
$i$ in the characteristic function (cf) can be treated in the same way as if it were a real constant. Now

$$
\phi_W(t) = E(e^{itW}) = E(\exp[it \sum_{j=1}^{n} (a_j + b_j Y_j)])
$$

$$
= \exp(it \sum_{j=1}^{n} a_j) E(\exp[\sum_{j=1}^{n} itb_j Y_j])
$$

$$
= \exp(it \sum_{j=1}^{n} a_j) E(\prod_{i=1}^{n} \exp[itb_j Y_j])
$$

$$
= \exp(it \sum_{j=1}^{n} a_j) \prod_{i=1}^{n} E[\exp(itb_j Y_j)]
$$

since by Theorem 2.5 the expected value of a product of independent random variables is the product of the expected values of the independent random variables. Now in the definition of a cf, the $t$ is a dummy variable as long as $t$ is real. Hence $\phi_Y(t) = E[\exp(itY)]$ and $\phi_Y(s) = E[\exp(isY)]$. Taking $s = tb_j$ gives $E[\exp(itb_j Y_j)] = \phi_{Y_j}(tb_j)$. Thus

$$
\phi_W(t) = \exp(it \sum_{j=1}^{n} a_j) \prod_{i=1}^{n} \phi_{Y_j}(tb_j). \ \text{QED}
$$

The distribution of $W = \sum_{i=1}^{n} Y_i$ is known as the convolution of $Y_1, ..., Y_n$. Even for $n = 2$ convolution formulas tend to be hard; however, the following two theorems suggest that to find the distribution of $W = \sum_{i=1}^{n} Y_i$, first find the mgf or characteristic function of $W$ using Theorem 2.16. If the mgf or cf is that of a brand name distribution, then $W$ has that distribution. For example, if the mgf of $W$ is a normal $(\nu, \tau^2)$ mgf, then $W$ has a normal $(\nu, \tau^2)$ distribution, written $W \sim N(\nu, \tau^2)$. This technique is useful for several brand name distributions. Chapter 10 will show that many of these distributions are exponential families.

**Theorem 2.17.** a) If $Y_1, ..., Y_n$ are independent binomial $\text{BIN}(k_i, \rho)$ random variables, then

$$
\sum_{i=1}^{n} Y_i \sim \text{BIN}(\sum_{i=1}^{n} k_i, \rho).
$$
Thus if $Y_1, ..., Y_n$ are iid BIN($k, \rho$) random variables, then $\sum_{i=1}^{n} Y_i \sim$ BIN(nk, $\rho$).

b) Denote a chi–square $\chi^2_p$ random variable by $\chi^2_p$. If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then
\[
\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).
\]

Thus if $Y_1, ..., Y_n$ are iid $\chi^2_p$, then
\[
\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.
\]

c) If $Y_1, ..., Y_n$ are iid exponential EXP($\lambda$), then
\[
\sum_{i=1}^{n} Y_i \sim G(n, \lambda).
\]

d) If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then
\[
\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).
\]

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then
\[
\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda).
\]

e) If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then
\[
\sum_{i=1}^{n} (a_i + b_i Y_i) \sim N(\sum_{i=1}^{n} (a_i + b_i \mu_i), \sum_{i=1}^{n} b_i^2 \sigma_i^2).
\]

Here $a_i$ and $b_i$ are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

f) If $Y_1, ..., Y_n$ are independent Poisson POIS($\theta_i$), then
\[
\sum_{i=1}^{n} Y_i \sim POIS(\sum_{i=1}^{n} \theta_i).
\]
Thus if \( Y_1, \ldots, Y_n \) are iid \( \text{POIS}(\theta) \), then

\[
\sum_{i=1}^{n} Y_i \sim \text{POIS}(n\theta).
\]

**Theorem 2.18.** a) If \( Y_1, \ldots, Y_n \) are independent Cauchy \( C(\mu_i, \sigma_i) \), then

\[
\sum_{i=1}^{n} (a_i + b_i Y_i) \sim C\left(\sum_{i=1}^{n} (a_i + b_i \mu_i), \sum_{i=1}^{n} |b_i| \sigma_i\right).
\]

Thus if \( Y_1, \ldots, Y_n \) are iid \( C(\mu, \sigma) \), then \( \overline{Y} \sim C(\mu, \sigma) \).

b) If \( Y_1, \ldots, Y_n \) are iid geometric \( \text{geom}(p) \), then

\[
\sum_{i=1}^{n} Y_i \sim \text{NB}(n, p).
\]

c) If \( Y_1, \ldots, Y_n \) are iid inverse Gaussian \( \text{IG}(\theta, \lambda) \), then

\[
\sum_{i=1}^{n} Y_i \sim \text{IG}(n\theta, n^2\lambda).
\]

Also

\[
\overline{Y} \sim \text{IG}(\theta, n\lambda).
\]

d) If \( Y_1, \ldots, Y_n \) are independent negative binomial \( \text{NB}(r_i, \rho) \), then

\[
\sum_{i=1}^{n} Y_i \sim \text{NB}\left(\sum_{i=1}^{n} r_i, \rho\right).
\]

Thus if \( Y_1, \ldots, Y_n \) are iid \( \text{NB}(r, \rho) \), then

\[
\sum_{i=1}^{n} Y_i \sim \text{NB}(nr, \rho).
\]

**Example 2.21: Common problem.** Given that \( Y_1, \ldots, Y_n \) are independent random variables from one of the distributions in Theorem 2.17, find the distribution of \( W = \sum_{i=1}^{n} Y_i \) or \( W = \sum_{i=1}^{n} b_i Y_i \) by finding the mgf or characteristic function of \( W \) and recognizing that it comes from a brand name distribution.
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

Tips: a) in the product, anything that does not depend on the product index \( i \) is treated as a constant.

b) \( \exp(a) = e^a \) and \( \log(y) = \ln(y) = \log_e(y) \) is the natural logarithm.

c) 
\[
\prod_{i=1}^{n} a^{b \theta_i} = a^{\sum_{i=1}^{n} b \theta_i} = a^{b \sum_{i=1}^{n} \theta_i}.
\]

In particular, 
\[
\prod_{i=1}^{n} \exp(b \theta_i) = \exp(\sum_{i=1}^{n} b \theta_i) = \exp(b \sum_{i=1}^{n} \theta_i).
\]

Example 2.22. Suppose \( Y_1, ..., Y_n \) are iid \( IG(\theta, \lambda) \) where the mgf
\[
m_{Y_i}(t) = m(t) = \exp \left[ \frac{\lambda}{\theta} \left( 1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}} \right) \right]
\]
for \( t < \lambda/(2\theta^2) \). Then
\[
m_{\sum_{i=1}^{n} Y_i}(t) = \prod_{i=1}^{n} m_{Y_i}(t) = [m(t)]^n = \exp \left[ \frac{n\lambda}{\theta} \left( 1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}} \right) \right]
\]
\[
= \exp \left[ \frac{n^2 \lambda}{n \theta} \left( 1 - \sqrt{1 - \frac{2(n\theta)^2 t}{n^2 \lambda}} \right) \right]
\]
which is the mgf of an \( IG(n\theta, n^2\lambda) \) RV. The last equality was obtained by multiplying \( \frac{n\lambda}{\theta} \) by \( 1 = n/n \) and by multiplying \( \frac{2\theta^2 t}{\lambda} \) by \( 1 = n^2/n^2 \). Hence \( \sum_{i=1}^{n} Y_i \sim IG(n\theta, n^2\lambda) \).

2.7 Random Vectors

Definition 2.21. \( Y = (Y_1, ..., Y_p) \) is a \( 1 \times p \) random vector if \( Y_i \) is a random variable for \( i = 1, ..., p \). \( Y \) is a discrete random vector if each \( Y_i \) is discrete, and \( Y \) is a continuous random vector if each \( Y_i \) is continuous. A random variable \( Y_1 \) is the special case of a random vector with \( p = 1 \).

In the previous sections each \( Y = (Y_1, ..., Y_n) \) was a random vector. In this section we will consider \( n \) random vectors \( Y_1, ..., Y_n \). Often double subscripts will be used: \( Y_i = (Y_{i,1}, ..., Y_{i,p_i}) \) for \( i = 1, ..., n \).
Notation. The notation for random vectors is rather awkward. In most of the statistical inference literature, $Y$ is a row vector, but in most of the multivariate analysis literature $Y$ is a column vector. In this text, if $X$ and $Y$ are both vectors, a phrase with $Y$ and $X^T$ means that $Y$ is a column vector and $X^T$ is a row vector where $T$ stands for transpose. Hence in the definition below, first $E(Y)$ is a $p \times 1$ row vector, but in the definition of Cov($Y$) below, $E(Y)$ and $Y - E(Y)$ are $p \times 1$ column vectors and $(Y - E(Y))^T$ is a $1 \times p$ row vector.

Definition 2.22. The population mean or expected value of a random $1 \times p$ random vector $(Y_1, \ldots, Y_p)$ is

$$E(Y) = (E(Y_1), \ldots, E(Y_p))$$

provided that $E(Y_i)$ exists for $i = 1, \ldots, p$. Otherwise the expected value does not exist. Now let $Y$ be a $p \times 1$ column vector. The $p \times p$ population covariance matrix

$$\text{Cov}(Y) = E(Y - E(Y))(Y - E(Y))^T = ((\sigma_{i,j}))$$

where the $ij$ entry of Cov($Y$) is Cov($Y_i, Y_j$) = $\sigma_{i,j}$ provided that each $\sigma_{i,j}$ exists. Otherwise Cov($Y$) does not exist.

The covariance matrix is also called the variance–covariance matrix and variance matrix. Sometimes the notation Var($Y$) is used. Note that Cov($Y$) is a symmetric positive semidefinite matrix. If $X$ and $Y$ are $p \times 1$ random vectors, $a$ a conformable constant vector and $A$ and $B$ are conformable constant matrices, then

$$E(a + X) = a + E(X) \quad \text{and} \quad E(X + Y) = E(X) + E(Y) \quad (2.24)$$

and

$$E(AX) = AE(X) \quad \text{and} \quad EAXB = AE(X)B. \quad (2.25)$$

Thus

$$\text{Cov}(a + AX) = \text{Cov}(AX) = ACov(X)A^T. \quad (2.26)$$

Definition 2.23. Let $Y_1, \ldots, Y_n$ be random vectors with joint pdf or pmf $f(y_1, \ldots, y_n)$. Let $f_{Y_1}(y_i)$ be the marginal pdf or pmf of $Y_i$. Then $Y_1, \ldots, Y_n$
are independent random vectors if
\[ f(y_1, \ldots, y_n) = f_{Y_1}(y_1) \cdots f_{Y_n}(y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i). \]

The following theorem is a useful generalization of Theorem 2.7.

**Theorem 2.19.** Let \( Y_1, \ldots, Y_n \) be independent random vectors where \( Y_i \) is a \( 1 \times p_i \) vector for \( i = 1, \ldots, n \). and let \( h_i : \mathbb{R}^{p_i} \to \mathbb{R}^{p_{ji}} \) be vector valued functions and suppose that \( h_i(y_i) \) is a function of \( y_i \) alone for \( i = 1, \ldots, n \). Then the random vectors \( X_i = h_i(Y_i) \) are independent. There are three important special cases.

i) If \( p_{ji} = 1 \) so that each \( h_i \) is a real valued function, then the random variables \( X_i = h_i(Y_i) \) are independent.

ii) If \( p_i = p_{ji} = 1 \) so that each \( Y_i \) and each \( X_i = h(Y_i) \) are random variables, then \( X_1, \ldots, X_n \) are independent.

iii) Let \( Y = (Y_1, \ldots, Y_n) \) and \( X = (X_1, \ldots, X_m) \) and assume that \( Y \perp X \). If \( h(Y) \) is a vector valued function of \( Y \) alone and if \( g(X) \) is a vector valued function of \( X \) alone, then \( h(Y) \) and \( g(X) \) are independent random vectors.

**Definition 2.24.** The characteristic function (cf) of a random vector \( Y \) is
\[ \phi_Y(t) = E(e^{it^TY}) \]
\( \forall t \in \mathbb{R}^n \) where the complex number \( i = \sqrt{-1} \).

**Definition 2.25.** The moment generating function (mgf) of a random vector \( Y \) is
\[ m_Y(t) = E(e^{t^TY}) \]
provided that the expectation exists for all \( t \) in some neighborhood of the origin 0.

**Theorem 2.20.** If \( Y_1, \ldots, Y_n \) have mgf \( m(t) \), then moments of all orders exist and
\[ E(Y_{i_1}^{k_1} \cdots Y_{i_j}^{k_j}) = \frac{\partial^{k_1+\cdots+k_j}}{\partial t_{i_1}^{k_1} \cdots \partial t_{i_j}^{k_j}} m(t) \bigg|_{t=0}. \]
In particular,
\[ E(Y_i) = \frac{\partial m(t)}{\partial t_i} \bigg|_{t=0} \]
and

\[ E(Y_i Y_j) = \frac{\partial^2 m(t)}{\partial t_i \partial t_j} \bigg|_{t=0}. \]

**Theorem 2.21.** If \( Y_1, \ldots, Y_n \) have a cf \( \phi_Y(t) \) and mgf \( m_Y(t) \) then the marginal cf and mgf for \( Y_{i_1}, \ldots, Y_{i_k} \) are found from the joint cf and mgf by replacing \( t_{i_j} \) by 0 for \( j = k + 1, \ldots, n \). In particular, if \( Y = (Y_1, Y_2) \) and \( t = (t_1, t_2) \), then

\[ \phi_{Y_1}(t_1) = \phi_Y(t_1, 0) \text{ and } m_{Y_1}(t_1) = m_Y(t_1, 0). \]

**Proof.** Use the definition of the cf and mgf. For example, if \( Y_1 = (Y_1, \ldots, Y_k) \) and \( s = t_1 \), then \( m(t_1, 0) = \)

\[ E[\exp(t_1 Y_1 + \cdots + t_k Y_k + 0 Y_{k+1} + \cdots + 0 Y_n)] = E[\exp(t_1 Y_1 + \cdots + t_k Y_k)] = E[\exp(s^T Y_1)] = m_Y(s), \text{ which is the mgf of } Y_1. \]

**QED**

**Theorem 2.22.** Partition the \( 1 \times n \) vectors \( Y \) and \( t \) as \( Y = (Y_1, Y_2) \) and \( t = (t_1, t_2) \). Then the random vectors \( Y_1 \) and \( Y_2 \) are independent iff their joint cf factors into the product of their marginal cfs:

\[ \phi_Y(t) = \phi_{Y_1}(t_1) \phi_{Y_2}(t_2) \quad \forall t \in \mathbb{R}^n. \]

If the joint mgf exists, then the random vectors \( Y_1 \) and \( Y_2 \) are independent iff their joint mgf factors into the product of their marginal mgfs:

\[ m_Y(t) = m_{Y_1}(t_1) m_{Y_2}(t_2) \]

\( \forall t \) in some neighborhood of \( 0 \).

### 2.8 The Multinomial Distribution

**Definition 2.26.** Assume that there are \( m \) iid trials with \( n \) outcomes. Let \( Y_i \) be the number of the \( m \) trials that resulted in the \( i \)th outcome and let \( \rho_i \) be the probability of the \( i \)th outcome for \( i = 1, \ldots, n \) where \( 0 \leq \rho_i \leq 1 \). Thus \( \sum_{i=1}^{n} Y_i = m \) and \( \sum_{i=1}^{n} \rho_i = 1 \). Then \( Y = (Y_1, \ldots, Y_n) \) has a multinomial
The support of $Y$ is $\mathcal{Y} = \{ \mathbf{y} : \sum_{i=1}^{n} y_i = m \text{ and } 0 \leq y_i \leq m \text{ for } i = 1, \ldots, n \}$. 

The multinomial theorem states that for real $x_i$ and positive integers $m$ and $n$,

\[(x_1 + \cdots + x_n)^m = \sum_{\mathbf{y} \in \mathcal{Y}} \frac{m!}{y_1! \cdots y_n!} x_1^{y_1} x_2^{y_2} \cdots x_n^{y_n}. \tag{2.28}\]

Taking $x_i = \rho_i$ shows that (2.27) is a pmf.

Since $Y_n$ and $\rho_n$ are known if $Y_1, \ldots, Y_{n-1}$ and $\rho_1, \ldots, \rho_{n-1}$ are known, it is convenient to act as if $n - 1$ of the outcomes $Y_1, \ldots, Y_{n-1}$ are important and the $n$th outcome means that none of the $n - 1$ important outcomes occurred. With this reasoning, suppose that $\{i_1, \ldots, i_{k-1}\} \subset \{1, \ldots, n\}$. Let $W_j = Y_{i_j}$, and let $W_k$ count the number of times that none of $Y_{i_1}, \ldots, Y_{i_{k-1}}$ occurred. Then $W_k = m - \sum_{j=1}^{k-1} Y_{i_j}$ and $P(W_k) = 1 - \sum_{j=1}^{k-1} \rho_{i_j}$. Here $W_k$ represents the unimportant outcomes and the joint distribution of $W_1, \ldots, W_{k-1}, W_k$ is multinomial $M_k(m, \rho_{i_1}, \ldots, \rho_{i_{k-1}}, 1 - \sum_{j=1}^{k-1} \rho_{i_j})$.

Notice that $\sum_{j=1}^{k} Y_{i_j}$ counts the number of times that the outcome “one of the outcomes $i_1, \ldots, i_k$ occurred,” an outcome with probability $\sum_{j=1}^{k} \rho_{i_j}$. Hence $\sum_{j=1}^{k} Y_{i_j} \sim \text{BIN}(m, \sum_{j=1}^{k} \rho_{i_j})$.

Now consider conditional distributions. If it is known that $Y_{i_j} = y_{i_j}$ for $j = k + 1, \ldots, n$, then there are $m - \sum_{j=k+1}^{n} y_{i_j}$ outcomes left to distribute among $Y_{i_1}, \ldots, Y_{i_k}$. The conditional probabilities of $Y_i$ remains proportional to $\rho_i$, but the conditional probabilities must sum to one. Hence the conditional distribution is again multinomial. These results prove the following theorem.

**Theorem 2.23.** Assume that $(Y_1, \ldots, Y_n)$ has an $M_n(m, \rho_1, \ldots, \rho_n)$ distribution and that $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ with $k < n$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq n$.

a) $(Y_{i_1}, \ldots, Y_{i_{k-1}}, m - \sum_{j=1}^{k-1} Y_{i_j})$ has an $M_k(m, \rho_{i_1}, \ldots, \rho_{i_{k-1}}, 1 - \sum_{j=1}^{k-1} \rho_{i_j})$ distribution.
2.9 The Multivariate Normal Distribution

**Definition 2.27:** Rao (1965, p. 437). A $p \times 1$ random vector $X$ has a $p$—dimensional multivariate normal distribution $N_p(\mu, \Sigma)$ iff $t^TX$ has a univariate normal distribution for any $p \times 1$ vector $t$. 
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

If $\Sigma$ is positive definite, then $X$ has a joint pdf

$$f(z) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}e^{-(1/2)(z-\mu)^T\Sigma^{-1}(z-\mu)} \quad (2.30)$$

where $|\Sigma|^{1/2}$ is the square root of the determinant of $\Sigma$. Note that if $p = 1$, then the quadratic form in the exponent is $(z - \mu)(\sigma^2)^{-1}(z - \mu)$ and $X$ has the univariate $N(\mu, \sigma^2)$ pdf. If $\Sigma$ is positive semidefinite but not positive definite, then $X$ has a degenerate distribution. For example, the univariate $N(0, 0^2)$ distribution is degenerate (the point mass at 0).

Some important properties of MVN distributions are given in the following three propositions. These propositions can be proved using results from Johnson and Wichern (1988, p. 127-132).

**Proposition 2.25.** a) If $X \sim N_p(\mu, \Sigma)$, then $E(X) = \mu$ and $\text{Cov}(X) = \Sigma$.

b) If $X \sim N_p(\mu, \Sigma)$, then any linear combination $t^T X = t_1 X_1 + \cdots + t_p X_p \sim N_1(t^T \mu, t^T \Sigma t)$. Conversely, if $t^T X \sim N_1(t^T \mu, t^T \Sigma t)$ for every $p \times 1$ vector $t$, then $X \sim N_p(\mu, \Sigma)$.

c) The joint distribution of independent normal random variables is MVN. If $X_1, ..., X_p$ are independent univariate normal $N(\mu_i, \sigma^2_i)$ random vectors, then $X = (X_1, ..., X_p)^T$ is $N_p(\mu, \Sigma)$ where $\mu = (\mu_1, ..., \mu_p)$ and $\Sigma = \text{diag}(\sigma^2_1, ..., \sigma^2_p)$ (so the off diagonal entries $\sigma_{i,j} = 0$ while the diagonal entries of $\Sigma$ are $\sigma^2_i$.)

d) If $X \sim N_p(\mu, \Sigma)$ and if $A$ is a $q \times p$ matrix, then $AX \sim N_q(A\mu, A\Sigma A^T)$. If $a$ is a $p \times 1$ vector of constants, then $a + X \sim N_p(a + \mu, \Sigma)$.

It will be useful to partition $X$, $\mu$, and $\Sigma$. Let $X_1$ and $\mu_1$ be $q \times 1$ vectors, let $X_2$ and $\mu_2$ be $(p - q) \times 1$ vectors, let $\Sigma_{11}$ be a $q \times q$ matrix, let $\Sigma_{12}$ be a $q \times (p - q)$ matrix, let $\Sigma_{21}$ be a $(p - q) \times q$ matrix, and let $\Sigma_{22}$ be a $(p - q) \times (p - q)$ matrix. Then

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

**Proposition 2.26.** a) All subsets of a MVN are MVN: $(X_{k_1}, ..., X_{k_q})^T \sim N_q(\mu, \Sigma)$ where $\mu = E(X_{k_i})$ and $\Sigma_{ij} = \text{Cov}(X_{k_i}, X_{k_j})$. In particular, $X_1 \sim N_q(\mu_1, \Sigma_{11})$ and $X_2 \sim N_{p-q}(\mu_2, \Sigma_{22})$. 
b) If $X_1$ and $X_2$ are independent, then $\text{Cov}(X_1, X_2) = \Sigma_{12} = E[(X_1 - E(X_1))(X_2 - E(X_2))^T] = 0$, a $q \times (p - q)$ matrix of zeroes.

c) If $X \sim N_p(\mu, \Sigma)$, then $X_1$ and $X_2$ are independent iff $\Sigma_{12} = 0$.

d) If $X_1 \sim N_q(\mu_1, \Sigma_{11})$ and $X_2 \sim N_{p-q}(\mu_2, \Sigma_{22})$ are independent, then

$$
\begin{pmatrix}
X_1 \\
X_2
\end{pmatrix} \sim N_p \left( \begin{pmatrix} \mu_1 \\
\mu_2
\end{pmatrix}, \begin{pmatrix}
\Sigma_{11} & 0 \\
0 & \Sigma_{22}
\end{pmatrix} \right).
$$

Proposition 2.27. The conditional distribution of a MVN is MVN. If $X \sim N_p(\mu, \Sigma)$, then the conditional distribution of $X_1$ given that $X_2 = x_2$ is multivariate normal with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. That is,

$$
X_1 | X_2 = x_2 \sim N_q(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).
$$

Example 2.23. Let $p = 2$ and let $(Y, X)^T$ have a bivariate normal distribution. That is,

$$
\begin{pmatrix}
Y \\
X
\end{pmatrix} \sim N_2 \left( \begin{pmatrix}
\mu_Y \\
\mu_X
\end{pmatrix}, \begin{pmatrix}
\sigma_Y^2 & \text{Cov}(Y, X) \\
\text{Cov}(X, Y) & \sigma_X^2
\end{pmatrix} \right)
$$

Also recall that the population correlation between $X$ and $Y$ is given by

$$
\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{VAR}(X)}\sqrt{\text{VAR}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}
$$

if $\sigma_X > 0$ and $\sigma_Y > 0$. Then $Y|X = x \sim N(E(Y|X = x), \text{VAR}(Y|X = x))$ where the conditional mean

$$
E(Y|X = x) = \mu_Y + \text{Cov}(Y, X)\frac{1}{\sigma_X^2}(x - \mu_x) = \mu_Y + \rho(X, Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}(x - \mu_X)
$$

and the conditional variance

$$
\text{VAR}(Y|X = x) = \sigma_Y^2 - \text{Cov}(X, Y)\frac{1}{\sigma_X^2}\text{Cov}(X, Y)
$$

$$
= \sigma_Y^2 - \rho(X, Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}\rho(X, Y)\sqrt{\frac{\sigma_Y^2}{\sigma_X^2}}.
$$
\[ = \sigma_Y^2 - \rho^2(X, Y)\sigma_Y^2 = \sigma_Y^2[1 - \rho^2(X, Y)]. \]

Also \( aX + bY \) is univariate normal with mean \( a\mu_X + b\mu_Y \) and variance
\[ a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab \text{Cov}(X, Y). \]

Remark 2.1. There are several common misconceptions. First, it is not true that every linear combination \( t^TX \) of normal random variables is a normal random variable, and it is not true that all uncorrelated normal random variables are independent. The key condition in Proposition 2.25b and Proposition 2.26c is that the joint distribution of \( X \) is MVN. It is possible that \( X_1, X_2, \ldots, X_p \) each has a marginal distribution that is univariate normal, but the joint distribution of \( X \) is not MVN. Examine the following example from Rohatgi (1976, p. 229). Suppose that the joint pdf of \( X \) and \( Y \) is a mixture of two bivariate normal distributions both with \( EX = EY = 0 \) and \( \text{VAR}(X) = \text{VAR}(Y) = 1 \), but \( \text{Cov}(X, Y) = \pm \rho \).

Hence
\[
\begin{align*}
\frac{1}{2} \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( \frac{-1}{2(1 - \rho^2)} \left( x^2 - 2\rho xy + y^2 \right) \right) &+ \\
\frac{1}{2} \frac{1}{2\pi \sqrt{1 - \rho^2}} \exp \left( \frac{-1}{2(1 - \rho^2)} \left( x^2 + 2\rho xy + y^2 \right) \right) &\equiv \frac{1}{2}f_1(x, y) + \frac{1}{2}f_2(x, y)
\end{align*}
\]
where \( x \) and \( y \) are real and \( 0 < \rho < 1 \). Since both marginal distributions of \( f_i(x, y) \) are N(0,1) for \( i = 1 \) and 2 by Proposition 2.26a, the marginal distributions of \( X \) and \( Y \) are N(0,1). Since \( \int \int xyf_i(x, y)dxdy = \rho \) for \( i = 1 \) and \( -\rho \) for \( i = 2 \), \( X \) and \( Y \) are uncorrelated, but \( X \) and \( Y \) are not independent since \( f(x, y) \neq f_X(x)f_Y(y) \).

Remark 2.2. In Proposition 2.27, suppose that \( X = (Y, X_2, \ldots, X_p)^T \). Let \( X_1 = Y \) and \( X_2 = (X_2, \ldots, X_p)^T \). Then \( E[Y|X_2] = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p \) and \( \text{VAR}[Y|X_2] \) is a constant that does not depend on \( X_2 \). Hence \( Y|X_2 = \beta_1 + \beta_2X_2 + \cdots + \beta_pX_p + e \) follows the multiple linear regression model.

### 2.10 Elliptically Contoured Distributions

Definition 2.28: Johnson (1987, p. 107-108). A \( p \times 1 \) random vector has an elliptically contoured distribution, also called an elliptically symmetric distribution, if \( X \) has joint pdf
\[
f(z) = k_p|\Sigma|^{-1/2}g[(z - \mu)^T \Sigma^{-1}(z - \mu)], \tag{2.31}
\]
and we say $X$ has an elliptically contoured $EC_p(\mu, \Sigma, g)$ distribution.

If $X$ has an elliptically contoured (EC) distribution, then the characteristic function of $X$ is

$$
\phi_X(t) = \exp(it^T\mu)\psi(t^T\Sigma t)
$$

for some function $\psi$. If the second moments exist, then

$$
E(X) = \mu \quad (2.33)
$$

and

$$
\text{Cov}(X) = c_X \Sigma \quad (2.34)
$$

where

$$
c_X = -2\psi'(0).
$$

**Definition 2.29.** The population squared Mahalanobis distance

$$
U \equiv D^2 = D^2(\mu, \Sigma) = (X - \mu)^T\Sigma^{-1}(X - \mu) \quad (2.35)
$$

has pdf

$$
h(u) = \frac{\pi^{p/2}}{\Gamma(p/2)}k_p u^{p/2-1}g(u). \quad (2.36)
$$

For $c > 0$, an $EC_p(\mu, cI, g)$ distribution is spherical about $\mu$ where $I$ is the $p \times p$ identity matrix. The multivariate normal distribution $N_p(\mu, \Sigma)$ has $k_p = (2\pi)^{-p/2}$, $\psi(u) = g(u) = \exp(-u/2)$, and $h(u)$ is the $\chi^2_p$ density.

The following lemma is useful for proving properties of EC distributions without using the characteristic function (2.32). See Eaton (1986) and Cook (1998, p. 57, 130).

**Lemma 2.28.** Let $X$ be a $p \times 1$ random vector with 1st moments; ie, $E(X)$ exists. Let $B$ be any constant full rank $p \times r$ matrix where $1 \leq r \leq p$. Then $X$ is elliptically contoured iff for all such conforming matrices $B$,

$$
E(X|B^T X) = \mu + M_B B^T (X - \mu) = a_B + M_B B^T X \quad (2.37)
$$

where the $p \times 1$ constant vector $a_B$ and the $p \times r$ constant matrix $M_B$ both depend on $B$.  

**Chapter 2. Multivariate Distributions**
A useful fact is that $a_B$ and $M_B$ do not depend on $g$:

$$a_B = \mu - M_B B^T \mu = (I_p - M_B B^T)\mu,$$

and

$$M_B = \Sigma B (B^T \Sigma B)^{-1}.$$\[Notice that in the formula for $M_B$, $\Sigma$ can be replaced by $c\Sigma$ where $c > 0$ is a constant. In particular, if the EC distribution has second moments, Cov($X$) can be used instead of $\Sigma$.\]

To use Lemma 2.28 to prove interesting properties, partition $X$, $\mu$, and $\Sigma$. Let $X_1$ and $\mu_1$ be $q \times 1$ vectors, let $X_2$ and $\mu_2$ be $(p-q) \times 1$ vectors. Let $\Sigma_{11}$ be a $q \times q$ matrix, let $\Sigma_{12}$ be a $q \times (p-q)$ matrix, let $\Sigma_{21}$ be a $(p-q) \times q$ matrix, and let $\Sigma_{22}$ be a $(p-q) \times (p-q)$ matrix. Then

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$\[Also assume that the $(p+1) \times 1$ vector $(Y, X^T)^T$ is $EC_{p+1}((\mu, \Sigma, g)$ where $Y$ is a random variable, $X$ is a $p \times 1$ vector, and use

$$\begin{pmatrix} Y \\ X \end{pmatrix}, \mu = \begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}.$$\[Proposition 2.29. Let $X \sim EC_p(\mu, \Sigma, g)$ and assume that $E(X)$ exists.\]

a) Any subset of $X$ is EC, in particular $X_1$ is EC.

b) (Cook 1998 p. 131, Kelker 1970). If Cov($X$) is nonsingular,

$$Cov(X|B^T X) = d_g(B^T X)\Sigma - \Sigma B (B^T \Sigma B)^{-1}B^T \Sigma$$

where the real valued function $d_g(B^T X)$ is constant iff $X$ is MVN.

Proof of a). Let $A$ be an arbitrary full rank $q \times r$ matrix where $1 \leq r \leq q$. Let

$$B = \begin{pmatrix} A \\ 0 \end{pmatrix}.$$\[Then $B^T X = A^T X_1$, and

$$E[X|B^T X] = E[X]$$
by Lemma 2.28. Hence $E[X_1|A^TX_1] = \mu_1 + M_{1B}A^T(X_1 - \mu_1)$. Since $A$ was arbitrary, $X_1$ is EC by Lemma 2.28. Notice that $M_B = \Sigma B(B^T \Sigma B)^{-1} =$

$$
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\begin{pmatrix}
A \\
0
\end{pmatrix}

\left[
\begin{pmatrix}
\Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & \Sigma_{22}
\end{pmatrix}
\begin{pmatrix}
A \\
0
\end{pmatrix}
\right]^{-1}
$$

Hence

$$M_{1B} = \Sigma_{11}A(A^T \Sigma_{11}A)^{-1}$$

and $X_1$ is EC with location and dispersion parameters $\mu_1$ and $\Sigma_{11}$. QED

**Proposition 2.30.** Let $(Y, X^T)^T$ be $EC_{p+1}(\mu, \Sigma, g)$ where $Y$ is a random variable.

a) Assume that $E[(Y, X^T)^T]$ exists. Then $E(Y|X) = \alpha + \beta^T X$ where

$$\alpha = \mu_Y - \beta^T \mu_X$$

and

$$\beta = \Sigma_{XX}^{-1} \Sigma_{XY}.$$

b) Even if the first moment does not exist, the conditional median

$$\text{MED}(Y|X) = \alpha + \beta^T X$$

where $\alpha$ and $\beta$ are given in a).

**Proof.** a) The trick is to choose $B$ so that Lemma 2.28 applies. Let

$$B = \begin{pmatrix} 0^T \\ I_p \end{pmatrix}.$$

Then $B^T \Sigma B = \Sigma_{XX}$ and

$$\Sigma B = \begin{pmatrix} \Sigma_{YY} \\ \Sigma_{XY} \end{pmatrix}.$$
\[ = \mu + \Sigma B(B^T \Sigma B)^{-1}B^T \left( Y - \mu_Y \right) \]

by Lemma 2.28. The right hand side of the last equation is equal to

\[ \mu + \left( \frac{\Sigma_{YX}}{\Sigma_{XX}} \right) \Sigma_{XX}^{-1}(X - \mu_X) = \left( \frac{\mu_Y - \Sigma_{YX} \Sigma_{XX}^{-1} \mu_X + \Sigma_{YX} \Sigma_{XX}^{-1} X}{X} \right) \]

and the result follows since

\[ \beta^T = \Sigma_{YX} \Sigma_{XX}^{-1}. \]

b) See Croux, Dehon, Rousseeuw and Van Aelst (2001) for references.

Example 2.24. This example illustrates another application of Lemma 2.28. Suppose that \( X \) comes from a mixture of two multivariate normals with the same mean and proportional covariance matrices. That is, let

\[ X \sim (1 - \gamma)N_p(\mu, \Sigma) + \gamma N_p(\mu, c\Sigma) \]

where \( c > 0 \) and \( 0 < \gamma < 1 \). Since the multivariate normal distribution is elliptically contoured (and see Proposition 1.14c),

\[ E(X|B^TX) = (1 - \gamma)[\mu + M_1B^T(X - \mu)] + \gamma[\mu + M_2B^T(X - \mu)] \]

\[ = \mu + [(1 - \gamma)M_1 + \gamma M_2]B^T(X - \mu) \equiv \mu + MB^T(X - \mu). \]

Since \( M_B \) only depends on \( B \) and \( \Sigma \), it follows that \( M_1 = M_2 = M = M_B \). Hence \( X \) has an elliptically contoured distribution by Lemma 2.28.

2.11 Summary

1) \( Y_1 \) and \( Y_2 \) are dependent if the support \( \mathcal{Y} = \{(y_1, y_2)|f(y_1, y_2) > 0\} \) is not a cross product.

2) If the support is a cross product, then \( Y_1 \) and \( Y_2 \) are independent iff \( f(y_1, y_2) = h_1(y_1)h_2(y_2) \) for all \( (y_1, y_2) \in \mathcal{Y} \) where \( h_i(y_i) \) is a positive function of \( y_i \) alone. If no such factorization exists, then \( Y_1 \) and \( Y_2 \) are dependent.

3) If \( Y_1, ..., Y_n \) are independent, then the functions \( h_1(Y_1), ..., h_n(Y_n) \) are independent.

4) Given \( f(y_1, y_2) \), find \( E[h(Y_i)] \) by finding the marginal pdf or pmf \( f_{Y_i}(y_i) \) and using the marginal distribution in the expectation.
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

5) \( E[Y] = E[E(Y|X)] \) and \( V(Y) = E[V(Y|X)] + V[E(Y|X)] \).

6) Find the pmf of \( Y = t(X) \) and the (sample space =) support \( Y \) given the pmf of \( X \) by collecting terms \( x : y = t(x) \).

7) For increasing or decreasing \( t \), the pdf of \( Y = t(X) \) is

\[
f_Y(y) = f_X(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right|
\]
for \( y \in \mathcal{Y} \). Also be able to find the support \( \mathcal{Y} \).

8) Find the joint pdf of \( Y_1 = t_1(X_1, X_2) \) and \( Y_2 = t_2(X_1, X_2) : f_{Y_1, Y_2}(y_1, y_2) = f_{X_1, X_2}(t_1^{-1}(y_1, y_2), t_2^{-1}(y_1, y_2))|J| \). Finding the support \( \mathcal{Y} \) is crucial. Using indicator functions can help. Know that \( \prod_{j=1}^k I_{A_j}(y) = I_{\bigcap_{j=1}^k A_j}(y) \).

The Jacobian of the bivariate transformation is

\[
J = \det \begin{bmatrix}
\frac{\partial t_1^{-1}}{\partial y_1} & \frac{\partial t_1^{-1}}{\partial y_2} \\
\frac{\partial t_2^{-1}}{\partial y_1} & \frac{\partial t_2^{-1}}{\partial y_2}
\end{bmatrix},
\]
and \( |J| \) is the absolute value of the determinant \( J \). Recall that

\[
\det \begin{bmatrix}
a & b \\
c & d
\end{bmatrix} = \frac{\partial t_1^{-1}}{\partial y_1} \frac{\partial t_2^{-1}}{\partial y_2} = ad - bc.
\]

To find \( t_i^{-1}(y_1, y_2) \), use \( y_i = t_i(x_1, x_2) \) and solve for \( x_1 \) and \( x_2 \) where \( i = 1, 2 \).

9) If \( Y_1, \ldots, Y_n \) are independent with mgfs \( m_{Y_i}(t) \), then the mgf of \( W = \sum_{i=1}^n Y_i \) is

\[
m_W(t) = \prod_{i=1}^n m_{Y_i}(t).
\]

10) If \( Y_1, \ldots, Y_n \) are iid with mgf \( m_Y(t) \), then the mgf of \( W = \sum_{i=1}^n Y_i \) is

\[
m_W(t) = [m_Y(t)]^n,
\]
and the mgf of \( \overline{Y} \) is

\[
m_{\overline{Y}}(t) = [m_Y(t/n)]^n.
\]

11) Know that if \( Y_1, \ldots, Y_n \) are iid with \( E(Y) = \mu \) and \( V(Y) = \sigma^2 \), then \( E(\overline{Y}) = \mu \) and \( V(\overline{Y}) = \sigma^2/n \).

12) Suppose \( W = \sum_{i=1}^n Y_i \) or \( W = \overline{Y} \) where \( Y_1, \ldots, Y_n \) are independent. For several distributions (especially \( Y_i \) iid gamma(\( \nu, \lambda \)) and \( Y_i \) independent
be able to find the distribution of $W$, the mgf of $W$, $E(W)$, 
$\text{Var}(W)$, and $E(W^2) = V(W) + [E(W)]^2$.

13) If $X \sim N_p(\mu, \Sigma)$, then $t^T X = t_1 X_1 + ... + t_p X_p \sim N(t^T \mu, t^T \Sigma t)$.

14) If $X \sim N_p(\mu, \Sigma)$ and if $A$ is a $q \times p$ matrix, then $AX \sim N_q(A\mu, A\Sigma A^T)$.
If $a$ is a $p \times 1$ vector of constants, then $a + X \sim N_p(a + \mu, \Sigma)$.

Suppose $X_1$ is $q \times 1$ and

$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$.

15) $X_1 \sim N_q(\mu_1, \Sigma_{11})$.

16) If $X \sim N_p(\mu, \Sigma)$, then the conditional distribution of $X_1$ given that $X_2 = x_2$ is multivariate normal with mean $\mu_1 + \Sigma_{12} \Sigma_{22}^{-1}(x_2 - \mu_2)$ and covariance matrix $\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$. That is,

$X_1 | X_2 = x_2 \sim N_q(\mu_1 + \Sigma_{12} \Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})$.

17) $\rho(X_i, X_j) = \frac{\sigma_{i,j}}{\sqrt{\sigma_{ii} \sigma_{jj}}} = \frac{\text{Cov}(X_i, X_j)}{\sqrt{\text{V}(X_i) \text{V}(X_j)}}$.

18) Know that $(X, Y)$ can have a joint distribution that is not multivariate normal, yet the marginal distributions of $X$ and $Y$ are both univariate normal. Hence $X$ and $Y$ can be normal, but $aX + bY$ is not normal. (Need the joint distribution of $(X, Y)$ to be MVN for all linear combinations to be univariate normal.)

### 2.12 Complements

Panjer (1969) provides generalizations of Steiner’s formula.

Johnson and Wichern (1988), Mardia, Kent and Bibby (1979) and Press (2005) are good references for multivariate statistical analysis based on the multivariate normal distribution. The elliptically contoured distributions generalize the multivariate normal distribution and are discussed (in increasing order of difficulty) in Johnson (1987), Fang, Kotz, and Ng (1990), Fang and Anderson (1990), and Gupta and Varga (1993). Fang, Kotz, and Ng (1990) sketch the history of elliptically contoured distributions while Gupta and Varga (1993) discuss matrix valued elliptically contoured distributions.
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS


Broffitt (1986), Kowalski (1973), Melnick and Tenebien (1982) and Seber and Lee (2003, p. 23) give examples of dependent marginally normal random variables that have 0 correlation. The example in Remark 2.1 appears in Rohatgi (1976, p. 229) and Lancaster (1959).

See Abuhassan (2007) for more information about the distributions in problems 2.52–2.59.

2.13 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

Theorem 2.16 is useful for Problems 2.1–2.7.

2.1*. Let $X_1, \ldots, X_n$ be independent Poisson($\lambda_i$). Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$.

2.2*. Let $X_1, \ldots, X_n$ be iid Bernoulli($\rho$). Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$.

2.3*. Let $X_1, \ldots, X_n$ be iid exponential ($\lambda$). Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$.

2.4*. Let $X_1, \ldots, X_n$ be independent $N(\mu_i, \sigma_i^2)$. Let $W = \sum_{i=1}^{n} (a_i + b_i X_i)$ where $a_i$ and $b_i$ are fixed constants. Find the mgf of $W$ and find the distribution of $W$.

2.5*. Let $X_1, \ldots, X_n$ be iid negative binomial (1, $\rho$). Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$.

2.6*. Let $X_1, \ldots, X_n$ be independent gamma ($\nu_i, \lambda$). Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$.

2.7*. Let $X_1, \ldots, X_n$ be independent $\chi^2_{p_i}$. Let $W = \sum_{i=1}^{n} X_i$. Find the mgf of $W$ and find the distribution of $W$. 
2.8. a) Let \( f_Y(y) \) be the pdf of \( Y \). If \( W = \mu + Y \) where \(-\infty < \mu < \infty\), show that the pdf of \( W \) is \( f_W(w) = f_Y(w - \mu) \).

b) Let \( f_Y(y) \) be the pdf of \( Y \). If \( W = \sigma Y \) where \( \sigma > 0 \), show that the pdf of \( W \) is \( f_W(w) = (1/\sigma) f_Y(w/\sigma) \).

c) Let \( f_Y(y) \) be the pdf of \( Y \). If \( W = \mu + \sigma Y \) where \(-\infty < \mu < \infty \) and \( \sigma > 0 \), show that the pdf of \( W \) is \( f_W(w) = (1/\sigma) f_Y((w - \mu)/\sigma) \).

2.9. a) If \( Y \) is lognormal \( LN(\mu, \sigma^2) \), show that \( W = \log(Y) \) is a normal \( N(\mu, \sigma^2) \) random variable.

b) If \( Y \) is a normal \( N(\mu, \sigma^2) \) random variable, show that \( W = e^Y \) is a lognormal \( LN(\mu, \sigma^2) \) random variable.

2.10. a) If \( Y \) is uniform \((0,1)\), show that \( W = -\log(Y) \) is exponential \( (1) \).

b) If \( Y \) is exponential \((1)\), show that \( W = \exp(-Y) \) is uniform \((0,1)\).

2.11. If \( Y \sim N(\mu, \sigma^2) \), find the pdf of

\[
W = \left( \frac{Y - \mu}{\sigma} \right)^2
\]

2.12. If \( Y \) has a half normal distribution, \( Y \sim HN(\mu, \sigma^2) \), show that \( W = (Y - \mu)^2 \sim G(1/2, 2\sigma^2) \).

2.13. a) Suppose that \( Y \) has a Weibull \((\phi, \lambda)\) distribution with pdf

\[
f(y) = \frac{\phi}{\lambda} y^{\phi-1} e^{-y^\phi/\lambda}
\]

where \( \lambda, y, \) and \( \phi \) are all positive. Show that \( W = \log(Y) \) has a smallest extreme value SEV\((\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)\) distribution.

b) If \( Y \) has a SEV\((\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)\) distribution, show that \( W = e^Y \) has a Weibull \((\phi, \lambda)\) distribution.

2.14. a) Suppose that \( Y \) has a Pareto\((\sigma, \lambda)\) distribution with pdf

\[
f(y) = \frac{1}{\lambda} \frac{\sigma^{1/\lambda}}{y^{1+1/\lambda}}
\]

where \( y \geq \sigma, \sigma > 0, \) and \( \lambda > 0 \). Show that \( W = \log(Y) \sim EXP(\theta = \log(\sigma), \lambda) \).
b) If $Y$ as an $\text{EXP}(\theta = \log(\sigma), \lambda)$ distribution, show that $W = e^Y$ has a Pareto$(\sigma, \lambda)$ distribution.

2.15. a) If $Y$ is chi $\chi_p$, then the pdf of $Y$ is

$$f(y) = \frac{y^{p-1}e^{-y^2/2}}{2^{p-1}\Gamma(p/2)}$$

where $y \geq 0$ and $p$ is a positive integer. Show that the pdf of $W = Y^2$ is the $\chi_p^2$ pdf.

b) If $Y$ is a chi-square $\chi_p^2$ random variable, show that $W = \sqrt{Y}$ is a chi $\chi_p$ random variable.

2.16. a) If $Y$ is power $\text{POW}(\lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda}y^{1/\lambda-1},$$

where $\lambda > 0$ and $0 < y < 1$. Show that $W = -\log(Y)$ is an exponential ($\lambda$) random variable.

b) If $Y$ is an exponential($\lambda$) random variable, show that $W = e^{-Y}$ is a power $\text{POW}(\lambda)$ random variable.

2.17. a) If $Y$ is truncated extreme value $\text{TEV}(\lambda)$ then the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda} \exp(y - \frac{e^y - 1}{\lambda})$$

where $y > 0$, and $\lambda > 0$. Show that $W = e^Y - 1$ is an exponential ($\lambda$) random variable.

b) If $Y$ is an exponential($\lambda$) random variable, show that $W = \log(Y + 1)$ is a truncated extreme value $\text{TEV}(\lambda)$ random variable.

c) If $Y$ has an inverse exponential distribution, $Y \sim \text{IEXP}(\theta)$, show that $W = 1/Y \sim \text{EXP}(1/\theta)$.

d) If $Y$ has an inverse Weibull distribution, $Y \sim IW(\phi, \lambda)$, show that $1/Y \sim W(\phi, \lambda)$, the Weibull distribution with parameters $\phi$ and $\lambda$.

e) If $Y$ has a log-gamma distribution, $Y \sim LG(\nu, \lambda)$, show that $W = e^Y \sim \text{gamma}(\nu, \lambda)$.

f) If $Y$ has a two parameter power distribution, $Y \sim \text{power}(\tau, \lambda)$, show that $W = -\log(Y) \sim \text{EXP}(-\log(\tau), \lambda)$. 
2.18. a) If \( Y \) is BurrXII(\( \phi, \lambda \)), show that \( W = \log(1 + Y^\phi) \) is an exponential(\( \lambda \)) random variable.

b) If \( Y \) is an exponential(\( \lambda \)) random variable, show that \( W = (e^Y - 1)^{1/\phi} \) is a BurrXII(\( \phi, \lambda \)) random variable.

2.19. a) If \( Y \) is Pareto \( PAR(\sigma, \lambda) \), show that \( W = \log(Y/\sigma) \) is an exponential(\( \lambda \)) random variable.

b) If \( Y \) is an exponential(\( \lambda \)) random variable, show that \( W = \sigma e^Y \) is a Pareto \( PAR(\sigma, \lambda) \) random variable.

2.20. a) If \( Y \) is Weibull \( W(\phi, \lambda) \), show that \( W = Y^{1/\phi} \) is an exponential(\( \lambda \)) random variable.

b) If \( Y \) is an exponential(\( \lambda \)) random variable, show that \( W = Y^{1/\phi} \) is a Weibull \( W(\phi, \lambda) \) random variable.

2.21. If \( Y \) is double exponential (\( \theta, \lambda \)), show that \( W = |Y - \theta| \sim \text{EXP}(\lambda) \).

2.22. If \( Y \) has a generalized gamma distribution, \( Y \sim G\text{G}(\nu, \lambda, \phi) \), show that \( W = Y^\phi \sim G(\nu, \lambda^\phi) \).

2.23. If \( Y \) has an inverted gamma distribution, \( Y \sim \text{INVG}(\nu, \lambda) \), show that \( W = 1/Y \sim G(\nu, \lambda) \).

2.24. a) If \( Y \) has a largest extreme value distribution \( Y \sim \text{LEV}(\theta, \sigma) \), show that \( W = \exp(-(Y - \theta)/\sigma) \sim \text{EXP}(1) \).

b) If \( Y \sim \text{EXP}(1) \), show that \( W = \theta - \sigma \log(Y) \sim \text{LEV}(\theta, \sigma) \).

2.25. a) If \( Y \) has a log–Cauchy distribution, \( Y \sim LC(\mu, \sigma) \), show that \( W = \log(Y) \) has a Cauchy(\( \mu, \sigma \)) distribution.

b) If \( Y \sim C(\mu, \sigma) \) show that \( W = e^Y \sim LC(\mu, \sigma) \).

2.26. a) If \( Y \) has a log–logistic distribution, \( Y \sim LL(\phi, \tau) \), show that \( W = \log(Y) \) has a logistic \( L(\mu = -\log(\phi), \sigma = 1/\tau) \) distribution.

b) If \( Y \sim L(\mu = -\log(\phi), \sigma = 1/\tau) \), show that \( W = e^Y \sim LL(\phi, \tau) \).

2.27. If \( Y \) has a Maxwell–Boltzmann distribution, \( Y \sim MB(\mu, \sigma) \), show that \( W = (Y - \mu)^2 \sim G(3/2, 2\sigma^2) \).

2.28. If \( Y \) has a one sided stable distribution, \( Y \sim OSS(\sigma) \), show that \( W = 1/Y \sim G(1/2, 2/\sigma) \).
2.29. a) If $Y$ has a Rayleigh distribution, $Y \sim R(\mu, \sigma)$, show that $W = (Y - \mu)^2 \sim \exp(2\sigma^2)$.

b) If $Y \sim \exp(2\sigma^2)$, show that $W = \sqrt{Y} + \mu \sim R(\mu, \sigma)$.

2.30. If $Y$ has a smallest extreme value distribution, $Y \sim SEV(\theta, \sigma)$, show that $W = -Y$ has an $LEV(-\theta, \sigma)$ distribution.

2.31. Let $Y \sim C(0, 1)$. Show that the Cauchy distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim C(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$.

2.32. Let $Y$ have a chi distribution, $Y \sim \chi(p, 1)$ where $p$ is known. Show that the $\chi(p, \sigma)$ distribution is a scale family for $p$ known by showing that $W = \sigma Y \sim \chi(p, \sigma)$ for $\sigma > 0$.

2.33. Let $Y \sim DE(0, 1)$. Show that the double exponential distribution is a location–scale family by showing that $W = \theta + \lambda Y \sim DE(\theta, \lambda)$ where $\theta$ is real and $\lambda > 0$.

2.34. Let $Y \sim \exp(1)$. Show that the exponential distribution is a scale family by showing that $W = \lambda Y \sim \exp(\lambda)$ for $\lambda > 0$.

2.35. Let $Y \sim \exp(0, 1)$. Show that the two parameter exponential distribution is a location–scale family by showing that $W = \theta + \lambda Y \sim \exp(\theta, \lambda)$ where $\theta$ is real and $\lambda > 0$.

2.36. Let $Y \sim LEV(0, 1)$. Show that the largest extreme value distribution is a location–scale family by showing that $W = \theta + \sigma Y \sim LEV(\theta, \sigma)$ where $\theta$ is real and $\sigma > 0$.

2.37. Let $Y \sim G(\nu, 1)$ where $\nu$ is known. Show that the gamma $(\nu, \lambda)$ distribution is a scale family for $\nu$ known by showing that $W = \lambda Y \sim G(\nu, \lambda)$ for $\lambda > 0$.

2.38. Let $Y \sim HC(0, 1)$. Show that the half Cauchy distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim HC(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$.

2.39. Let $Y \sim HL(0, 1)$. Show that the half logistic distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim HL(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$. 

2.40. Let $Y \sim HN(0, 1)$. Show that the half normal distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim HN(\mu, \sigma^2)$ where $\mu$ is real and $\sigma > 0$.

2.41. Let $Y \sim L(0, 1)$. Show that the logistic distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim L(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$.

2.42. Let $Y \sim MB(0, 1)$. Show that the Maxwell–Boltzmann distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim MB(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$.

2.43. Let $Y \sim N(0, 1)$. Show that the normal distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim N(\mu, \sigma^2)$ where $\mu$ is real and $\sigma > 0$.

2.44. Let $Y \sim OSS(1)$. Show that the one sided stable distribution is a scale family by showing that $W = \sigma Y \sim OSS(\sigma)$ for $\sigma > 0$.

2.45. Let $Y \sim PAR(1, \lambda)$ where $\lambda$ is known. Show that the Pareto ($\sigma, \lambda$) distribution is a scale family for $\lambda$ known by showing that $W = \sigma Y \sim PAR(\sigma, \lambda)$ for $\sigma > 0$.

2.46. Let $Y \sim R(0, 1)$. Show that the Rayleigh distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim R(\mu, \sigma)$ where $\mu$ is real and $\sigma > 0$.

2.47. Let $Y \sim U(0, 1)$. Show that the uniform distribution is a location–scale family by showing that $W = \mu + \sigma Y \sim U(\theta_1, \theta_2)$ where $\mu = \theta_1$ is real and $\sigma = \theta_2 - \theta_1 > 0$.

2.48. Examine the proof of Theorem 2.2b for a joint pdf and prove the result for a joint pmf by replacing the integrals by appropriate sums.

2.49. Examine the proof of Theorem 2.3 for a joint pdf and prove the result for a joint pmf by replacing the integrals by appropriate sums.

2.50. Examine the proof of Theorem 2.4 for a joint pdf and prove the result for a joint pmf by replacing the integrals by appropriate sums.

2.51. Examine the proof of Theorem 2.5 for a joint pdf and prove the result for a joint pmf by replacing the integrals by appropriate sums.
2.52. If $Y \sim hburr(\phi, \lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{2}{\lambda \sqrt{2\pi}} \frac{\phi y^{\phi - 1}}{(1 + y^\phi)} \exp \left( -\frac{\log(1 + y^\phi)}{2\lambda^2} \right) I(y > 0)$$

where $\phi$ and $\lambda$ are positive.

a) Show that $W = \log(1 + Y^\phi) \sim HN(0, \lambda)$, the half normal distribution with parameters 0 and $\lambda$.

b) If $W \sim HN(0, \lambda)$, then show $Y = [e^W - 1]^{1/\phi} \sim hburr(\phi, \lambda)$.

2.53. If $Y \sim hlev(\theta, \lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{2}{\lambda \sqrt{2\pi}} \exp \left( -\frac{(y - \theta)}{\lambda} \right) \exp \left[ \frac{1}{2} \left( \exp \left( \frac{(y - \theta)}{\lambda} \right) \right)^2 \right]$$

where $y$ and $\theta$ are real and $\lambda > 0$.

a) Show that $W = \exp(-(Y - \theta)/\lambda) \sim HN(0, 1)$, the half normal distribution with parameters 0 and 1.

b) If $W \sim HN(0, 1)$, then show $Y = -\lambda \log(W) + \theta \sim hlev(\theta, \lambda)$.

2.54. If $Y \sim hpar(\theta, \lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{2}{\lambda \sqrt{2\pi}} \frac{1}{y} I[y \geq \theta] \exp \left[ \frac{-(\log(y) - \log(\theta))^2}{2\lambda^2} \right]$$

where $\theta > 0$ and $\lambda > 0$.

a) Show that $W = \log(Y) \sim HN(\mu = \log(\theta), \sigma = \lambda)$. (See the half normal distribution in Chapter 10.)

b) If $W \sim HN(\mu, \sigma)$, then show $Y = e^W \sim hpar(\theta = e^\mu, \lambda = \sigma)$.

2.55. If $Y \sim hpow(\lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{2}{\lambda \sqrt{2\pi}} \frac{1}{y} I_{[0,1]}(y) \exp \left[ \frac{-(\log(y))^2}{2\lambda^2} \right]$$

where $\lambda > 0$.

a) Show that $W = -\log(Y) \sim HN(0, \sigma = \lambda)$, the half normal distribution with parameters 0 and $\lambda$.

b) If $W \sim HN(0, \sigma)$, then show $Y = e^{-W} \sim hpow(\lambda = \sigma)$. 
2.56. If \( Y \sim hray(\theta, \lambda) \), then the pdf of \( Y \) is
\[
f(y) = \frac{4}{\lambda \sqrt{2\pi}} (y - \theta) I[y \geq \theta] \exp \left[ -\frac{(y - \theta)^4}{2\lambda^2} \right]
\]
where \( \lambda > 0 \) and \( \theta \) is real.

a) Show that \( W = (Y - \theta)^2 \sim HN(0, \sigma = \lambda) \), the half normal distribution with parameters 0 and \( \lambda \).

b) If \( W \sim HN(0, \sigma) \), then show \( Y = \sqrt{W} + \theta \sim hray(\theta, \lambda = \sigma) \).

2.57. If \( Y \sim hsev(\theta, \lambda) \), then the pdf of \( Y \) is
\[
f(y) = \frac{2}{\lambda \sqrt{2\pi}} \exp\left(\frac{y - \theta}{\lambda}\right) \exp\left(-\frac{1}{2} \left[ \exp\left(\frac{y - \theta}{\lambda}\right) \right]^2\right)
\]
where \( y \) and \( \theta \) are real and \( \lambda > 0 \).

a) Show that \( W = \exp[(y - \theta)/\lambda] \sim HN(0, 1) \).

b) If \( W \sim HN(0, 1) \), then show \( Y = \lambda \log(W) + \theta \sim hsev(\theta, \lambda) \).

2.58. If \( Y \sim htev(\lambda) \), then the pdf of \( Y \) is
\[
f(y) = \frac{2}{\lambda \sqrt{2\pi}} \exp\left(\frac{y - (e^y - 1)^2}{2\lambda^2}\right) = \frac{2}{\lambda \sqrt{2\pi}} e^y \exp\left(-\frac{(e^y - 1)^2}{2\lambda^2}\right)
\]
where \( y > 0 \) and \( \lambda > 0 \).

a) Show that \( W = e^{Y} - 1 \sim HN(0, \sigma = \lambda) \), the half normal distribution with parameters 0 and \( \lambda \).

b) If \( W \sim HN(0, \sigma) \), then show \( Y = \log(W + 1) \sim htev(\lambda = \sigma) \).

2.59. If \( Y \sim hweib(\phi, \lambda) \), then the pdf of \( Y \) is
\[
f(y) = \frac{2}{\lambda \sqrt{2\pi}} \phi y^{\phi-1} I[y > 0] \exp\left(-\frac{y^{2\phi}}{2\lambda^2}\right)
\]
where \( \lambda \) and \( \phi \) are positive.

a) Show that \( W = Y^\phi \sim HN(0, \sigma = \lambda) \), the half normal distribution with parameters 0 and \( \lambda \).

b) If \( W \sim HN(0, \sigma) \), then show \( Y = W^{1/\phi} \sim hweib(\phi, \lambda = \sigma) \).

Problems from old quizzes and exams.
2.60. If \( Y \) is a random variable with pdf
\[
f(y) = \lambda y^{\lambda-1} \quad \text{for } 0 < y < 1
\]
where \( \lambda > 0 \), show that \( W = -\log(Y) \) is an exponential\((1/\lambda)\) random variable.

2.61. If \( Y \) is an exponential\((1/\lambda)\) random variable, show that \( W = e^{-Y} \) has pdf
\[
f_W(w) = \lambda w^{\lambda-1} \quad \text{for } 0 < w < 1.
\]

2.62. If \( Y \sim \text{EXP}(\lambda) \), find the pdf of \( W = 2 \lambda Y \).

2.63*. (Mukhopadhyay 2000, p. 113): Suppose that \( X|Y \sim N(\beta_0 + \beta_1 Y, Y^2) \), and that \( Y \sim N(3, 10) \). That is, the conditional distribution of \( X \) given that \( Y = y \) is normal with mean \( \beta_0 + \beta_1 y \) and variance \( y^2 \) while the (marginal) distribution of \( Y \) is normal with mean 3 and variance 10.

a) Find \( EX. \)

b) Find \( \text{Var} X. \)

c) Which pairs of random variables \( X_i \) and \( X_j \) are independent?

d) Find the correlation \( \rho(X_1, X_3). \)

2.64*. Suppose that
\[
\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim N_4 \left( \begin{pmatrix} 49 \\ 100 \\ 17 \\ 7 \end{pmatrix}, \begin{pmatrix} 3 & 1 & -1 & 0 \\ 1 & 6 & 1 & -1 \\ -1 & 1 & 4 & 0 \\ 0 & -1 & 0 & 2 \end{pmatrix} \right).
\]

a) Find the distribution of \( X_2. \)

b) Find the distribution of \( (X_1, X_3)^T. \)

c) Which pairs of random variables \( X_i \) and \( X_j \) are independent?

d) Find the correlation \( \rho(X_1, X_3). \)

2.65*. Recall that if \( \mathbf{X} \sim N_p(\mu, \Sigma) \), then the conditional distribution of \( \mathbf{X}_1 \) given that \( \mathbf{X}_2 = \mathbf{x}_2 \) is multivariate normal with mean
\[
\mu_1 + \Sigma_{12} \Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2)
\]
and covariance matrix
\[
\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\]

Let \( \sigma_{12} = \text{Cov}(Y, X) \) and suppose \( Y \) and \( X \) follow a bivariate normal distribution
\[
\begin{pmatrix} Y \\ X \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 49 \\ 100 \end{pmatrix}, \begin{pmatrix} 16 & \sigma_{12} \\ \sigma_{12} & 25 \end{pmatrix} \right).
\]
a) If $\sigma_{12} = 0$, find $Y|X$. Explain your reasoning.

b) If $\sigma_{12} = 10$ find $E(Y|X)$.

c) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

2.66. Let $\sigma_{12} = \text{Cov}(Y, X)$ and suppose $Y$ and $X$ follow a bivariate normal distribution

\[
\begin{pmatrix}
  Y \\
  X
\end{pmatrix} \sim N_2 \left( \begin{pmatrix}
  15 \\
  20
\end{pmatrix}, \begin{pmatrix}
  64 & \sigma_{12} \\
  \sigma_{12} & 81
\end{pmatrix} \right).
\]

a) If $\sigma_{12} = 10$ find $E(Y|X)$.

b) If $\sigma_{12} = 10$, find $\text{Var}(Y|X)$.

c) If $\sigma_{12} = 10$, find $\rho(Y, X)$, the correlation between $Y$ and $X$.

2.67*. (Mukhopadhyay 2000, p. 197): Suppose that $X_1$ and $X_2$ have a joint pdf given by

\[
f(x_1, x_2) = 3(x_1 + x_2)I(0 < x_1 < 1)I(0 < x_2 < 1)I(0 < x_1 + x_2 < 1).
\]

Consider the transformation $Y_1 = X_1 + X_2$ and $Y_2 = X_1 - X_2$.

a) Find the Jacobian $J$ for the transformation.

b) Find the support $Y$ of $Y_1$ and $Y_2$.

c) Find the joint density $f_{Y_1, Y_2}(y_1, y_2)$.

d) Find the marginal pdf $f_{Y_1}(y_1)$.

e) Find the marginal pdf $f_{Y_2}(y_2)$.

Hint for d) and e): $I_{A_1}(\mathbf{y})I_{A_2}(\mathbf{y})I_{A_3}(\mathbf{y}) = I_{\cap_{j=1}^3 A_j}(\mathbf{y}) = I_Y(\mathbf{y})$ where $Y$ is a triangle.
2.68*. (Aug. 2000 Qual): The number of defects per yard, $Y$ of a certain fabric is known to have a Poisson distribution with parameter $\lambda$. However, $\lambda$ is a random variable with pdf $f(\lambda) = e^{-\lambda}I(\lambda > 0)$.

a) Find $E(Y)$.

b) Find $\text{Var}(Y)$.

2.69. Let $A$ and $B$ be positive integers. A hypergeometric random variable $X = W_1 + W_2 + \cdots + W_n$ where the random variables $W_i$ are identically distributed random variables with $P(W_i = 1) = A/(A + B)$ and $P(W_i = 0) = B/(A + B)$. You may use the fact that $E(W_i) = A/(A + B)$ and that $E(X) = nA/(A + B)$.

a) Find $\text{Var}(W_1)$.

b) If $i \neq j$, then $\text{Cov}(W_i, W_j) = -AB/(A + B)^2(A + B - 1)$. Find $\text{Var}(X)$ using the formula

$$Var\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} Var(W_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{Cov}(W_i, W_j).$$

(Hint: the sum $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n}$ has $(n-1)n/2$ terms.)

2.70. Let $X = W_1 + W_2 + \cdots + W_n$ where the joint distribution of the random variables $W_i$ is an $n$-dimensional multivariate normal distribution with $E(W_i) = 1$ and $\text{Var}(W_i) = 100$ for $i = 1, \ldots, n$.

a) Find $E(X)$.

b) Suppose that if $i \neq j$, then $\text{Cov}(W_i, W_j) = 10$. Find $\text{Var}(X)$ using the formula

$$Var\left(\sum_{i=1}^{n} W_i\right) = \sum_{i=1}^{n} Var(W_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \text{Cov}(W_i, W_j).$$

(Hint: the sum $\sum_{i=1}^{n-1} \sum_{j=i+1}^{n}$ has $(n-1)n/2$ terms.)
CHAPTER 2. MULTIVARIATE DISTRIBUTIONS

2.71. Find the moment generating function for $Y_1$ if the joint probability mass function $f(y_1, y_2)$ of $Y_1$ and $Y_2$ is tabled as shown.

<table>
<thead>
<tr>
<th>$f(y_1, y_2)$</th>
<th>$y_2$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>0</td>
<td>0.38</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.17</td>
<td>0.02</td>
<td>0.05</td>
</tr>
</tbody>
</table>

2.72. Suppose that the joint pdf of $X$ and $Y$ is $f(x, y) =$

$$
\frac{1}{2} \frac{1}{2\pi \sqrt{1-\rho^2}} \exp\left( \frac{-1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2) \right) + \frac{1}{2} \frac{1}{2\pi \sqrt{1-\rho^2}} \exp\left( \frac{-1}{2(1-\rho^2)}(x^2 + 2\rho xy + y^2) \right)
$$

where $x$ and $y$ are real and $0 < \rho < 1$. It can be shown that the marginal pdfs are

$$
f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x^2 \right)
$$

for $x$ real and

$$
f_Y(y) = \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} y^2 \right)
$$

for $y$ real. Are $X$ and $Y$ independent? Explain briefly.

2.73*. Suppose that the conditional distribution of $Y|P = \rho$ is the binomial($k, \rho$) distribution and that the random variable $P$ has a beta($\delta = 4, \nu = 6$) distribution.

a) Find $E(Y)$.

b) Find $\text{Var}(Y)$.

2.74*. Suppose that the joint probability mass function $f(y_1, y_2)$ of $Y_1$ and $Y_2$ is given in the following table.

<table>
<thead>
<tr>
<th>$f(y_1, y_2)$</th>
<th>$y_2$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>0</td>
<td>0.38</td>
<td>0.14</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.17</td>
<td>0.02</td>
<td>0.05</td>
</tr>
</tbody>
</table>

a) Find the marginal probability function $f_{Y_2}(y_2)$ for $Y_2$. 
b) Find the conditional probability function \( f(y_1|y_2) \) of \( Y_1 \) given \( Y_2 = 2 \).

\[ 2.75^* \]. Find the pmf of \( Y = X^2 + 4 \) where the pmf of \( X \) is given below.

\[
\begin{array}{c|ccccc}
X & -2 & -1 & 0 & 1 & 2 \\
\hline
\text{probability} & 0.1 & 0.2 & 0.4 & 0.2 & 0.1 \\
\end{array}
\]

\[ 2.76 \]. Suppose that \( X_1 \) and \( X_2 \) are independent with \( X_1 \sim N(0, 1) \) and \( X_2 \sim N(0, 4) \) so \( \text{Var}(X_2) = 4 \). Consider the transformation \( Y_1 = X_1 + X_2 \) and \( Y_2 = X_1 - X_2 \).

a) Find the Jacobian \( J \) for the transformation.

b) Find the joint pdf \( f(y_1, y_2) \) of \( Y_1 \) and \( Y_2 \).

c) Are \( Y_1 \) and \( Y_2 \) independent? Explain briefly.

Hint: can you factor the joint pdf so that \( f(y_1, y_2) = g(y_1)h(y_2) \) for every real \( y_1 \) and \( y_2 \)?
Chapter 3

Exponential Families

3.1 Regular Exponential Families

The theory of exponential families will be used in the following chapters to study some of the most important topics in statistical inference such as minimal and complete sufficient statistics, maximum likelihood estimators (MLEs), uniform minimum variance estimators (UMVUEs) and the Fréchet Cramér Rao lower bound (FCRLB), uniformly most powerful (UMP) tests and large sample theory.

Often a “brand name distribution” such as the normal distribution will have three useful parameterizations: the usual parameterization with parameter space $\Theta_U$ is simply the formula for the probability distribution or mass function (pdf or pmf, respectively) given when the distribution is first defined. The $k$-parameter exponential family parameterization with parameter space $\Theta$, given in Definition 3.1 below, provides a simple way to determine if the distribution is an exponential family while the natural parameterization with parameter space $\Omega$, given in Definition 3.2 below, is used for theory that requires a complete sufficient statistic.

**Definition 3.1.** A family of joint pdfs or joint pmfs $\{f(y|\theta) : \theta = (\theta_1,...,\theta_j) \in \Theta\}$ for a random vector $Y$ is an exponential family if

$$f(y|\theta) = h(y)c(\theta)\exp\left[\sum_{i=1}^{k} w_i(\theta)t_i(y)\right]$$

(3.1)

for all $y$ where $c(\theta) \geq 0$ and $h(y) \geq 0$. The functions $c, h, t_i$, and $w_i$ are real valued functions. The parameter $\theta$ can be a scalar and $y$ can be a scalar.
It is crucial that $c, w_1, ..., w_k$ do not depend on $y$ and that $h, t_1, ..., t_k$ do not depend on $\theta$. The support of the distribution is $\mathcal{Y}$ and the parameter space is $\Theta$. The family is a $k$-parameter exponential family if $k$ is the smallest integer where (3.1) holds.

Notice that the distribution of $Y$ is an exponential family if

$$f(y|\theta) = h(y)c(\theta) \exp\left[\sum_{i=1}^{k} w_i(\theta)t_i(y)\right]$$

and the distribution is a one parameter exponential family if

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)].$$

The parameterization is not unique since, for example, $w_i$ could be multiplied by a nonzero constant $a$ if $t_i$ is divided by $a$. Many other parameterizations are possible. If $h(y) = g(y)I_y(y)$, then usually $c(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = \exp\left[\sum_{i=1}^{k} w_i(\theta)t_i(y) + d(\theta) + S(y)\right] I_y(y)$$

where $S(y) = \log(g(y))$, $d(\theta) = \log(c(\theta))$, and $\mathcal{Y}$ does not depend on $\theta$.

To demonstrate that $\{f(y|\theta) : \theta \in \Theta\}$ is an exponential family, find $h(y), c(\theta), w_i(\theta)$ and $t_i(y)$ such that (3.1), (3.2), (3.3) or (3.4) holds.

**Theorem 3.1.** Suppose that $Y_1, ..., Y_n$ are iid random vectors from an exponential family. Then the joint distribution of $Y_1, ..., Y_n$ follows an exponential family.

**Proof.** Suppose that $f_{Y_i}(y_i)$ has the form of (3.1). Then by independence,

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} h(y_i)c(\theta) \exp\left[\sum_{j=1}^{k} w_j(\theta)t_j(y_i)\right]$$

$$= \prod_{i=1}^{n} h(y_i)[c(\theta)]^{n} \prod_{i=1}^{n} \exp\left[\sum_{j=1}^{k} w_j(\theta)t_j(y_i)\right]$$
= \left[ \prod_{i=1}^{n} h(y_i) \right] \left[ c(\theta) \right]^{n} \exp \left( \sum_{i=1}^{n} \left[ \sum_{j=1}^{k} w_j(\theta) t_j(y_i) \right] \right)

= \left[ \prod_{i=1}^{n} h(y_i) \right] \left[ c(\theta) \right]^{n} \exp \left( \sum_{i=1}^{n} \left( \sum_{j=1}^{k} w_j(\theta) t_j(y_i) \right) \right).

To see that this has the form (3.1), take
\[ h^\ast(y_1, \ldots, y_n) = \prod_{i=1}^{n} h(y_i), \quad c^\ast(\theta) = \left[ c(\theta) \right]^{n}, \quad w_j^\ast(\theta) = w_j(\theta) \quad \text{and} \quad t_j^\ast(y_1, \ldots, y_n) = \sum_{i=1}^{n} t_j(y_i). \]

QED

The parameterization that uses the natural parameter \( \eta \) is especially useful for theory. See Definition 3.3 for the natural parameter space \( \Omega \).

**Definition 3.2.** Let \( \Omega \) be the natural parameter space for \( \eta \). The natural parameterization for an exponential family is

\[
f(y|\eta) = h(y) b(\eta) \exp \left( \sum_{i=1}^{k} \eta_i t_i(y) \right)
\]  

(3.5)

where \( h(y) \) and \( t_i(y) \) are the same as in Equation (3.1) and \( \eta \in \Omega \). The natural parameterization for a random variable \( Y \) is

\[
f(y|\eta) = h(y) b(\eta) \exp \left( \sum_{i=1}^{k} \eta_i t_i(y) \right)
\]  

(3.6)

where \( h(y) \) and \( t_i(y) \) are the same as in Equation (3.2) and \( \eta \in \Omega \). Again, the parameterization is not unique. If \( a \neq 0 \), then \( a\eta_i \) and \( t_i(y)/a \) would also work.

Notice that the natural parameterization (3.6) has the same form as (3.2) with \( \theta^\ast = \eta \), \( c^\ast(\theta^\ast) = b(\eta) \) and \( w_i(\theta^\ast) = w_i(\eta) = \eta_i \). In applications often \( \eta \) and \( \Omega \) are of interest while \( b(\eta) \) is not computed.

The next important idea is that of a regular exponential family (and of a full exponential family). Let \( d_i(x) \) denote \( t_i(y) \), \( w_i(\theta) \) or \( \eta_i \). A linearity constraint is satisfied by \( d_1(x), \ldots, d_k(x) \) if \( \sum_{i=1}^{k} a_i d_i(x) = c \) for some constants \( a_i \) and \( c \) and for all \( x \) (or \( \eta_i \)) in the sample or parameter space where not all of the \( a_i = 0 \). If \( \sum_{i=1}^{k} a_i d_i(x) = c \) for all \( x \) only if \( a_1 = \cdots = a_k = 0 \), then the \( d_i(x) \) do not satisfy a linearity constraint. In linear algebra, we would say that the \( d_i(x) \) are linearly independent if they do not satisfy a linearity constraint.
For $k = 2$, a linearity constraint is satisfied if a plot of $d_1(x)$ versus $d_2(x)$ falls on a line as $x$ varies. If the parameter space for the $\eta_1$ and $\eta_2$ is a nonempty open set, then the plot of $\eta_1$ versus $\eta_2$ is that nonempty open set, and the $\eta_i$ cannot satisfy a linearity constraint since the plot is not a line.

Let $\Omega$ be the set where the integral of the kernel function is finite:

$$\tilde{\Omega} = \{ \eta = (\eta_1, ..., \eta_k) : \frac{1}{b(\eta)} \equiv \int_{-\infty}^{\infty} h(y) \exp[\sum_{i=1}^{k} \eta_i t_i(y)]dy < \infty \}.$$

Replace the integral by a sum for a pmf. An interesting fact is that $\tilde{\Omega}$ is a convex set.

**Definition 3.3.** Condition E1: the natural parameter space $\Omega = \tilde{\Omega}$.
Condition E2: assume that in the natural parameterization, neither the $\eta_i$ nor the $t_i$ satisfy a linearity constraint.
Condition E3: $\Omega$ is a $k$-dimensional nonempty open set.
If conditions E1), E2) and E3) hold then the exponential family is a $k$-parameter regular exponential family (REF).
If conditions E1) and E2) hold then the exponential family is a $k$-parameter full exponential family.

**Notation.** A $k$P–REF is a $k$-parameter regular exponential family. So a 1P–REF is a 1-parameter REF and a 2P–REF is a 2-parameter REF.

Notice that every REF is full. Any $k$-dimensional open set will contain a $k$-dimensional rectangle. A $k$-fold cross product of nonempty open intervals is a $k$-dimensional open set. For a one parameter exponential family, a one dimensional rectangle is just an interval, and the only type of function of one variable that satisfies a linearity constraint is a constant function. In the definition of an exponential family and in the usual parameterization, $\theta$ is a $j \times 1$ vector. Typically $j = k$ if the family is a kP–REF. If $j < k$ and $k$ is as small as possible, the family will usually not be regular. For example, a $N(\theta, \theta^2)$ family has $\theta = \theta$ with $j = 1 < 2 = k$, and is not regular. See Example 3.8 for more details.

Some care has to be taken with the definitions of $\Theta$ and $\Omega$ since formulas (3.1) and (3.6) need to hold for every $\theta \in \Theta$ and for every $\eta \in \Omega$. For a continuous random variable or vector, the pdf needs to exist. Hence all degenerate distributions need to be deleted from $\Theta_U$ to form $\Theta$ and $\Omega$. For
continuous and discrete distributions, the natural parameter needs to exist (and often does not exist for discrete degenerate distributions). As a rule of thumb, remove values from $\Theta_U$ that cause the pmf to have the form $0^0$. For example, for the binomial$(k, \rho)$ distribution with $k$ known, the natural parameter $\eta = \log(\rho/(1-\rho))$. Hence instead of using $\Theta_U = [0, 1]$, use $\rho \in \Theta = (0, 1)$, so that $\eta \in \Omega = (-\infty, \infty)$.

These conditions have some redundancy. If $\Omega$ contains a $k$-dimensional rectangle (eg if the family is a $k$P-REF, then $\Omega$ is a $k$-dimensional open set and contains a $k$-dimensional open ball which contains a $k$-dimensional rectangle), no $\eta_i$ is completely determined by the remaining $\eta_j's$. In particular, the $\eta_i$ cannot satisfy a linearity constraint. If the $\eta_i$ do satisfy a linearity constraint, then the $\eta_i$ lie on a hyperplane of dimension at most $k$, and such a surface cannot contain a $k$-dimensional rectangle. For example, if $k = 2$, a line cannot contain an open box. If $k = 2$ and $\eta_2 = \eta_1^2$, then the parameter space is not a 2-dimensional open set and does not contain a 2-dimensional rectangle. Thus the family is not a 2P–REF although $\eta_1$ and $\eta_2$ do not satisfy a linearity constraint. Again, see Example 3.8.

The most important 1P–REFs are the binomial $(k, \rho)$ distribution with $k$ known, the exponential $(\lambda)$ distribution, and the Poisson $(\theta)$ distribution. Other 1P–REFs include the Burr type III $(\lambda, \phi)$ distribution with $\phi$ known, the Burr Type X $(\tau)$ distribution, the Burr type XII $(\phi, \lambda)$ distribution with $\phi$ known, the double exponential $(\theta, \lambda)$ distribution with $\theta$ known, the two parameter exponential $(\theta, \lambda)$ distribution with $\theta$ known, the generalized negative binomial $(\mu, \kappa)$ distribution if $\kappa$ is known, the geometric $(\rho)$ distribution, the Gompertz $(\theta, \nu)$ distribution with $\theta > 0$ known, the half normal $(\mu, \sigma^2)$ distribution with $\mu$ known, the inverse exponential $(\theta)$ distribution, the inverse Weibull $(\phi, \lambda)$ distribution with $\phi$ known, the largest extreme value $(\theta, \sigma)$ distribution if $\sigma$ is known, the smallest extreme value $(\theta, \sigma)$ distribution if $\sigma$ is known, the inverted gamma $(\nu, \lambda)$ distribution if $\nu$ is known, the logarithmic $(\theta)$ distribution, the Maxwell–Boltzmann $(\mu, \sigma)$ distribution if $\mu$ is known, the modified DeMoivre’s law $(\theta, \phi)$ distribution if $\theta$ is known, the negative binomial $(r, \rho)$ distribution if $r$ is known, the one sided stable $(\sigma)$ distribution, the Pareto $(\sigma, \lambda)$ distribution if $\sigma$ is known, the power $(\lambda)$ distribution, the Rayleigh $(\mu, \sigma)$ distribution if $\mu$ is known, the Topp-Leone $(\nu)$ distribution, the two parameter power $(\tau, \lambda)$ distribution with $\tau$ known, the truncated extreme value $(\lambda)$ distribution, the Weibull $(\phi, \lambda)$ distribution if $\phi$ is known, the zero truncated Poisson $(\theta)$ distribution,
the Zeta ($\nu$) distribution and the Zipf ($\nu$) distribution. A one parameter exponential family can often be obtained from a $k$–parameter exponential family by holding $k - 1$ of the parameters fixed. Hence a normal ($\mu, \sigma^2$) distribution is a 1P–REF if $\sigma^2$ is known. Usually assuming scale, location or shape parameters are known is a bad idea.

The most important 2P–REFs are the beta ($\delta, \nu$) distribution, the gamma ($\nu, \lambda$) distribution and the normal ($\mu, \sigma^2$) distribution. The chi ($p, \sigma$) distribution, the log-gamma ($\nu, \lambda$) distribution and the lognormal ($\mu, \sigma^2$) distribution are also 2P-REFs. Example 3.9 will show that the inverse Gaussian distribution is full but not regular. The two parameter Cauchy distribution is not an exponential family because its pdf cannot be put into the form of Equation (3.1).

The natural parameterization can result in a family that is much larger than the family defined by the usual parameterization. See the definition of $\Omega = \tilde{\Omega}$ given by Equation (3.7). Casella and Berger (2002, p. 114) remarks that

$$\{\eta : \eta = (w_1(\theta), ..., w_k(\theta)) | \theta \in \Theta\} \subseteq \Omega,$$  

(3.8)

but often $\Omega$ is a strictly larger set.

**Remark 3.1.** For the families in Chapter 10 other than the $\chi^2_p$ and inverse Gaussian distributions, make the following assumptions if $\dim(\Theta) = k = \dim(\Omega)$. Assume that $\eta_i = w_i(\theta)$. Assume the usual parameter space $\Theta_U$ is as big as possible (replace the integral by a sum for a pmf):

$$\Theta_U = \{\theta \in \mathbb{R}^k : \int f(y|\theta)dy = 1\},$$

and let

$$\Theta = \{\theta \in \Theta_U : w_1(\theta), ..., w_k(\theta) \text{ are defined }\}.$$ 

Then assume that the natural parameter space satisfies condition E1) with

$$\Omega = \{(\eta_1, ..., \eta_k) : \eta_i = w_i(\theta) \text{ for } \theta \in \Theta\}.$$

In other words, simply define $\eta_i = w_i(\theta)$. For many common distributions, $\eta$ is a one to one function of $\theta$, and the above map is correct, especially if $\Theta_U$ is an open interval or cross product of open intervals.

**Remark 3.2.** Chapter 10 has many examples showing that a distribution is a 1P-REF or 2P-REF.
Example 3.1. Let \( f(x|\mu, \sigma) \) be the \( N(\mu, \sigma^2) \) family of pdfs. Then \( \theta = (\mu, \sigma) \) where \(-\infty < \mu < \infty \) and \( \sigma > 0 \). Recall that \( \mu \) is the mean and \( \sigma \) is the standard deviation (SD) of the distribution. The usual parameterization is

\[
f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)I_{\mathbb{R}}(x)
\]

where \( \mathbb{R} = (-\infty, \infty) \) and the indicator \( I_A(x) = 1 \) if \( x \in A \) and \( I_A(x) = 0 \) otherwise. Notice that \( I_{\mathbb{R}}(x) = 1 \) \( \forall \ x \). Since

\[
f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-\mu^2}{2\sigma^2}\right) \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right) + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i \right) I_{\mathbb{R}}(x),
\]

this family is a 2-parameter exponential family. Hence \( \eta_1 = -0.5/\sigma^2 \) and \( \eta_2 = \mu/\sigma^2 \) if \( \sigma > 0 \), and \( \Omega = (-\infty, 0) \times (-\infty, \infty) \). Plotting \( \eta_1 \) on the horizontal axis and \( \eta_2 \) on the vertical axis yields the left half plane which certainly contains a 2-dimensional rectangle. Since \( t_1 \) and \( t_2 \) lie on a quadratic rather than a line, the family is a 2P–REF. Notice that if \( X_1, \ldots, X_n \) are iid \( N(\mu, \sigma^2) \) random variables, then the joint pdf \( f(x|\theta) = f(x_1, \ldots, x_n|\mu, \sigma) = \]

\[
\left[\frac{1}{\sqrt{2\pi\sigma}} \exp\left(\frac{-\mu^2}{2\sigma^2}\right)\right]^n \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right) + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i \right) I_{\mathbb{R}}(x),
\]

and is thus a 2P–REF.

Example 3.2. The \( \chi^2_p \) distribution is not a 1P-REF since the usual parameter space \( \Theta_U \) for the \( \chi^2_p \) distribution is the set of integers, which is neither an open set nor a convex set. Nevertheless, the natural parameterization is the gamma(\( \nu, \lambda = 2 \)) family which is a 1P-REF. Note that this family has uncountably many members while the \( \chi^2_p \) family does not.

Example 3.3. The binomial(\( k, \rho \)) pmf is

\[
f(x|\rho) = \binom{k}{x} \rho^x (1-\rho)^{k-x} I_{\{0, \ldots, k\}}(x)
\]

\[
= \binom{k}{x} I_{\{0, \ldots, k\}}(x) \left(1-\rho\right)^{k-x} \exp\left[\log\left(\frac{\rho}{1-\rho}\right) x \right]
\]

\[
= \binom{k}{x} I_{\{0, \ldots, k\}}(x) \left(1-\rho\right)^{k-x} \exp\left[\log\left(\frac{\rho}{1-\rho}\right) x \right]
\]
where \( \Theta_U = [0, 1] \). Since the pmf and \( \eta = \log(\rho/(1 - \rho)) \) is undefined for \( \rho = 0 \) and \( \rho = 1 \), we have \( \Theta = (0, 1) \). Notice that \( \Omega = (-\infty, \infty) \).

**Example 3.4.** The uniform(0,\( \theta \)) family is not an exponential family since the support \( Y_{\theta} = (0, \theta) \) depends on the unknown parameter \( \theta \).

**Example 3.5.** If \( Y \) has a half normal distribution, \( Y \sim \text{HN}(\mu, \sigma) \), then the pdf of \( Y \) is

\[
f(y) = \frac{2}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)
\]

where \( \sigma > 0 \) and \( y \geq \mu \) and \( \mu \) is real. Notice that \( f(y) = \frac{2}{\sqrt{2\pi} \sigma} I(y \geq \mu) \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right] \)

is a 1P–REF if \( \mu \) is known. Hence \( \Theta = (0, \infty) \), \( \eta = -1/(2\sigma^2) \) and \( \Omega = (-\infty, 0) \). Notice that a different 1P–REF is obtained for each value of \( \mu \) when \( \mu \) is known with support \( Y'_\mu = [\mu, \infty) \). If \( \mu \) is not known, then this family is not an exponential family since the support depends on \( \mu \).

The following two examples are important examples of REFs where \( \dim(\Theta) > \dim(\Omega) \).

**Example 3.6.** If the \( t_i \) or \( \eta_i \) satisfy a linearity constraint, then the number of terms in the exponent of Equation (3.1) can be reduced. Suppose that \( Y_1, \ldots, Y_n \) follow the multinomial \( \text{M}_n(m, \rho_1, \ldots, \rho_n) \) distribution which has \( \dim(\Theta) = n \) if \( m \) is known. Then \( \sum_{i=1}^n Y_i = m \), \( \sum_{i=1}^n \rho_i = 1 \) and the joint pmf of \( Y \) is

\[
f(y) = m! \prod_{i=1}^n \frac{\rho_i^{y_i}}{y_i!}.
\]

The support of \( Y \) is \( \mathcal{Y} = \{y : \sum_{i=1}^n y_i = m \text{ and } 0 \leq y_i \leq m \text{ for } i = 1, \ldots, n\} \).

Since \( Y_n \) and \( \rho_n \) are known if \( Y_1, \ldots, Y_{n-1} \) and \( \rho_1, \ldots, \rho_{n-1} \) are known, we can use an equivalent joint pmf \( f_{EF} \) in terms of \( Y_1, \ldots, Y_{n-1} \). Let

\[
h(y_1, \ldots, y_{n-1}) = \left[\frac{m!}{\prod_{i=1}^n y_i!}\right] I[(y_1, \ldots, y_{n-1}, y_n) \in \mathcal{Y}].
\]

(This is a function of \( y_1, \ldots, y_{n-1} \) since \( y_n = m - \sum_{i=1}^{n-1} y_i \).) Then \( Y_1, \ldots, Y_{n-1} \)
have a \( M_n(m, \rho_1, \ldots, \rho_n) \) distribution if the joint pmf of \( Y_1, \ldots, Y_{n-1} \) is
\[
f_{\text{EF}}(y_1, \ldots, y_{n-1}) = \exp\left[ \sum_{i=1}^{n-1} y_i \log(\rho_i) + (m - \sum_{i=1}^{n-1} y_i) \log(\rho_n) \right] h(y_1, \ldots, y_{n-1})
\]
\[
= \exp[m \log(\rho_n)] \exp\left[ \sum_{i=1}^{n-1} y_i \log(\rho_i/\rho_n) \right] h(y_1, \ldots, y_{n-1}). \tag{3.9}
\]
Since \( \rho_n = 1 - \sum_{j=1}^{n-1} \rho_j \), this is an \( n-1 \) dimensional REF with
\[
\eta_i = \log(\rho_i/\rho_n) = \log \left( \frac{\rho_i}{1 - \sum_{j=1}^{n-1} \rho_j} \right)
\]
and \( \Omega = \mathbb{R}^{n-1} \).

**Example 3.7.** Similarly, let \( \mu \) be a \( 1 \times j \) row vector and let \( \Sigma \) be a \( j \times j \) positive definite matrix. Then the usual parameterization of the multivariate normal \( \text{MVN}_j(\mu, \Sigma) \) distribution has \( \dim(\Theta) = j + j^2 \) but is a \( j + j(j+1)/2 \) parameter REF.

A **curved exponential family** is a \( k \)-parameter exponential family where the elements of \( \theta = (\theta_1, \ldots, \theta_k) \) are completely determined by \( d < k \) of the elements. For example if \( \theta = (\theta, \theta^2) \) then the elements of \( \theta \) are completely determined by \( \theta_1 = \theta \). A curved exponential family is not regular since it places a restriction on the parameter space \( \Omega \) resulting in a new “natural parameter space” \( \Omega_C \) where \( \Omega_C \) does not contain a \( k \)-dimensional rectangle and is not a \( k \)-dimensional open set.

**Example 3.8.** The \( N(\theta, \theta^2) \) distribution is a 2-parameter exponential family with \( \eta_1 = -1/(2\theta^2) \) and \( \eta_2 = 1/\theta \). To see this, note that
\[
f(y|\theta) = \frac{1}{\theta \sqrt{2\pi}} \exp \left( \frac{-(y - \theta)^2}{2\theta^2} \right) = \frac{1}{\theta \sqrt{2\pi}} \exp \left[ -\frac{1}{2\theta^2} (y^2 - 2\theta y + \theta^2) \right] = \frac{1}{\theta \sqrt{2\pi}} \exp(-1/2) \exp \left[ -\frac{1}{2\theta^2} y^2 + \frac{1}{\theta} y \right].
\]
Thus the “natural parameter space” is
\[
\Omega_C = \{ (\eta_1, \eta_2) | \eta_1 = -0.5\eta_2^2, -\infty < \eta_1 < 0, -\infty < \eta_2 < \infty, \eta_2 \neq 0 \}.
\]
Figure 3.1: The parameter space is a quadratic, not a 2-dimensional open set.
CHAPTER 3. EXPONENTIAL FAMILIES

To be a 2P-REF, the parameter space needs to be a 2-dimensional open set. A $k$-dimensional open set contains a $k$-dimensional open ball which contains a $k$-dimensional open set. The graph of the “natural parameter space” is a quadratic and cannot contain a 2-dimensional rectangle. (Any rectangle will contain points that are not on the quadratic, so $\Omega_C$ is not a 2-dimensional open set.) See Figure 3.1 where the small rectangle centered at (-2,2) contains points that are not in the parameter space. Hence this 2-parameter curved exponential family is not a 2P-REF.

3.2 Properties of $(t_1(Y), \ldots, t_k(Y))$

This section follows Lehmann (1983, p. 29-35) closely. Write the natural parameterization for the exponential family as

$$f(y|\eta) = h(y)b(\eta) \exp \left[ \sum_{i=1}^{k} \eta_i t_i(y) \right]$$

$$= h(y) \exp \left[ \sum_{i=1}^{k} \eta_i t_i(y) - a(\eta) \right]$$

(3.10)

where $a(\eta) = -\log(b(\eta))$. The kernel function of this pdf or pmf is

$$h(y) \exp \left[ \sum_{i=1}^{k} \eta_i t_i(y) \right].$$

Lemma 3.2. Suppose that $Y$ comes from an exponential family (3.10) and that $g(y)$ is any function with $E_{\eta}[|g(Y)|] < \infty$. Then for any $\eta$ in the interior of $\Omega$, the integral $\int g(y) f(y|\theta)dy$ is continuous and has derivatives of all orders. These derivatives can be obtained by interchanging the derivative and integral operators. If $f$ is a pmf, replace the integral by a sum.


Hence

$$\frac{\partial}{\partial \eta_i} \int g(y) f(y|\eta)dy = \int g(y) \frac{\partial}{\partial \eta_i} f(y|\eta)dy$$

(3.11)

if $f$ is a pdf and

$$\frac{\partial}{\partial \eta_i} \sum g(y) f(y|\eta) = \sum g(y) \frac{\partial}{\partial \eta_i} f(y|\eta)$$

(3.12)
CHAPTER 3. EXPONENTIAL FAMILIES 103

if \( f \) is a pmf.

**Remark 3.2.** If \( Y \) comes from an exponential family (3.1), then the derivative and integral (or sum) operators can be interchanged. Hence

\[
\frac{\partial}{\partial \theta_i} \int \ldots \int g(y) f(y|\theta) dy = \int \ldots \int g(y) \frac{\partial}{\partial \theta_i} f(y|\theta) dx
\]

for any function \( g(y) \) with \( E_{\theta}[g(Y)] < \infty \).

The behavior of \((t_1(Y), \ldots, t_k(Y))\) will be of considerable interest in later chapters. The following result is in Lehmann (1983, p. 29-30). Also see Johnson, Ladella, and Liu (1979).

**Theorem 3.3.** Suppose that \( Y \) comes from an exponential family (3.10). Then a)

\[
E(t_i(Y)) = \frac{\partial}{\partial \eta_i} a(\eta) = - \frac{\partial}{\partial \eta_i} \log(b(\eta))
\]

and b)

\[
\text{Cov}(t_i(Y), t_j(Y)) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\eta) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\eta)).
\]

Notice that \( i = j \) gives the formula for \( \text{VAR}(t_i(Y)) \).

**Proof.** The proof will be for pdfs. For pmfs replace the integrals by sums. Use Lemma 3.2 with \( g(y) = 1 \ \forall y \). a) Since \( 1 = \int f(y|\eta) dy \),

\[
0 = \frac{\partial}{\partial \eta_i} 1 = \frac{\partial}{\partial \eta_i} \int h(y) \exp \left[ \sum_{m=1}^{k} \eta_m t_m(y) - a(\eta) \right] dy
\]

\[
= \int h(y) \frac{\partial}{\partial \eta_i} \exp \left[ \sum_{m=1}^{k} \eta_m t_m(y) - a(\eta) \right] dy
\]

\[
= \int h(y) \exp \left[ \sum_{m=1}^{k} \eta_m t_m(y) - a(\eta) \right] (t_i(y) - \frac{\partial}{\partial \eta_i} a(\eta)) dy
\]

\[
= \int (t_i(y) - \frac{\partial}{\partial \eta_i} a(\eta)) f(y|\eta) dy
\]

\[
= E(t_i(Y)) - \frac{\partial}{\partial \eta_i} a(\eta).
\]
b) Similarly,
\[ 0 = \int h(y) \frac{\partial^2}{\partial \eta_i \partial \eta_j} \exp \left[ \sum_{m=1}^{k} \eta_m t_m(y) - a(\eta) \right] dy. \]

From the proof of a),
\[ 0 = \int h(y) \frac{\partial}{\partial \eta_j} \left[ \exp \left[ k \sum_{i=1}^{m=1} \eta_i t_i(y) - a(\eta) \right] \left( \frac{\partial}{\partial \eta_i} a(\eta) \right) \right] dy \]
\[ = \int h(y) \exp \left[ k \sum_{m=1}^{m=1} \eta_m t_m(y) - a(\eta) \right] \left( \frac{\partial}{\partial \eta_i} a(\eta) \right) dy \]
\[ - \int h(y) \exp \left[ k \sum_{m=1}^{m=1} \eta_m t_m(y) - a(\eta) \right] \left( \frac{\partial}{\partial \eta_j} a(\eta) \right) dy \]
\[ = \text{Cov}(t_i(Y), t_j(Y)) - \frac{\partial^2}{\partial \eta_i \partial \eta_j} a(\eta) \]
since \( \frac{\partial}{\partial \eta_j} a(\eta) = E(t_j(Y)) \) by a). QED

**Theorem 3.4.** Suppose that \( Y \) comes from an exponential family (3.10), and let \( T = (t_1(Y), ..., t_k(Y)) \). Then for any \( \eta \) in the interior of \( \Omega \), the moment generating function of \( T \) is
\[ m_T(s) = \exp[a(\eta + s) - a(\eta)] = \exp[a(\eta + s)]/\exp[a(\eta)]. \]

**Proof.** The proof will be for pdfs. For pmfs replace the integrals by sums. Since \( \eta \) is in the interior of \( \Omega \) there is a neighborhood of \( \eta \) such that if \( s \) is in that neighborhood, then \( \eta + s \in \Omega \). (Hence there exists a \( \delta > 0 \) such that if \( ||s|| < \delta \), then \( \eta + s \in \Omega \).) For such \( s \) (see Definition 2.25),
\[ m_T(s) = E[\exp(\sum_{i=1}^{k} s_i t_i(Y))] \equiv E(g(Y)). \]

It is important to notice that we are finding the mgf of \( T \), not the mgf of \( Y \). Hence we can use the kernel method of Section 1.5 to find \( E(g(Y)) = \int g(y) f(y) dy \) without finding the joint distribution of \( T \). So
\[ m_T(s) = \int \exp(\sum_{i=1}^{k} s_i t_i(y)) h(y) \exp \left[ \sum_{i=1}^{k} \eta_i t_i(y) - a(\eta) \right] dy \]
\[
= \int h(y) \exp \left[ \sum_{i=1}^{k} (\eta_i + s_i) t_i(y) - a(\eta + s) + a(\eta + s) - a(\eta) \right] dy
\]

\[
= \exp[a(\eta + s) - a(\eta)] \int h(y) \exp \left[ \sum_{i=1}^{k} (\eta_i + s_i) t_i(y) - a(\eta + s) \right] dy
\]

\[
= \exp[a(\eta + s) - a(\eta)] \int f(y|\eta + s) dy = \exp[a(\eta + s) - a(\eta)]
\]
since the pdf \( f(y|\eta + s) \) integrates to one. QED

**Theorem 3.5.** Suppose that \( Y \) comes from an exponential family (3.10), and let \( T = (t_1(Y), ..., t_k(Y)) = (T_1, ..., T_k) \). Then the distribution of \( T \) is an exponential family with

\[
f(t|\eta) = h^*(t) \exp \left[ \sum_{i=1}^{k} \eta_i t_i - a(\eta) \right]
\]

**Proof.** See Lehmann (1986, p. 58).

The main point of this section is that \( T \) is well behaved even if \( Y \) is not. For example, if \( Y \) follows a one sided stable distribution, then \( Y \) is from an exponential family, but \( E(Y) \) does not exist. However the mgf of \( T \) exists, so all moments of \( T \) exist. If \( Y_1, ..., Y_n \) are iid from a one parameter exponential family, then \( T \equiv T_n = \sum_{i=1}^{n} t(Y_i) \) is from a one parameter exponential family. One way to find the distribution function of \( T \) is to find the distribution of \( t(Y) \) using the transformation method, then find the distribution of \( \sum_{i=1}^{n} t(Y_i) \) using moment generating functions or Theorems 2.17 and 2.18. This technique results in the following two theorems. Notice that \( T \) often has a gamma distribution.

**Theorem 3.6.** Let \( Y_1, ..., Y_n \) be iid from the given one parameter exponential family and let \( T \equiv T_n = \sum_{i=1}^{n} t(Y_i) \).

a) If \( Y_i \) is from a binomial \((k, \rho)\) distribution with \( k \) known, then \( t(Y) = Y \sim \text{BIN}(k, \rho) \) and \( T_n = \sum_{i=1}^{n} Y_i \sim \text{BIN}(nk, \rho) \).

b) If \( Y \) is from an exponential \((\lambda)\) distribution then, \( t(Y) = Y \sim \text{EXP}(\lambda) \) and \( T_n = \sum_{i=1}^{n} Y_i \sim G(n, \lambda) \).

c) If \( Y \) is from a gamma \((\nu, \lambda)\) distribution with \( \nu \) known, then \( t(Y) = Y \sim G(\nu, \lambda) \) and \( T_n = \sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda) \).
d) If $Y$ is from a geometric ($\rho$) distribution, then $t(Y) = Y \sim \text{geom}(\rho)$ and $T_n = \sum_{i=1}^n Y_i \sim \text{NB}(n, \rho)$ where NB stands for negative binomial.

e) If $Y$ is from a negative binomial ($r, \rho$) distribution with $r$ known, then $t(Y) = Y \sim \text{NB}(r, \rho)$ and $T_n = \sum_{i=1}^n Y_i \sim \text{NB}(nr, \rho)$.

f) If $Y$ is from a normal ($\mu, \sigma^2$) distribution with $\sigma^2$ known, then $t(Y) = Y \sim N(\mu, \sigma^2)$ and $T_n = \sum_{i=1}^n Y_i \sim N(n\mu, n\sigma^2)$.

g) If $Y$ is from a normal ($\mu, \sigma^2$) distribution with $\mu$ known, then $t(Y) = (Y - \mu)^2 \sim G(1/2, 2\sigma^2)$ and $T_n = \sum_{i=1}^n (Y_i - \mu)^2 \sim G(n/2, 2\sigma^2)$.

h) If $Y$ is from a Poisson ($\theta$) distribution, then $t(Y) = Y \sim \text{POIS}(\theta)$ and $T_n = \sum_{i=1}^n Y_i \sim \text{POIS}(n\theta)$.

**Theorem 3.7.** Let $Y_1, ..., Y_n$ be iid from the given one parameter exponential family and let $T \equiv T_n = \sum_{i=1}^n t(Y_i)$.

a) If $Y_i$ is from a Burr type XII ($\phi, \lambda$) distribution with $\phi$ known, then $t(Y) = \log(1 + Y^\phi) \sim \text{EXP}(\lambda)$ and $T_n = \sum \log(1 + Y_i^\phi) \sim G(n, \lambda)$.

b) If $Y$ is from a chi($p, \sigma$) distribution with $p$ known, then $t(Y) = Y^2 \sim G(p/2, 2\sigma^2)$ and $T_n = \sum Y_i^2 \sim G(np/2, 2\sigma^2)$.

c) If $Y$ is from a double exponential ($\theta, \lambda$) distribution with $\theta$ known, then $t(Y) = |Y - \theta| \sim \text{EXP}(\lambda)$ and $T_n = \sum |Y_i - \theta| \sim G(n, \lambda)$.

d) If $Y$ is from a two parameter exponential ($\theta, \lambda$) distribution with $\theta$ known, then $t(Y) = Y_i - \theta \sim \text{EXP}(\lambda)$ and $T_n = \sum_{i=1}^n (Y_i - \theta) \sim G(n, \lambda)$.

e) If $Y$ is from a generalized negative binomial GNB($\mu, \kappa$) distribution with $\kappa$ known, then $T_n = \sum_{i=1}^n Y_i \sim \text{GNB}(n\mu, n\kappa)$.

f) If $Y$ is from a half normal ($\mu, \sigma^2$) distribution with $\mu$ known, then $t(Y) = (Y - \mu)^2 \sim G(1/2, 2\sigma^2)$ and $T_n = \sum_{i=1}^n (Y_i - \mu)^2 \sim G(n/2, 2\sigma^2)$.

g) If $Y$ is from an inverse Gaussian IG($\theta, \lambda$) distribution with $\lambda$ known, then $T_n = \sum_{i=1}^n Y_i \sim IG(n\theta, n^2\lambda)$.
h) If $Y$ is from an inverted gamma ($\nu, \lambda$) distribution with $\nu$ known, then $t(Y) = 1/Y \sim G(\nu, \lambda)$ and $T_n = \sum_{i=1}^n 1/Y_i \sim G(n\nu, \lambda)$.

i) If $Y$ is from a lognormal ($\mu, \sigma^2$) distribution with $\mu$ known, then $t(Y) = (\log(Y) - \mu)^2 \sim G(1/2, 2\sigma^2)$ and $T_n = \sum_{i=1}^n (\log(Y_i) - \mu)^2 \sim G(n/2, 2\sigma^2)$.

j) If $Y$ is from a lognormal ($\mu, \sigma^2$) distribution with $\sigma^2$ known, then $t(Y) = \log(Y) \sim N(\mu, \sigma^2)$ and $T_n = \sum_{i=1}^n \log(Y_i) \sim N(n\mu, n\sigma^2)$.

k) If $Y$ is from a Maxwell-Boltzmann ($\mu, \sigma$) distribution with $\mu$ known, then $t(Y) = (Y - \mu)^2 \sim G(3/2, 2\sigma^2)$ and $T_n = \sum_{i=1}^n (Y_i - \mu)^2 \sim G(3n/2, 2\sigma^2)$.

l) If $Y$ is from a one sided stable ($\sigma$) distribution, then $t(Y) = 1/Y \sim G(1/2, 2/\sigma)$ and $T_n = \sum_{i=1}^n 1/Y_i \sim G(n/2, 2/\sigma)$.  

m) If $Y$ is from a Pareto $(\sigma, \lambda)$ distribution with $\sigma$ known, then $t(Y) = \log(Y/\sigma) \sim \text{EXP}(\lambda)$ and $T_n = \sum_{i=1}^{n} \log(Y_i/\sigma) \sim G(n, \lambda)$.

n) If $Y$ is from a power $(\lambda)$ distribution, then $t(Y) = -\log(Y) \sim \text{EXP}(\lambda)$ and $T_n = \sum_{i=1}^{n} [-\log(Y_i)] \sim G(n, \lambda)$.

o) If $Y$ is from a Rayleigh $(\mu, \sigma)$ distribution with $\mu$ known, then $t(Y) = (Y - \mu)^2 \sim \text{EXP}(2\sigma^2)$ and $T_n = \sum_{i=1}^{n} (Y_i - \mu)^2 \sim G(n, 2\sigma^2)$.

p) If $Y$ is from a Topp-Leone $(\nu)$ distribution, then $t(Y) = -\log(2Y - Y^2) \sim \text{EXP}(1/\nu)$ and $T_n = \sum_{i=1}^{n} [-\log(2Y_i - Y_i^2)] \sim G(n, 1/\nu)$.

q) If $Y$ is from a truncated extreme value $(\lambda)$ distribution, then $t(Y) = e^{Y - 1} \sim \text{EXP}(\lambda)$ and $T_n = \sum_{i=1}^{n} (e^{Y_i} - 1) \sim G(n, \lambda)$.

r) If $Y$ is from a Weibull $(\phi, \lambda)$ distribution with $\phi$ known, then $t(Y) = Y^\phi \sim \text{EXP}(\lambda)$ and $T_n = \sum_{i=1}^{n} Y_i^\phi \sim G(n, \lambda)$.

3.3 Summary

1) Given the pmf or pdf of $Y$ or that $Y$ is a brand name distribution from Section 1.7, know how to show whether $Y$ belongs to an exponential family or not using

$$f(y|\theta) = h(y)c(\theta)\exp[\sum_{i=1}^{k} w_i(\theta)t_i(y)]$$

or

$$f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)].$$

Tips: a) The F, Cauchy, logistic, t and uniform distributions can not be put in form 1) and so are not exponential families.

b) If the support depends on an unknown parameter, then the family is not an exponential family.

2) If $Y$ belongs to an exponential family, set $\eta_i = w_i(\theta)$ and typically set the natural parameter space $\Omega$ equal to the cross product of the ranges of $\eta_i$. For a kP-REF, $\Omega$ is an open set that is typically a cross product of $k$ open intervals. For a 1P-REF, $\Omega$ is an open interval.

3) If $Y$ belongs to an exponential family, know how to show whether $Y$ belongs to a $k$-dimensional regular exponential family. Suppose $\Theta$ is found after deleting values of $\theta$ from $\Theta_U$ such that the pdf or pmf is undefined and such that $w_i(\theta)$ is undefined. Assume $\dim(\Theta) = k = \dim(\Omega)$. Typically we assume that condition E1) is true: that is, $\Omega$ is given by (3.7). Then
typically $\Omega$ is the cross product of the ranges of $\eta_i = w_i(\theta)$. If $\Omega$ contains a $k$-dimensional rectangle, e.g. if $\Omega$ is a $k$-dimensional open set, then $\eta_1,...,\eta_k$ do not satisfy a linearity constraint. In particular, if $p = 2$ you should be able to show whether $\eta_1$ and $\eta_2$ satisfy a linearity constraint and whether $t_1$ and $t_2$ satisfy a linearity constraint, to plot $\Omega$ and to determine whether the natural parameter space $\Omega$ contains a 2-dimensional rectangle.

Tips: a) If $\dim(\Theta) = k = \dim(\Omega)$, then the usual parameterization has $k$ parameters, and so does the kP-REF parameterization.

b) If one of the two parameters is known for the usual parameterization, then the family will often be a 1P-REF with $\eta = w(\theta)$ where $\theta$ is the unknown parameter.

c) If the family is a 2 dimensional exponential family with natural parameters $\eta_1$ and $\eta_2$, but the usual parameterization is determined by 1 parameter $\theta$, then the family is probably not regular. The $N(a\mu,b\mu^2)$ and $N(a\sigma,b\sigma^2)$ families are typical examples. If $\dim(\Theta) = j < k = \dim(\Omega)$, the family is usually not regular. If $\eta_1$ is a simple function of $\eta_2$, then the “natural parameter space” is not a cross product of the ranges of $\eta_i$. For example, if $\eta_1 = \eta_2^2$, then the “natural parameter space” is a parabola, and is not a 2-dimensional open set, and does not contain a 2-dimensional rectangle.

3.4 Complements

Theorem 3.8. Suppose $Y$ has a pdf that is a kP-REF, and that $X = t_0(Y)$ where $t_0(y)$ is a parameter free monotone differentiable transformation. Then $X$ has a pdf that is a kP-REF.

Proof. Let $f_Y(y|\eta) = h(y)b(\eta)\exp[\sum_{i=1}^k \eta_i t_i(y)]$ where the $\eta_i$ and the $t_i(y)$ do not satisfy a linearity constraint. Then

$$f_X(x|\eta) = h(t_0^{-1}(x)) \left| \frac{dt_0^{-1}(x)}{dx} \right| b(\eta) \exp[\sum_{i=1}^k \eta_i t_i(t_0^{-1}(x))]$$

$$= h^*(x)b(\eta)\exp[\sum_{i=1}^k \eta_i t_i^*(x)]$$

which is an exponential family. The family is a kP-REF since the $\eta_i$ and natural parameter space $\Omega$ are the same as for $Y$, and neither the $\eta_i$ nor the $t_i^*(x)$ satisfy a linearity constraint. To see this, note that if the $t_i^*(x)$ satisfy a linearity constraint, then $\sum_{i=1}^k a_i t_i(t_0^{-1}(x)) = \sum_{i=1}^k a_i t_i(y) = c$ where not all $a_i = 0$. But this contradicts the fact that the $t_i(y)$ do not satisfy a linearity constraint. QED
CHAPTER 3. EXPONENTIAL FAMILIES

If \( Y \) is a \( k \)-parameter exponential family, then \( X = t_0(Y) \) is an exponential family of dimension no larger than \( k \) since
\[
    f_X(x|\theta) = h(t_0^{-1}(x)) \left| \frac{dh_0^{-1}(x)}{dx} \right| c(\theta) \exp \left\{ \sum_{i=1}^{k} w_i(\theta)t_i(t_0^{-1}(x)) \right\}.
\]

Notice that for a 1-parameter exponential family with \( t_0(y) \equiv t(y) \), the above result implies that \( t(Y) \) has a 1-parameter exponential family. This result is a special case of Theorem 3.5. The chi distribution and log gamma distribution are 2P-REFs since they are transformations of the gamma distribution. The lognormal distribution is a 2P-REF since it is a transformation of the normal distribution. The power and truncated extreme value distributions are 1P-REFs since they are transformations of the exponential distributions.

**Example 3.9.** Following Barndorff-Nielsen (1978, p. 117), if \( Y \) has an inverse Gaussian distribution, \( Y \sim IG(\theta, \lambda) \), then the pdf of \( Y \) is
\[
    f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp \left\{ -\frac{(y-\theta)^2}{2\theta^2y} \right\}
\]
where \( y, \theta, \lambda > 0 \).

Notice that
\[
    f(y) = \sqrt{\frac{\lambda}{2\pi}} e^{\lambda/y} \sqrt{\frac{1}{y^3}} I(y > 0) \exp \left\{ -\frac{\lambda}{2\theta^2y} - \frac{\lambda}{2y} \right\}
\]
is a two parameter exponential family.

Another parameterization of the inverse Gaussian distribution takes \( \theta = \sqrt{\lambda/\psi} \) so that
\[
    f(y) = \sqrt{\frac{\lambda}{2\pi}} e^{\sqrt{\lambda/\psi}} \sqrt{\frac{1}{y^3}} I(y > 0) \exp \left\{ -\frac{\psi}{2y} - \frac{\lambda}{2y} \right\},
\]
where \( \lambda > 0 \) and \( \psi \geq 0 \). Here \( \Theta = (0, \infty) \times [0, \infty), \eta_1 = -\psi/2, \eta_2 = -\lambda/2 \) and \( \Omega = (-\infty, 0) \times (-\infty, 0) \). Since \( \Omega \) is not an open set, this is a 2 parameter full exponential family that is not regular. If \( \psi \) is known then \( Y \) is a 1P–REF, but if \( \lambda \) is known then \( Y \) is a one parameter full exponential family. When \( \psi = 0 \), \( Y \) has a one sided stable distribution.

The following chapters show that exponential families can be used to simplify the theory of sufficiency, MLEs, UMVUEs, UMP tests and large sample
CHAPTER 3. EXPONENTIAL FAMILIES

theory. Barndorff-Nielsen (1982) and Olive (2008) are useful introductions to exponential families. Also see Bühler and Sehr (1987). Interesting subclasses of exponential families are given by Rahman and Gupta (1993) and Sankaran and Gupta (2005). Most statistical inference texts at the same level as this text also cover exponential families. History and references for additional topics (such as finding conjugate priors in Bayesian statistics) can be found in Lehmann (1983, p. 70), Brown (1986) and Barndorff-Nielsen (1978, 1982).


3.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

3.1*. Show that each of the following families is a 1P–REF by writing the pdf or pmf as a one parameter exponential family, finding \( \eta = w(\theta) \) and by showing that the natural parameter space \( \Omega \) is an open interval.

a) The binomial \((k; \rho)\) distribution with \( k \) known and \( \rho \in \Theta = (0, 1) \).
b) The exponential \((\lambda)\) distribution with \( \lambda \in \Theta = (0, \infty) \).
c) The Poisson \((\theta)\) distribution with \( \theta \in \Theta = (0, \infty) \).
d) The half normal \((\mu, \sigma^2)\) distribution with \( \mu \) known and \( \sigma^2 \in \Theta = (0, \infty) \).

3.2*. Show that each of the following families is a 2P–REF by writing the pdf or pmf as a two parameter exponential family, finding \( \eta_i = w_i(\theta) \) for \( i = 1, 2 \) and by showing that the natural parameter space \( \Omega \) is a cross product of two open intervals.

a) The beta \((\delta, \nu)\) distribution with \( \Theta = (0, \infty) \times (0, \infty) \).
b) The chi \((p, \sigma)\) distribution with \( \Theta = (0, \infty) \times (0, \infty) \).
c) The gamma \((\nu, \lambda)\) distribution with \( \Theta = (0, \infty) \times (0, \infty) \).
d) The lognormal \((\mu, \sigma^2)\) distribution with \( \Theta = (-\infty, \infty) \times (0, \infty) \).
3.3. Show that each of the following families is a 1P–REF by writing the
pdf or pmf as a one parameter exponential family, finding \( \eta = w(\theta) \) and by
showing that the natural parameter space \( \Omega \) is an open interval.

a) The generalized negative binomial \((\mu, \kappa)\) distribution if \( \kappa \) is known.
b) The geometric \((\rho)\) distribution.
c) The logarithmic \((\theta)\) distribution.
d) The negative binomial \((r, \rho)\) distribution if \( r \) is known.
e) The one sided stable \((\sigma)\) distribution.
f) The power \((\lambda)\) distribution.
g) The truncated extreme value \((\lambda)\) distribution.
h) The Zeta \((\nu)\) distribution.

3.4*. Show that each of the following families is a 1P–REF by writing
the pdf or pmf as a one parameter exponential family, finding \( \eta = w(\theta) \) and
by showing that the natural parameter space \( \Omega \) is an open interval.

a) The \( N(\mu, \sigma^2) \) family with \( \sigma > 0 \) known.
b) The \( N(\mu, \sigma^2) \) family with \( \mu \) known and \( \sigma > 0 \).
(See Problem 3.12 for a common error.)
c) The gamma \((\nu, \lambda)\) family with \( \nu \) known.
d) The gamma \((\nu, \lambda)\) family with \( \lambda \) known.
e) The beta \((\delta, \nu)\) distribution with \( \delta \) known.
f) The beta \((\delta, \nu)\) distribution with \( \nu \) known.

3.5. Show that each of the following families is a 1P–REF by writing the
pdf or pmf as a one parameter exponential family, finding \( \eta = w(\theta) \) and by
showing that the natural parameter space \( \Omega \) is an open interval.

a) The Burr Type XII \((\phi, \lambda)\) distribution with \( \phi \) known.
b) The double exponential \((\theta, \lambda)\) distribution with \( \theta \) known.
c) The two parameter exponential \((\theta, \lambda)\) distribution with \( \theta \) known.
d) The largest extreme value \((\theta, \sigma)\) distribution if \( \sigma \) is known.
e) The smallest extreme value \((\theta, \sigma)\) distribution if \( \sigma \) is known.
f) The inverted gamma \((\nu, \lambda)\) distribution if \( \nu \) is known.
g) The Maxwell–Boltzmann \((\mu, \lambda)\) distribution if \( \mu \) is known.
h) The Pareto \((\sigma, \lambda)\) distribution if \( \sigma \) is known.
i) The Rayleigh \((\mu, \sigma)\) distribution if \( \mu \) is known.
j) The Weibull \((\phi, \lambda)\) distribution if \( \phi \) is known.
3.6. Determine whether the Pareto \((\sigma, \lambda)\) distribution is an exponential family or not.

3.7. Following Kotz and van Dorp (2004, p. 35-36), if \(Y\) has a Topp-Leone distribution, \(Y \sim TL(\nu)\), then the cdf of \(Y\) is \(F(y) = (2y - y^2)^{\nu}\) for \(\nu > 0\) and \(0 < y < 1\). The pdf of \(Y\) is

\[
f(y) = \nu(2 - 2y)(2y - y^2)^{\nu-1}
\]

for \(0 < y < 1\). Determine whether this distribution is an exponential family or not.

3.8. In Spiegel (1975, p. 210), \(Y\) has pdf

\[
f_Y(y) = \frac{2\gamma^{3/2}}{\sqrt{\pi}} y^2 \exp(-\gamma y^2)
\]

where \(\gamma > 0\) and \(y\) is real. Is \(Y\) a 1P-REF?

3.9. Let \(Y\) be a (one sided) truncated exponential \(T\text{EXP}(\lambda, b)\) random variable. Then the pdf of \(Y\) is

\[
f_Y(y|\lambda, b) = \frac{\frac{1}{\lambda} e^{-y/\lambda}}{1 - \exp(-\frac{b}{\lambda})}
\]

for \(0 < y \leq b\) where \(\lambda > 0\). If \(b\) is known, is \(Y\) a 1P-REF? (Also see O’Reilly and Rueda (2007).)

3.10. Suppose \(Y\) has a Birnbaum Saunders distribution. If \(t(Y) = \frac{Y}{\theta} + \frac{\theta}{Y} - 2\), then \(t(Y) \sim \nu^2 \chi^2_1\). If \(\theta\) is known, is this distribution a regular exponential family?

3.11. If \(Y\) has a Burr Type X distribution, then the pdf of \(Y\) is

\[
f(y) = I(y > 0) 2 \tau y e^{-y^2} (1 - e^{-y^2})^{\tau-1} =
\]

\[
I(y > 0) 2y e^{-y^2} \tau \exp[(1 - \tau)(-\log(1 - e^{-y^2}))]
\]

where \(\tau > 0\) and \(-\log(1 - e^{-y^2}) \sim \text{EXP}(1/\tau)\). Is this distribution a regular exponential family?
Problems from old quizzes and exams.

3.12*. Suppose that \( X \) has a \( N(\mu, \sigma^2) \) distribution where \( \sigma > 0 \) and \( \mu \) is known. Then

\[
 f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}x^2} \exp\left[-\frac{1}{2\sigma^2}x^2 + \frac{1}{\sigma^2}\mu x\right].
\]

Let \( \eta_1 = -1/(2\sigma^2) \) and \( \eta_2 = 1/\sigma^2 \). Why is this parameterization not the regular exponential family parameterization? (Hint: show that \( \eta_1 \) and \( \eta_2 \) satisfy a linearity constraint.)

3.13. Let \( X_1, \ldots, X_n \) be iid \( N(\mu, \gamma_o^2 \mu^2) \) random variables where \( \gamma_o^2 > 0 \) is known and \( \mu > 0 \).

a) Find the distribution of \( \sum_{i=1}^n X_i \).

b) Find \( E[(\sum_{i=1}^n X_i)^2] \).

c) The pdf of \( X \) is

\[
 f_X(x|\mu) = \frac{1}{\gamma_o^2 \mu \sqrt{2\pi}} \exp \left[ -\frac{(x-\mu)^2}{2\gamma_o^2 \mu^2} \right].
\]

Show that the family \( \{f(x|\mu) : \mu > 0\} \) is a two parameter exponential family.

d) Show that the “natural parameter space” is a parabola. You may assume that \( \eta_i = w_i(\mu) \). Is this family a regular exponential family?

3.14. Let \( X_1, \ldots, X_n \) be iid \( N(\alpha \sigma, \sigma^2) \) random variables where \( \alpha \) is a known real number and \( \sigma > 0 \).

a) Find \( E[\sum_{i=1}^n X_i^2] \).

b) Find \( E[(\sum_{i=1}^n X_i)^2] \).

c) Show that the family \( \{f(x|\sigma) : \sigma > 0\} \) is a two parameter exponential family.

d) Show that the “natural parameter space” is a parabola. You may assume that \( \eta_i = w_i(\sigma) \). Is this family a regular exponential family?
Chapter 4

Sufficient Statistics

4.1 Statistics and Sampling Distributions

Suppose that the data $Y_1, ..., Y_n$ is drawn from some population. The observed data is $Y_1 = y_1, ..., Y_n = y_n$ where $y_1, ..., y_n$ are numbers. Let $y = (y_1, ..., y_n)$. Real valued functions $T(y_1, ..., y_n) = T(y)$ are of interest as are vector valued functions $T(y) = (T_1(y), ..., T_k(y))$. Sometimes the data $Y_1, ..., Y_n$ are random vectors. Again interest is in functions of the data. Typically the data has a joint pdf or pmf $f(y_1, ..., y_n | \theta)$ where the vector of unknown parameters is $\theta = (\theta_1, ..., \theta_k)$. (In the joint pdf or pmf, the $y_1, ..., y_n$ are dummy variables, not the observed data.)

**Definition 4.1.** A statistic is a function of the data that does not depend on any unknown parameters. The probability distribution of the statistic is called the **sampling distribution** of the statistic.

Let the data $Y = (Y_1, ..., Y_n)$ where the $Y_i$ are random variables. If $T(y_1, ..., y_n)$ is a real valued function whose domain includes the sample space $\mathcal{Y}$ of $Y$, then $W = T(Y_1, ..., Y_n)$ is a statistic provided that $T$ does not depend on any unknown parameters. The data comes from some probability distribution and the statistic is a random variable and hence also comes from some probability distribution. To avoid confusing the distribution of the statistic with the distribution of the data, the distribution of the statistic is called the sampling distribution of the statistic. If the observed data is $Y_1 = y_1, ..., Y_n = y_n$, then the observed value of the statistic is $W = w = T(y_1, ..., y_n)$. Similar remarks apply when the statistic $T$ is vector valued and
when the data $Y_1, \ldots, Y_n$ are random vectors.

Often $Y_1, \ldots, Y_n$ will be iid and statistics of the form

$$\sum_{i=1}^{n} a_i Y_i \quad \text{and} \quad \sum_{i=1}^{n} t(Y_i)$$

are especially important. Chapter 10 and Theorems 2.17, 2.18, 3.6 and 3.7 are useful for finding the sampling distributions of some of these statistics when the $Y_i$ are iid from a given brand name distribution that is usually an exponential family. The following example lists some important statistics.

**Example 4.1.** Let the $Y_1, \ldots, Y_n$ be the data.

a) The *sample mean*

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \quad (4.1)$$

b) The *sample variance*

$$S^2 \equiv S_n^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n - 1} = \frac{\sum_{i=1}^{n} Y_i^2 - n(\bar{Y})^2}{n - 1}. \quad (4.2)$$

c) The *sample standard deviation* $S \equiv S_n = \sqrt{S_n^2}$.

d) If the data $Y_1, \ldots, Y_n$ is arranged in ascending order from smallest to largest and written as $Y_{(1)} \leq \cdots \leq Y_{(n)}$, then $Y_{(i)}$ is the $i$th order statistic and the $Y_{(i)}$’s are called the *order statistics*.

e) The *sample median*

$$\text{MED}(n) = Y_{((n+1)/2)} \quad \text{if } n \text{ is odd}, \quad (4.3)$$

$$\text{MED}(n) = \frac{Y_{(n/2)} + Y_{(n/2)+1}}{2} \quad \text{if } n \text{ is even.}$$

f) The *sample median absolute deviation* is

$$\text{MAD}(n) = \text{MED}(|Y_i - \text{MED}(n)|, \ i = 1, \ldots, n). \quad (4.4)$$

g) The *sample maximum*

$$\max(n) = Y_{(n)} \quad (4.5)$$

and the observed max $y_{(n)}$ is the largest value of the observed data.
h) The sample minimum

\[ \min(n) = Y_{(1)} \] (4.6)

and the observed \( y_{(1)} \) is the smallest value of the observed data.

**Example 4.2.** Usually the term “observed” is dropped. Hence below “data” is “observed data”, “observed order statistics” is “order statistics” and “observed value of MED(n)” is “MED(n).”

Let the data be 9, 2, 7, 4, 1, 6, 3, 8, 5 (so \( Y_1 = y_1 = 9, \ldots, Y_9 = y_9 = 5 \)). Then the order statistics are 1, 2, 3, 4, 5, 6, 7, 8, 9. Then MED(n) = 5 and MAD(n) = 2 = MED{0, 1, 1, 2, 2, 3, 3, 4, 4}.

**Example 4.3.** Let the \( Y_1, \ldots, Y_n \) be iid \( N(\mu, \sigma^2) \). Then

\[ T_n = \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{n} \]

is a statistic iff \( \mu \) is known.

The following theorem is extremely important and the proof follows Rice (1988, p. 171-173) closely.

**Theorem 4.1.** Let the \( Y_1, \ldots, Y_n \) be iid \( N(\mu, \sigma^2) \).

a) The sample mean \( \overline{Y} \sim N(\mu, \sigma^2/n) \).

b) \( \overline{Y} \) and \( S^2 \) are independent.

c) \((n-1)S^2/\sigma^2 \sim \chi_{n-1}^2\). Hence \( \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \sim \sigma^2 \chi_{n-1}^2 \).

d) \( \frac{\sqrt{n}(\overline{Y} - \mu)}{S/\sqrt{n}} = \frac{(\overline{Y} - \mu)}{S/\sqrt{n}} \sim t_{n-1} \).

**Proof.** a) This result follows from Theorem 2.17e.

b) The moment generating function of \((\overline{Y}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})\) is

\[ m(s, t_1, \ldots, t_n) = E(\exp[s\overline{Y} + t_1(Y_1 - \overline{Y}) + \cdots + t_n(Y_n - \overline{Y})]) \]

By Theorem 2.22, \( \overline{Y} \) and \((Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})\) are independent if

\[ m(s, t_1, \ldots, t_n) = m_{\overline{Y}}(s) m(t_1, \ldots, t_n) \]

where \( m_{\overline{Y}}(s) \) is the mgf of \( \overline{Y} \) and \( m(t_1, \ldots, t_n) \) is the mgf of \((Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})\).

Now

\[ \sum_{i=1}^{n} t_i(Y_i - \overline{Y}) = \sum_{i=1}^{n} t_iY_i - \overline{Y} t = \sum_{i=1}^{n} t_iY_i - \overline{Y} \sum_{i=1}^{n} t_i \]
and thus
\[ s\bar{Y} + \sum_{i=1}^{n} t_i (Y_i - \bar{Y}) = \sum_{i=1}^{n} \left( \frac{s}{n} + (t_i - \bar{t}) \right)Y_i = \sum_{i=1}^{n} a_i Y_i. \]

Now \( \sum_{i=1}^{n} a_i = \sum_{i=1}^{n} \left[ \frac{s}{n} + (t_i - \bar{t}) \right] = s \) and
\[ \sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n} \left[ \frac{s^2}{n^2} + 2 \frac{s}{n} (t_i - \bar{t}) + (t_i - \bar{t})^2 \right] = \frac{s^2}{n} + \sum_{i=1}^{n} (t_i - \bar{t})^2. \]

Hence
\[ m(s, t_1, ..., t_n) = E(\exp[s\bar{Y} + \sum_{i=1}^{n} t_i (Y_i - \bar{Y})]) = E[\exp(\sum_{i=1}^{n} a_i Y_i)] \]
\[ = m_{Y_1, ..., Y_n}(a_1, ..., a_n) = \prod_{i=1}^{n} m_{Y_i}(a_i) \]

since the \( Y_i \) are independent. Now
\[ \prod_{i=1}^{n} m_{Y_i}(a_i) = \prod_{i=1}^{n} \exp \left( \mu a_i + \frac{\sigma^2}{2} a_i^2 \right) = \exp \left( \mu \sum_{i=1}^{n} a_i + \frac{\sigma^2}{2} \sum_{i=1}^{n} a_i^2 \right) \]
\[ = \exp \left[ \mu s + \frac{\sigma^2}{2} s_n^2 + \frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2 \right] \]
\[ = \exp \left[ \mu s + \frac{\sigma^2}{2} s_n^2 \right] \exp \left[ \frac{\sigma^2}{2} \sum_{i=1}^{n} (t_i - \bar{t})^2 \right]. \]

Now the first factor is the mgf of \( \bar{Y} \) and the second factor is \( m(t_1, ..., t_n) = m(0, t_1, ..., t_n) \) since the mgf of the marginal is found from the mgf of the joint distribution by setting all terms not in the marginal to 0 (ie set \( s = 0 \) in \( m(s, t_1, ..., t_n) \) to find \( m(t_1, ..., t_n) \)). Hence the mgf factors and
\[ \bar{Y} \perp (Y_1 - \bar{Y}, ..., Y_n - \bar{Y}). \]

Since \( S^2 \) is a function of \((Y_1 - \bar{Y}, ..., Y_n - \bar{Y})\), it is also true that \( \bar{Y} \perp S^2 \).
c) \((Y_i - \mu)/\sigma \sim N(0, 1)\) so \((Y_i - \mu)^2/\sigma^2 \sim \chi_1^2\) and
\[
\frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 \sim \chi_n^2.
\]

Now
\[
\sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y} + \bar{Y} - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2.
\]

Hence
\[
W = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \mu)^2 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 = U + V.
\]

Since \(U \perp V\) by b), \(m_W(t) = m_U(t) m_V(t)\). Since \(W \sim \chi_n^2\) and \(V \sim \chi_1^2\),
\[
m_U(t) = \left(1 - 2t\right)^{-n/2} \quad m_V(t) = \left(1 - 2t\right)^{-1/2} = (1 - 2t)^{-(n-1)/2}
\]
which is the mgf of a \(\chi_{n-1}^2\) distribution.

\[d\]
\[
Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1),
\]

and
\[
S^2/\sigma^2 = \frac{(n-1)S^2}{n-1} \sim \frac{1}{n-1} \chi_{n-1}^2.
\]

Suppose \(Z \sim N(0, 1), X \sim \chi_n^2\) and \(Z \perp X\). Then \(Z/\sqrt{X/(n-1)} \sim t_{n-1}\).

Hence
\[
\frac{\sqrt{n}(\bar{Y} - \mu)}{S} = \frac{\sqrt{n}(\bar{Y} - \mu)/\sigma}{\sqrt{S^2/\sigma^2}} \sim t_{n-1}.
\]

QED

**Theorem 4.2.** Let the \(Y_1, ..., Y_n\) be iid with cdf \(F_Y\) and pdf \(f_Y\).

a) The pdf of \(T = Y_{(n)}\) is
\[
f_{Y_{(n)}}(t) = n[F_Y(t)]^{n-1}f_Y(t).
\]
b) The pdf of \( T = Y_1 \) is

\[
f_{Y_1}(t) = n(1 - F_Y(t))^{n-1}f_Y(t).
\]

c) Let \( 2 \leq r \leq n \). Then the joint pdf of \( Y_1, Y_2, ..., Y_r \) is

\[
f_{Y_1, \ldots, Y_r}(t_1, ..., t_r) = \frac{n!}{(n-r)!} [1 - F_Y(t_r)]^{n-r} \prod_{i=1}^{r} f_Y(t_i).
\]

Proof of a) and b). a) The cdf of \( Y_n \) is

\[
F_{Y_n}(t) = P(Y_n \leq t) = P(Y_1 \leq t, ..., Y_n \leq t) = \prod_{i=1}^{n} P(Y_i \leq t) = [F_Y(t)]^n.
\]

Hence the pdf of \( Y_n \) is

\[
\frac{d}{dt} F_{Y_n}(t) = \frac{d}{dt} [F_Y(t)]^n = n[F_Y(t)]^{n-1}f_Y(t).
\]

b) The cdf of \( Y_1 \) is

\[
F_{Y_1}(t) = P(Y_1 \leq t) = 1 - P(Y_1 > t) = 1 - P(Y_1 > t, ..., Y_n > t)
\]

\[
= 1 - \prod_{i=1}^{n} P(Y_i > t) = 1 - [1 - F_Y(t)]^n.
\]

Hence the pdf of \( Y_n \) is

\[
\frac{d}{dt} F_{Y_n}(t) = \frac{d}{dt} (1 - [1 - F_Y(t)]^n) = n[1 - F_Y(t)]^{n-1}f_Y(t). \quad \text{QED}
\]

To see that c) may be true, consider the following argument adapted from Mann, Schafer and Singpurwalla (1974, p. 93). Let \( \Delta t_i \) be a small positive number and notice that \( P(E) \equiv \)

\[
P(t_1 < Y_1 < t_1 + \Delta t_1, t_2 < Y_2 < t_2 + \Delta t_2, ..., t_r < Y_r < t_r + \Delta t_r)
\]

\[
= \int_{t_r}^{t_r+\Delta t_r} \cdots \int_{t_1}^{t_1+\Delta t_1} f_{Y_1, \ldots, Y_r}(w_1, ..., w_r)dw_1 \cdots dw_r
\]

\[
\approx f_{Y_1, \ldots, Y_r}(t_1, ..., t_r) \prod_{i=1}^{r} \Delta t_i.
\]
CHAPTER 4. SUFFICIENT STATISTICS

120

Since the event $E$ denotes the occurrence of no observations before $t_1$, exactly one occurrence between $t_1$ and $t_1 + \Delta t_1$, no observations between $t_1 + \Delta t_1$ and $t_2$ and so on, and finally the occurrence of $n - r$ observations after $t_r + \Delta t_r$, using the multinomial pmf shows that

$$P(E) = \frac{n!}{0!1! \cdots 0!(n-r)!} \rho_1^0 \rho_2^1 \rho_3^0 \cdots \rho_{2r-1}^0 \rho_{2r}^1 \rho_{2r+1}^{n-r}$$

where

$$\rho_{2i} = P(t_i < Y < t_i + \Delta t_i) \approx f(t_i) \Delta t_i$$

for $i = 1, \ldots, r$ and

$$\rho_{2r+1} = P(n - r \text{ Y's} > t_r + \Delta t_r) \approx (1 - F(t_r))^{n-r}.$$

Hence

$$P(E) \approx \frac{n!}{(n-r)!} (1 - F(t_r))^{n-r} \prod_{i=1}^{r} f(t_i) \prod_{i=1}^{r} \Delta t_i$$

$$\approx f_{Y_1, \ldots, Y_n}(t_1, \ldots, t_r) \prod_{i=1}^{r} \Delta t_i,$$

and result c) seems reasonable.

Example 4.4. Let $Y_1, \ldots, Y_n$ be iid from the following distributions.

a) Bernoulli$(\rho)$: Then $Y_{(1)} \sim$ Bernoulli$(\rho^n)$.

b) Geometric$(\rho)$: Then $Y_{(1)} \sim$ Geometric$(1 - (1 - \rho)^n)$.

c) Burr Type XII $(\phi, \lambda)$: Then $Y_{(1)} \sim$ Burr Type XII $(\phi, \lambda/n)$.

d) EXP$(\lambda)$: Then $Y_{(1)} \sim$ EXP$(\lambda/n)$.

e) EXP$(\theta, \lambda)$: Then $Y_{(1)} \sim$ EXP$(\theta, \lambda/n)$.

f) Pareto PAR$(\sigma, \lambda)$: Then $Y_{(1)} \sim$ PAR$(\sigma, \lambda/n)$.

g) Gompertz Gomp$(\theta, \nu)$: Then $Y_{(1)} \sim$ Gomp$(\theta, \nu)$.

h) Rayleigh $R(\mu, \sigma)$: Then $Y_{(1)} \sim$ $R(\mu, \sigma/\sqrt{n})$.

i) Truncated Extreme Value TEV$(\lambda)$: Then $Y_{(1)} \sim$ TEV$(\lambda/n)$.

j) Weibull $W(\phi, \lambda)$: Then $Y_{(1)} \sim W(\phi, \lambda/n)$.

Proof: a) $Y_i \in \{0, 1\}$ so $Y_{(1)} \in \{0, 1\}$. Hence $Y_{(1)}$ is Bernoulli, and $P(Y_{(1)} = 1) = P(Y_1 = 1, \ldots, Y_n = 1) = [P(Y_1 = 1)]^n = \rho^n$.

b) $P(Y_{(1)} \leq y) = 1 - P(Y_{(1)} > y) = 1 - [1 - F(y)]^n = 1 - [1 - (1 - (1 - \rho)^n)]^y = 1 - [(1 - \rho)^n]^y = 1 - [1 - \rho^{n+y}]$ for $y \geq 0$ which is the cdf of a Geometric$(1 - (1 - \rho)^n)$ random variable.
Parts c)-j) follow from Theorem 4.2b. For example, suppose \( Y_1, \ldots, Y_n \) are iid \( \text{EXP}(\lambda) \) with cdf \( F(y) = 1 - \exp(-y/\lambda) \) for \( y > 0 \). Then \( F_{Y^{(1)}}(t) = 1 - [1 - (1 - \exp(-t/\lambda))]^n = 1 - [\exp(-t/\lambda)]^n = 1 - \exp[-t/(\lambda/n)] \) for \( t > 0 \). Hence \( Y^{(1)} \sim \text{EXP}(\lambda/n) \).

4.2 Minimal Sufficient Statistics

For parametric inference, the pmf or pdf of a random variable \( Y \) is \( f_{\theta}(y) \) where \( \theta \in \Theta \) is unknown. Hence \( Y \) comes from a family of distributions indexed by \( \theta \), and quantities such as \( E_{\theta}(g(Y)) \) depend on \( \theta \). Since the parametric distribution is completely specified by \( \theta \), an important goal of parametric inference is finding good estimators of \( \theta \).

For example, if \( Y_1, \ldots, Y_n \) are iid \( \text{N}(\mu, \sigma^2) \), then \( \theta = (\mu, \sigma) \) is fixed but unknown, \( \theta \in \Theta = (-\infty, \infty) \times (0, \infty) \) and \( E_{\theta}(\overline{Y}) \equiv E_{(\mu, \sigma)}(\overline{Y}) = \mu \). Since \( V_{\theta}(\overline{Y}) \equiv V_{(\mu, \sigma)}(\overline{Y}) = \sigma^2/n \), \( \overline{Y} \) is a good estimator for \( \mu \) if \( n \) is large. The notation \( f_{\theta}(y) \equiv f(y|\theta) \) is also used.

The basic idea of a sufficient statistic \( T(Y) \) for \( \theta \) is that all of the information needed for inference from the data \( Y_1, \ldots, Y_n \) about the parameter \( \theta \) is contained in the statistic \( T(Y) \). For example, suppose that \( Y_1, \ldots, Y_n \) are iid binomial(1, \( \rho \)) random variables. Hence each observed \( Y_i \) is a 0 or a 1 and the observed data is an \( n \)–tuple of 0’s and 1’s, eg 0,0,1,...,0,0,1. It will turn out that \( \sum_{i=1}^n Y_i \), the number of 1’s in the \( n \)–tuple, is a sufficient statistic for \( \rho \). From Theorem 2.17a, \( \sum_{i=1}^n Y_i \sim \text{BIN}(n, \rho) \). The importance of a sufficient statistic is dimension reduction: the statistic \( \sum_{i=1}^n Y_i \) has all of the information from the data needed to perform inference about \( \rho \), and the statistic is one dimensional and thus much easier to understand than the \( n \) dimensional \( n \)–tuple of 0’s and 1’s. Also notice that all \( n \)–tuples with the same number of 1’s have the same amount of information needed for inference about \( \rho \): the \( n \)–tuples 1,1,1,0,0,0,0 and 0,1,0,0,1,0,1 both give \( \sum_{i=1}^n Y_i = 3 \).

Definition 4.2. Suppose that \( (Y_1, \ldots, Y_n) \) have a joint distribution that depends on a vector of parameters \( \theta \) for \( \theta \in \Theta \) where \( \Theta \) is the parameter space. A statistic \( T(Y_1, \ldots, Y_n) \) is a sufficient statistic for \( \theta \) if the conditional distribution of \( (Y_1, \ldots, Y_n) \) given \( T = t \) does not depend on \( \theta \) for any value of \( t \) in the support of \( T \).

Example 4.5. Suppose \( T(y) \equiv 7 \forall y \). Then \( T \) is a constant and any constant is independent of a random vector \( Y \). Hence the conditional distri-
CHAPTER 4. SUFFICIENT STATISTICS

function \( f_\theta(y|T) = f_\theta(y) \) is not independent of \( \theta \). Thus \( T \) is not a sufficient statistic.

Often \( T \) and \( Y_i \) are real valued. Then \( T(Y_1, ..., Y_n) \) is a sufficient statistic if the conditional distribution of \( Y = (Y_1, ..., Y_n) \) given \( T = t \) does not depend on \( \theta \). The following theorem provides such an effective method for showing that a statistic is a sufficient statistic that the definition should rarely be used to prove that the statistic is a sufficient statistic.

Regularity Condition F.1: If \( f(y|\theta) \) is a family of pmfs for \( \theta \in \Theta \), assume that there exists a set \( \{y_i\}_{i=1}^\infty \) that does not depend on \( \theta \in \Theta \) such that \( \sum_{i=1}^\infty f(y_i|\theta) = 1 \) for all \( \theta \in \Theta \). (This condition is usually satisfied. For example, F.1 holds if the support \( Y \) is free of \( \theta \) or if \( y = (y_1, ..., y_n) \) and \( y_i \) takes on values on a lattice such as \( y_i \in \{1, ..., \theta \} \) for \( \theta \in \{1, 2, 3, \ldots\} \).)

Theorem 4.3: Factorization Theorem. Let \( f(y|\theta) \) for \( \theta \in \Theta \) denote a family of pdfs or pmfs for \( Y \). For a family of pmfs, assume condition F.1 holds. A statistic \( T(Y) \) is a sufficient statistic for \( \theta \) iff for all sample points \( y \) and for all \( \theta \) in the parameter space \( \Theta \),

\[
f(y|\theta) = g(T(y)|\theta) \cdot h(y)
\]

where both \( g \) and \( h \) are nonnegative functions. The function \( h \) does not depend on \( \theta \) and the function \( g \) depends on \( y \) only through \( T(y) \).

Proof for pmfs. If \( T(Y) \) is a sufficient statistic, then the conditional distribution of \( Y \) given \( T(Y) = t \) does not depend on \( \theta \) for any \( t \) in the support of \( T \). Taking \( t = T(y) \) gives

\[
P_\theta(Y = y|T(Y) = T(y)) \equiv P(Y = y|T(Y) = T(y))
\]

for all \( \theta \) in the parameter space. Now

\[
\{Y = y\} \subseteq \{T(Y) = T(y)\} \quad (4.7)
\]

and \( P(A) = P(A \cap B) \) if \( A \subseteq B \). Hence

\[
f(y|\theta) = P_\theta(Y = y) = P_\theta(Y = y \text{ and } T(Y) = T(y)) = P_\theta(T(Y) = T(y)) \cdot P(Y = y|T(Y) = T(y)) = g(T(y)|\theta) h(y).
\]

Now suppose

\[
f(y|\theta) = g(T(y)|\theta) h(y)
\]
for all \( y \) and for all \( \theta \in \Theta \). Now
\[
P_{\theta}(T(Y) = t) = \sum_{\{y : T(y) = t\}} f(y|\theta) \sum_{\{y : T(y) = t\}} h(y).
\]
If \( Y = y \) and \( T(Y) = t \), then \( T(y) = t \) and \( \{Y = y\} \subseteq \{T(Y) = t\} \). Thus
\[
P_{\theta}(Y = y|T(Y) = t) = \frac{P_{\theta}(Y = y, T(Y) = t)}{P_{\theta}(T(Y) = t)} = \frac{P_{\theta}(Y = y)}{P_{\theta}(T(Y) = t)}
\]
which does not depend on \( \theta \) since the terms in the sum do not depend on \( \theta \) by condition F.1. Hence \( T \) is a sufficient statistic. QED

**Remark 4.1.** If no such factorization exists for \( T \), then \( T \) is not a sufficient statistic.

**Example 4.6.** To use factorization to show that the data \( Y = (Y_1, \ldots, Y_n) \) is a sufficient statistic, take \( T(Y) = Y \), \( g(T(y)|\theta) = f(y|\theta) \), and \( h(y) = 1 \forall y \).

**Example 4.7.** Let \( X_1, \ldots, X_n \) be iid \( N(\mu, \sigma^2) \). Then
\[
f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i) = \left[ \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \right] \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i \right)
\]
where \( \theta = (\mu, \sigma) \) and \( h(x) = 1 \). Hence \( T(X) = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i) \) is a sufficient statistic for \( (\mu, \sigma) \) or equivalently for \( (\mu, \sigma^2) \) by the factorization theorem.

**Example 4.8.** Let \( Y_1, \ldots, Y_n \) be iid binomial\((k, \rho)\) with \( k \) known and pmf
\[
f(y|\rho) = \binom{k}{y} \rho^y (1-\rho)^{k-y} I_{\{0,\ldots,k\}}(y).
\]
Then
\[
f(y|\rho) = \prod_{i=1}^{n} f(y_i|\rho) = \prod_{i=1}^{n} \left[ \binom{k}{y_i} I_{\{0,\ldots,k\}}(y_i) \right] (1-\rho)^{nk} \left( \frac{\rho}{1-\rho} \right)^{\sum_{i=1}^{n} y_i}.
\]
Hence by the factorization theorem, $\sum_{i=1}^{n} Y_i$ is a sufficient statistic.

**Example 4.9.** Suppose $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta + 1)$, $-\infty < \theta < \infty$. Notice that

$$\prod_{i=1}^{n} I_{A_i}(x_i) = I(\text{all } x_i \in A) \quad \text{and} \quad \prod_{i=1}^{n} I_{A_i}(x) = I_{\cap_{i=1}^{n} A_i}(x)$$

where the latter holds since both terms are 1 if $x$ is in all sets $A_i$ for $i = 1, \ldots, n$ and both terms are 0 otherwise. Hence $f(x|\theta) =$

$$\prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} 1 I(x_i \geq \theta) I(x_i \leq \theta + 1) = 1 I(\min(x_i) \geq \theta) I(\max(x_i) \leq \theta + 1).$$

Then $h(x) \equiv 1$ and $g(T(x)|\theta) = I(\min(x_i) \geq \theta) I(\max(x_i) \leq \theta + 1)$, and $T(x) = (X_1, X_n)$ is a sufficient statistic by the factorization theorem.

**Remark 4.2.** i) Suppose $Y_1, \ldots, Y_n$ are iid from a distribution with support $\mathcal{Y}_i \equiv \mathcal{Y}^*$ and pdf or pmf $f(y|\theta) = k(y|\theta)I(y \in \mathcal{Y}^*)$. Then $f(y|\theta) = \prod_{i=1}^{n} k(y_i|\theta) \prod_{i=1}^{n} I(y_i \in \mathcal{Y}^*)$. Now the support of $\mathbf{Y}$ is the $n$-fold cross product $\mathcal{Y} = \mathcal{Y}^* \times \cdots \times \mathcal{Y}^*$, and $I(y \in \mathcal{Y}) = \prod_{i=1}^{n} I(y_i \in \mathcal{Y}^*) = I(\text{all } y_i \in \mathcal{Y}^*)$. Thus $f(y|\theta) = \prod_{i=1}^{n} k(y_i|\theta) I(\text{all } y_i \in \mathcal{Y}^*)$.

ii) If $\mathcal{Y}^*$ does not depend on $\theta$, then $I(\text{all } y_i \in \mathcal{Y}^*)$ is part of $h(y)$. If $\mathcal{Y}^*$ does depend on unknown $\theta$, then $I(\text{all } y_i \in \mathcal{Y}^*)$ could be placed in $g(T(y)|\theta)$.

Typically $\mathcal{Y}^*$ is an interval with endpoints $a$ and $b$, not necessarily finite. For pdfs, $\prod_{i=1}^{n} I(y_i \in [a, b]) = I(a \leq y(1) < y(n) \leq b) = I[a \leq y(1)] I[y(n) \leq b]$. If both $a$ and $b$ are unknown parameters, put the middle term in $g(T(y)|\theta)$. If both $a$ and $b$ are known, put the middle term in $h(y)$. If $a$ is an unknown parameter and $b$ is known, put $I[a \leq y(1)]$ in $g(T(y)|\theta)$ and $I[y(n) \leq b]$ in $h(y)$.

iii) $\prod_{i=1}^{n} I(y_i \in (-\infty, b)) = I(y(n) < b)$.

$\prod_{i=1}^{n} I(y_i \in [a, \infty)) = I(a \leq y(1))$ et cetera.

iv) Another useful fact is that $\prod_{j=1}^{k} I(y \in A_j) = I(y \in \cap_{j=1}^{k} A_j)$.

**Example 4.10.** Try to place any part of $f(y|\theta)$ that depends on $y$ but not on $\theta$ into $h(y)$. For example, if $Y_1, \ldots, Y_n$ are iid $U(\theta_1, \theta_2)$, then $f(y|\theta) =$

$$\prod_{i=1}^{n} f(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y_i \leq \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 \leq y(1) < y(n) \leq \theta_2).$$
Then \( h(y) \equiv 1 \) and \( T(y) = (Y_1, Y_n) \) is a sufficient statistic for \((\theta_1, \theta_2)\) by factorization.

If \( \theta_1 \) or \( \theta_2 \) is known, then the above factorization works, but it is better to make the dimension of the sufficient statistic as small as possible. If \( \theta_1 \) is known, then

\[
f(y|\theta) = \frac{1}{(\theta_2 - \theta_1)^n} I(y(n) \leq \theta_2)I(\theta_1 \leq y(1))
\]

where the first two terms are \( g(T(y)|\theta_2) \) and the third term is \( h(y) \). Hence \( T(Y) = Y(n) \) is a sufficient statistic for \( \theta_2 \) by factorization. If \( \theta_2 \) is known, then

\[
f(y|\theta) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 \leq y(1))I(y(n) \leq \theta_2)
\]

where the first two terms are \( g(T(y)|\theta_1) \) and the third term is \( h(y) \). Hence \( T(Y) = Y(1) \) is a sufficient statistic for \( \theta_1 \) by factorization.

There are infinitely many sufficient statistics (see Theorem 4.8 below), but typically we want the dimension of the sufficient statistic to be as small as possible since lower dimensional statistics are easier to understand and to use for inference than higher dimensional statistics. Data reduction is extremely important and the following definition is useful.

**Definition 4.3.** Suppose that \( Y_1, ..., Y_n \) have a joint distribution that depends on a vector of parameters \( \theta \) for \( \theta \in \Theta \) where \( \Theta \) is the parameter space. A sufficient statistic \( T(Y) \) for \( \theta \) is a **minimal sufficient statistic** for \( \theta \) if \( T(Y) \) is a function of \( S(Y) \) for any other sufficient statistic \( S(Y) \) for \( \theta \).

**Remark 4.3.** A useful mnemonic is that \( S = Y \) is a sufficient statistic, and \( T \equiv T(Y) \) is a function of \( S \).

**Warning:** Complete sufficient statistics, defined below, are primarily used for the theory of uniformly minimum variance estimators, which are rarely used in applied work unless they are nearly identical to the corresponding maximum likelihood estimators.

**Definition 4.4.** Suppose that a statistic \( T(Y) \) has a pmf or pdf \( f(t|\theta) \). Then \( T(Y) \) is a **complete sufficient statistic** for \( \theta \) if \( E_{\theta}[g(T(Y))] = 0 \) for all \( \theta \) implies that \( P_{\theta}[g(T(Y)) = 0] = 1 \) for all \( \theta \). The function \( g \) can not depend on any unknown parameters.
The following two theorems are useful for finding minimal sufficient statistics.

**Theorem 4.4: Lehmann-Scheffé Theorem for Minimal Sufficient Statistics (LSM).** Let \( f(y|\theta) \) be the pmf or pdf of an iid sample \( Y \). Let \( c_{x,y} \) be a constant. Suppose there exists a function \( T(y) \) such that for any two sample points \( x \) and \( y \), the ratio \( R_{x,y}(\theta) = f(x|\theta)/f(y|\theta) = c_{x,y} \) for all \( \theta \) in \( \Theta \) iff \( T(x) = T(y) \). Then \( T(Y) \) is a minimal sufficient statistic for \( \theta \).

In the Lehmann-Scheffé Theorem, for \( R \) to be constant as a function of \( \theta \), define \( 0/0 = c_{x,y} \). Alternatively, replace \( R_{x,y}(\theta) = f(x|\theta)/f(y|\theta) = c_{x,y} \) by \( f(x|\theta) = c_{x,y}f(y|\theta) \) in the above definition.

Finding sufficient, minimal sufficient, and complete sufficient statistics is often simple for a kP-REF (k-parameter regular exponential family). If the family given by Equation (4.8) is a kP-REF, then the conditions for Theorem 4.5abcd are satisfied as are the conditions for e) if \( \eta \) is a one to one function of \( \theta \). In a), \( k \) does not need to be as small as possible. In Corollary 4.6 below, assume that both Equation (4.8) and (4.9) hold.

Note that any one to one function is onto its range. Hence if \( \eta = \tau(\theta) \) for any \( \eta \in \Omega \) where \( \tau \) is a one to one function, then \( \tau: \Theta \rightarrow \Omega \) is one to one and onto. Thus there is a one to one (and onto) inverse function \( \tau^{-1} \) such that \( \theta = \tau^{-1}(\eta) \) for any \( \theta \in \Theta \).

**Theorem 4.5: Sufficiency, Minimal Sufficiency, and Completeness of Exponential Families.** Suppose that \( Y_1, ..., Y_n \) are iid from an exponential family

\[
f(y|\theta) = h(y)c(\theta) \exp [w_1(\theta)t_1(y) + \cdots + w_k(\theta)t_k(y)]
\]

with the natural parameterization

\[
f(y|\eta) = h(y)b(\eta) \exp [\eta_1t_1(y) + \cdots + \eta_kt_k(y)]
\]

so that the joint pdf or pmf is given by

\[
f(y_1, ..., y_n|\eta) = \left( \prod_{j=1}^{n} h(y_j) \right) [b(\eta)]^n \exp [\eta \sum_{j=1}^{n} t_1(y_j) + \cdots + \eta_k \sum_{j=1}^{n} t_k(y_j)]
\]
which is a $k$-parameter exponential family. Then

$$ T(Y) = \left( \sum_{j=1}^{n} t_1(Y_j), ..., \sum_{j=1}^{n} t_k(Y_j) \right) $$

is a) a sufficient statistic for $\theta$ and for $\eta$,
b) a minimal sufficient statistic for $\eta$ if $\eta_1, ..., \eta_k$ do not satisfy a linearity constraint,
c) a minimal sufficient statistic for $\theta$ if $w_1(\theta), ..., w_k(\theta)$ do not satisfy a linearity constraint,
d) a complete sufficient statistic for $\eta$ if $\Omega$ contains a $k$–dimensional rectangle,
e) a complete sufficient statistic for $\theta$ if $\eta$ is a one to one function of $\theta$ and if $\Omega$ contains a $k$–dimensional rectangle.

**Proof.**
a) Use the factorization theorem.
b) The proof expands on remarks given in Johanson (1979, p. 3) and Lehmann (1983, p. 44). The ratio

$$ \frac{f(x|\eta)}{f(y|\eta)} = \frac{\prod_{j=1}^{n} h(x_j)}{\prod_{j=1}^{n} h(y_j)} \exp\left[\sum_{i=1}^{k} \eta_i (T_i(x) - T_i(y))\right] $$

is equal to a constant with respect to $\eta$ iff

$$ \sum_{i=1}^{k} \eta_i [T_i(x) - T_i(y)] = \sum_{i=1}^{k} \eta_i a_i = d $$

for all $\eta_i$ where $d$ is some constant and where $a_i = T_i(x) - T_i(y)$ and $T_i(x) = \sum_{j=1}^{n} t_i(x_j)$. Since the $\eta_i$ do not satisfy a linearity constraint, $\sum_{i=1}^{k} \eta_i a_i = d$ for all $\eta$ iff all of the $a_i = 0$. Hence

$$ T(Y) = (T_1(Y), ..., T_k(Y)) $$

is a minimal sufficient statistic by the Lehmann-Scheffé LSM theorem.
c) Use almost the same proof as b) with $w_i(\theta)$ in the place of $\eta_i$ and $\theta$ in the place of $\eta$. (In particular, the result holds if $\eta_i = w_i(\theta)$ for $i = 1, ..., k$ provided that the $\eta_i$ do not satisfy a linearity constraint.)
e) If $\eta = \tau(\theta)$ then $\theta = \tau^{-1}(\eta)$ and the parameters have just been renamed.
Hence $E_{\theta}[g(T)] = 0$ for all $\theta$ implies that $E_{\eta}[g(T)] = 0$ for all $\eta$, and thus $P_{\eta}[g(T(Y)) = 0] = 1$ for all $\eta$ since $T$ is a complete sufficient statistic for $\eta$ by d). Thus $P_{\theta}[g(T(Y)) = 0] = 1$ for all $\theta$, and $T$ is a complete sufficient statistic for $\theta$.

**Corollary 4.6: Completeness of a kP–REF.** Suppose that $Y_1, \ldots, Y_n$ are iid from a kP–REF ($k$-parameter regular exponential family)

$$f(y|\theta) = h(y)c(\theta)\exp\left[w_1(\theta)t_1(y) + \cdots + w_k(\theta)t_k(y)\right]$$

with $\theta \in \Theta$ and natural parameterization given by (4.9) with $\eta \in \Omega$. Then

$$T(Y) = \left(\sum_{j=1}^n t_1(Y_j), \ldots, \sum_{j=1}^n t_k(Y_j)\right)$$

a) a minimal sufficient statistic for $\theta$ and for $\eta$,

b) a complete sufficient statistic for $\theta$ and for $\eta$ if $\eta$ is a one to one function of $\theta$.

**Proof.** The result follows by Theorem 4.5 since for a kP–REF, the $w_i(\theta)$ and $\eta_i$ do not satisfy a linearity constraint and $\Omega$ contains a $k$–dimensional rectangle. QED

**Theorem 4.7: Bahadur’s Theorem.** A finite dimensional complete sufficient statistic is also minimal sufficient.

**Theorem 4.8.** A one to one function of a sufficient, minimal sufficient, or complete sufficient statistic is sufficient, minimal sufficient, or complete sufficient respectively.

Note that in a kP–REF, the statistic $T$ is $k$–dimensional and thus $T$ is minimal sufficient by Theorem 4.7 if $T$ is complete sufficient. Corollary 4.6 is useful because often you know or can show that the given family is a REF. The theorem gives a particularly simple way to find complete sufficient statistics for one parameter exponential families and for any family that is known to be REF. If it is known that the distribution is regular, find the exponential family parameterization given by Equation (4.8) or (4.9). These parameterizations give $t_1(y), \ldots, t_k(y)$. Then $T(Y) = (\sum_{j=1}^n t_1(Y_j), \ldots, \sum_{j=1}^n t_k(Y_j))$. 
Example 4.11. Let \( X_1, \ldots, X_n \) be iid \( N(\mu, \sigma^2) \). Then the \( N(\mu, \sigma^2) \) pdf is
\[
f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(-\frac{1}{2\sigma^2} \frac{\sigma^2}{t_1(x)} \frac{x}{x} + \mu \frac{\sigma^2}{t_2(x)} h(x) \right) I_{\mathbb{R}}(x),
\]
with \( \eta_1 = -0.5/\sigma^2 \) and \( \eta_2 = \mu/\sigma^2 \) if \( \sigma > 0 \). As shown in Example 3.1, this is a 2P–REF. By Corollary 4.6, \( T = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2) \) is a complete sufficient statistic for \((\mu, \sigma^2)\). The one to one functions
\[
T_2 = (\overline{X}, S^2) \quad \text{and} \quad T_3 = (\overline{X}, S)
\]
of \( T \) are also complete sufficient where \( \overline{X} \) is the sample mean and \( S \) is the sample standard deviation. \( T, T_2 \) and \( T_3 \) are minimal sufficient by Corollary 4.6 or by Theorem 4.7 since the statistics are 2 dimensional.

Example 4.12. Let \( Y_1, \ldots, Y_n \) be iid binomial\((k, \rho)\) with \( k \) known and pmf
\[
f(y|\rho) = \binom{k}{y} \rho^y (1-\rho)^{k-y} I_{\{0,\ldots,k\}}(y)
\]
\[
= \binom{k}{y} I_{\{0,\ldots,k\}}(y) \frac{(1-\rho)^k}{c(\rho) \geq 0} \exp\left[\log\left(\frac{\rho}{1-\rho}\right) y \right] I_{\mathbb{R}}(y)
\]
where \( \Theta = (0, 1) \) and \( \Omega = (-\infty, \infty) \). Notice that \( \eta = \log(\frac{\rho}{1-\rho}) \) is an increasing and hence one to one function of \( \rho \). Since this family is a 1P–REF, \( T_n = \sum_{i=1}^n t(Y_i) = \sum_{i=1}^n Y_i \) is complete sufficient statistic for \( \rho \).

Compare Examples 4.7 and 4.8 with Examples 4.11 and 4.12. The exponential family theorem gives more powerful results than the factorization theorem, but often the factorization theorem is useful for suggesting a potential minimal sufficient statistic.

Example 4.13. In testing theory, a single sample is often created by combining two independent samples of iid data. Let \( X_1, \ldots, X_n \) be iid exponential \((\theta)\) and \( Y_1, \ldots, Y_m \) iid exponential \((\theta/2)\). If the two samples are independent, then the joint pdf \( f(x, y|\theta) \) belongs to a regular one parameter exponential family with complete sufficient statistic \( T = \sum_{i=1}^n X_i + 2 \sum_{i=1}^m Y_i \). (Let \( W_i = 2Y_i \). Then the \( W_i \) and \( X_i \) are iid and Corollary 4.6 applies.)
Chapter 4. Sufficient Statistics

Rule of thumb 4.1: A k–parameter minimal sufficient statistic for a d–dimensional parameter where d < k will not be complete. In the following example d = 1 < 2 = k. (A rule of thumb is something that is frequently true but can not be used to rigorously prove something. Hence this rule of thumb can not be used to prove that the minimal sufficient statistic is not complete.)

Warning: Showing that a minimal sufficient statistic is not complete is of little applied interest since complete sufficient statistics are rarely used in applications; nevertheless, many qualifying exams in statistical inference contain such a problem.

Example 4.14, Cox and Hinckley (1974, p. 31). Let $X_1, \ldots, X_n$ be iid $N(\mu, \gamma_o^2 \mu^2)$ random variables where $\gamma_o^2 > 0$ is known and $\mu > 0$. Then this family has a one dimensional parameter $\mu$, but

$$f(x|\mu) = \frac{1}{\sqrt{2\pi \gamma_o^2 \mu^2}} \exp\left(-\frac{1}{2\gamma_o^2} \right) \exp\left(-\frac{1}{2\gamma_o^2 \mu^2} x^2 + \frac{1}{\gamma_o^2 \mu} x\right)$$

is a two parameter exponential family with $\Theta = (0, \infty)$ (which contains a one dimensional rectangle), and $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic. (Theorem 4.5 applies since the functions $1/\mu$ and $1/\mu^2$ do not satisfy a linearity constraint.) However, since $E_{\mu}(X^2) = \gamma_o^2 \mu^2 + \mu^2$ and $\sum_{i=1}^n X_i \sim N(n\mu, n\gamma_o^2 \mu^2)$ implies that

$$E_{\mu}\left[(\sum_{i=1}^n X_i)^2\right] = n\gamma_o^2 \mu^2 + n^2 \mu^2,$$

we find that

$$E_{\mu}\left[\frac{n + \gamma_o^2}{1 + \gamma_o^2} \sum_{i=1}^n X_i^2 - \sum_{i=1}^n X_i^2\right] = \frac{n + \gamma_o^2}{1 + \gamma_o^2} n\mu^2 (1 + \gamma_o^2) - (n\gamma_o^2 \mu^2 + n^2 \mu^2) = 0$$

for all $\mu$ so the minimal sufficient statistic is not complete. Notice that

$$\Omega = \{ (\eta_1, \eta_2) : \eta_1 = \frac{-1}{2} \gamma_o^2 \eta_2 \}$$

and a plot of $\eta_1$ versus $\eta_2$ is a quadratic function which can not contain a 2–dimensional rectangle. Notice that $(\eta_1, \eta_2)$ is a one to one function of $\mu$,
and thus this example illustrates that the rectangle needs to be contained in \( \Omega \) rather than \( \Theta \).

**Example 4.15.** The theory does not say that any sufficient statistic from a REF is complete. Let \( Y \) be a random variable from a normal \( N(0, \sigma^2) \) distribution with \( \sigma^2 > 0 \). This family is a REF with complete minimal sufficient statistic \( Y^2 \). The data \( Y \) is also a sufficient statistic, but \( Y \) is not a function of \( Y^2 \). Hence \( Y \) is not minimal sufficient and (by Bahadur’s theorem) not complete. Alternatively \( E_{\sigma^2}(Y) = 0 \) but \( P_{\sigma^2}(Y = 0) = 0 < 1 \), so \( Y \) is not complete.

**Theorem 4.9.** Let \( Y_1, \ldots, Y_n \) be iid.

a) If \( Y_i \sim U(\theta_1, \theta_2) \), then \((Y_{(1)}, Y_{(n)})\) is a complete sufficient statistic for \((\theta_1, \theta_2)\). See David (1981, p. 123.)

b) If \( Y_i \sim U(\theta_1, \theta_2) \) with \( \theta_1 \) known, then \( Y_{(n)} \) is a complete sufficient statistic for \( \theta_2 \).

c) If \( Y_i \sim U(\theta_1, \theta_2) \) with \( \theta_2 \) known, then \( Y_{(1)} \) is a complete sufficient statistic for \( \theta_1 \).

d) If \( Y_i \sim U(-\theta, \theta) \), then \( \max(|Y_i|) \) is a complete sufficient statistic for \( \theta \).

e) If \( Y_i \sim EXP(\theta, \lambda) \), then \((Y_{(1)}, \overline{Y})\) is a complete sufficient statistic for \((\theta, \lambda)\). See David (1981, p. 153-154.)

f) If \( Y_i \sim EXP(\theta, \lambda) \) with \( \lambda \) known, then \( Y_{(1)} \) is a complete sufficient statistic for \( \theta \).

g) If \( Y_i \sim \text{Cauchy}(\mu, \sigma) \) with \( \sigma \) known, then the order statistics are minimal sufficient.

h) If \( Y_i \sim \text{Double Exponential}(\theta, \lambda) \) with \( \lambda \) known, then the order statistics \((Y_{(1)}, \ldots, Y_{(n)})\) are minimal sufficient.

i) If \( Y_i \sim \text{logistic}(\mu, \sigma) \), then the order statistics are minimal sufficient.

j) If \( Y_i \sim \text{Weibull}(\phi, \lambda) \), then the order statistics \((Y_{(1)}, \ldots, Y_{(n)})\) are minimal sufficient.

**A common midterm, final and qual question** takes \( X_1, \ldots, X_n \) iid \( U(h_l(\theta), h_u(\theta)) \) where \( h_l \) and \( h_u \) are functions of \( \theta \) such that \( h_l(\theta) < h_u(\theta) \). The function \( h_l \) and \( h_u \) are chosen so that the min = \( X_{(1)} \) and the max = \( X_{(n)} \) form the 2-dimensional minimal sufficient statistic by the LSM theorem. Since \( \theta \) is one dimensional, the rule of thumb suggests that the minimal sufficient statistic is not complete. State this fact, but if you have time find \( E_\theta[X_{(1)}] \) and \( E_\theta[X_{(n)}] \). Then show that \( E_\theta[ax_{(1)} + bx_{(n)} + c] \equiv 0 \) so that \( T = (X_{(1)}, X_{(n)}) \) is not complete.
Example 4.16. The uniform distribution is tricky since usually $(X(1), X(n))$ is minimal sufficient by the LSM theorem, since

$$f(x) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 < x(1) < x(n) < \theta_2)$$

if $n > 1$. But occasionally $\theta_1$ and $\theta_2$ are functions of $\theta$ such that the indicator can be written as $I(\theta > T)$ or $I(\theta < T)$ where the minimal sufficient statistic $T$ is a one dimensional function of $(X(1), X(n))$. If $X \sim U(c_1 + d_1 \theta, c_2 + d_2 \theta)$ where $d_1, d_2, c_1, c_2$ are known and $d_1 < 0$ and $d_2 > \theta$, then

$$T = \max\left(\frac{X(1) - c_1}{d_1}, \frac{X(n) - c_2}{d_2}\right)$$

is minimal sufficient.

Let $X_1, ..., X_n$ be iid $U(1 - \theta, 1 + \theta)$ where $\theta > 0$ is unknown. Hence

$$f_X(x) = \frac{1}{2\theta} I(1 - \theta < x < 1 + \theta)$$

and

$$\frac{f(x)}{f(y)} = \frac{I(1 - \theta < x(1) \leq x(n) < 1 + \theta)}{I(1 - \theta < y(1) \leq y(n) < 1 + \theta)}.$$ 

This ratio may look to be constant for all $\theta > 0$ iff $(x(1), x(n)) = (y(1), y(n))$, but it is not. To show that $T_1 = (X(1), X(n))$ is not a minimal sufficient statistic, note that

$$f_X(x) = \frac{1}{2\theta} I(\theta > 1 - x) I(\theta > x - 1).$$

Hence

$$\frac{f(x)}{f(y)} = \frac{I(\theta > \max(1 - x(1), x(n) - 1))}{I(\theta > \max(1 - y(1), y(n) - 1))}$$

which is constant for all $\theta > 0$ iff $T_2(x) = T_2(y)$ where $T_2(x) = \max(1 - x(1), x(n) - 1)$. Hence $T_2 = T_2(X) = \max(1 - X(1), X(n) - 1)$ is minimal sufficient by the LSM theorem. Thus $T_1$ is not a minimal sufficient statistic (and so not complete) since $T_1$ is not a function of $T_2$.

To show that $T_1$ is not complete using the definition of complete statistic, first find $E(T_1)$. Now

$$F_X(t) = \int_{1-\theta}^{t} \frac{1}{2\theta} dx = \frac{t + \theta - 1}{2\theta}$$
for $1 - \theta < t < 1 + \theta$. Hence by Theorem 4.2a),

$$f_{X(n)}(t) = \frac{n}{2\theta} \left( \frac{t + \theta - 1}{2\theta} \right)^{n-1}$$

for $1 - \theta < t < 1 + \theta$ and

$$E_\theta(X(n)) = \int_{1-\theta}^{1+\theta} x f_{X(n)}(x) dx = \int_{1-\theta}^{1+\theta} x \frac{n}{2\theta} \left( \frac{x + \theta - 1}{2\theta} \right)^{n-1} dx.$$  

Use u–substitution with $u = (x + \theta - 1)/2\theta$ and $x = 2\theta u + 1 - \theta$. Hence $x = 1 + \theta$ implies $u = 1$, and $x = 1 - \theta$ implies $u = 0$ and $dx = 2\theta du$. Thus

$$E_\theta(X(n)) = n \int_0^1 \frac{2\theta u + 1 - \theta}{2\theta} u^{n-1} 2\theta du =$$

$$= n \int_0^1 [2\theta u + 1 - \theta] u^{n-1} du = 2\theta n \int_0^1 u^n du + (n - n\theta) \int_0^1 u^{n-1} du =$$

$$= 2\theta n \frac{u^{n+1}}{n+1} \bigg|_0^1 + n(1 - \theta) \frac{u^n}{n} \bigg|_0^1 =$$

$$= 2\theta \frac{n}{n+1} + n(1 - \theta) \frac{n}{n} = 1 - \theta + 2\theta \frac{n}{n+1}.$$  

Note that $E_\theta(X(n)) \approx 1 + \theta$ as you should expect.

By Theorem 4.2b),

$$f_{X(1)}(t) = \frac{n}{2\theta} \left( \frac{\theta - t + 1}{2\theta} \right)^{n-1}$$

for $1 - \theta < t < 1 + \theta$ and thus

$$E_\theta(X(1)) = \int_{1-\theta}^{1+\theta} x \frac{n}{2\theta} \left( \frac{\theta - x + 1}{2\theta} \right)^{n-1} dx.$$  

Use u–substitution with $u = (\theta - x + 1)/2\theta$ and $x = \theta + 1 - 2\theta u$. Hence $x = 1 + \theta$ implies $u = 0$, and $x = 1 - \theta$ implies $u = 1$ and $dx = -2\theta du$. Thus

$$E_\theta(X(1)) = \int_0^1 \frac{n}{2\theta} (\theta + 1 - 2\theta u) u^{n-1} (-2\theta) du = n \int_0^1 (\theta + 1 - 2\theta u) u^{n-1} du =$$
CHAPTER 4. SUFFICIENT STATISTICS

\[ n(\theta + 1) \int_0^1 u^{n-1} du - 2\theta n \int_0^1 u^n du = (\theta + 1)n/n - 2\theta n/(n+1) = \theta + 1 - 2\theta \frac{n}{n+1}. \]

To show that \( T_1 \) is not complete try showing \( E_\theta(aX_{(1)} + bX_{(n)} + c) = 0 \) for some constants \( a, b \) and \( c \). Note that \( a = b = 1 \) and \( c = -2 \) works. Hence \( E_\theta(X_{(1)} + X_{(n)} - 2) = 0 \) for all \( \theta > 0 \) but \( P_\theta(g(T_1) = 0) = P_\theta(X_{(1)} + X_{(n)} - 2 = 0) = 0 < 1 \) for all \( \theta > 0 \). Hence \( T_1 \) is not complete. (To see that \( P_\theta(g(T_1) = 0) = 0 \), note that if random variables \( Y_1, \ldots, Y_n \) have a joint pdf, then a nonzero linear combination of the random variables is a random variable \( W \) with a pdf by Theorem 2.14. So \( P(W = 0) = 0 \) since the probability that a random variable \( W \) is equal to a constant is 0 if \( W \) has a pdf. The order statistics have a joint pdf by Theorem 4.2 c). Thus the linear combination \( W = g(T_1) \) has a pdf and \( P(W = 0) = 0 \).

**Definition 4.5.** Let \( Y_1, \ldots, Y_n \) have pdf or pmf \( f(y|\theta) \). A statistic \( W(Y) \) whose distribution does not depend on \( \theta \) is called an **ancillary statistic**.

**Theorem 4.10, Basu’s Theorem.** Let \( Y_1, \ldots, Y_n \) have pdf or pmf \( f(y|\theta) \). If \( T(Y) \) is a \( k \)-dimensional complete sufficient statistic, then \( T(Y) \) is independent of every ancillary statistic.

**Remark 4.4.** Basu’s Theorem says that if \( T \) is minimal sufficient and complete, then \( T \perp R \) if \( R \) is ancillary. Application: If \( T \) is minimal sufficient, \( R \) ancillary and \( R \) is a function of \( T \) (so \( R = h(T) \) is not independent of \( T \)), then \( T \) is not complete. Since \( \theta \) is a scalar, usually need \( k = 1 \) for \( T = T(Y) = T(Y) = T \) to be complete.

**Example 4.17.** Suppose \( X_1, \ldots, X_n \) are iid uniform observations on the interval \((\theta, \theta + 1), -\infty < \theta < \infty \). Let \( X_{(1)} = \min(X_1, \ldots, X_n), X_{(n)} = \max(X_1, \ldots, X_n) \) and \( T(X) = (X_{(1)}, X_{(n)}) \) be a minimal sufficient statistic. Then \( R = X_{(n)} - X_{(1)} \) is ancillary since \( R = \max(X_1 - \theta, \ldots, X_n - \theta) + \theta - \min(X_1 - \theta, \ldots, X_n - \theta) \) is a distribution that does not depend on \( \theta \). \( R \) is not independent of \( T \), so \( T \) is not complete.

**Example 4.18.** Let \( Y_1, \ldots, Y_n \) be iid from a location family with pdf \( f_Y(y|\theta) = f_X(y - \theta) \) where \( Y = X + \theta \) and \( f_X(y) \) is the standard pdf for the location family (and thus the distribution of \( X \) does not depend on \( \theta \)). Claim: \( W = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y}) \) is ancillary.
Proof: Since $Y_i = X_i + \theta$,

$$W = \left( X_1 + \theta - \frac{1}{n} \sum_{i=1}^{n} (X_i + \theta), \ldots, X_n + \theta - \frac{1}{n} \sum_{i=1}^{n} (X_i + \theta) \right)$$

$$= (X_1 - \bar{X}, \ldots, X_n - \bar{X})$$

and the distribution of the final vector is free of $\theta$. QED

Application: Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$. For any fixed $\sigma^2$, this is a location family with $\theta = \mu$ and complete sufficient statistic $T(Y) = \bar{Y}$. Thus $\bar{Y} \perp W$ by Basu’s Theorem. Hence $\bar{Y} \perp S^2$ for any known $\sigma^2 > 0$ since

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

is a function of $W$. Thus $\bar{Y} \perp S^2$ even if $\sigma^2 > 0$ is not known.

### 4.3 Summary

1) A statistic is a function of the data that does not depend on any unknown parameters.

2) For parametric inference, the data $Y_1, \ldots, Y_n$ comes from a family of parametric distributions $f(y|\theta)$ for $\theta \in \Theta$. Often the data are iid and $f(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$. The parametric distribution is completely specified by the unknown parameters $\theta$. The statistic is a random vector or random variable and hence also comes from some probability distribution. The distribution of the statistic is called the sampling distribution of the statistic.

3) For iid $N(\mu, \sigma^2)$ data, $\bar{Y} \perp S^2$, $\bar{Y} \sim N(\mu, \sigma^2/n)$ and $\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_{n-1}$.

4) For iid data with cdf $F_Y$ and pdf $f_Y$, $f_{Y_{(n)}}(t) = n[F_Y(t)]^{n-1}f_Y(t)$ and $f_{Y_{(1)}}(t) = n[1 - F_Y(t)]^{n-1}f_Y(t)$.

5) A statistic $T(Y_1, \ldots, Y_n)$ is a sufficient statistic for $\theta$ if the conditional distribution of $(Y_1, \ldots, Y_n)$ given $T$ does not depend on $\theta$.

6) A sufficient statistic $T(Y)$ is a minimal sufficient statistic if for any other sufficient statistic $S(Y)$, $T(Y)$ is a function of $S(Y)$. 
CHAPTER 4. SUFFICIENT STATISTICS

7) Suppose that a statistic $T(Y)$ has a pmf or pdf $f(t|\theta)$. Then $T(Y)$ is a complete statistic if $E_{\theta}[g(T(Y))] = 0$ for all $\theta \in \Theta$ implies that $P_{\theta}[g(T(Y)) = 0] = 1$ for all $\theta \in \Theta$.

8) A one to one function of a sufficient, minimal sufficient, or complete sufficient statistic is sufficient, minimal sufficient, or complete sufficient respectively.

9) Factorization Theorem. Let $f(y|\theta)$ denote the pdf or pmf of a sample $Y$. A statistic $T(Y)$ is a sufficient statistic for $\theta$ iff for all sample points $y$ and for all $\theta$ in the parameter space $\Theta$,

$$f(y|\theta) = g(T(y)|\theta) \cdot h(y)$$

where both $g$ and $h$ are nonnegative functions.

Tips: i) for iid data with marginal support $\mathcal{Y} \equiv \mathcal{Y}^*$, $I_{\mathcal{Y}}(y) = I(\text{all } y_i \in \mathcal{Y}^*)$. If $\mathcal{Y}^* = (a, b)$, then $I_{\mathcal{Y}}(y) = I(a < y_{(1)} < y_{(n)} < b) = I(a < y_{(1)})I(y_{(n)} < b)$. Put $I(a < y_{(1)})$ in $g(T(y)|\theta)$ if $a$ is an unknown parameter but put $I(a < y_{(1)})$ in $h(y)$ if $a$ is known. If both $a$ and $b$ are unknown parameters, put $I(a < y_{(1)} < y_{(n)} < b)$ in $g(T(y)|\theta)$. If $b = \infty$, then $I_{\mathcal{Y}}(y) = I(a < y_{(1)})$. If $\mathcal{Y}^* = [a, b]$, then $I_{\mathcal{Y}}(y) = I(a \leq y_{(1)} < y_{(n)} \leq b) = I(a \leq y_{(1)})I(y_{(n)} \leq b)$.

i) Try to make the dimension of $T(y)$ as small as possible. Put anything that depends on $y$ but not $\theta$ into $h(y)$.

10) Minimal and complete sufficient statistics for $k$-parameter exponential families: Let $Y_1, ..., Y_n$ be iid from an exponential family $f(y|\eta) = h(y) c(\eta) \exp[\sum_{j=1}^{k} \eta_j t_j(y)]$ with the natural parameterization $f(y|\eta) = h(y) b(\eta) \exp[\sum_{j=1}^{k} \eta_j t_j(y)]$. Then

$$T(Y) = \left(\sum_{i=1}^{n} t_1(Y_i), ..., \sum_{i=1}^{n} t_k(Y_i)\right)$$

a) a minimal sufficient statistic for $\eta$ if the $\eta_i$ do not satisfy a linearity constraint and for $\theta$ if the $w_j(\theta)$ do not satisfy a linearity constraint.

b) a complete sufficient statistic for $\theta$ and for $\eta$ if $\eta$ is a one to one function of $\theta$ and if $\Omega$ contains a $k$-dimensional rectangle.

11) Completeness of REFs: Suppose that $Y_1, ..., Y_n$ are iid from a kP–REF

$$f(y|\theta) = h(y) c(\theta) \exp\left[w_1(\theta)t_1(y) + \cdots + w_k(\theta)t_k(y)\right]$$

with $\theta \in \Theta$ and natural parameter $\eta \in \Omega$. Then

$$T(Y) = \left(\sum_{i=1}^{n} t_1(Y_i), ..., \sum_{i=1}^{n} t_k(Y_i)\right)$$
a) a minimal sufficient statistic for \( \eta \) and for \( \theta \),
b) a complete sufficient statistic for \( \theta \) and for \( \eta \) if \( \eta \) is a one to one function of \( \theta \) and if \( \Omega \) contains a \( k \)-dimensional rectangle.

12) For a 2-parameter exponential family \( (k = 2) \), \( \eta_1 \) and \( \eta_2 \) satisfy a linearity constraint if the plotted points fall on a line in a plot of \( \eta_1 \) versus \( \eta_2 \). If the plotted points fall on a nonlinear curve, then \( T \) is minimal sufficient but \( \Omega \) does not contain a 2-dimensional rectangle.

13) LSM Theorem: Let \( f(y|\theta) \) be the pmf or pdf of a sample \( Y \). Let \( c_{x,y} \) be a constant. Suppose there exists a function \( T(y) \) such that for any two sample points \( x \) and \( y \), the ratio \( R_{x,y}(\theta) = f(x|\theta)/f(y|\theta) = c_{x,y} \) for all \( \theta \) in \( \Theta \) iff \( T(x) = T(y) \). Then \( T(Y) \) is a minimal sufficient statistic for \( \theta \). (Define \( 0/0 \equiv c_{x,y} \).)

14) Tips for finding sufficient, minimal sufficient and complete sufficient statistics. a) Typically \( Y_1, \ldots, Y_n \) are iid so the joint distribution \( f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i) \) where \( f(y_i) \) is the marginal distribution. Use the factorization theorem to find the candidate sufficient statistic \( T \).
b) Use factorization to find candidates \( T \) that might be minimal sufficient statistics. Try to find \( T \) with as small a dimension \( k \) as possible. If the support of the random variable depends on \( \theta \), often \( Y_1 \) or \( Y_n \) will be a component of the minimal sufficient statistic. To prove that \( T \) is minimal sufficient, use the LSM theorem. Alternatively prove or recognize that \( Y \) comes from a regular exponential family. \( T \) will be minimal sufficient for \( \theta \) if \( Y \) comes from an exponential family as long as the \( w_i(\theta) \) do not satisfy a linearity constraint.
c) To prove that the statistic is complete, prove or recognize that \( Y \) comes from a regular exponential family. Check whether \( \dim(\Theta) = k \). If \( \dim(\Theta) < k \), then the family is usually not a \( kP-REF \) and Theorem 4.5 and Corollary 4.6 do not apply. The uniform distribution where one endpoint is known also has a complete sufficient statistic.
d) Let \( k \) be free of the sample size \( n \). Then a \( k \)-dimensional complete sufficient statistic is also a minimal sufficient statistic (Bahadur’s theorem).
e) To show that a statistic \( T \) is not a sufficient statistic, either show that factorization fails or find a minimal sufficient statistic \( S \) and show that \( S \) is not a function of \( T \).
f) To show that \( T \) is not minimal sufficient, first try to show that \( T \) is not a sufficient statistic. If \( T \) is sufficient, find a minimal sufficient statistic \( S \) and
show that $T$ is not a function of $S$. (Of course $S$ will be a function of $T$.) The Lehmann-Scheffé (LSM) theorem cannot be used to show that a statistic is not minimal sufficient.

g) To show that a sufficient statistics $T$ is not complete, find a function $g(T)$ such that $E_{\theta}(g(T)) = 0$ for all $\theta$ but $g(T)$ is not equal to the zero with probability one. Finding such a $g$ is often hard, unless there are clues. For example, if $T = (X, Y, \ldots)$ and $\mu_1 = \mu_2$, try $g(T) = X - Y$. As a rule of thumb, a $k$-dimensional minimal sufficient statistic will generally not be complete if $k > \dim(\Theta)$. In particular, if $T$ is $k$-dimensional and $\theta$ is $j$-dimensional with $j < k$ (especially $j = 1 < 2 = k$) then $T$ will generally not be complete. If you can show that a $k$-dimensional sufficient statistic $T$ is not minimal sufficient (often hard), then $T$ is not complete by Bahadur’s Theorem. Basu’s Theorem can sometimes be used to show that a minimal sufficient statistic is not complete. See Remark 4.4 and Example 4.18.

15) A common question takes $Y_1, \ldots, Y_n$ iid $U(h_l(\theta), h_u(\theta))$ where the min $= Y_{(1)}$ and the max $= Y_{(n)}$ form the 2-dimensional minimal sufficient statistic. Since $\theta$ is one dimensional, the minimal sufficient statistic is probably not complete. Find $E_{\theta}[Y_{(1)}]$ and $E_{\theta}[Y_{(n)}]$. Then show that $E_{\theta}[aY_{(1)} + bY_{(n)} + c] \equiv 0$ so that $T = (Y_{(1)}, Y_{(n)})$ is not complete.

## 4.4 Complements

Some minimal sufficient statistics and complete sufficient statistics are given in Theorem 4.9 for distributions that are not exponential families.

Stigler (1984) presents Kruskal’s proof that $\overline{Y} \not\perp S^2$ when the data are iid $N(\mu, \sigma^2)$, but Zehna (1991) states that there is a flaw in the proof.

The Factorization Theorem was developed with increasing generality by Fisher, Neyman and by Halmos and Savage (1949).

Bahadur’s Theorem is due to Bahadur (1958) and Lehmann and Scheffé (1950).

Basu’s Theorem is due to Basu (1959). Also see Koehn and Thomas (1975) and Boos and Hughes-Oliver (1998). An interesting alternative method for proving independence between two statistics that works for some important examples is given in Datta and Sarker (2008).

Some techniques for showing whether a statistic is minimal sufficient are illustrated in Sampson and Spencer (1976).
4.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

4.1. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, which is an exponential family. Show that the sample space of $(T_1, T_2)$ contains an open subset of $\mathbb{R}^2$, if $n \geq 2$ but not if $n = 1$.

Hint: Show that if $n \geq 2$, then $T_1 = \sum_{i=1}^n X_i$ and $T_2 = \sum_{i=1}^n X_i^2$. Then $T_2 = aT_1^2 + b(X_1, \ldots, X_n)$ for some constant $a$ where $b(X_1, \ldots, X_n) = \sum_{i=1}^n (X_i - \bar{X})^2 \in (0, \infty)$. So range$(T_1, T_2) = \{(t_1, t_2) | t_2 \geq at_1^2 \}$. Find $a$. If $n = 1$ then $b(X_1) \equiv 0$ and the curve can not contain an open 2-dimensional rectangle.

4.2. Let $X_1, \ldots, X_n$ be iid exponential($\lambda$) random variables. Use the Factorization Theorem to show that $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for $\lambda$.

4.3. Let $X_1, \ldots, X_n$ be iid from a regular exponential family with pdf

$$f(x|\eta) = h(x)b(\eta)\exp[\sum_{i=1}^k \eta_i t_i(x)].$$

Let $T(X) = (T_1(X), \ldots, T_k(X))$ where $T_i(X) = \sum_{j=1}^n t_i(X_j)$.

a) Use the factorization theorem to show that $T(X)$ is a $k$-dimensional sufficient statistic for $\eta$.

b) Use the Lehmann Scheffé LSM theorem to show that $T(X)$ is a minimal sufficient statistic for $\eta$.

(Hint: in a regular exponential family, if $\sum_{i=1}^k a_i \eta_i = c$ for all $\eta$ in the natural parameter space for some fixed constants $a_1, \ldots, a_k$ and $c$, then $a_1 = \cdots = a_k = 0$.)

4.4. Let $X_1, \ldots, X_n$ be iid $N(\mu, \gamma^2_o \mu^2)$ random variables where $\gamma^2_o > 0$ is known and $\mu > 0$.

a) Find a sufficient statistic for $\mu$.

b) Show that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a minimal sufficient statistic.

c) Find $E_{\mu}[\sum_{i=1}^n X_i^2]$.

d) Find $E_{\mu}[(\sum_{i=1}^n X_i)^2]$. 

e) Find
\[ E_{\mu}[\frac{n + \gamma_0^2}{1 + \gamma_0^2} \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2]. \]
(Hint: use c) and d).)

f) Is the minimal sufficient statistic given in b) complete? Explain.

4.5. If \( X_1, \ldots, X_n \) are iid with \( f(x|\theta) = \exp[-(x - \theta)] \) for \( x > \theta \), then the joint pdf can be written as
\[ f(x|\theta) = e^{n\theta} \exp(-\sum x_i) I[\theta < x_{(1)}]. \]
By the factorization theorem, \( T(X) = (\sum X_i, X_{(1)}) \) is a sufficient statistic. Show that \( R(\theta) = f(x|\theta)/f(y|\theta) \) can be constant even though \( T(x) \neq T(y) \). Hence the Lehmann-Scheffé theorem does not imply that \( T(X) \) is a minimal sufficient statistic.

4.6. Find a complete minimal sufficient statistic if \( Y_1, \ldots, Y_n \) are iid from the following 1P-REF's.

a) \( Y \sim \text{binomial}(k, \rho) \) with \( k \) known.
b) \( Y \sim \text{exponential}(\lambda) \).
c) \( Y \sim \text{gamma}(\nu, \lambda) \) with \( \nu \) known.
d) \( Y \sim \text{geometric}(\rho) \).
e) \( Y \sim \text{negative binomial}(r, \rho) \) with \( r \) known.
f) \( Y \sim \text{normal}(\mu, \sigma^2) \) with \( \sigma^2 \) known.
g) \( Y \sim \text{normal}(\mu, \sigma^2) \) with \( \mu \) known.
h) \( Y \sim \text{Poisson}(\theta) \).
4.7. Find a complete minimal sufficient statistic if \( Y_1, ..., Y_n \) are iid from the following 1P-REFs.
   a) \( Y \sim \text{Burr Type XII } (\phi, \lambda) \) with \( \phi \) known.
   b) \( Y \sim \text{chi} (p, \sigma) \) with \( p \) known
   c) \( Y \sim \text{double exponential } (\theta, \lambda) \) with \( \theta \) known.
   d) \( Y \sim \text{two parameter exponential } (\theta, \lambda) \) with \( \theta \) known.
   e) \( Y \sim \text{generalized negative binomial } (\mu, \kappa) \) with \( \kappa \) known.
   f) \( Y \sim \text{half normal } (\mu, \sigma^2) \) with \( \mu \) known.
   g) \( Y \sim \text{inverse Gaussian } (\theta, \lambda) \) with \( \lambda \) known.
   h) \( Y \sim \text{inverted gamma } (\nu, \lambda) \) with \( \nu \) known.
   i) \( Y \sim \text{lognormal } (\mu, \sigma^2) \) with \( \mu \) known.
   j) \( Y \sim \text{lognormal } (\mu, \sigma^2) \) with \( \sigma^2 \) known.
   k) \( Y \sim \text{Maxwell-Boltzmann } (\mu, \sigma) \) with \( \mu \) known.
   l) \( Y \sim \text{one sided stable } (\sigma) \).
   m) \( Y \sim \text{Pareto } (\sigma, \lambda) \) with \( \sigma \) known.
   n) \( Y \sim \text{power } (\lambda) \).
   o) \( Y \sim \text{Rayleigh } (\mu, \sigma) \) with \( \mu \) known.
   p) \( Y \sim \text{Topp-Leone } (\nu) \).
   q) \( Y \sim \text{truncated extreme value } (\lambda) \).
   r) \( Y \sim \text{Weibull } (\phi, \lambda) \) with \( \phi \) known.

4.8. Find a complete minimal sufficient statistic \( T \) if \( Y_1, ..., Y_n \) are iid from the following 2P-REFs.
   a) The beta \((\delta, \nu)\) distribution.
   b) The chi \((p, \sigma)\) distribution.
   c) The gamma \((\nu, \lambda)\) distribution.
   d) The lognormal \((\mu, \sigma^2)\) distribution.
   e) The normal \((\mu, \sigma^2)\) distribution.

4.9. i) Show that each of the following families is a 1P-REF. ii) Find a complete minimal sufficient statistic if \( Y_1, ..., Y_n \) are iid from the 1P-REF.
   a) Let
   \[
   f(y) = \frac{\log(\theta)}{\theta - 1} \theta^y
   \]
   where \( 0 < y < 1 \) and \( \theta > 1 \).
   Comment:
   \[
   F(y) = \frac{\theta^y - 1}{\theta - 1}
   \]
for $0 < y < 1$, and the mgf

$$m(t) = \frac{\log(\theta) e^{(t+\log(\theta))} - 1}{\theta - 1 \ t + \log(\theta)}.$$  

(b) $Y$ has an inverse Weibull distribution.

c) $Y$ has a Zipf distribution.

**4.10.** Suppose $Y$ has a log-gamma distribution, $Y \sim LG(\nu, \lambda)$.

i) Show the $Y$ is a 2P-REF.

ii) If $Y_1, \ldots, Y_n$ are iid $LG(\nu, \lambda)$, find a complete minimal sufficient statistic.

iii) Show $W = e^Y \sim$ gamma $(\nu, \lambda)$.

**Problems from old quizzes and exams.**

**4.11.** Suppose that $X_1, \ldots, X_m; Y_1, \ldots, Y_n$ are iid $N(\mu, 1)$ random variables. Find a minimal sufficient statistic for $\mu$.

**4.12.** Let $X_1, \ldots, X_n$ be iid from a uniform $U(\theta - 1, \theta + 2)$ distribution. Find a sufficient statistic for $\theta$.

**4.13.** Let $Y_1, \ldots, Y_n$ be iid with a distribution that has pmf $P_\theta(X = x) = \theta(1 - \theta)^{x-1}$, $x = 1, 2, \ldots$, where $0 < \theta < 1$. Find a minimal sufficient statistic for $\theta$.

**4.14.** Let $Y_1, \ldots, Y_n$ be iid Poisson($\lambda$) random variables. Find a minimal sufficient statistic for $\lambda$ using the fact that the Poisson distribution is a regular exponential family (REF).

**4.15.** Suppose that $X_1, \ldots, X_n$ are iid from a REF with pdf (with respect to the natural parameterization)

$$f(x) = h(x) c^*(\eta) \exp[\sum_{i=1}^{4} \eta_i t_i(x)].$$

Assume $\text{dim}(\Theta) = 4$. Find a complete minimal sufficient statistic $T(X)$ in terms of $n, t_1, t_2, t_3,$ and $t_4$.

**4.16.** Let $X$ be a uniform $U(-\theta, \theta)$ random variable (sample size $n = 1$).

a) Find $E_\theta X$. b) Is $T(X) = X$ a complete sufficient statistic? c) Show that $|X| = \max(-X, X)$ is a minimal sufficient statistic.
4.17. A fact from mathematics is that if the polynomial
\[ P(w) = a_n w^n + a_{n-1} w^{n-1} + \cdots + a_2 w^2 + a_1 w + a_0 \equiv 0 \]
for all \( w \) in a domain that includes an open interval, then \( a_n = \cdots = a_1 = a_0 = 0 \). Suppose that
you are trying to use the Lehmann Scheffé (LSM) theorem to show that
\( (\sum X_i, \sum X_i^2) \) is a minimal sufficient statistic and that you have managed to
show that
\[ \frac{f(x|\mu)}{f(y|\mu)} \equiv c \]
iff
\[ -\frac{1}{2\gamma^2 \mu^2} [\sum x_i^2 - \sum y_i^2] + \frac{1}{\gamma^2 \mu^2} \sum x_i - \sum y_i \equiv d \quad (4.11) \]
for all \( \mu > 0 \). Parts a) and b) give two different ways to proceed.

a) Let \( w = 1/\mu \) and assume that \( \gamma \) is known. Identify \( a_2, a_1, \) and \( a_0 \) and
show that \( a_i = 0 \) implies that \( (\sum X_i, \sum X_i^2) \) is a minimal sufficient statistic.

b) Let \( \eta_1 = 1/\mu^2 \) and \( \eta_2 = 1/\mu \). Since (4.11) is a polynomial in \( 1/\mu \),
can \( \eta_1 \) and \( \eta_2 \) satisfy a linearity constraint? If not, why is \( (\sum X_i, \sum X_i^2) \) a
minimal sufficient statistic?

4.18 Let \( X_1, \ldots, X_n \) be iid Exponential(\( \lambda \)) random variables and \( Y_1, \ldots, Y_m \)
iid Exponential(\( \lambda/2 \)) random variables. Assume that the \( Y_i \)'s and \( X_j \)'s are
independent. Show that the statistic \( (\sum_{i=1}^n X_i, \sum_{i=1}^m Y_i) \) is not a complete
sufficient statistic.

4.19. Let \( X_1, \ldots, X_n \) be iid gamma(\( \nu, \lambda \)) random variables. Find a com-
plete, minimal sufficient statistic \( (T_1(X), T_2(X)) \). (Hint: recall a theorem for
exponential families. The gamma pdf is (for \( x > 0 \))
\[ f(x) = \frac{x^{\nu-1} e^{-x/\lambda}}{\lambda^\nu \Gamma(\nu)} \]

4.20. Let \( X_1, \ldots, X_n \) be iid uniform(\( \theta - 1, \theta + 1 \)) random variables. The
following expectations may be useful:
\[ E_{\theta} \overline{X} = \theta, \quad E_{\theta} X_{(1)} = 1 + \theta - 2\theta \frac{n}{n+1}, \quad E_{\theta} X_{(n)} = 1 - \theta + 2\theta \frac{n}{n+1}. \]

a) Find a minimal sufficient statistic for \( \theta \).

b) Show whether the minimal sufficient statistic is complete or not.
**CHAPTER 4. SUFFICIENT STATISTICS**

4.21. Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf

$$f(x) = \sqrt{\frac{\sigma}{2\pi x^3}} \exp\left(-\frac{\sigma}{2x}\right)$$

where $x$ and $\sigma$ are both positive. Find a sufficient statistic $T(X)$ for $\sigma$.

4.22. Suppose that $X_1, ..., X_n$ are iid beta($\delta, \nu$) random variables. Find a minimal sufficient statistic for ($\delta, \nu$). Hint: write as a 2 parameter REF.

4.23. Let $X_1, ..., X_n$ be iid from a distribution with pdf

$$f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.$$ 

Find a sufficient statistic for $\theta$.

4.24. Let $X_1, ..., X_n$ be iid with a distribution that has pdf

$$f(x) = \frac{x}{\sigma^2} \exp\left(-\frac{x}{2\sigma^2}\right)$$

for $x > 0$ and $\sigma^2 > 0$. Find a minimal sufficient statistic for $\sigma^2$ using the Lehmann-Scheffé theorem.

4.25. Let $X_1, ..., X_n$ be iid exponential ($\lambda$) random variables. Find a minimal sufficient statistic for $\lambda$ using the fact that the exponential distribution is a 1P–REF.

4.26. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$. Find a complete sufficient statistic for ($\mu, \sigma^2$).

4.27. (Jan. 2003 Qual) Let $X_1$ and $X_2$ be iid Poisson ($\lambda$) random variables. Show that $T = X_1 + 2X_2$ is not a sufficient statistic for $\lambda$. (Hint: the Factorization Theorem uses the word iff. Alternatively, find a minimal sufficient statistic $S$ and show that $S$ is not a function of $T$.)

4.28. (Aug. 2002 Qual): Suppose that $X_1, ..., X_n$ are iid $N(\sigma, \sigma)$ where $\sigma > 0$.

a) Find a minimal sufficient statistic for $\sigma$.

b) Show that $(\bar{X}, S^2)$ is a sufficient statistic but is not a complete sufficient statistic for $\sigma$.

4.29. Let $X_1, ..., X_n$ be iid binomial($k = 1, \theta$) random variables and $Y_1, ..., Y_m$ iid binomial($k = 1, \theta/2$) random variables. Assume that the $Y_i$'s
and $X_j$'s are independent. Show that the statistic $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{m} Y_i)$ is not a complete sufficient statistic.

4.30. Suppose that $X_1, ..., X_n$ are iid Poisson($\lambda$) where $\lambda > 0$. Show that $(\bar{X}, S^2)$ is not a complete sufficient statistic for $\lambda$.

4.31. (Aug. 2004 Qual): Let $X_1, ..., X_n$ be iid beta($\theta, \theta$). (Hence $\delta = \nu = \theta$.)
   a) Find a minimal sufficient statistic for $\theta$.
   b) Is the statistic found in a) complete? (prove or disprove)

4.32. (Sept. 2005 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables with probability mass function
   \[ f(x) = P(X = x) = \frac{1}{x^\nu \zeta(\nu)} \]
   where $\nu > 1$ and $x = 1, 2, 3, ....$ Here the zeta function
   \[ \zeta(\nu) = \sum_{x=1}^{\infty} \frac{1}{x^\nu} \]
   for $\nu > 1$.
   a) Find a minimal sufficient statistic for $\nu$.
   b) Is the statistic found in a) complete? (prove or disprove)
   c) Give an example of a sufficient statistic that is strictly not minimal.

4.33. Let $X_1, ..., X_n$ be a random sample from a half normal distribution with pdf
   \[ f(x) = \frac{2}{\sqrt{2\pi} \sigma} \exp \left( \frac{-(x - \mu)^2}{2\sigma^2} \right) \]
   where $\sigma > 0$ and $x > \mu$ and $\mu$ is real.
   Find a sufficient statistic $T = (T_1, T_2, ..., T_k)$ for $(\mu, \sigma)$ with dimension $k \leq 3$.

4.34. Let $X_{(1)} = \min_{1 \leq i \leq n} X_i$. If $X_1, ..., X_n$ are iid exponential(1) random variables, find $E(X_{(1)})$.

4.35. Let $X_{(n)} = \max_{1 \leq i \leq n} X_i$. If $X_1, ..., X_n$ are iid uniform(0,1) random variables, find $E(X_{(n)})$. 

4.36. (Aug. 2009 Qual): Let $X_1, ..., X_n$ be iid uniform($\theta, \theta + 1$) random variables where $\theta$ is real.

a) Find a minimal sufficient statistic for $\theta$.

b) Show whether the minimal sufficient statistic is complete or not.

4.37. (Sept. 2010 Qual): Let $Y_1, ..., Y_n$ be iid from a distribution with pdf

$$f(y) = 2 \tau y e^{-y^2} (1 - e^{-y^2})^{\tau-1}$$

for $y > 0$ and $f(y) = 0$ for $y \leq 0$ where $\tau > 0$.

a) Find a minimal sufficient statistic for $\tau$.

b) Is the statistic found in a) complete? Prove or disprove.
Chapter 5

Point Estimation

5.1 Maximum Likelihood Estimators

A point estimator gives a single value as an estimate of a parameter. For example, \( \bar{Y} = 10.54 \) is a point estimate of the population mean \( \mu \). An interval estimator gives a range \((L_n, U_n)\) of reasonable values for the parameter. Confidence intervals, studied in Chapter 9, are interval estimators. The most widely used point estimators are the maximum likelihood estimators.

**Definition 5.1.** Let \( f(y|\theta) \) be the pmf or pdf of a sample \( Y \) with parameter space \( \Theta \). If \( Y = y \) is observed, then the **likelihood function** \( L(\theta) \equiv L(\theta|y) = f(y|\theta) \). For each sample point \( y = (y_1, \ldots, y_n) \), let \( \hat{\theta}(y) \in \Theta \) be a parameter value at which \( L(\theta) \equiv L(\theta|y) \) attains its maximum as a function of \( \theta \) with \( y \) held fixed. Then a maximum likelihood estimator (MLE) of the parameter \( \theta \) based on the sample \( Y \) is \( \hat{\theta}(Y) \).

The following remarks are important. I) It is crucial to observe that the likelihood function is a function of \( \theta \) (and that \( y_1, \ldots, y_n \) act as fixed constants). Note that the pdf or pmf \( f(y|\theta) \) is a function of \( n \) variables while \( L(\theta) \) is a function of \( k \) variables if \( \theta \) is a \( 1 \times k \) vector. Often \( k = 1 \) or \( k = 2 \) while \( n \) could be in the hundreds or thousands.

II) If \( Y_1, \ldots, Y_n \) is an independent sample from a population with pdf or pmf \( f(y|\theta) \), then the likelihood function

\[
L(\theta) \equiv L(\theta|y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i|\theta). \tag{5.1}
\]
CHAPTER 5. POINT ESTIMATION

L(θ) = \prod_{i=1}^{n} f_i(y_i|\theta)

if the Y_i are independent but have different pdfs or pmfs.

III) If the MLE \hat{\theta} exists, then \hat{\theta} \in \Theta. Hence if \hat{\theta} is not in the parameter space \Theta, then \hat{\theta} is not the MLE of \theta.

IV) If the MLE is unique, then the MLE is a function of the minimal sufficient statistic. See Levy (1985) and Moore (1971). This fact is useful since exponential families tend to have a tractable log likelihood and an easily found minimal sufficient statistic.

Theorem 5.1: Invariance Principle. If \hat{\theta} is the MLE of \theta, then \hat{h}(\hat{\theta}) is the MLE of h(\theta) where h is a function with domain \Theta.

This theorem will be proved in Section 5.4. Really just need \Theta \subseteq \text{dom}(h) so h(\hat{\theta}) is well defined: can’t have log(-7.89) or \sqrt{-1.57}.

There are four commonly used techniques for finding the MLE.

- Potential candidates can be found by differentiating log L(\theta), the log likelihood.
- Potential candidates can be found by differentiating the likelihood L(\theta).
- The MLE can sometimes be found by direct maximization of the likelihood L(\theta).
- Invariance Principle: If \hat{\theta} is the MLE of \theta, then \hat{h}(\hat{\theta}) is the MLE of h(\theta).

The one parameter case can often be solved by hand with the following technique. To show that \hat{\theta} is the MLE of \theta is equivalent to showing that \hat{\theta} is the global maximizer of log L(\theta) on \Theta where \Theta is an interval with endpoints a and b, not necessarily finite. Suppose that log L(\theta) is continuous on \Theta. Show that log L(\theta) is differentiable on (a, b). Then show that \hat{\theta} is the unique solution to the equation \frac{d}{d\theta} \log L(\theta) = 0 and that the 2nd derivative evaluated at \hat{\theta} is negative: \frac{d^2}{d\theta^2} \log L(\theta) \bigg|_{\hat{\theta}} < 0. See Remark 5.1V below.
Remark 5.1. From calculus, recall the following facts. I) If the function $h$ is continuous on an interval $[a, b]$ then both the max and min of $h$ exist. Suppose that $h$ is continuous on an interval $[a, b]$ and differentiable on $(a, b)$. Solve $h'(\theta) \equiv 0$ and find the places where $h'(\theta)$ does not exist. These values are the critical points. Evaluate $h$ at $a$, $b$, and the critical points. One of these values will be the min and one the max.

II) Assume $h$ is continuous. Then a critical point $\theta_o$ is a local max of $h(\theta)$ if $h$ is increasing for $\theta < \theta_o$ in a neighborhood of $\theta_o$ and if $h$ is decreasing for $\theta > \theta_o$ in a neighborhood of $\theta_o$ (and $\theta_o$ is a global max if you can remove the phrase “in a neighborhood of $\theta_o$”). The first derivative test is often used.

III) If $h$ is strictly concave ($\frac{d^2}{d\theta^2}h(\theta) < 0$ for all $\theta \in \Theta$), then any local max of $h$ is a global max.

IV) Suppose $h'(\theta_o) = 0$. The 2nd derivative test states that if $\frac{d^2}{d\theta^2}h(\theta_o) < 0$, then $\theta_o$ is a local max.

V) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on $(a, b)$ and if the critical point is unique,
then the critical point is a **global maximum** if it is a local maximum. To see this claim, note that if the critical point is not the global max then there would be a local minimum and the critical point would not be unique. Also see Casella and Berger (2002, p. 317). Let $a = -2$ and $b = 4$. In Figure 5.1 a), the critical point for $g(\theta) = -\theta^2 + 25$ is at $\theta = 0$, is unique, and is both a local and global maximum. In Figure 5.1 b), $h(\theta) = \theta^3 - 1.5\theta^2 - 6\theta + 11$, the critical point $\theta = -1$ is not unique and is a local max but not a global max.

VI) If $h$ is strictly convex ($\frac{d^2}{d\theta^2} h(\theta) > 0$ for all $\theta \in \Theta$), then any local min of $h$ is a global min. If $h'(\theta_o) = 0$, then the 2nd derivative test states that if $\frac{d^2}{d\theta^2} h(\theta_o) > 0$, then $\theta_o$ is a local min.

VII) If $h(\theta)$ is a continuous function on an interval with endpoints $a < b$ (not necessarily finite), differentiable on $(a, b)$ and if the **critical point** is unique, then the critical point is a **global minimum** if it is a local minimum. To see this claim, note that if the critical point is not the global min then there would be a local maximum and the critical point would not be unique.

Tips: a) $\exp(a) = e^a$ and $\log(y) = \ln(y) = \log_e(y)$ is the **natural logarithm**.
   b) $\log(a^b) = b \log(a)$ and $\log(e^b) = b$.
   c) $\log(\prod_{i=1}^{n} a_i) = \sum_{i=1}^{n} \log(a_i)$.
   d) $\log L(\theta) = \log(\prod_{i=1}^{n} f(y_i|\theta)) = \sum_{i=1}^{n} \log(f(y_i|\theta))$.
   e) If $t$ is a differentiable function and $t(\theta) \neq 0$, then $\frac{d}{d\theta} \log(|t(\theta)|) = \frac{t'(\theta)}{t(\theta)}$ where $t'(\theta) = \frac{d}{d\theta} t(\theta)$. In particular, $\frac{d}{d\theta} \log(\theta) = 1/\theta$.
   f) Anything that does not depend on $\theta$ is treated as a constant with respect to $\theta$ and hence has derivative 0 with respect to $\theta$.

Showing that $\hat{\theta}$ is the global maximum of $\log(L(\theta))$ is much more difficult in the multiparameter case. To show that $\hat{\theta}$ is a local max often involves using a Hessian matrix of second derivatives. Calculations involving the Hessian matrix are often too difficult for exams. Often there is no closed form solution for the MLE and a computer needs to be used. For hand calculations, Remark 5.2 and Theorem 5.2 can often be used to avoid using the Hessian matrix.

**Definition 5.2.** Let the data be $Y_1, \ldots, Y_n$ and suppose that the pa-
CHAPTER 5. POINT ESTIMATION

parameter $\theta$ has components $(\theta_1, \ldots, \theta_k)$. Then $\hat{\theta}_i$ will be called the MLE of $\theta_i$. Without loss of generality, assume that $\theta = (\theta_1, \theta_2)$, that the MLE of $\theta$ is $(\hat{\theta}_1, \hat{\theta}_2)$ and that $\theta_2$ is known. The profile likelihood function is $L_P(\theta_1) = L(\theta_1, \hat{\theta}_2(y))$ with domain $\{\theta_1 : (\theta_1, \hat{\theta}_2) \in \Theta\}$.

Remark 5.2. Since $L(\theta_1, \theta_2)$ is maximized over $\Theta$ by $(\hat{\theta}_1, \hat{\theta}_2)$, the maximizer of the profile likelihood function and of the log profile likelihood function is $\hat{\theta}_1$. The log profile likelihood function can often be maximized using calculus if $\theta_1 = \theta_1$ is a scalar.

Theorem 5.2: Existence of the MLE for a REF (Barndorff-Nielsen 1982): Assume that the natural parameterization of the kP–REF ($k$–parameter regular exponential family) is used so that $\Omega$ is an open $k$-dimensional convex set (usually an open interval or cross product of open intervals). Then the log likelihood function $\log[L(\eta)]$ is a strictly concave function of $\eta$. Hence if $\hat{\eta}$ is a critical point of $\log[L(\eta)]$ and if $\hat{\eta} \in \Omega$, then $\hat{\eta}$ is the unique MLE of $\eta$. Hence the Hessian matrix of 2nd derivatives does not need to be checked! Moreover, the critical point $\hat{\eta}$ is the solution to the equations $T_j(y) = \sum_{m=1}^{n} t_j(y_m) = \sum_{m=1}^{n} E[t_j(Y_m)] = E[T_j(Y)]$ for $j = 1, \ldots, k$.

Proof Sketch. The proof needs several results from nonlinear programming = optimization theory. Suppose that $h(\eta)$ is a function with continuous first and second derivatives on an open convex set $\Omega$. Suppose the $h(\eta)$ has Hessian (or Jacobian) matrix $H$ with $ij$ entry

$$H_{ij}(\eta) = \frac{\partial^2}{\partial \eta_i \partial \eta_j} h(\eta).$$

Let the critical point $\hat{\eta}$ be the solution to

$$\frac{\partial}{\partial \eta_i} h(\eta) \bigg|_{\eta = \hat{\eta}} = 0$$

for $i = 1, \ldots, k$. If $H(\eta)$ is positive definite on $\Omega$, then $h$ is strictly convex and $\hat{\eta}$ is a global minimizer of $h$. Hence $-h$ is strictly concave and $\hat{\eta}$ is a global maximizer of $h$. See Peressini, Sullivan and Uhl (1988, p. 10-13, 54).

Suppose $Y$ comes from a $k$–parameter exponential family. Then the distribution of $(t_1(Y), \ldots, t_k(Y))$ is a $k$–parameter exponential family by Theorem 3.5 that has a covariance matrix $\Sigma_t$ by Theorem 3.4. An important fact is that $\Sigma_t$ is positive definite.
Let $Y_1, ..., Y_n$ be iid from a kP-REF with pdf or pmf

\[ f(y|\eta) = h(y)b(\eta) \exp \left[ \sum_{i=1}^{k} \eta_i t_i(y) \right] . \]

As stated above Definition 3.3, $\Omega$ is an open convex set. Now the likelihood function\n
\[ L(\eta) = \prod_{m=1}^{n} f(y_m|\eta) = \left[ \prod_{m=1}^{n} h(y_m) \right]^{n} \exp[\eta_1 \sum_{m=1}^{n} t_1(y_m) + \cdots + \eta_k \sum_{m=1}^{n} t_k(y_m)] . \]

Hence the log likelihood function

\[ \log[L(\eta)] = d + n \log(b(\eta)) + \sum_{j=1}^{k} \eta_j T_j \]

where $T_j \equiv T_j(y) = \sum_{m=1}^{n} t_j(y_m)$.

Now

\[ \frac{\partial}{\partial \eta_i} \log[L(\eta)] = n \frac{\partial}{\partial \eta_i} \log(b(\eta)) + T_i = -n E[t_i(Y)] + T_i = \]

\[ - \sum_{m=1}^{n} E[t_i(Y_m)] + \sum_{m=1}^{n} t_i(y_m) \]

for $i = 1, ..., k$ by Theorem 3.3a. Since the critical point $\hat{\eta}$ is the solution to the $k$ equations $\frac{\partial}{\partial \eta_i} \log[L(\eta)] = 0$, the critical point $\hat{\eta}$ is also a solution to $T_i(y) = \sum_{m=1}^{n} t_i(y_m) = \sum_{m=1}^{n} E[t_i(Y_m)] = E[T_i(Y)]$ for $i = 1, ..., k$.

Now

\[ \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log[L(\eta)] = n \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\eta)) = -n \text{Cov}(t_i(Y), t_j(Y)) \]

by Theorem 3.3b, and the covariance matrix $\Sigma_t$ with $ij$ entry $\text{Cov}(t_i(Y), t_j(Y))$ is positive definite. Thus $h(\eta) = -\log[L(\eta)]$ has a positive definite Hessian and is strictly convex with global min $\hat{\eta}$ if $\hat{\eta} \in \Omega$. So $\log[L(\eta)]$ is strictly concave with global max $\hat{\eta}$ if $\hat{\eta} \in \Omega$.

**Remark 5.3.** For $k$-parameter exponential families with $k > 1$, it is usually easier to verify that the family is regular than to calculate the Hessian
matrix. For 1P–REFs, check that the critical point is a global maximum using standard calculus techniques such as calculating the second derivative of the log likelihood \( \log L(\theta) \). For a 1P–REF (one parameter regular exponential family), verifying that the family is regular is often more difficult than using calculus. Also, often the MLE is desired for a parameter space \( \Theta_U \) which is not an open set (eg for \( \Theta_U = [0, 1] \) instead of \( \Theta = (0, 1) \)).

**Remark 5.4, (Barndorff–Nielsen 1982).** The MLE does not exist if \( \hat{\eta} \) is not in \( \Omega \), an event that occurs with positive probability for discrete distributions. If \( T \) is the complete sufficient statistic and \( C \) is the closed convex hull of the support of \( T \), then the MLE exists iff \( T \in \text{int } C \) where \( \text{int } C \) is the interior of \( C \).

**Remark 5.5.** Suppose \( L(\theta) = f(y|\theta) = g(y|\theta)I(y \in \mathcal{Y}_\theta) = g(y|\theta)I(\theta \in A_y) = g(y|\theta)I(\theta \in A_y) + 0I(\theta \in A^c_y) \) where \( A^c_y \) is the complement of \( A_y \). Then \( \text{log}(L(\theta)) = \text{log}[g(y|\theta)]I(\theta \in A_y) + (-\infty)I(\theta \in A^c_y) \). Neither \( L(\theta) \) nor \( \text{log}(L(\theta)) \) is maximized for \( \theta \in A^c_y \), and for \( \theta \in A_y \), the log likelihood \( \text{log}(L(\theta)) = \text{log}[g(y|\theta)]I(y \in \mathcal{Y}_\theta) = \text{log}[g(y|\theta)]I(\theta \in A_y) \). Thus if \( L(\theta) = g(y|\theta)I(y \in \mathcal{Y}_\theta) \), do not do anything to the indicator when finding \( \text{log}(L(\theta)) \), but only consider values of \( \theta \) for which the indicator is equal to one when maximizing \( \text{log}(L(\theta)) \).

**Remark 5.6.** As illustrated in the following examples, the 2nd derivative is evaluated at \( \hat{\theta}(y) \). The MLE is a statistic and \( T_n(y) = \hat{\theta}(y) \) is the observed value of the MLE \( T_n(Y) = \hat{\theta}(Y) \). Often \( y \) and \( Y \) are suppressed.

**Example 5.1.** Suppose that \( Y_1, ..., Y_n \) are iid Poisson \((\theta)\). This distribution is a 1P–REF with \( \Theta = (0, \infty) \). The likelihood

\[
L(\theta) = c \ e^{-n\theta} \exp[\text{log}(\theta) \sum y_i]
\]

where the constant \( c \) does not depend on \( \theta \), and the log likelihood

\[
\text{log}(L(\theta)) = d - n\theta + \text{log}(\theta) \sum y_i
\]

where \( d = \text{log}(c) \) does not depend on \( \theta \). Hence

\[
\frac{d}{d\theta} \text{log}(L(\theta)) = -n + \frac{1}{\theta} \sum y_i \overset{\text{set}}{=} 0,
\]

or \( \sum y_i = n\theta \), or

\[
\hat{\theta} = \bar{y}.
\]
Notice that \( \hat{\theta} \) is the unique solution and
\[
\frac{d^2}{d\theta^2} \log(L(\theta)) = -\frac{\sum y_i}{\theta^2} < 0
\]
unless \( \sum y_i = 0 \). Hence for \( \sum y_i > 0 \) the log likelihood is strictly concave and \( \bar{Y} \) is the MLE of \( \theta \). The MLE does not exist if \( \sum_{i=1}^n Y_i = 0 \) since 0 is not in \( \Theta \).

Now suppose that \( \Theta = [0, \infty) \). This family is not an exponential family since the same formula for the pmf needs to hold for all values of \( \theta \in \Theta \) and \( 0^0 \) is not defined. Notice that
\[
f(y|\theta) = e^{-\theta y} I[\theta > 0] + 1 I[\theta = 0, y = 0].
\]

Now
\[
I_A(\theta)I_B(\theta) = I_{A \cap B}(\theta)
\]
and \( I_{\emptyset}(\theta) = 0 \) for all \( \theta \). Hence the likelihood
\[
L(\theta) = e^{-n\theta} \exp[\log(\theta) \sum_{i=1}^n y_i] \frac{1}{\prod_{i=1}^n y_i} I[\theta > 0] + 1 I[\theta = 0, \sum_{i=1}^n y_i = 0].
\]

If \( \sum y_i \neq 0 \), then \( \bar{y} \) maximizes \( L(\theta) \) by the work above. If \( \sum y_i = 0 \), then \( L(\theta) = e^{-n\theta} I(\theta > 0) + I(\theta = 0) = e^{-n\theta} I(\theta \geq 0) \) which is maximized by \( \theta = 0 = \bar{y} \). Hence \( \bar{Y} \) is the MLE of \( \theta \) if \( \Theta = [0, \infty) \).

By invariance, \( t(\bar{Y}) \) is the MLE of \( t(\theta) \). Hence \( \bar{Y}^2 \) is the MLE of \( \theta^2 \). \( \sin(\bar{Y}) \) is the MLE of \( \sin(\theta) \), etcetera.

**Example 5.2.** Suppose that \( Y_1, \ldots, Y_n \) are iid \( N(\mu, \sigma^2) \) where \( \sigma^2 > 0 \) and \( \mu \in \mathcal{R} = (-\infty, \infty) \). Then
\[
L(\mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{(\sigma^2)^{n/2}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 \right].
\]

Notice that
\[
\frac{d}{d\mu} \sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n -2(y_i - \mu) \text{ \text{set}} = 0
\]
or \( \sum_{i=1}^n y_i = n\mu \) or \( \mu = \bar{y} \). Since \( \hat{\mu} \) is the unique solution and
\[
\frac{d^2}{d\mu^2} \sum_{i=1}^n (y_i - \mu)^2 = 2n > 0,
\]
\( \hat{\mu} = \overline{y} \) is the minimizer of \( h(\mu) = \sum_{i=1}^{n}(y_i - \mu)^2 \). Hence \( \overline{y} \) is the maximizer of

\[
\exp \left[ \frac{-1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2 \right]
\]

regardless of the value of \( \sigma^2 > 0 \). Hence \( \hat{\mu} = \overline{Y} \) is the MLE of \( \mu \) and the MLE of \( \sigma^2 \) can be found by maximizing the profile likelihood

\[
L_P(\sigma^2) = L(\hat{\mu}(\overline{y}), \sigma^2) = \left( \frac{1}{\sqrt{2\pi}} \right)^n \frac{1}{(\sigma^2)^{n/2}} \exp \left[ \frac{-1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \overline{y})^2 \right].
\]

Writing \( \tau = \sigma^2 \) often helps prevent calculus errors. Then

\[
\log(L_p(\tau)) = d - \frac{n}{2} \log(\tau) + \frac{-1}{2\tau} \sum_{i=1}^{n}(y_i - \overline{y})^2
\]

where the constant \( d \) does not depend on \( \tau \). Hence

\[
\frac{d}{d\tau} \log(L_p(\tau)) = -\frac{n}{2 \tau} + \frac{1}{2\tau^2} \sum_{i=1}^{n}(y_i - \overline{y})^2 \text{ set } 0,
\]

or

\[
n\tau = \sum_{i=1}^{n}(y_i - \overline{y})^2
\]

or

\[
\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n}(y_i - \overline{y})^2
\]

and the solution \( \hat{\tau} \) is the unique critical point. Note that

\[
\frac{d^2}{d\mu^2} \log(L_P(\tau)) = \frac{n}{2(\tau)^2} - \frac{\sum(y_i - \overline{y})^2}{(\tau)^3} \Bigg|_{\tau = \hat{\tau}} = \frac{n}{2(\hat{\tau})^2} - \frac{n\hat{\tau}}{(\hat{\tau})^3/2}
\]

\[
= -\frac{n}{2(\hat{\tau})^2} < 0.
\]

Hence \( \hat{\sigma}^2 = \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \) is the MLE of \( \sigma^2 \) by Remark 5.1 V). Thus \((\overline{Y}, \frac{1}{n} \sum_{i=1}^{n}(Y_i - \overline{Y})^2) \) is the MLE of \((\mu, \sigma^2)\).
Example 5.3. Following Pewsey (2002), suppose that $Y_1, ..., Y_n$ are iid $HN(\mu, \sigma^2)$ where $\mu$ and $\sigma^2$ are both unknown. Let the $i$th order statistic $Y_{(i)} \equiv Y_{(i:n)}$. Then the likelihood

$$L(\mu, \sigma^2) = c I[y_{1:n} \geq \mu] \frac{1}{\sigma^n} \exp \left( \frac{-1}{2\sigma^2} \sum (y_i - \mu)^2 \right).$$

For any fixed $\sigma^2 > 0$, this likelihood is maximized by making $\sum (y_i - \mu)^2$ as small as possible subject to the constraint $y_{1:n} \geq \mu$. Notice that for any $\mu_o < y_{1:n}$, the terms $(y_i - y_{1:n})^2 < (y_i - \mu_o)^2$. Hence the MLE of $\mu$ is

$$\hat{\mu} = Y_{1:n}$$

and the MLE of $\sigma^2$ is found by maximizing the log profile likelihood

$$\log(L_P(\sigma^2)) = \log(L(y_{1:n}, \sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - y_{1:n})^2,$$

and

$$\frac{d}{d(\sigma^2)} \log(L(y_{1:n}, \sigma^2)) = -\frac{n}{2(\sigma^2)} + \frac{1}{2(\sigma^2)^2} \sum (y_i - y_{1:n})^2 \overset{set}{=} 0.$$

Or $\sum (y_i - y_{1:n})^2 = n\sigma^2$. So

$$\hat{\sigma}^2 \equiv w_n = \frac{1}{n} \sum (y_i - y_{1:n})^2.$$

Since the solution $\hat{\sigma}^2$ is unique and

$$\frac{d^2}{d(\sigma^2)^2} \log(L(y_{1:n}, \sigma^2)) =$$

$$\frac{n}{2(\sigma^2)^2} - \frac{\sum (y_i - \mu)^2}{(\sigma^2)^3} \bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{(\hat{\sigma}^2)^3/2} = -\frac{n}{2\hat{\sigma}^2} < 0,$$

$(\hat{\mu}, \hat{\sigma}^2) = (Y_{1:n}, W_n)$ is MLE of $(\mu, \sigma^2)$.

Example 5.4. Suppose that the random vectors $X_1, ..., X_n$ are iid from a multivariate normal $N_p(\mu, \Sigma)$ distribution where $\Sigma$ is a positive definite matrix. To find the MLE of $(\mu, \Sigma)$ we will use three results proved in Anderson (1984, p. 62).

1) $\sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) = \text{tr}(\Sigma^{-1} A) + n(\overline{x} - \mu)^T \Sigma^{-1} (\overline{x} - \mu)$
CHAPTER 5. POINT ESTIMATION

where

\[ A = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T. \]

ii) Let \( C \) and \( D \) be positive definite matrices. Then \( C = \frac{1}{n} D \) maximizes

\[ h(C) = -n \log(|C|) - tr(C^{-1} D) \]

with respect to positive definite matrices.

iii) Since \( \Sigma^{-1} \) is positive definite, \( (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \geq 0 \) as a function of \( \mu \) with equality iff \( \mu = \bar{x} \).

Since

\[ f(x|\mu, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right], \]

the likelihood function

\[ L(\mu, \Sigma) = \prod_{i=1}^{n} f(x_i|\mu, \Sigma) \]

\[ = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \right], \]

and the log likelihood log\( (L(\mu, \Sigma)) = \)

\[ -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) \]

\[ = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log(|\Sigma|) - \frac{1}{2} tr(\Sigma^{-1} A) - \frac{n}{2} (\bar{x} - \mu)^T \Sigma^{-1} (\bar{x} - \mu) \]

by i). Now the last term is maximized by \( \mu = \bar{x} \) by iii) and the middle two terms are maximized by \( \frac{1}{n} A \) by ii) since \( \Sigma \) and \( A \) are both positive definite. Hence the MLE of \( (\mu, \Sigma) \) is

\[ (\hat{\mu}, \hat{\Sigma}) = (\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T). \]

**Example 5.5.** Let \( X_1, \ldots, X_n \) be independent identically distributed random variables from a lognormal \( (\mu, \sigma^2) \) distribution with pdf

\[ f(x) = \frac{1}{x \sqrt{2\pi \sigma^2}} \exp \left( \frac{-(\log(x) - \mu)^2}{2\sigma^2} \right) \]
where $\sigma > 0$ and $x > 0$ and $\mu$ is real. **Assume that $\sigma$ is known.**

a) Find the maximum likelihood estimator of $\mu$.

b) What is the maximum likelihood estimator of $\mu^3$? Explain.

Solution: a)

$$\hat{\mu} = \frac{\sum \log(X_i)}{n}$$

To see this note that

$$L(\mu) = \left( \prod \frac{1}{x_i \sqrt{2\pi \sigma^2}} \right) \exp\left( -\frac{\sum (\log(x_i) - \mu)^2}{2\sigma^2} \right).$$

So

$$\log(L(\mu)) = \log(c) - \frac{\sum (\log(x_i) - \mu)^2}{2\sigma^2}$$

and the derivative of the log likelihood wrt $\mu$ is

$$\frac{\sum 2(\log(x_i) - \mu)}{2\sigma^2}.$$ Setting this quantity equal to 0 gives $n\mu = \sum \log(x_i)$ and the solution $\hat{\mu}$ is unique. The second derivative is $-n/\sigma^2 < 0$, so $\hat{\mu}$ is indeed the global maximum.

b) 

$$\left( \frac{\sum \log(X_i)}{n} \right)^3$$ by invariance.

**Example 5.6.** Suppose that the joint probability distribution function of $X_1, ..., X_k$ is

$$f(x_1, x_2, ..., x_k|\theta) = \frac{n!}{(n-k)!\theta^k} \exp\left( -\frac{\left[ (\sum_{i=1}^{k} x_i) + (n-k)x_k \right]}{\theta} \right)$$

where $0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$ and $\theta > 0$.

a) Find the maximum likelihood estimator (MLE) for $\theta$.

b) What is the MLE for $\theta^2$? Explain briefly.

Solution: a) Let $t = [\sum_{i=1}^{k} x_i] + (n-k)x_k$. $L(\theta) = f(\mathbf{x}|\theta)$ and $\log(L(\theta)) = \log(f(\mathbf{x}|\theta)) = d - k \log(\theta) - \frac{t}{\theta}$. 

Hence
\[ \frac{d}{d\theta} \log(L(\theta)) = -\frac{k}{\theta} + \frac{t}{\theta^2} = 0. \]

Hence
\[ k\theta = t \]
or
\[ \hat{\theta} = \frac{t}{k}. \]

This is a unique solution and
\[ \frac{d^2}{d\theta^2} \log(L(\theta)) = \frac{k}{\theta^2} - \frac{2t}{\theta^3} \bigg|_{\theta=\hat{\theta}} = \frac{k}{\hat{\theta}^2} - \frac{2k\hat{\theta}}{\hat{\theta}^3} = -\frac{k}{\hat{\theta}^2} < 0. \]

Hence \( \hat{\theta} = T/k \) is the MLE where \( T = [(\sum_{i=1}^k X_i) + (n - k)X_k] \).

b) \( \hat{\theta}^2 \) by the invariance principle.

**Example 5.7.** Let \( X_1, \ldots, X_n \) be independent identically distributed random variables with pdf
\[ f(x) = \frac{1}{\lambda} x^{\frac{1}{\lambda} - 1}, \]
where \( \lambda > 0 \) and \( 0 < x \leq 1 \).

a) Find the maximum likelihood estimator of \( \lambda \).

b) What is the maximum likelihood estimator of \( \lambda^3 \)? Explain.

**Solution:**

a) For \( 0 < x \leq 1 \)
\[ f(x) = \frac{1}{\lambda} \exp \left( \frac{1}{\lambda} - 1 \right) \log(x) \right). \]

Hence the likelihood
\[ L(\lambda) = \frac{1}{\lambda^n} \exp \left( \frac{1}{\lambda} - 1 \right) \sum \log(x_i) \right), \]
and the log likelihood
\[ \log(L(\lambda)) = -n \log(\lambda) + \left( \frac{1}{\lambda} - 1 \right) \sum \log(x_i). \]

Hence
\[ \frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} - \frac{\sum \log(x_i)}{\lambda^2} \frac{\text{set}}{\lambda^2} = 0, \]
or \(-\sum \log(x_i) = n\lambda\), or

\[
\hat{\lambda} = \frac{-\sum \log(x_i)}{n}.
\]

Notice that \(\hat{\lambda}\) is the unique solution and that

\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} + \frac{2\sum \log(x_i)}{\lambda^3} \bigg|_{\lambda = \hat{\lambda}}
\]

\[
= \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\hat{\lambda}^3} = \frac{-n}{\lambda^2} < 0.
\]

Hence \(\hat{\lambda} = -\sum \log(X_i)/n\) is the MLE of \(\lambda\).

b) By invariance, \(\hat{\lambda}^3\) is the MLE of \(\lambda\).

**Example 5.8.** Suppose \(Y_1, ..., Y_n\) are iid \(U(\theta - 1, \theta + 1)\). Then

\[
L(\theta) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \frac{1}{2} I(\theta - 1 \leq y_i \leq \theta + 1) = \frac{1}{2^n} I(\theta - 1 \leq y_i \leq \theta + 1)
\]

\[
= \frac{1}{2^n} I(\theta - 1 \leq y_{(1)} \leq y_{(n)} \leq \theta + 1) = \frac{1}{2^n} I(y_{(n)} - 1 \leq \theta \leq y_{(1)} + 1).
\]
Let $0 \leq c \leq 1$. Then any estimator of the form $\hat{\theta}_c = c[Y_{(n)} - 1] + (1-c)[Y_{(1)} + 1]$ is an MLE of $\theta$. Figure 5.2 shows $L(\theta)$ for $U(2, 4)$ data with $n = 10$, $y_{(1)} = 2.0375$ and $y_{(n)} = 3.7383$.

**Remark 5.7.** Chapter 10 has many MLE examples.

### 5.2 Method of Moments Estimators

The method of moments is another useful way for obtaining point estimators. Let $Y_1, ..., Y_n$ be an iid sample and let

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y_i^j$$

and $\mu_j \equiv \mu_j(\theta) = E_\theta(Y^j)$ (5.2)

for $j = 1, ..., k$. So $\hat{\mu}_j$ is the $j$th sample moment and $\mu_j$ is the $j$th population moment. Fix $k$ and assume that $\mu_j = \mu_j(\theta_1, ..., \theta_k)$. Solve the system

$$\begin{align*}
\hat{\mu}_1 &= \mu_1(\theta_1, ..., \theta_k) \\
\vdots \\
\hat{\mu}_k &= \mu_k(\theta_1, ..., \theta_k)
\end{align*}$$

for $\tilde{\theta}$.

**Definition 5.3.** The solution $\tilde{\theta} = (\tilde{\theta}_1, ..., \tilde{\theta}_k)$ is the **method of moments estimator** of $\theta$. If $g$ is a continuous function of the first $k$ moments and $h(\theta) = g(\mu_1(\theta), ..., \mu_k(\theta))$, then the method of moments estimator of $h(\theta)$ is

$$g(\hat{\mu}_1, ..., \hat{\mu}_k).$$

Sometimes the notation $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MM}$ will be used to denote the MLE and method of moments estimators of $\theta$, respectively.

**Example 5.9.** Let $Y_1, ..., Y_n$ be iid from a distribution with a given pdf or pmf $f(y|\theta)$.

- a) If $E(Y) = h(\theta)$, then $\hat{\theta}_{MM} = h^{-1}(\bar{y})$.
- b) The method of moments estimator of $E(Y) = \mu_1$ is $\hat{\mu}_1 = \bar{y}$. 
c) The method of moments estimator of \( \text{VAR}_\theta(Y) = \mu_2(\theta) - [\mu_1(\theta)]^2 \) is

\[
\hat{\sigma}_{MM}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2 - (\bar{Y})^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \equiv S_M^2.
\]

Method of moments estimators need not be unique. For example both \( \bar{Y} \) and \( S_M^2 \) are method of moment estimators of \( \theta \) for iid Poisson(\( \theta \)) data. Generally the method of moments estimators that use small \( j \) for \( \hat{\mu}_j \) are preferred, so use \( \bar{Y} \) for Poisson data.

**Proposition 5.3.** Let \( S_M^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) and suppose that \( E(Y) = h_1(\theta_1, \theta_2) \) and \( V(Y) = h_2(\theta_1, \theta_2) \). Then solving

\[
\begin{align*}
\bar{Y} &\overset{\text{set}}{=} h_1(\theta_1, \theta_2) \\
S_M^2 &\overset{\text{set}}{=} h_2(\theta_1, \theta_2)
\end{align*}
\]

for \( \tilde{\theta} \) is a method of moments estimator.

**Proof.** Notice that \( \mu_1 = h_1(\theta_1, \theta_2) = \mu_1(\theta_1, \theta_2) \) while \( \mu_2 - [\mu_1]^2 = h_2(\theta_1, \theta_2) \). Hence \( \mu_2 = h_2(\theta_1, \theta_2) + [h_1(\theta_1, \theta_2)]^2 = \mu_2(\theta_1, \theta_2) \). Hence the method of moments estimator is a solution to \( \bar{Y} \overset{\text{set}}{=} \mu_1(\theta_1, \theta_2) \) and \( \frac{1}{n} \sum_{i=1}^{n} Y_i^2 \overset{\text{set}}{=} h_2(\theta_1, \theta_2) + [\mu_1(\theta_1, \theta_2)]^2 \). Equivalently, solve \( \bar{Y} \overset{\text{set}}{=} h_1(\theta_1, \theta_2) \) and

\[
\frac{1}{n} \sum_{i=1}^{n} Y_i^2 - (\bar{Y})^2 = S_M^2 \overset{\text{set}}{=} h_2(\theta_1, \theta_2). \]

**QED**

**Example 5.10.** Suppose that \( Y_1, \ldots, Y_n \) be iid gamma (\( \nu, \lambda \)). Then \( \hat{\mu}_1 \overset{\text{set}}{=} E(Y) = \nu \lambda \) and \( \hat{\mu}_2 \overset{\text{set}}{=} E(Y^2) = \text{VAR}(Y) + [E(Y)]^2 = \nu \lambda^2 + \nu^2 \lambda^2 = \nu \lambda^2 (1 + \nu) \). Substitute \( \nu = \hat{\mu}_1 / \lambda \) into the 2nd equation to obtain

\[
\hat{\mu}_2 = \frac{\hat{\mu}_1 \lambda^2 (1 + \frac{\hat{\mu}_1}{\lambda})}{\lambda} = \lambda \hat{\mu}_1 + \hat{\mu}_1^2.
\]

Thus

\[
\tilde{\lambda} = \frac{\hat{\mu}_2 - \hat{\mu}_1^2}{\hat{\mu}_1} = \frac{S_M^2}{\bar{Y}} \quad \text{and} \quad \tilde{\nu} = \frac{\hat{\mu}_1}{\lambda} = \frac{\hat{\mu}_1^2}{\hat{\mu}_2 - \hat{\mu}_1^2} = \frac{(\bar{Y})^2}{S_M^2}.
\]

Alternatively, solve \( \bar{Y} \overset{\text{set}}{=} \nu \lambda \) and \( S_M^2 \overset{\text{set}}{=} \nu \lambda^2 = (\nu \lambda) \lambda \). Hence \( \tilde{\lambda} = S_M^2 / \bar{Y} \) and

\[
\tilde{\nu} = \frac{\bar{Y}}{\lambda} = \frac{(\bar{Y})^2}{S_M^2}.
\]
5.3 Summary

Let \( f(y|\theta) \) be the pmf or pdf of a sample \( Y \). If \( Y = y \) is observed, then the likelihood function \( L(\theta) = f(y|\theta) \). For each sample point \( y = (y_1, ..., y_n) \), let \( \hat{\theta}(y) \) be a parameter value at which \( L(\theta|y) \) attains its maximum as a function of \( \theta \) with \( y \) held fixed. Then a maximum likelihood estimator (MLE) of the parameter \( \theta \) based on the sample \( Y \) is \( \hat{\theta}(Y) \).

Note: it is crucial to observe that the likelihood function is a function of \( \theta \) (and that \( y_1, ..., y_n \) act as fixed constants).

Note: If \( Y_1, ..., Y_n \) is an independent sample from a population with pdf or pmf \( f(y|\theta) \) then the likelihood function

\[
L(\theta) = L(\theta|y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i|\theta).
\]

Note: If the MLE \( \hat{\theta} \) exists, then \( \hat{\theta} \in \Theta \).

Potential candidates for the MLE can sometimes be found by differentiating \( L(\theta) \), by direct maximization, or using the profile likelihood. Typically the log likelihood is used as in A) below.

A) Let \( Y_1, ..., Y_n \) be iid with pdf or pmf \( f(y|\theta) \). Then \( L(\theta) = \prod_{i=1}^{n} f(y_i|\theta) \).

To find the MLE,

i) find \( L(\theta) \) and then find the log likelihood \( \log L(\theta) \).

ii) Find the derivative \( \frac{d}{d\theta} \log L(\theta) \), set the derivative equal to zero and solve for \( \theta \). The solution is a candidate for the MLE.

iii) Invariance Principle: If \( \hat{\theta} \) is the MLE of \( \theta \), then \( \tau(\hat{\theta}) \) is the MLE of \( \tau(\theta) \).

iv) Show that \( \hat{\theta} \) is the MLE by showing that \( \hat{\theta} \) is the global maximizer of \( \log L(\theta) \). Often this is done by noting that \( \hat{\theta} \) is the unique solution to the equation \( \frac{d}{d\theta} \log L(\theta) = 0 \) and that the 2nd derivative evaluated at \( \hat{\theta} \) is negative: \( \frac{d^2}{d\theta^2} \log L(\theta)|_{\hat{\theta}} < 0 \).

B) If \( \log L(\theta) \) is strictly concave (\( \frac{d^2}{d\theta^2} \log L(\theta) < 0 \) for all \( \theta \in \Theta \)), then any local max of \( \log L(\theta) \) is a global max.

C) Know how to find the MLE for the normal distribution (including
when \( \mu \) or \( \sigma^2 \) is known). Memorize the MLEs

\[
\bar{Y}, \quad S^2_M = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2
\]

for the normal and for the uniform distribution. Also \( \bar{Y} \) is the MLE for \( V(Y) \) and is the MLE for \( V(Y) \) if the data are iid \( N(\mu, \sigma^2) \).

D) **On qualifying exams**, the \( N(\mu, \mu) \) and \( N(\mu, \mu^2) \) distributions are common. See Problem 5.35.

E) **Indicators** are useful. For example, \( \prod_{i=1}^{n} I_{A_i}(y_i) = I(\text{all } y_i \in A) \) and \( \prod_{j=1}^{k} I_{A_j}(y) = I(\cap_{j=1}^{k} A_j(y)) \). Hence \( I(0 \leq y \leq \theta) = I(0 \leq y)I(y \leq \theta) \), and \( \prod_{i=1}^{n} I(\theta_1 \leq y_i \leq \theta_2) = I(\theta_1 \leq y(1) \leq y(n) \leq \theta_2) = I(\theta_1 \leq y(1))I(y(n) \leq \theta_2) \).

F) Suppose \( X_1, \ldots, X_n \) are iid with pdf or pmf \( f(x|\lambda) \) and \( Y_1, \ldots, Y_n \) are iid with pdf or pmf \( g(y|\mu) \). Suppose that the \( X \)'s are independent of the \( Y \)'s. Then

\[
\sup_{(\lambda, \mu) \in \Theta} L(\lambda, \mu|x, y) \leq \sup_{\lambda} L_x(\lambda) \sup_{\mu} L_y(\mu)
\]

where \( L_x(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) \). Hence if \( \hat{\lambda} \) is the marginal MLE of \( \lambda \) and \( \hat{\mu} \) is the marginal MLE of \( \mu \), then \( (\hat{\lambda}, \hat{\mu}) \) is the MLE of \( (\lambda, \mu) \) provided that \( (\hat{\lambda}, \hat{\mu}) \) is in the parameter space \( \Theta \).

G) Let \( \hat{\mu}_j = \frac{1}{n} \sum_{i=1}^{n} Y^j_i \), let \( \mu_j = E(Y^j) \) and assume that \( \mu_j = \mu_j(\theta_1, \ldots, \theta_k) \). Solve the system

\[
\hat{\mu}_1^{\text{set}} = \mu_1(\theta_1, \ldots, \theta_k) \\
\vdots \\
\hat{\mu}_k^{\text{set}} = \mu_k(\theta_1, \ldots, \theta_k)
\]

for the method of moments estimator \( \hat{\theta} \).

H) If \( g \) is a continuous function of the first \( k \) moments and \( h(\theta) = g(\mu_1(\theta), \ldots, \mu_k(\theta)) \), then the method of moments estimator of \( h(\theta) \) is \( g(\hat{\mu}_1, \ldots, \hat{\mu}_k) \).

**5.4 Complements**

Optimization theory is also known as nonlinear programming and shows how to find the global max and min of a multivariate function. Peressini, Sullivan
CHAPTER 5. POINT ESTIMATION

and Uhl (1988) is an undergraduate text. Also see Sundaram (1996) and Bertsekas (1999).

Maximum likelihood estimation is widely used in statistical models. See Pawitan (2001) and texts for Categorical Data Analysis, Econometrics, Multiple Linear Regression, Generalized Linear Models, Multivariate Analysis and Survival Analysis.

Suppose that \( Y = t(W) \) and \( W = t^{-1}(Y) \) where \( W \) has a pdf with parameters \( \theta \), the transformation \( t \) does not depend on any unknown parameters, and the pdf of \( Y \) is

\[
    f_Y(y) = f_W(t^{-1}(y)) \left| \frac{dt^{-1}(y)}{dy} \right|.
\]

If \( W_1, ..., W_n \) are iid with pdf \( f_W(w) \), assume that the MLE of \( \theta \) is \( \hat{\theta}_W(w) \) where the \( w_i \) are the observed values of \( W_i \) and \( w = (w_1, ..., w_n) \). If \( Y_1, ..., Y_n \) are iid and the \( y_i \) are the observed values of \( Y_i \), then the likelihood is

\[
    L_Y(\theta) = \left( \prod_{i=1}^{n} \left| \frac{dt^{-1}(y_i)}{dy} \right| \right) \prod_{i=1}^{n} f_W(t^{-1}(y_i)|\theta) = c \prod_{i=1}^{n} f_W(t^{-1}(y_i)|\theta).
\]

Hence the log likelihood is \( \log(L_Y(\theta)) = \)

\[
    d + \sum_{i=1}^{n} \log[f_W(t^{-1}(y_i)|\theta)] = d + \sum_{i=1}^{n} \log[f_W(w_i|\theta)] = d + \log[L_W(\theta)]
\]

where \( w_i = t^{-1}(y_i) \). Hence maximizing the \( \log(L_Y(\theta)) \) is equivalent to maximizing \( \log(L_W(\theta)) \) and

\[
    \hat{\theta}_Y(y) = \hat{\theta}_W(w) = \hat{\theta}_W(t^{-1}(y_1), ..., t^{-1}(y_n)). \tag{5.3}
\]


**Example 5.11.** Suppose \( Y_1, ..., Y_n \) are iid lognormal \((\mu, \sigma^2)\). Then \( W_i = \log(Y_i) \sim N(\mu, \sigma^2) \) and the MLE \( (\hat{\mu}, \hat{\sigma}^2) = (\bar{W}, \frac{1}{n} \sum_{i=1}^{n}(W_i - \bar{W})^2) \).

One of the most useful properties of the maximum likelihood estimator is the invariance property: if \( \hat{\theta} \) is the MLE of \( \theta \), then \( \tau(\hat{\theta}) \) is the MLE of \( \tau(\theta) \). Olive (2004) is a good discussion of the MLE invariance principle. Also see Pal and Berry (1992). Many texts either define the MLE of \( \tau(\theta) \) to be \( \tau(\hat{\theta}) \),
say that the property is immediate from the definition of the MLE, or quote Zehna (1966). A little known paper, Berk (1967), gives an elegant proof of the invariance property that can be used in introductory statistical courses. The next subsection will show that Berk (1967) answers some questions about the MLE which can not be answered using Zehna (1966).

5.4.1 Two “Proofs” of the Invariance Principle

“Proof” I) The following argument of Zehna (1966) also appears in Casella and Berger (2002, p. 320). Let \( \theta \in \Theta \) and let \( h : \Theta \to \Lambda \) be a function. Since the MLE 

\[
\hat{\theta} \in \Theta, \quad h(\hat{\theta}) = \hat{\lambda} \in \Lambda.
\]

If \( h \) is not one to one, then many values of \( \theta \) may be mapped to \( \lambda \). Let

\[
\Theta_\lambda = \{ \theta : h(\theta) = \lambda \}
\]

and define the induced likelihood function \( M(\lambda) \) by

\[
M(\lambda) = \sup_{\theta \in \Theta_\lambda} L(\theta).
\]

(5.4)

Then for any \( \lambda \in \Lambda \),

\[
M(\lambda) = \sup_{\theta \in \Theta_\lambda} L(\theta) \leq \sup_{\theta \in \Theta} L(\theta) = L(\hat{\theta}) = M(\hat{\lambda}).
\]

(5.5)

Hence \( h(\hat{\theta}) = \hat{\lambda} \) maximizes the induced likelihood \( M(\lambda) \). Zehna (1966) says that since \( h(\hat{\theta}) \) maximizes the induced likelihood, we should call \( h(\hat{\theta}) \) the MLE of \( h(\theta) \), but the definition of MLE says that we should be maximizing a genuine likelihood.

This argument raises two important questions.

- If we call \( h(\hat{\theta}) \) the MLE of \( h(\theta) \) and \( h \) is not one to one, does \( h(\hat{\theta}) \) maximize a likelihood or should \( h(\hat{\theta}) \) be called a maximum induced likelihood estimator?

- If \( h(\hat{\theta}) \) is an MLE, what is the likelihood function \( K(h(\theta)) \)?

Some examples might clarify these questions.
• If the population come from a $N(\mu, \sigma^2)$ distribution, the invariance principle says that the MLE of $\mu/\sigma$ is $X/S_M$ where
\[
X = \frac{1}{n} \sum_{i=1}^{n} X_i
\]
and
\[
S_M^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - X)^2
\]
are the MLEs of $\mu$ and $\sigma^2$. Since the function $h(x, y) = x/\sqrt{y}$ is not one to one (eg $h(x, y) = 1$ if $x = \sqrt{y}$), what is the likelihood $K(h(\mu, \sigma^2)) = K(\mu/\sigma)$ that is being maximized?

• If $X_i$ comes from a Bernoulli($\rho$) population, why is $X_n(1 - X_n)$ the MLE of $\rho(1 - \rho)$?

Proof II) Examining the invariance principle for one to one functions $h$ is also useful. When $h$ is one to one, let $\eta = h(\theta)$. Then the inverse function $h^{-1}$ exists and $\theta = h^{-1}(\eta)$. Hence
\[
f(x|\theta) = f(x|h^{-1}(\eta))
\] is the joint pdf or pmf of $x$. So the likelihood function of $h(\theta) = \eta$ is
\[
L^*(\eta) \equiv K(\eta) = L(h^{-1}(\eta)).
\] (5.7)
Also note that
\[
\sup_{\eta} K(\eta|x) = \sup_{\eta} L(h^{-1}(\eta)|x) = L(\hat{\theta}|x).
\] (5.8)
Thus
\[
\hat{\eta} = h(\hat{\theta})
\] (5.9)
is the MLE of $\eta = h(\theta)$ when $h$ is one to one.

If $h$ is not one to one, then the new parameters $\eta = h(\theta)$ do not give enough information to define $f(x|\eta)$. Hence we cannot define the likelihood. That is, a $N(\mu, \sigma^2)$ density cannot be defined by the parameter $\mu/\sigma$ alone. Before concluding that the MLE does not exist if $h$ is not one to one, note that if $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$ then $X_1, ..., X_n$ remain iid $N(\mu, \sigma^2)$ even
though the investigator did not rename the parameters wisely or is interested in a function $h(\mu, \sigma) = \mu / \sigma$ that is not one to one. Berk (1967) said that if $h$ is not one to one, define \[ w(\theta) = (h(\theta), u(\theta)) = (\eta, \gamma) = \xi \] (5.10) such that $w(\theta)$ is one to one. Note that the choice \[ w(\theta) = (h(\theta), \theta) \] works. In other words, we can always take $u$ to be the identity function.

The choice of $w$ is not unique, but the inverse function \[ w^{-1}(\xi) = \theta \] is unique. Hence the likelihood is well defined, and $w(\hat{\theta})$ is the MLE of $\xi$.

**QED**

**Example 5.12.** Following Lehmann (1999, p. 466), let \[ f(x|\sigma) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \] where $x$ is real and $\sigma > 0$. Let $\eta = \sigma^k$ so $\sigma = \eta^{1/k} = h^{-1}(\eta)$. Then \[ f^*(x|\eta) = \frac{1}{\sqrt{2\pi} \eta^{1/k}} \exp\left(-\frac{x^2}{2\eta^{2/k}}\right) = f(x|\sigma = h^{-1}(\eta)). \]

Notice that calling $h(\hat{\theta})$ the MLE of $h(\theta)$ is analogous to calling $\overline{X}_n$ the MLE of $\mu$ when the data are from a $N(\mu, \sigma^2)$ population. It is often possible to choose the function $u$ so that if $\theta$ is a $p \times 1$ vector, then so is $\xi$. For the $N(\mu, \sigma^2)$ example with $h(\mu, \sigma^2) = h(\theta) = \mu / \sigma$ we can take $u(\theta) = \mu$ or $u(\theta) = \sigma^2$. For the Ber($\rho$) example, $w(\rho) = (\rho(1 - \rho), \rho)$ is a reasonable choice.

To summarize, Berk’s proof should be widely used to prove the invariance principle, and

1) changing the names of the parameters does not change the distribution of the sample, eg, if $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then $X_1, \ldots, X_n$ remain iid
CHAPTER 5. POINT ESTIMATION

$N(\mu, \sigma^2)$ regardless of the function $h(\mu, \sigma^2)$ that is of interest to the investigator.

II) The invariance principle holds as long as $h(\hat{\theta})$ is a random variable or random vector; $h$ does not need to be a one to one function. If there is interest in $\eta = h(\theta)$ where $h$ is not one to one, then additional parameters $\gamma = u(\theta)$ need to be specified so that $w(\theta) = \xi = (\eta, \gamma) = (h(\theta), u(\theta))$ has a well defined likelihood $K(\xi) = L(w^{-1}(\xi))$. Then by Definition 5.2, the MLE is $\hat{\xi} = w(\hat{\theta}) = w(h(\hat{\theta}), u(\hat{\theta}))$ and the MLE of $\eta = h(\theta)$ is $\hat{\eta} = h(\hat{\theta})$.

III) Using the identity function $\gamma = u(\theta) = \theta$ always works since $\xi = w(\theta) = (h(\theta), \theta)$ is a one to one function of $\theta$. However, using $u(\theta)$ such that $\xi$ and $\theta$ have the same dimension is often useful.

5.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

5.1*. Let $Y_1, ..., Y_n$ be iid binomial $(k = 1, \rho)$.

a) Assume that $\rho \in \Theta = (0, 1)$ and that $0 < \sum_{i=1}^{n} y_i < n$. Show that the MLE of $\rho$ is $\hat{\rho} = \bar{Y}$.

b) Now assume that $\rho \in \Theta = [0, 1]$. Show that $f(y|\rho) = \rho^y (1-\rho)^{1-y} I(0 < \rho < 1) + I(\rho = 0, y = 0) + I(\rho = 1, y = 1)$. Then show that

$L(\rho) = \rho^\sum_y (1-\rho)^{n-\sum_y} I(0 < \rho < 1) + I(\rho = 0, \sum y = 0) + I(\rho = 1, \sum y = n)$. If $\sum y = 0$ show that $\hat{\rho} = 0 = \bar{y}$. If $\sum y = n$ show that $\hat{\rho} = 1 = \bar{y}$. Then explain why $\hat{\rho} = \bar{Y}$ if $\Theta = [0, 1]$.

5.2. (1989 Univ. of Minn. and Aug. 2000 SIU Qual): Let $(X, Y)$ have the bivariate density

$$f(x, y) = \frac{1}{2\pi} \exp\left( -\frac{1}{2} \left[ (x - \rho \cos \theta)^2 + (y - \rho \sin \theta)^2 \right] \right).$$

Suppose that there are $n$ independent pairs of observations $(X_i, Y_i)$ from the above density and that $\rho$ is known. Assume that $0 \leq \theta \leq 2\pi$. Find a candidate for the maximum likelihood estimator $\hat{\theta}$ by differentiating the
CHAPTER 5. POINT ESTIMATION

log likelihood \( \log(L(\theta)) \). (Do not show that the candidate is the MLE, it is difficult to tell whether the candidate, 0 or 2\( \pi \) is the MLE without the actual data.)

5.3*. Suppose a single observation \( X = x \) is observed where \( X \) is a random variable with pmf given by the table below. Assume \( 0 \leq \theta \leq 1 \), and find the MLE \( \hat{\theta}_{MLE}(x) \). (Hint: drawing \( L(\theta) = L(\theta|x) \) for each of the four values of \( x \) may help.)

<table>
<thead>
<tr>
<th>( x )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(x</td>
<td>\theta) )</td>
<td>1/4</td>
<td>1/4</td>
<td>1+\theta</td>
</tr>
</tbody>
</table>

5.4. Let \( X_1, ..., X_n \) be iid normal \( N(\mu, \gamma^2 \sigma^2) \) random variables where \( \gamma^2 > 0 \) is known and \( \mu > 0 \). Find the log likelihood \( \log(L(\mu|x_1, ..., x_n)) \) and solve

\[
\frac{d}{d\mu} \log(L(\mu|x_1, ..., x_n)) = 0
\]

for \( \hat{\mu}_o \), a potential candidate for the MLE of \( \mu \).

5.5. Suppose that \( X_1, ..., X_n \) are iid uniform \( U(0, \theta) \). Use the factorization theorem to write \( f(x|\theta) = g(T(x)|\theta)I[x(1) \geq 0] \) where \( T(x) \) is a one dimensional sufficient statistic. Then plot the likelihood function \( L(\theta) = g(T(x)|\theta) \) and find the MLE of \( \theta \).

5.6. Let \( Y_1, ..., Y_n \) be iid Burr Type XII\( (\lambda, \phi) \) with \( \phi \) known. Find the MLE of \( \lambda \).

5.7. Let \( Y_1, ..., Y_n \) be iid chi\( (p, \sigma) \) with \( p \) known. Find the MLE of \( \sigma^2 \).

5.8. Let \( Y_1, ..., Y_n \) iid double exponential \( DE(\theta, \lambda) \) with \( \theta \) known. Find the MLE of \( \lambda \).

5.9. Let \( Y_1, ..., Y_n \) be iid exponential \( EXP(\lambda) \). Find the MLE of \( \lambda \).

5.10. If \( Y_1, ..., Y_n \) are iid gamma \( G(\nu, \lambda) \) with \( \nu \) known, find the MLE of \( \lambda \).

5.11. If \( Y_1, ..., Y_n \) are iid geometric \( geom(\rho) \), find the MLE of \( \rho \).

5.12. If \( Y_1, ..., Y_n \) are iid inverse Gaussian \( IG(\theta, \lambda) \) with \( \lambda \) known, find the MLE of \( \theta \).

5.13. If \( Y_1, ..., Y_n \) are iid inverse Gaussian \( IG(\theta, \lambda) \) with \( \theta \) known, find the MLE of \( \lambda \).
5.14. If $Y_1, \ldots, Y_n$ are iid largest extreme value $\text{LEV} (\theta, \sigma)$ where $\sigma$ is known, find the MLE of $\theta$.

5.15. If $Y_1, \ldots, Y_n$ are iid negative binomial $\text{NB}(r, \rho)$ with $r$ known, find the MLE of $\rho$.

5.16. If $Y_1, \ldots, Y_n$ are iid Rayleigh $\text{R}(\mu, \sigma)$ with $\mu$ known, find the MLE of $\sigma^2$.

5.17. If $Y_1, \ldots, Y_n$ are iid Weibull $\text{W}(\phi, \lambda)$ with $\phi$ known, find the MLE of $\lambda$.

5.18. If $Y_1, \ldots, Y_n$ are iid binomial $\text{BIN}(k, \rho)$ with $k$ known, find the MLE of $\rho$.

5.19. Suppose $Y_1, \ldots, Y_n$ are iid two parameter exponential $\text{EXP}(\theta, \lambda)$.
   a) Show that for any fixed $\lambda > 0$, the log likelihood is maximized by $y(1)$. Hence the MLE $\hat{\theta} = Y(1)$.
   b) Find $\hat{\lambda}$ by maximizing the profile likelihood.

5.20. Suppose $Y_1, \ldots, Y_n$ are iid truncated extreme value $\text{TEV}(\lambda)$. Find the MLE of $\lambda$.

**Problems from old quizzes and exams.**

Note: Problem 5.21 would be better if it replaced “$\lambda \geq 0$” by “$\lambda > 0$, and assume $\sum x_i > 0$.” But problems like 5.21 are extremely common on exams and in texts.

5.21. Suppose that $X_1, \ldots, X_n$ are iid Poisson with pmf

$$f(x|\lambda) = P(X = x|\lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$$

where $x = 0, 1, \ldots$ and $\lambda \geq 0$.
   a) Find the MLE of $\lambda$. (Make sure that you prove that your estimator maximizes the likelihood).
   b) Find the MLE of $(1 - \lambda)^2$.

5.22. Suppose that $X_1, \ldots, X_n$ are iid $U(0, \theta)$. Make a plot of $L(\theta|x_1, \ldots, x_n)$. 
CHAPTER 5. POINT ESTIMATION

a) If the uniform density is \( f(x) = \frac{1}{\theta} I(0 \leq x \leq \theta) \), find the MLE of \( \theta \) if it exists.

b) If the uniform density is \( f(x) = \frac{1}{\theta} I(0 < x < \theta) \), find the MLE of \( \theta \) if it exists.

5.23. (Jan. 2001 Qual): Let \( X_1, \ldots, X_n \) be a random sample from a normal distribution with known mean \( \mu \) and unknown variance \( \tau \).

a) Find the maximum likelihood estimator of the variance \( \tau \).

b) Find the maximum likelihood estimator of the standard deviation \( \sqrt{\tau} \).

Explain how the MLE was obtained.

5.24. Suppose a single observation \( X = x \) is observed where \( X \) is a random variable with pmf given by the table below. Assume \( 0 \leq \theta \leq 1 \) and find the MLE \( \hat{\theta}_{MLE}(x) \). (Hint: drawing \( L(\theta) = L(\theta|x) \) for each of the values of \( x \) may help.)

| \( x \) | \( f(x|\theta) \) |
|------|----------------|
| 0    | \( \frac{1+\theta}{2} \) |
| 1    | \( \frac{1-\theta}{2} \) |

5.25. Suppose that \( X \) is a random variable with pdf \( f(x|\theta) = (x-\theta)^2/3 \) for \( \theta - 1 \leq x \leq 2 + \theta \). Hence \( L(\theta) = (x-\theta)^2/3 \) for \( x - 2 \leq \theta \leq x + 1 \). Suppose that one observation \( X = 7 \) was observed. Find the MLE \( \hat{\theta} \) for \( \theta \).

(Hint: evaluate the likelihood at the critical value and the two endpoints. One of these three values has to be the MLE.)

5.26. Let \( X_1, \ldots, X_n \) be iid from a distribution with pdf

\[
f(x|\theta) = \theta x^{-2}, \quad 0 < \theta \leq x < \infty.
\]

a) Find a minimal sufficient statistic for \( \theta \).

b) Find the MLE for \( \theta \).

5.27. Let \( Y_1, \ldots, Y_n \) be iid from a distribution with probability mass function

\[
f(y|\theta) = \theta(1-\theta)^y, \quad \text{where } y = 0, 1, \ldots \text{ and } 0 < \theta < 1.
\]

Assume \( 0 < \sum y_i < n \).

a) Find the MLE of \( \theta \). (Show that it is the global maximizer.)
c) What is the MLE of $1/\theta^2$? Explain.

5.28. (Aug. 2002 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables from a half normal $HN(\mu, \sigma^2)$ distribution with pdf

$$f(x) = \frac{2}{\sqrt{2\pi \sigma}} \exp \left( \frac{-(x - \mu)^2}{2\sigma^2} \right)$$

where $\sigma > 0$ and $x > \mu$ and $\mu$ is real. **Assume that $\mu$ is known.**

a) Find the maximum likelihood estimator of $\sigma^2$.

b) What is the maximum likelihood estimator of $\sigma$? Explain.

5.29. (Jan. 2003 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables from a lognormal $(\mu, \sigma^2)$ distribution with pdf

$$f(x) = \frac{1}{x \sqrt{2\pi \sigma^2}} \exp \left( \frac{-(\log(x) - \mu)^2}{2\sigma^2} \right)$$

where $\sigma > 0$ and $x > 0$ and $\mu$ is real. **Assume that $\sigma$ is known.**

a) Find the maximum likelihood estimator of $\mu$.

b) What is the maximum likelihood estimator of $\mu^3$? Explain.

5.30. (Aug. 2004 Qual): Let $X$ be a single observation from a normal distribution with mean $\theta$ and with variance $\theta^2$, where $\theta > 0$. Find the maximum likelihood estimator of $\theta^2$.

5.31. (Sept. 2005 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables with probability density function

$$f(x) = \frac{\sigma^{1/\lambda}}{\lambda} \exp \left[ -(1 + \frac{1}{\lambda}) \log(x) \right] I[x \geq \sigma]$$

where $x \geq \sigma$, $\sigma > 0$, and $\lambda > 0$. The indicator function $I[x \geq \sigma] = 1$ if $x \geq \sigma$ and 0, otherwise. Find the maximum likelihood estimator (MLE) $(\hat{\sigma}, \hat{\lambda})$ of $(\sigma, \lambda)$ with the following steps.

a) Explain why $\hat{\sigma} = X_{(1)} = \min(X_1, \ldots, X_n)$ is the MLE of $\sigma$ regardless of the value of $\lambda > 0$.

b) Find the MLE $\hat{\lambda}$ of $\lambda$ if $\sigma = \hat{\sigma}$ (that is, act as if $\sigma = \hat{\sigma}$ is known).
5.32. (Aug. 2003 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables with pdf
\[ f(x) = \frac{1}{\lambda} \exp \left[ -(1 + \frac{1}{\lambda}) \log(x) \right] \]
where $\lambda > 0$ and $x \geq 1$.
   a) Find the maximum likelihood estimator of $\lambda$.
   b) What is the maximum likelihood estimator of $\lambda^8$? Explain.

5.33. (Jan. 2004 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables with probability mass function
\[ f(x) = e^{-2\theta} \frac{1}{x!} \exp[\log(2\theta)x], \]
for $x = 0, 1, \ldots$, where $\theta > 0$. Assume that at least one $X_i > 0$.
   a) Find the maximum likelihood estimator of $\theta$.
   b) What is the maximum likelihood estimator of $(\theta)^4$? Explain.

5.34. (Jan. 2006 Qual): Let $X_1, \ldots, X_n$ be iid with one of two probability density functions. If $\theta = 0$, then
\[ f(x|\theta) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \]
If $\theta = 1$, then
\[ f(x|\theta) = \begin{cases} \frac{1}{\sqrt{\pi x}}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise.} \end{cases} \]
Find the maximum likelihood estimator of $\theta$.

Warning: Variants of the following question often appear on qualifying exams.

5.35. (Aug. 2006 Qual): Let $Y_1, \ldots, Y_n$ denote a random sample from a $N(a\theta, \theta)$ population.
   a) Find the MLE of $\theta$ when $a = 1$.
   b) Find the MLE of $\theta$ when $a$ is known but arbitrary.
5.36. Suppose that $X_1,\ldots,X_n$ are iid random variable with pdf
\[ f(x|\theta) = \frac{(x - \theta)^2}{3} \]
for $\theta - 1 \leq x \leq 2 + \theta$.

a) Assume that $n = 1$ and that $X = 7$ was observed. Sketch the log likelihood function $L(\theta)$ and find the maximum likelihood estimator (MLE) $\hat{\theta}$.

b) Again assume that $n = 1$ and that $X = 7$ was observed. Find the MLE of $h(\theta) = 2\theta - \exp(-\theta^2)$.

5.37. (Aug. 2006 Qual): Let $X_1,\ldots,X_n$ be independent identically distributed (iid) random variables with probability density function
\[ f(x) = \frac{2}{\lambda\sqrt{2\pi}} e^x \exp\left( -\frac{(e^x - 1)^2}{2\lambda^2} \right) \]
where $x > 0$ and $\lambda > 0$.

a) Find the maximum likelihood estimator (MLE) $\hat{\lambda}$ of $\lambda$.

b) What is the MLE of $\lambda^2$? Explain.

5.38. (Jan. 2007 Qual): Let $X_1,\ldots,X_n$ be independent identically distributed random variables from a distribution with pdf
\[ f(x) = \frac{2}{\lambda\sqrt{2\pi}} \frac{1}{x} \exp\left[ -\frac{(\log(x))^2}{2\lambda^2} \right] \]
where $\lambda > 0$ where and $0 \leq x \leq 1$.

a) Find the maximum likelihood estimator (MLE) of $\lambda$.

b) Find the MLE of $\lambda^2$.

5.39. Suppose that $X_1,\ldots,X_n$ are iid $U(\theta,\theta+1)$ so that
\[ L(\theta) = 1^n I[x(1) \geq \theta]I[x(n) \leq \theta + 1] = I[x(n) - 1 \leq \theta \leq x(1)]. \]

a) Sketch $L(\theta)$.

b) An MLE of $\theta$ is $\hat{\theta}_{MLE}(x) = t$ for some fixed $t \in [c,d]$. Find $[c,d]$. 
5.40. Let $Y_1, ..., Y_n$ be independent identically distributed random variables with pdf
\[ f_Y(y) = \frac{2\gamma^{3/2}}{\sqrt{\pi}} y^2 \exp(-\gamma y^2) \]
where $\gamma > 0$ and $y$ is real.

a) Find the maximum likelihood estimator of $\gamma$. (Make sure that you prove that your answer is the MLE.)
b) What is the maximum likelihood estimator of $1/\gamma$? Explain.

5.41. (Jan. 2009 Qual): Suppose that $X$ has probability density function
\[ f_X(x) = \frac{\theta}{x^{1+\theta}}, \quad x \geq 1 \]
where $\theta > 0$.

a) If $U = X^2$, derive the probability density function $f_U(u)$ of $U$.
b) Find the method of moments estimator of $\theta$.
c) Find the method of moments estimator of $\theta^2$.

5.42. (Jan. 2009 Qual): Suppose that the joint probability distribution function of $X_1, ..., X_k$ is
\[ f(x_1, x_2, ..., x_k|\theta) = \frac{n!}{(n-k)!\theta^k} \exp\left(-\frac{1}{\theta}\sum_{i=1}^{k} x_i + (n-k)x_k\right) \]
where $0 \leq x_1 \leq x_2 \leq \cdots \leq x_k$ and $\theta > 0$.

a) Find the maximum likelihood estimator (MLE) for $\theta$.
b) What is the MLE for $\theta^2$? Explain briefly.

5.43. (Jan. 2010 Qual): Let $X_1, ..., X_n$ be iid with pdf
\[ f(x) = \frac{\cos(\theta)}{2 \cosh(\pi x/2)} \exp(\theta x) \]
where $x$ is real and $|\theta| < \pi/2$.

a) Find the maximum likelihood estimator (MLE) for $\theta$.
b) What is the MLE for $\tan(\theta)$? Explain briefly.

5.44. (Aug. 2009 Qual): Let $X_1, ..., X_n$ be a random sample from a population with pdf
\[ f(x) = \frac{1}{\sigma} \exp\left(-\frac{x - \mu}{\sigma}\right), \quad x \geq \mu, \]
where $-\infty < \mu < \infty$, $\sigma > 0$.

a) Find the maximum likelihood estimator of $\mu$ and $\sigma$.

b) Evaluate $\tau(\mu, \sigma) = P_{\mu, \sigma}[X_1 \geq t]$ where $t > \mu$. Find the maximum likelihood estimator of $\tau(\mu, \sigma)$.

5.45. (Sept. 2010 Qual): Let $Y_1, ..., Y_n$ be independent identically distributed (iid) random variables from a distribution with probability density function (pdf)

$$f(y) = \frac{1}{2\sqrt{2\pi}} \left( \frac{1}{\theta} \sqrt{\frac{x}{y}} + \frac{\theta}{\sqrt{2\pi}} \right) \frac{1}{\nu} \exp \left[ -\frac{1}{2\nu^2} \left( \frac{y}{\theta} + \frac{\theta}{y} - 2 \right) \right]$$

where $y > 0$, $\theta > 0$ is known and $\nu > 0$.

a) Find the maximum likelihood estimator (MLE) of $\nu$.

b) Find the MLE of $\nu^2$.

5.46. (Sept. 2011 Qual): Let $Y_1, ..., Y_n$ be independent identically distributed (iid) random variables from a distribution with probability density function (pdf)

$$f(y) = \phi y^{-(\phi+1)} \frac{1}{1+y^{-\phi}} \frac{1}{\lambda} \exp \left[ -\frac{1}{\lambda} \log(1 + y^{-\phi}) \right]$$

where $y > 0$, $\phi > 0$ is known and $\lambda > 0$.

a) Find the maximum likelihood estimator (MLE) of $\lambda$.

b) Find the MLE of $\lambda^2$. 
Chapter 6
UMVUEs and the FCRLB

Warning: UMVUE theory is rarely used in practice unless the UMVUE $U_n$ of $\theta$ satisfies $U_n = a_n \hat{\theta}_{MLE}$ where $a_n$ is a constant that could depend on the sample size $n$. UMVUE theory tends to be somewhat useful if the data is iid from a 1P–REF.

6.1 MSE and Bias

Definition 6.1. Let the sample $Y = (Y_1, ..., Y_n)$ where $Y$ has a pdf or pmf $f(y|\theta)$ for $\theta \in \Theta$. Assume all relevant expectations exist. Let $\tau(\theta)$ be a real valued function of $\theta$, and let $T \equiv T(Y_1, ..., Y_n)$ be an estimator of $\tau(\theta)$. The bias of the estimator $T$ for $\tau(\theta)$ is

$$B(T) \equiv B_{\tau(\theta)}(T) \equiv \text{Bias}(T) \equiv \text{Bias}_{\tau(\theta)}(T) = E_{\theta}(T) - \tau(\theta).$$

(6.1)

The mean squared error (MSE) of an estimator $T$ for $\tau(\theta)$ is

$$\text{MSE}(T) \equiv \text{MSE}_{\tau(\theta)}(T) = E_{\theta}[(T - \tau(\theta))^2]$$

$$= \text{Var}_{\theta}(T) + |\text{Bias}_{\tau(\theta)}(T)|^2.$$  

(6.2)

$T$ is an unbiased estimator of $\tau(\theta)$ if

$$E_{\theta}(T) = \tau(\theta)$$

(6.3)

for all $\theta \in \Theta$. Notice that $\text{Bias}_{\tau(\theta)}(T) = 0$ for all $\theta \in \Theta$ if $T$ is an unbiased estimator of $\tau(\theta)$.
Notice that the bias and MSE are functions of $\theta$ for $\theta \in \Theta$. If $MSE_{\tau;\theta_1}(T_1) < MSE_{\tau;\theta_2}(T_2)$ for all $\theta \in \Theta$, then $T_1$ is “a better estimator” of $\tau(\theta)$ than $T_2$. So estimators with small MSE are judged to be better than ones with large MSE. Often $T_1$ has smaller MSE than $T_2$ for some $\theta$ but larger MSE for other values of $\theta$.

Often $\theta$ is real valued. A common problem considers a class of estimators $T_k(Y)$ of $\tau(\theta)$ where $k \in \Lambda$. Find the MSE as a function of $k$ and then find the value $k_o \in \Lambda$ that is the global minimizer of $MSE(k) \equiv MSE(T_k)$. This type of problem is a lot like the MLE problem except you need to find the global min rather than the global max.

This type of problem can often be done if $T_k = kW_1(X) + (1-k)W_2(X)$ where both $W_1$ and $W_2$ are unbiased estimators of $\tau(\theta)$ and $0 \leq k \leq 1$.

Example 6.1. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then $k_o = n+1$ will minimize the MSE for estimators of $\sigma^2$ of the form
\[
S^2(k) = \frac{1}{k} \sum_{i=1}^{n} (X_i - \overline{X})^2
\]
where $k > 0$. See Problem 6.2.

Example 6.2. Find the bias and MSE (as a function of $n$ and $c$) of an estimator $T = c \sum_{i=1}^{n} Y_i$ or $(T = b\overline{Y})$ of $\theta$ when $Y_1, \ldots, Y_n$ are iid with $E(Y_1) = \mu = h(\theta)$ and $V(Y_i) = \sigma^2$.
Solution: $E(T) = c \sum_{i=1}^{n} E(Y_i) = n c \mu$, $V(T) = c^2 \sum_{i=1}^{n} V(Y_i) = n c^2 \sigma^2$, $B(T) = E(T) - \theta$ and $MSE(T) = V(T) + [B(T)]^2$. (For $T = b\overline{Y}$, use $c = b/n$.)

Example 6.3. Suppose that $Y_1, \ldots, Y_n$ are independent binomial$(m_i, \rho)$ where the $m_i \geq 1$ are known constants. Let
\[
T_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} m_i} \quad \text{and} \quad T_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{m_i}
\]
be estimators of $\rho$.

a) Find $MSE(T_1)$.
b) Find $MSE(T_2)$.
c) Which estimator is better?
Hint: by the arithmetic–geometric–harmonic mean inequality,

\[ \frac{1}{n} \sum_{i=1}^{n} m_i \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{m_i}}. \]

Solution: a)

\[ E(T_1) = \frac{\sum_{i=1}^{n} E(Y_i)}{\sum_{i=1}^{n} m_i} = \frac{\sum_{i=1}^{n} m_i \rho}{\sum_{i=1}^{n} m_i} = \rho, \]

so \( \text{MSE}(T_1) = V(T_1) = \)

\[ \frac{1}{(\sum_{i=1}^{n} m_i)^2} V(\sum_{i=1}^{n} Y_i) = \frac{1}{(\sum_{i=1}^{n} m_i)^2} \sum_{i=1}^{n} V(Y_i) = \frac{1}{(\sum_{i=1}^{n} m_i)^2} \sum_{i=1}^{n} m_i \rho (1 - \rho) = \frac{\rho (1 - \rho)}{\sum_{i=1}^{n} m_i}. \]

b) \[ E(T_2) = \frac{1}{n} \sum_{i=1}^{n} \frac{E(Y_i)}{m_i} = \frac{1}{n} \sum_{i=1}^{n} \frac{m_i \rho}{m_i} = \frac{1}{n} \sum_{i=1}^{n} \rho = \rho, \]

so \( \text{MSE}(T_2) = V(T_2) = \)

\[ \frac{1}{n^2} V(\sum_{i=1}^{n} Y_i = \frac{1}{n^2} \sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{V(Y_i)}{(m_i)^2} = \frac{1}{n^2} \sum_{i=1}^{n} \frac{m_i \rho (1 - \rho)}{(m_i)^2} = \frac{\rho (1 - \rho)}{n^2} \sum_{i=1}^{n} \frac{1}{m_i}. \]

c) The hint

\[ \frac{1}{n} \sum_{i=1}^{n} m_i \geq \frac{n}{\sum_{i=1}^{n} \frac{1}{m_i}} \]

implies that

\[ \frac{n}{\sum_{i=1}^{n} m_i} \leq \frac{\sum_{i=1}^{n} \frac{1}{m_i}}{n} \quad \text{and} \quad \frac{1}{\sum_{i=1}^{n} m_i} \leq \frac{\sum_{i=1}^{n} \frac{1}{m_i}}{n^2}. \]

Hence \( \text{MSE}(T_1) \leq \text{MSE}(T_2), \) and \( T_1 \) is better.
6.2 Exponential Families, UMVUEs and the FCRLB.

**Definition 6.2.** Let the sample \( Y = (Y_1, ..., Y_n) \) where \( Y \) has a pdf or pmf \( f(y|\theta) \) for \( \theta \in \Theta \). Assume all relevant expectations exist. Let \( \tau(\theta) \) be a real valued function of \( \theta \), and let \( U \equiv U(Y_1, ..., Y_n) \) be an estimator of \( \tau(\theta) \). Then \( U \) is the **uniformly minimum variance unbiased estimator (UMVUE)** of \( \tau(\theta) \) if \( U \) is an unbiased estimator of \( \tau(\theta) \) and if \( \text{Var}_\theta(U) \leq \text{Var}_\theta(W) \) for all \( \theta \in \Theta \) where \( W \) is any other unbiased estimator of \( \tau(\theta) \).

The following theorem is the most useful method for finding UMVUEs since if \( Y_1, ..., Y_n \) are iid from a 1P–REF \( f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)] \) where \( \eta = w(\theta) \in \Omega = (a,b) \) and \( a < b \) are not necessarily finite, then \( T(Y) = \sum_{i=1}^n t(Y_i) \) is a complete sufficient statistic. It will turn out that \( E_\theta[W(Y)|T(Y)] \equiv E[W(Y)|T(Y)] \) does not depend on \( \theta \). Hence \( U = E[W(Y)|T(Y)] \) is a statistic.

**Theorem 6.1, Lehmann-Scheffé UMVUE (LSU) Theorem:** If \( T(Y) \) is a complete sufficient statistic for \( \theta \), then
\[
U = g(T(Y))
\] (6.4)
is the UMVUE of its expectation \( E_\theta(U) = E_\theta[g(T(Y))] \). In particular, if \( W(Y) \) is any unbiased estimator of \( \tau(\theta) \), then
\[
U \equiv g(T(Y)) = E[W(Y)|T(Y)]
\] (6.5)
is the UMVUE of \( \tau(\theta) \). If \( V_\theta(U) < \infty \) for all \( \theta \in \Theta \), then \( U \) is the unique UMVUE of \( \tau(\theta) = E_\theta[g(T(Y))] \).

The process (6.5) is called Rao-Blackwellization because of the following theorem. The theorem is also called the Rao-Blackwell-Lehmann-Scheffé theorem.

**Theorem 6.2, Rao-Blackwell Theorem.** Let \( W \equiv W(Y) \) be an unbiased estimator of \( \tau(\theta) \) and let \( T \equiv T(Y) \) be a sufficient statistic for \( \theta \). Then \( \phi(T) = E[W|T] \) is an unbiased estimator of \( \tau(\theta) \) and \( \text{VAR}_\theta[\phi(T)] \leq \text{VAR}_\theta(W) \) for all \( \theta \in \Theta \).

**Proof.** Notice that \( \phi(T) \) does not depend on \( \theta \) by the definition of a sufficient statistic, and that \( \phi(T) \) is an unbiased estimator for \( \tau(\theta) \) since
$\tau(\theta) = E_\theta(W) = E_\theta(E(W|T)) = E_\theta(\phi(T))$ by iterated expectations (Theorem 2.10). By Steiner’s formula (Theorem 2.11), $\text{VAR}_\theta(W) = E_\theta[\text{VAR}(W|T)] + \text{VAR}_\theta[E(W|T)] \geq \text{VAR}_\theta[E(W|T)] = \text{VAR}_\theta(\phi(T))$. QED

**Tips for finding the UMVUE:**

i) From the LSU Theorem, if $T(Y)$ is complete sufficient statistic and $g(T(Y))$ is a real valued function, then $U = g(T(Y))$ is the UMVUE of its expectation $E_\theta[g(T(Y))]$.

ii) Given a complete sufficient statistic $T(Y)$ (eg $T(Y) = \sum_{i=1}^n t(Y_i)$ if the data are iid from a 1P–REF), the first method for finding the UMVUE of $\tau(\theta)$ is to guess $g$ and show that $E_\theta[g(T(Y))] = \tau(\theta)$ for all $\theta$.

iii) If $T(Y)$ is complete, the second method is to find any unbiased estimator $W(Y)$ of $\tau(\theta)$. Then $U(Y) = E[W(Y)|T(Y)]$ is the UMVUE of $\tau(\theta)$.

This problem is often very hard because guessing $g$ or finding an unbiased estimator $W(Y)$ tends to be difficult. Write down the two methods for finding the UMVUE and simplify $E[W(Y)|T(Y)]$ as far as you can for partial credit. If you are asked to find the UMVUE of $\tau(\theta)$, see if an unbiased estimator $W(Y)$ is given in the problem. Also check whether you are asked to compute $E[W(Y)|T(Y) = t]$ anywhere.

iv) The following facts can be useful for computing the conditional expectation via Rao-Blackwellization (see problems 6.7, 6.10 and 6.12). Suppose $Y_1, ..., Y_n$ are iid with finite expectation.

a) Then $E[Y_1|\sum_{i=1}^n Y_i = x] = x/n$.

b) If the $Y_i$ are iid Poisson($\theta$), then $(Y_1|\sum_{i=1}^n Y_i = x) \sim \text{bin}(x, 1/n)$.

c) If the $Y_i$ are iid Bernoulli Ber($\rho$), then $(Y_1|\sum_{i=1}^n Y_i = x) \sim \text{Ber}(x/n)$.

d) If the $Y_i$ are iid $N(\mu, \sigma^2)$, then $(Y_1|\sum_{i=1}^n Y_i = x) \sim N[x/n, \sigma^2(1 - 1/n)]$.

Often students will be asked to compute a lower bound on the variance of unbiased estimators of $\eta = \tau(\theta)$ when $\theta$ is a scalar.

**Definition 6.3.** Let $Y = (Y_1, ..., Y_n)$ have a pdf or pmf $f(y|\theta)$. Then the **information number** or **Fisher Information** is

$$I_Y(\theta) \equiv I_n(\theta) = E_\theta \left( \left[ \frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right]^2 \right). \quad (6.6)$$
Let \( \eta = \tau(\theta) \) where \( \tau'(\theta) \neq 0 \). Then
\[
I_n(\eta) \equiv I_n(\tau(\theta)) = \frac{I_n(\theta)}{[\tau'(\theta)]^2}.
\] (6.7)

**Theorem 6.3.** a) Equations (6.6) and (6.7) agree if \( \tau'(\theta) \) is continuous, \( \tau'(\theta) \neq 0 \), and \( \tau(\theta) \) is one to one and onto so that an inverse function exists such that \( \theta = \tau^{-1}(\eta) \).

b) If the \( Y_1 \equiv Y \) is from a 1P–REF, then the Fisher information in a sample of size one is
\[
I_1(\theta) = -E_{\theta} \left[ \frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta)) \right].
\] (6.8)

c) If the \( Y_1, \ldots, Y_n \) are iid from a 1P–REF, then
\[
I_n(\theta) = nI_1(\theta).
\] (6.9)

Hence if \( \tau'(\theta) \) exists and is continuous and if \( \tau'(\theta) \neq 0 \), then
\[
I_n(\tau(\theta)) = \frac{nI_1(\theta)}{[\tau'(\theta)]^2}.
\] (6.10)


b) The proof will be for a pdf. For a pmf replace the integrals by sums. By Remark 3.2, the integral and differentiation operators of all orders can be interchanged. Note that
\[
0 = E \left[ \frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right]
\] (6.11)
since
\[
\frac{\partial}{\partial \theta}^1 = 0 = \frac{\partial}{\partial \theta} \int f(y|\theta)dy = \int \frac{\partial}{\partial \theta} f(y|\theta)dy = \int \frac{\partial}{\partial \theta} f(y|\theta) f(y|\theta)dy
\]
or
\[
0 = \frac{\partial}{\partial \theta} \int f(y|\theta)dy = \int \left[ \frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta)dy
\]
which is (6.11). Taking 2nd derivatives of the above expression gives

\[ 0 = \frac{\partial^2}{\partial \theta^2} \int f(y|\theta) dy = \frac{\partial}{\partial \theta} \int \left[ \frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) dy = \]

\[ \int \frac{\partial}{\partial \theta} \left( \left[ \frac{\partial}{\partial \theta} \log(f(y|\theta)) \right] f(y|\theta) \right) dy = \]

\[ \int \left[ \frac{\partial^2}{\partial \theta^2} \log(f(y|\theta)) \right] f(y|\theta) dy + \int \left[ \frac{\partial}{\partial \theta} \log(f(y|\theta)) \left( \frac{\partial}{\partial \theta} f(y|\theta) \right) \right] f(y|\theta) dy \]

or

\[ I_1(\theta) = E_\theta [(\frac{\partial}{\partial \theta} \log f(Y|\theta))^2] = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta)) \right]. \]

c) By independence,

\[ I_n(\theta) = E_\theta \left[ \left( \frac{\partial}{\partial \theta} \log \left( \prod_{i=1}^{n} f(Y_i|\theta) \right) \right)^2 \right] = E_\theta \left[ \left( \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log(f(Y_i|\theta)) \right)^2 \right] = \]

\[ E_\theta \left[ \left( \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log(f(Y_i|\theta)) \right) \left( \frac{\partial}{\partial \theta} \sum_{j=1}^{n} \log(f(Y_j|\theta)) \right) \right] = \]

\[ E_\theta \left[ \left( \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(f(Y_i|\theta)) \right) \left( \sum_{j=1}^{n} \frac{\partial}{\partial \theta} \log(f(Y_j|\theta)) \right) \right] = \]

\[ \sum_{i=1}^{n} E_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(f(Y_i|\theta)) \right)^2 \right] + \]

\[ \sum_{i \neq j} E_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(f(Y_i|\theta)) \right) \left( \frac{\partial}{\partial \theta} \log(f(Y_j|\theta)) \right) \right]. \]

Hence

\[ I_n(\theta) = nI_1(\theta) + \sum_{i \neq j} E_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(f(Y_i|\theta)) \right) \right] E_\theta \left[ \left( \frac{\partial}{\partial \theta} \log(f(Y_j|\theta)) \right) \right] \]
CHAPTER 6. UMVUES AND THE FCRLB

by independence. Hence

\[ I_n(\theta) = nI_1(\theta) + n(n-1) \left[ E_\theta \left( \frac{\partial}{\partial \theta} \log(f(Y_j|\theta)) \right) \right]^2 \]

since the \( Y_i \) are iid. Thus \( I_n(\theta) = nI_1(\theta) \) by Equation (6.11) which holds since the \( Y_i \) are iid from a 1P–REF. QED

**Definition 6.4.** Let \( Y = (Y_1, ..., Y_n) \) be the data, and consider \( \tau(\theta) \) where \( \tau'(\theta) \neq 0 \). The quantity

\[ \text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} \]

is called the **Fréchet Cramér Rao lower bound** (FCRLB) for the variance of unbiased estimators of \( \tau(\theta) \). In particular, if \( \tau(\theta) = \theta \), then \( \text{FCRLB}_n(\theta) = \frac{1}{I_n(\theta)} \). The FCRLB is often called the Cramér Rao lower bound (CRLB).

**Theorem 6.4, Fréchet Cramér Rao Lower Bound or Information Inequality.** Let \( Y_1, ..., Y_n \) be iid from a 1P–REF with pdf or pmf \( f(y|\theta) \). Let \( W(Y_1, ..., Y_n) = W(Y) \) be any unbiased estimator of \( \tau(\theta) \equiv E_\theta W(Y) \). Then

\[ \text{VAR}_\theta(W(Y)) \geq \text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}. \]

**Proof.** By Definition 6.4 and Theorem 6.3c,

\[ \text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)} = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}. \]

Since the \( Y_i \) are iid from a 1P–REF, by Remark 3.2 the derivative and integral or sum operators can be interchanged when finding the derivative of \( E_\theta h(Y) \) if \( E_\theta |h(Y)| < \infty \). The following argument will be for pdfs. For pmfs, replace the integrals by appropriate sums. Following Casella and Berger (2002, p. 335-8), the Cauchy Schwarz Inequality is

\[ [\text{Cov}(X,Y)]^2 \leq V(X)V(Y), \quad \text{or} \quad V(X) \geq \frac{[\text{Cov}(X,Y)]^2}{V(Y)}. \]

Hence

\[ V_\theta(W(Y)) \geq \frac{[\text{Cov}_\theta[W(Y), \frac{\partial}{\partial \theta} \log(f(Y|\theta))]]^2}{V_\theta[\frac{\partial}{\partial \theta} \log(f(Y|\theta))]} \].

(6.12)
Now
\[ E_\theta \left[ \frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right] = E_\theta \left[ \frac{\partial}{\partial \theta} \frac{f(Y|\theta)}{f(Y|\theta)} \right] \]
since the derivative of \( \log(h(t)) \) is \( h'(t)/h(t) \). By the definition of expectation,
\[ E_\theta \left[ \frac{\partial}{\partial \theta} \log(f(Y|\theta)) \right] = \int Y \cdots \int \frac{\partial}{\partial \theta} f(y|\theta) \frac{f(y|\theta)}{f(y|\theta)} dy = \frac{d}{d\theta} \int Y \cdots \int f(y|\theta) dy = \frac{d}{d\theta} 1 = 0. \]

Notice that \( f(y|\theta) > 0 \) on the support \( Y \), that the \( f(y|\theta) \) cancelled in the 2nd term, that the derivative was moved outside of the integral by Remark 3.2, and that the integral of \( f(y|\theta) \) on the support \( Y \) is equal to 1.

This result implies that
\[ \text{Cov}_\theta[W(Y), \frac{\partial}{\partial \theta} \log(f(Y|\theta))] = E_\theta[W(Y) \frac{\partial}{\partial \theta} \log(f(Y|\theta))] \]
since the derivative of \( \log(h(t)) \) is \( h'(t)/h(t) \). By the definition of expectation, the right hand side is equal to
\[ \int Y \cdots \int W(y) \frac{\partial f(y|\theta)}{\partial \theta} f(y|\theta) dy = \frac{d}{d\theta} \int Y \cdots \int W(y) f(y|\theta) dy \]
\[ = \frac{d}{d\theta} E_\theta W(Y) = \tau'(\theta) = \text{Cov}_\theta[W(Y), \frac{\partial}{\partial \theta} \log(f(Y|\theta))]. \quad (6.13) \]
Since
\[ E_\theta[\frac{\partial}{\partial \theta} \log f(Y|\theta)] = 0, \]
\[ V_\theta[\frac{\partial}{\partial \theta} \log(f(Y|\theta))] = E_\theta[(\frac{\partial}{\partial \theta} \log(f(Y|\theta)))^2] = I_n(\theta) \quad (6.14) \]
by Definition 6.3. Plugging (6.13) and (6.14) into (6.12) gives the result.

QED

Theorem 6.4 is not very useful in applications. If the data are iid from a 1P--REF then \( FCRLB_n(\tau(\theta)) = [\tau'(\theta)]^2/[n I_1(\theta)] \) by Theorem 6.4. Notice
that $W(Y)$ is an unbiased estimator of $\tau(\theta)$ since $E_\theta W_Y(Y) = \tau(\theta)$. Hence if the data are iid from a 1P–REF and if $\text{VAR}_\theta(W(Y)) = FCRLB_n(\tau(\theta))$ for all $\theta \in \Theta$ then $W(Y)$ is the UMVUE of $\tau(\theta)$; however, this technique for finding a UMVUE rarely works since typically equality holds only if 1) the data come from a 1P–REF with complete sufficient statistic $T$, and 2) $W = a + bT$ is a linear function of $T$.

The FCRLB inequality will typically be strict for nonlinear functions of $T$ if the data is iid from a 1P–REF. If $T$ is complete, $g(T)$ is the UMVUE of its expectation, and determining that $T$ is the complete sufficient statistic from a 1P–REF is simpler than computing $\text{VAR}_\theta(W)$ and FCRLB$_n(\tau(\theta))$.

If the family is not an exponential family, the FCRLB may not be a lower bound on the variance of unbiased estimators of $\tau(\theta)$.

Example 6.4. Let $Y_1, ..., Y_n$ be iid random variables with pdf

$$f(y) = \frac{2}{\sqrt{2\pi}\lambda} y^{I_{[0,1]}(y)} \exp\left[-\frac{[\log(y)]^2}{2\lambda^2}\right]$$

where $\lambda > 0$. Then $[\log(Y_i)]^2 \sim G(1/2, 2\lambda^2) \sim \lambda^2 \chi^2_1$.

a) Find the uniformly minimum variance estimator (UMVUE) of $\lambda^2$.
b) Find the information number $I_1(\lambda)$.
c) Find the Fréchet Cramér Rao lower bound (FCRLB) for estimating $\tau(\lambda) = \lambda^2$.

Solution. a) This is a one parameter exponential family with complete sufficient statistic $T_n = \sum_{i=1}^n [\log(Y_i)]^2$. Now $E(T_n) = nE([\log(Y_i)]^2) = n\lambda^2$. Hence $E(T_n/n) = \lambda^2$ and $T_n/n$ is the UMVUE of $\lambda^2$ by the LSU Theorem.
b) Now

$$\log(f(y|\lambda)) = \log(2/\sqrt{2\pi}) - \log(\lambda) - \log(y) - \frac{[\log(y)]^2}{2\lambda^2}.$$ 

Hence

$$\frac{d}{d\lambda} \log(f(y|\lambda)) = \frac{-1}{\lambda^2} + \frac{[\log(y)]^2}{\lambda^3},$$

and

$$\frac{d^2}{d\lambda^2} \log(f(y|\lambda)) = \frac{1}{\lambda^2} - \frac{3[\log(y)]^2}{\lambda^4}.$$ 

Thus

$$I_1(\lambda) = -E \left[ \frac{1}{\lambda^2} - \frac{3[\log(Y)]^2}{\lambda^4} \right] = -\frac{1}{\lambda^2} + \frac{3\lambda^2}{\lambda^4} = \frac{2}{\lambda^2}. $$
CHAPTER 6. UMVUES AND THE FCRLB

188

c) 

\[ FCRLB_n(\tau(\lambda)) = \frac{[\tau'(\lambda)]^2}{nI_1(\lambda)}. \]

Now \( \tau(\lambda) = \lambda^2 \) and \( \tau'(\lambda) = 2\lambda. \) So

\[ FCRLB_n(\tau(\lambda)) = \frac{4\lambda^2}{n^2/\lambda^2} = \frac{2\lambda^4}{n}. \]

**Example 6.5.** Suppose that \( X_1, \ldots, X_n \) are iid Bernoulli(\( p \)) where \( n \geq 2 \) and \( 0 < p < 1 \) is the unknown parameter.

a) Derive the UMVUE of \( \tau(p) \), where \( \tau(p) = e^{2(p - (1 - p))} \).

b) Find the FCRLB for estimating \( \tau(p) = e^{2(p(1 - p))} \).

**Solution:**

a) Consider the statistic \( W = X_1(1 - X_2) \) which is an unbiased estimator of \( \tau(p) = p(1 - p) \). The statistic \( T = \sum_{i=1}^{n} X_i \) is both complete and sufficient. The possible values of \( W \) are 0 or 1. Then

\[ \phi(t) = E[ X_1(1 - X_2) | T = t] = \frac{P[X_1 = 1, X_2 = 0 \text{ and } \sum_{i=1}^{n} X_i = t]}{P[\sum_{i=1}^{n} X_i = t]} \]

Now \( \sum_{i=3}^{n} X_i \) is Bin\((n - 2, p)\) and \( \sum_{i=1}^{n} X_i \) is Bin\((n, p)\). Thus

\[ \phi(t) = \frac{p(1-p)[\binom{n-2}{t-2}p^{t-1}(1-p)^{n-t-1}]}{\binom{n}{t}p^t(1-p)^n} \]

\[ = \binom{n-2}{t-1} \binom{n}{t} \frac{(n-2)!}{(t-1)!(n-2-t+1)!} \frac{t(t-1)!(n-t)(n-t-1)!}{n(n-1)(n-2)!} = \frac{t(n-t)}{n(n-1)} \]

\[ = \frac{\binom{n}{t} n(n-1)}{n(n-1)} = \frac{\binom{n}{t} n(1 - \frac{t}{n})}{n-1} = \frac{n}{n-1} \tau(1 - \tau). \]

Thus \( \frac{n}{n-1} \overline{X}(1 - \overline{X}) \) is the UMVUE of \( p(1-p) \) and \( U = e^{2 \frac{n}{n-1} \overline{X}(1 - \overline{X})} \) is the UMVUE of \( \tau(p) = e^{2p(1 - p)} \).
Alternatively, $\overline{X}$ is a complete sufficient statistic, so try an estimator of the form $U = a(\overline{X})^2 + b\overline{X} + c$. Then $U$ is the UMVUE if $E_p(U) = e^2p(1-p) = e^2(p - p^2)$. Now $E(\overline{X}) = E(X_1) = p$ and $V(\overline{X}) = V(X_1)/n = p(1-p)/n$ since $\sum X_i \sim Bin(n,p)$. So $E[(\overline{X})^2] = V(\overline{X}) + [E(\overline{X})]^2 = p(1-p)/n + p^2$. So $E_p(U) = a[p(1-p)/n] + ap^2 + bp + c$.

$U$ is the UMVUE if $E_p(U) = e^2p(1-p) = e^2p - e^2p^2$.

Now $E(X) = E(X_1) = p$ and $V(X) = V(X_1)/n = p(1-p)/n$ since $\sum X_i \sim Bin(n,p)$. So $E(X^2) = V(X) + [E(X)]^2 = p(1-p)/n + p^2$.

So $E_p(U) = ap^2 - ap^2/n + ap^2 + bp + c = (a/n + b)p + (a - ap^2/n + bp + c$.

So $e = 0$ and $a - ap^2/n = -e^2$ or $a = -n e^2$.

Hence $a - a/n = a(n-1)/n = -e^2$ or $a = -n e^2$.

So $U = -n e^2(\overline{X})^2 + n(n-1) e^2\overline{X} = n e^2\overline{X}(1 - \overline{X})$.

b) The FCRLB for $\tau(p)$ is $[\tau'(p)]^2/nI_1(p)$. Now $f(x) = p^x(1-p)^{1-x}$, so $log f(x) = x \log(p) + (1-x) \log(1-p)$. Hence

$$\frac{\partial \log f}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p}$$

and

$$\frac{\partial^2 \log f}{\partial p^2} = \frac{-x}{p^2} - \frac{1-x}{(1-p)^2}.$$ 

So $I_1(p) = -E(\frac{\partial^2 \log f}{\partial p^2}) = -\frac{-p}{p^2} - \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}$.

So $FCRLB_n = \frac{[e^2(1-2p)]^2}{p(1-p)/n} = \frac{e^4(1-2p)^2p(1-p)}{n}.$

Example 6.6. Let $X_1, ..., X_n$ be iid random variables with pdf

$$f(x) = \frac{1}{\lambda} \phi x^{\phi-1} \frac{1}{1+x^{\phi}} \exp \left[-\frac{1}{\lambda} \log(1 + x^{\phi}) \right]$$
where $x, \phi,$ and $\lambda$ are all positive. If $\phi$ is known, find the uniformly minimum unbiased estimator of $\lambda$ using the fact that $\log(1 + X_i^\phi) \sim \text{Gamma}(\nu = 1, \lambda)$.

Solution: This is a regular one parameter exponential family with complete sufficient statistic $T_n = \sum_{i=1}^n \log(1 + X_i^\phi) \sim G(n, \lambda)$. Hence $E(T_n) = n\lambda$ and $T_n/n$ is the UMVUE of $\lambda$.

Example 6.7. If $\sum_{i=1}^n Y_i$ is a complete sufficient statistic for $\theta$, then by the LSU theorem, $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $E[e^{t\sum_{i=1}^n Y_i}] = m_{\sum_{i=1}^n Y_i}(t)$, the mgf of $\sum_{i=1}^n Y_i$. Refer to Theorems 2.17 and 2.18 for the following special cases.

a) If $Y_1, ..., Y_n$ are iid $\text{BIN}(k, \rho)$ where $k$ is known, then $\sum_{i=1}^n Y_i \sim \text{BIN}(nk, \rho)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\left[(1 - \rho) + \rho e^{tk}\right]^n$.

b) If $Y_1, ..., Y_n$ are iid $\text{EXP}(\lambda)$, then $\sum_{i=1}^n Y_i \sim G(n, \lambda)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\left(\frac{1}{1 - \lambda t}\right)^n$ for $t < 1/\lambda$.

c) If $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$ where $\nu$ is known, then $\sum_{i=1}^n Y_i \sim G(n\nu, \lambda)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\left(\frac{1}{1 - \lambda t}\right)^{n\nu}$ for $t < 1/\lambda$.

d) If $Y_1, ..., Y_n$ are iid $\text{N}(\mu, \sigma^2)$ where $\sigma^2$ is known, then $\sum_{i=1}^n Y_i \sim \text{N}(n\mu, n\sigma^2)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\exp(tn\mu + t^2n\sigma^2/2)$.

e) If $Y_1, ..., Y_n$ are iid Poisson($\theta$), then $\sum_{i=1}^n Y_i \sim \text{Poisson}(n\theta)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\exp(n\theta(e^t - 1))$.

f) If $Y_1, ..., Y_n$ are iid $\text{NB}(r, \rho)$ where $r$ is known, then $\sum_{i=1}^n Y_i \sim \text{NB}(nr, \rho)$, and $e^{t\sum_{i=1}^n Y_i}$ is the UMVUE of $\left[\frac{\rho}{1 - (1 - \rho)e^t}\right]^{nr}$.
for \( t < -\log(1 - \rho) \).

The following theorem compares the UMVUE with the estimator that minimizes the MSE for one parameter exponential families. Note that the constant \( c \) can not depend on the unknown parameter \( \theta \) since \( cT(Y) \) needs to be a statistic. Often \( \theta X \sim \theta G(1,1) \sim G(1,\theta) \). Note \( cM/cU \to 1 \) as \( n \to \infty \). Hence the UMVUE and the estimator that minimizes the MSE behave similarly in terms of MSE for large \( n \).

**Theorem 6.5.** Let \( Y_1, ..., Y_n \) be iid from a one parameter exponential family with pdf or pmf \( f(y|\theta) \) with complete sufficient statistic \( T(Y) = \sum_{i=1}^{n} t(Y_i) \) where \( t(Y_i) \sim \theta X \) and \( X \) has a know distribution with known mean \( E(X) \) and known variance \( V(X) \). Let \( W_n = cT(Y) \) be an estimator of \( \theta \).

a) The value \( c \) that minimizes the MSE is
\[
c_M = \frac{E(X)}{V(X) + n[E(X)]^2}.
\]
b) The UMVUE of \( \theta \) is \( \frac{T(Y)}{nE(X)} \) which uses \( c_U = \frac{1}{nE(X)} \).

**Proof.** a) \( E(W_n) = c \sum_{i=1}^{n} E(t(Y_i)) = cn\theta E(X) \), and \( V(W_n) = c^2 \sum_{i=1}^{n} V(t(Y_i)) = c^2 n\theta^2 V(X) \). Hence \( MSE(c) \equiv MSE(W_n) = V(W_n) + [E(W_n) - \theta]^2 = c^2 n\theta^2 V(X) + (cn\theta E(X) - \theta)^2 \). Thus
\[
\frac{d}{dc} MSE(c) = 2cn\theta^2 V(X) + 2(cn\theta E(X) - \theta)n\theta E(X) \overset{set}{=} 0,
\]
or\[
c(n\theta^2 V(X) + n^2\theta^2[E(X)]^2) = n\theta^2 E(X),
\]
or\[
c = \frac{E(X)}{V(X) + n[E(X)]^2},
\]
which is unique. Now
\[
\frac{d^2}{dc^2} MSE(c) = 2[n\theta^2 V(X) + n^2\theta^2[E(X)]^2] > 0.
\]
So \( MSE(c) \) is convex and \( c = c_M \) is the minimizer.

b) \( E[c_U T(Y)] = \theta \), hence \( c_U T(Y) \) is the UMVUE of \( \theta \) by the LSU theorem. \( \square \)

**Remark 6.1.** Chapter 10 has several UMVUE examples.
6.3 Summary

1) The bias of the estimator $T$ for $\tau(\theta)$ is

$$B(T) \equiv B_{\tau(\theta)}(T) \equiv \text{Bias}_{\tau(\theta)}(T) = E_{\theta}T - \tau(\theta)$$

and the MSE is

$$\text{MSE}_{\tau(\theta)}(T) = E_{\theta}[(T - \tau(\theta))^2] = V_{\theta}(T) + \left[B_{\tau(\theta)}(T)\right]^2.$$  

2) $T$ is an unbiased estimator of $\tau(\theta)$ if $E_{\theta}[T] = \tau(\theta)$ for all $\theta \in \Theta$.

3) Let $U \equiv U(Y_1, \ldots, Y_n)$ be an estimator of $\tau(\theta)$. Then $U$ is the UMVUE of $\tau(\theta)$ if $U$ is an unbiased estimator of $\tau(\theta)$ and if $\text{VAR}_{\theta}(U) \leq \text{VAR}_{\theta}(W)$ for all $\theta \in \Theta$ where $W$ is any other unbiased estimator of $\tau(\theta)$.

4) If $Y_1, \ldots, Y_n$ are iid from a 1P–REF $f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$ where $\eta = w(\theta) \in \Omega = (a, b)$, and if $T \equiv T(Y) = \sum_{i=1}^{n} t(Y_i)$, then by the LSU Theorem, $g(T)$ is the UMVUE of its expectation $\tau(\theta) = E_{\theta}(g(T))$.

5) Given a complete sufficient statistic $T(Y)$ and any unbiased estimator $W(Y)$ of $\tau(\theta)$, then $U(Y) = E[W(Y)|T(Y)]$ is the UMVUE of $\tau(\theta)$.

7) $I_n(\theta) = E_{\theta}\left[\left(\frac{\partial}{\partial \theta} \log f(Y|\theta)\right)^2\right]$.

8) $\text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_n(\theta)}$.

9) If $Y_1, \ldots, Y_n$ are iid from a 1P–REF $f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$, then a) $I_1(\theta) = -E_{\theta}\left[\frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta))\right]$. b) $I_n(\tau(\theta)) = \frac{nI_1(\theta)}{[\tau'(\theta)]^2}$. c) $\text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}$. d) Information inequality: Let $Y_1, \ldots, Y_n$ be iid from a 1P–REF and let $W(Y)$ be any unbiased estimator of $\tau(\theta) \equiv E_{\theta}[W(Y)]$. Then $\text{VAR}_{\theta}(W(Y)) \geq \text{FCRLB}_n(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{nI_1(\theta)}$. 
e) Rule of thumb for a 1P–REF: Let $T(Y) = \sum_{i=1}^n t(Y_i)$ and $\tau(\theta) = E_\theta(g(T(Y)))$. Then $g(T(Y))$ is the UMVUE of $\tau(\theta)$ by LSU, but the information inequality is strict for nonlinear functions $g(T(Y))$. Expect the equality

$$\text{VAR}_\theta(g(T(Y))) = \frac{[\tau'(\theta)]^2}{n I_1(\theta)}$$

only if $g$ is a linear function, i.e., $g(T) = a + bT$ for some fixed constants $a$ and $b$.

10) If the family is not an exponential family, the FCRLB may not be a lower bound on the variance of unbiased estimators of $\tau(\theta)$.

### 6.4 Complements

For a more precise statement of when the FCRLB is achieved and for some counterexamples, see Wijsman (1973) and Joshi (1976). Although the FCRLB is not very useful for finding UMVUEs, similar ideas are useful for finding the asymptotic variances of UMVUEs and MLEs. See Chapter 8 and Portnoy (1977).

Karakostas (1985) has useful references for UMVUEs. Also see Guenther (1978).

### 6.5 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.**

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

**6.1*.** Let $W$ be an estimator of $\tau(\theta)$. Show that

$$MSE_{\tau(\theta)}(W) = \text{Var}_\theta(W) + [\text{Bias}_{\tau(\theta)}(W)]^2.$$ 

**6.2.** (Aug. 2002 QUAL): Let $X_1, \ldots, X_n$ be independent identically distributed random variable from a $N(\mu, \sigma^2)$ distribution. Hence $E(X_1) = \mu$
and \( VAR(X_1) = \sigma^2 \). Consider estimators of \( \sigma^2 \) of the form

\[
S^2(k) = \frac{1}{k} \sum_{i=1}^{n} (X_i - \overline{X})^2
\]

where \( k > 0 \) is a constant to be chosen. Determine the value of \( k \) which gives the smallest mean square error. (Hint: Find the MSE as a function of \( k \), then take derivatives with respect to \( k \). Also, use Theorem 4.1c and Remark 5.1 VII.)

6.3. Let \( X_1, \ldots, X_n \) be iid \( N(\mu, 1) \) random variables. Find \( \tau(\mu) \) such that 
\[
T(X_1, \ldots, X_n) = (\sum_{i=1}^{n} X_i)^2
\]
is the UMVUE of \( \tau(\mu) \).

6.4. Let \( X \sim N(\mu, \sigma^2) \) where \( \sigma^2 \) is known. Find the Fisher information \( I_1(\mu) \).

6.5. Let \( X \sim N(\mu, \sigma^2) \) where \( \mu \) is known. Find the Fisher information \( I_1(\sigma^2) \).

6.6. Let \( X_1, \ldots, X_n \) be iid \( N(\mu, \sigma^2) \) random variables where \( \mu \) is known and \( \sigma^2 > 0 \). Then 
\[
W = \sum_{i=1}^{n} (X_i - \mu)^2
\]
is a complete sufficient statistic and 
\[
W \sim \sigma^2 \chi^2_n.
\]
From Chapter 10,
\[
EY^k = \frac{2^k \Gamma(k + n/2)}{\Gamma(n/2)}
\]
if \( Y \sim \chi^2_n \). Hence

\[
T_k(X_1, \ldots, X_n) \equiv \frac{\Gamma(n/2)W^k}{2^k \Gamma(k + n/2)}
\]
is the UMVUE of \( \tau_k(\sigma^2) = \sigma^{2k} \) for \( k > 0 \). Note that \( \tau_k(\theta) = (\theta)^k \) and \( \theta = \sigma^2 \).

a) Show that 
\[
Var\theta T_k(X_1, \ldots, X_n) = \sigma^{4k} \left[ \frac{\Gamma(n/2)\Gamma(2k+n/2)}{\Gamma(k+n/2)\Gamma(k+n/2)} - 1 \right] \equiv c_k \sigma^{4k}.
\]

b) Let \( k = 2 \) and show that \( Var\theta[T_2] - FCRLB(\tau_2(\theta)) > 0 \) where \( FCRLB(\tau_2(\theta)) \) is for estimating \( \tau_2(\sigma^2) = \sigma^4 \) and \( \theta = \sigma^2 \).
6.7. (Jan. 2001 Qual): Let $X_1, \ldots, X_n$ be independent, identically distributed $N(\mu, 1)$ random variables where $\mu$ is unknown and $n \geq 2$. Let $t$ be a fixed real number. Then the expectation

$$E_\mu(I_{(-\infty,t]}(X_1)) = P_\mu(X_1 \leq t) = \Phi(t - \mu)$$

for all $\mu$ where $\Phi(x)$ is the cumulative distribution function of a $N(0,1)$ random variable.

a) Show that the sample mean $\overline{X}$ is a sufficient statistic for $\mu$.

b) Explain why (or show that) $\overline{X}$ is a complete sufficient statistic for $\mu$.

c) Using the fact that the conditional distribution of $X_1$ given $\overline{X} = \overline{\tau}$ is the $N(\overline{\tau}, 1 - 1/n)$ distribution where the second parameter $1 - 1/n$ is the variance of conditional distribution, find

$$E_\mu(I_{(-\infty,t]}(X_1)|\overline{X} = \overline{\tau}) = E_\mu[I_{(-\infty,t]}(W)]$$

where $W \sim N(\overline{\tau}, 1 - 1/n)$. (Hint: your answer should be $\Phi(g(\overline{\tau}))$ for some function $g$.)

d) What is the uniformly minimum variance unbiased estimator for $\Phi(t - \mu)$?

Problems from old quizzes and exams.

6.8. Suppose that $X$ is Poisson with pmf

$$f(x|\lambda) = P(X = x|\lambda) = \frac{e^{-\lambda \lambda^x}}{x!}$$

where $x = 0, 1, \ldots$ and $\lambda > 0$. Find the Fisher information $I_1(\lambda)$.

6.9. Let $X_1, \ldots, X_n$ be iid Exponential($\beta$) random variables and $Y_1, \ldots, Y_m$ iid Exponential($\beta/2$) random variables. Assume that the $Y_i$’s and $X_j$’s are independent.

a) Find the joint pdf $f(x_1, \ldots, x_n, y_1, \ldots, y_m)$ and show that this pdf is a regular exponential family with complete sufficient statistic $T = \sum_{i=1}^n X_i + 2 \sum_{i=1}^m Y_i$.

b) Find the function $\tau(\beta)$ such that $T$ is the UMVUE of $\tau(\beta)$. (Hint: find $E_\beta[T]$. The theorems of this chapter apply since $X_1, \ldots, X_n, 2Y_1, \ldots, 2Y_m$ are iid.)
CHAPTER 6. UMVUES AND THE FCRLB

6.10. Let $X_1, ..., X_n$ be independent, identically distributed $N(\mu, 1)$ random variables where $\mu$ is unknown.

a) Find $E_{\mu}[X_1^2]$.

b) Using the fact that the conditional distribution of $X_1$ given $\overline{X} = \overline{x}$ is the $N(\overline{x}, 1 - 1/n)$ distribution where the second parameter $1 - 1/n$ is the variance of conditional distribution, find

$$E_{\mu}(X_1^2 | \overline{X} = \overline{x}).$$

[Hint: this expected value is equal to $E(W^2)$ where $W \sim N(\overline{x}, 1 - 1/n)$.

c) What is the MLE for $\mu^2 + 1$? (Hint: you may use the fact that the MLE for $\mu$ is $\overline{X}$.)

d) What is the uniformly minimum variance unbiased estimator for $\mu^2 + 1$? Explain.

6.11. Let $X_1, ..., X_n$ be a random sample from a Poisson($\lambda$) population.

a) Find the Fréchet Cramér Rao lower bound $FCRLB_n(\lambda^2)$ for the variance of an unbiased estimator of $\tau(\lambda) = \lambda^2$.

b) The UMVUE for $\lambda^2$ is $T(X_1, ..., X_n) = (\overline{X})^2 - \overline{X}/n$. Will $Var_{\lambda}[T] = FCRLB_n(\lambda^2)$ or will $Var_{\lambda}[T] > FCRLB_n(\lambda^2)$? Explain. (Hint: use the rule of thumb 9e from Section 6.3.)

6.12. Let $X_1, ..., X_n$ be independent, identically distributed Poisson($\lambda$) random variables where $\lambda > 0$ is unknown.

a) Find $E_{\lambda}[X_1^2]$.

b) Using the fact that the conditional distribution of $X_1$ given $\sum_{i=1}^n X_i = y$ is the Binomial($y, 1/n$) distribution, find

$$E_{\lambda}(X_1^2 | \sum_{i=1}^n X_i = y).$$

c) Find $\tau(\lambda)$ such that $E_{\lambda}(X_1^2 | \sum_{i=1}^n X_i)$ is the uniformly minimum variance unbiased estimator for $\tau(\lambda)$.

6.13. Let $X_1, ..., X_n$ be iid Bernoulli($\rho$) random variables.
a) Find the Fisher information $I_1(\rho)$.

b) Find the Fréchet Cramér Rao lower bound for unbiased estimators of $\tau(\rho) = \rho$.

c) The MLE for $\rho$ is $\bar{X}$. Find $\text{Var}(\bar{X})$.

d) Does the MLE achieve the FCRLB? Is this surprising? Explain.

6.14. (Jan. 2003 Qual): Let $X_1, ..., X_n$ be independent, identically distributed exponential$(\theta)$ random variables where $\theta > 0$ is unknown. Consider the class of estimators of $\theta$

\[ \{ T_n(c) = c \sum_{i=1}^{n} X_i \mid c > 0 \}. \]

Determine the value of $c$ that minimizes the mean square error MSE. Show work and prove that your value of $c$ is indeed the global minimizer.

6.15. Let $X_1, ..., X_n$ be iid from a distribution with pdf

\[ f(x|\theta) = \theta x^{\theta-1} I(0 < x < 1), \quad \theta > 0. \]

a) Find the MLE of $\theta$.

b) What is the MLE of $1/\theta^2$? Explain.

c) Find the Fisher information $I_1(\theta)$. You may use the fact that $-\log(X) \sim$ exponential$(1/\theta)$.

d) Find the Fréchet Cramér Rao lower bound for unbiased estimators of $\tau(\theta) = 1/\theta^2$.

6.16. Let $X_1, ..., X_n$ be iid random variables with $E(X) = \mu$ and $\text{Var}(X) = 1$. Suppose that $T = \sum_{i=1}^{n} X_i$ is a complete sufficient statistic. Find the UMVUE of $\mu^2$.

6.17. Let $X_1, ..., X_n$ be iid exponential$(\lambda)$ random variables.

a) Find $I_1(\lambda)$.

b) Find the FCRLB for estimating $\tau(\lambda) = \lambda^2$.

c) If $T = \sum_{i=1}^{n} X_i$, it can be shown that the UMVUE of $\lambda^2$ is

\[ W = \frac{\Gamma(n)}{\Gamma(2+n)} T^2. \]
Do you think that $\text{Var}_\lambda(W)$ is equal to the FCRLB in part b)? Explain briefly.

6.18. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ where $\mu$ is known and $n > 1$. Suppose interest is in estimating $\theta = \sigma^2$. You should have memorized the fact that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}. $$

a) Find the MSE of $S^2$ for estimating $\sigma^2$.

b) Find the MSE of $T$ for estimating $\sigma^2$ where

$$T = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2.$$

6.19. (Aug. 2000 SIU, 1995 Univ. Minn. Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables from a $N(\mu, \sigma^2)$ distribution. Hence $E(X_1) = \mu$ and $\text{VAR}(X_1) = \sigma^2$. Suppose that $\mu$ is known and consider estimates of $\sigma^2$ of the form

$$S^2(k) = \frac{1}{k} \sum_{i=1}^{n} (X_i - \mu)^2$$

where $k$ is a constant to be chosen. Note: $E(X_{m}^2) = m$ and $\text{VAR}(X_{m}^2) = 2m$. Determine the value of $k$ which gives the smallest mean square error. (Hint: Find the MSE as a function of $k$, then take derivatives with respect to $k$.)

6.20. (Aug. 2001 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables with pdf

$$f(x|\theta) = \frac{2x}{\theta} e^{-x^2/\theta}, \quad x > 0$$

and $f(x|\theta) = 0$ for $x \leq 0$.

a) Show that $X_1^2$ is an unbiased estimator of $\theta$. (Hint: use the substitution $W = X^2$ and find the pdf of $W$ or use u-substitution with $u = x^2/\theta$.)

b) Find the Cramer-Rao lower bound for the variance of an unbiased estimator of $\theta$. 

c) Find the uniformly minimum variance unbiased estimator (UMVUE) of $\theta$.

6.21. (Aug. 2001 Qual): See Mukhopadhyay (2000, p. 377). Let $X_1, \ldots, X_n$ be iid $N(\theta, \theta^2)$ normal random variables with mean $\theta$ and variance $\theta^2$. Let

$$T_1 = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and let

$$T_2 = c_n S = c_n \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

where the constant $c_n$ is such that $E_\theta[c_n S] = \theta$. You do not need to find the constant $c_n$. Consider estimators $W(\alpha)$ of $\theta$ of the form.

$$W(\alpha) = \alpha T_1 + (1 - \alpha) T_2$$

where $0 \leq \alpha \leq 1$.

a) Find the variance

$$\text{Var}_\theta[W(\alpha)] = \text{Var}_\theta(\alpha T_1 + (1 - \alpha) T_2).$$

b) Find the mean square error of $W(\alpha)$ in terms of $\text{Var}_\theta(T_1), \text{Var}_\theta(T_2)$ and $\alpha$.

c) Assume that

$$\text{Var}_\theta(T_2) \approx \frac{\theta^2}{2n}.$$ Determine the value of $\alpha$ that gives the smallest mean square error. (Hint: Find the MSE as a function of $\alpha$, then take the derivative with respect to $\alpha$. Set the derivative equal to zero and use the above approximation for $\text{Var}_\theta(T_2)$. Show that your value of $\alpha$ is indeed the global minimizer.)

6.22. (Aug. 2003 Qual): Suppose that $X_1, \ldots, X_n$ are iid normal distribution with mean 0 and variance $\sigma^2$. Consider the following estimators: $T_1 = \frac{1}{2}|X_1 - X_2|$ and $T_2 = \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_i^2}$.

a) Is $T_1$ unbiased for $\sigma$? Evaluate the mean square error (MSE) of $T_1$. 
b) Is $T_2$ unbiased for $\sigma$? If not, find a suitable multiple of $T_2$ which is unbiased for $\sigma$.

**6.23.** (Aug. 2003 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf (probability density function)

$$f(x) = \frac{1}{\lambda} \exp \left( -\frac{x}{\lambda} \right)$$

where $x$ and $\lambda$ are both positive. Find the uniformly minimum variance unbiased estimator (UMVUE) of $\lambda^2$.

**6.24.** (Jan. 2004 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables with pdf (probability density function)

$$f(x) = \sqrt{\frac{\sigma}{2\pi x^3}} \exp \left( -\frac{\sigma}{2x} \right)$$

where $x$ and $\sigma$ are both positive. Then $X_i = \frac{\sigma}{W_i}$ where $W_i \sim \chi^2_1$. Find the uniformly minimum variance unbiased estimator (UMVUE) of $\frac{1}{\sigma}$.

**6.25.** (Jan. 2004 Qual): Let $X_1, ..., X_n$ be a random sample from the distribution with density

$$f(x) = \begin{cases} \frac{2x}{\theta^2}, & 0 < x < \theta \\ 0, & \text{elsewhere} \end{cases}$$

Let $T = \max(X_1, ..., X_n)$. To estimate $\theta$ consider estimators of the form $CT$. Determine the value of $C$ which gives the smallest mean square error.

**6.26.** (Aug. 2004 Qual): Let $X_1, ..., X_n$ be a random sample from a distribution with pdf

$$f(x) = \frac{2x}{\theta^2}, \quad 0 < x < \theta.$$ 

Let $T = c\bar{X}$ be an estimator of $\theta$ where $c$ is a constant.

a) Find the mean square error (MSE) of $T$ as a function of $c$ (and of $\theta$ and $n$).

b) Find the value $c$ that minimizes the MSE. Prove that your value is the minimizer.
6.27. (Aug. 2004 Qual): Suppose that $X_1, ..., X_n$ are iid Bernoulli($p$) where $n \geq 2$ and $0 < p < 1$ is the unknown parameter.

a) Derive the UMVUE of $\nu(p)$, where $\nu(p) = e^2(p(1-p))$.

b) Find the Cramér Rao lower bound for estimating $\nu(p) = e^2(p(1-p))$.

6.28. Let $X_1, ..., X_n$ be independent identically distributed Poisson($\lambda$) random variables. Find the UMVUE of

$$\frac{\lambda}{n} + \lambda^2.$$

6.29. Let $Y_1, ..., Y_n$ be iid Poisson($\theta$) random variables.

a) Find the UMVUE for $\theta$.

b) Find the Fisher information $I_1(\theta)$.

c) Find the FCRLB for unbiased estimators of $\tau(\theta) = \theta$.

d) The MLE for $\theta$ is $\overline{Y}$. Find $\text{Var}(\overline{Y})$.

e) Does the MLE achieve the FCRLB? Is this surprising? Explain.

6.30. (Jan. 2009 Qual): Suppose that $Y_1, ..., Y_n$ are independent binomial($m_i, \rho$) where the $m_i \geq 1$ are known constants. Let

$$T_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} m_i} \quad \text{and} \quad T_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{m_i}$$

be estimators of $\rho$.

a) Find $\text{MSE}(T_1)$.

b) Find $\text{MSE}(T_2)$.

c) Which estimator is better?

Hint: by the arithmetic–geometric–harmonic mean inequality,

$$\frac{1}{n} \sum_{i=1}^{n} m_i \geq \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i}.$$

6.31. (Sept. 2010 Qual): Let $Y_1, ..., Y_n$ be iid gamma($\alpha = 10, \beta$) random variables. Let $T = c\overline{Y}$ be an estimator of $\beta$ where $c$ is a constant.

a) Find the mean square error (MSE) of $T$ as a function of $c$ (and of $\beta$ and $n$).
b) Find the value $c$ that minimizes the MSE. Prove that your value is the minimizer.

6.32. (Jan. 2011 Qual): Let $Y_1, ..., Y_n$ be independent identically distributed random variables with pdf (probability density function)

$$f(y) = (2 - 2y)I_{(0,1)}(y) \nu \exp[(1 - \nu)(-\log(2y - y^2))]$$

where $\nu > 0$ and $n > 1$. The indicator $I_{(0,1)}(y) = 1$ if $0 < y < 1$ and $I_{(0,1)}(y) = 0$, otherwise.

a) Find a complete sufficient statistic.

b) Find the Fisher information $I_1(\nu)$ if $n = 1$.

c) Find the Cramer Rao lower bound (CRLB) for estimating $1/\nu$.

d) Find the uniformly minimum unbiased estimator (UMVUE) of $\nu$.

Hint: You may use the fact that $T_n = -\sum_{i=1}^n \log(2Y_i - Y_i^2) \sim G(n, 1/\nu)$, and

$$E(T_n^r) = \frac{1}{\nu^r} \frac{\Gamma(r + n)}{\Gamma(n)}$$

for $r > -n$. Also $\Gamma(1 + x) = x\Gamma(x)$ for $x > 0$.

6.33. (Sept. 2011 Qual): Let $Y_1, ..., Y_n$ be iid random variables from a distribution with pdf

$$f(y) = \frac{\theta}{2(1 + |y|)^{\theta+1}}$$

where $\theta > 0$ and $y$ is real. Then $W = \log(1 + |Y|)$ has pdf $f(w) = \theta e^{-w\theta}$ for $w > 0$.

a) Find a complete sufficient statistic.

b) Find the (Fisher) information number $I_1(\theta)$.

c) Find the uniformly minimum variance unbiased estimator (UMVUE) for $\theta$.

6.34. (Similar to Sept. 2010 Qual): Suppose that $X_1, X_2, ..., X_n$ are independent identically distributed random variables from normal distribution with unknown mean $\mu$ and known variance $\sigma^2$. Consider the parametric function $g(\mu) = e^{2\mu}$.

a) Derive the uniformly minimum variance unbiased estimator (UMVUE) of $g(\mu)$.

b) Find the Cramer-Rao lower bound (CRLB) for the variance of an unbiased estimator of $g(\mu)$.

c) Is the CRLB attained by the variance of the UMVUE of $g(\mu)$?
Chapter 7

Testing Statistical Hypotheses

A hypothesis is a statement about a population parameter $\theta$, and in hypothesis testing there are two competing hypotheses called the null hypothesis $H_0 \equiv H_0$ and the alternative hypothesis $H_1 \equiv H_A$. Let $\Theta_1$ and $\Theta_0$ be disjoint sets with $\Theta_1 \subset \Theta$ where $\Theta$ is the parameter space. Then $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$.

When a researcher wants strong evidence about a hypothesis, usually this hypothesis is $H_1$. For example, if Ford claims that their latest car gets 30 mpg on average, then $H_0 : \mu = 30$ and $H_1 : \mu > 30$ are reasonable hypotheses where $\theta = \mu$ is the population mean mpg of the car.

Definition 7.1. Assume that the data $Y = (Y_1, \ldots, Y_n)$ has pdf or pmf $f(y|\theta)$ for $\theta \in \Theta$. A hypothesis test is a rule for rejecting $H_0$.

Definition 7.2. A type I error is rejecting $H_0$ when $H_0$ is true. A type II error is failing to reject $H_0$ when $H_0$ is false. $P_{\theta}(\text{reject } H_0) = P_{\theta}(\text{type I error})$ if $\theta \in \Theta_0$ while $P_{\theta}(\text{reject } H_0) = 1 - P_{\theta}(\text{type II error})$ if $\theta \in \Theta_1$.

Definition 7.3. The power function of a hypothesis test is

$$\beta(\theta) = P_{\theta}(\text{Ho is rejected})$$

for $\theta \in \Theta$.

Often there is a rejection region $R$ and an acceptance region. Reject $H_0$ if the observed statistic $T(y) \in R$, otherwise fail to reject $H_0$. Then $\beta(\theta) = P_{\theta}(T(Y) \in R) = P_{\theta}(\text{reject } H_0)$. 

203
Definition 7.4. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a size $\alpha$ test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \alpha$$

and a level $\alpha$ test if

$$\sup_{\theta \in \Theta_0} \beta(\theta) \leq \alpha.$$ 

Notice that for $\theta \in \Theta_0$, $\beta(\theta) = P_{\theta}(\text{type I error})$ and for $\theta \in \Theta_1$, $\beta(\theta) = 1 - P_{\theta}(\text{type II error})$. We would like $\beta(\theta) \approx 0$ for $\theta \in \Theta_0$ and $\beta(\theta) \approx 1$ for $\theta \in \Theta_1$, but this may not be possible even if the sample size $n$ is large. The tradeoff is that decreasing the probability of a type I error increases the probability of a type II error while decreasing the probability of a type II error increases the probability of a type I error. The size or level of the test gives an upper bound $\alpha$ on the probability of the type I error. Typically the level is fixed, e.g. $\alpha = 0.05$, and then we attempt to find tests that have a small probability of type II error. The following example is a level 0.07 and size 0.0668 test.

Example 7.1. Suppose that $Y \sim N(\mu, 1/9)$ where $\mu \in \{0, 1\}$. Let $H_0 : \mu = 0$ and $H_1 : \mu = 1$. Let $T(Y) = Y$ and suppose that we reject $H_0$ if $Y \geq 0.5$. Let $Z \sim N(0, 1)$ and $\sigma = 1/3$. Then

$$\beta(0) = P_0(Y \geq 0.5) = P_0\left(\frac{Y - 0}{1/3} \geq \frac{0.5}{1/3}\right) = P(Z \geq 1.5) \approx 0.0668.$$ 

$$\beta(1) = P_1(Y \geq 0.5) = P_1\left(\frac{Y - 1}{1/3} \geq \frac{0.5 - 1}{1/3}\right) = P(Z \geq -1.5) \approx 0.9332.$$ 

7.1 Exponential Families, the Neyman Pearson Lemma, and UMP Tests

Definition 7.5. Consider all level $\alpha$ tests of $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. A uniformly most powerful (UMP) level $\alpha$ test is a level $\alpha$ test with power function $\beta_{UMP}(\theta)$ such that $\beta_{UMP}(\theta) \geq \beta(\theta)$ for every $\theta \in \Theta_1$ where $\beta$ is the power function for any level $\alpha$ test of $H_0$ vs $H_1$.

The following three theorems can be used to find a UMP test that is both a level $\alpha$ and a size $\alpha$ test.
Theorem 7.1, The Neyman Pearson Lemma (NPL). Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$ where the pdf or pmf corresponding to $\theta_i$ is $f(y|\theta_i)$ for $i = 0, 1$. Suppose the test rejects $H_0$ if $f(y|\theta_1) > kf(y|\theta_0)$, and rejects $H_0$ with probability $\gamma$ if $f(y|\theta_1) = kf(y|\theta_0)$ for some $k \geq 0$. If

$$\alpha = \beta(\theta_0) = P_{\theta_0}[f(Y|\theta_1) > kf(Y|\theta_0)] + \gamma P_{\theta_0}[f(Y|\theta_1) = kf(Y|\theta_0)],$$

then this test is a UMP level $\alpha$ test.

**Proof.** The proof is for pdfs. Replace the integrals by sums for pmfs. Following Ferguson (1967, p. 202), a test can be written as a test function $\psi(y) \in [0, 1]$ where $\psi(y)$ is the probability that the test rejects $H_0$ when $Y = y$. The Neyman Pearson (NP) test function is

$$\phi(y) = \begin{cases} 
1, & f(y|\theta_1) > kf(y|\theta_0) \\
\gamma, & f(y|\theta_1) = kf(y|\theta_0) \\
0, & f(y|\theta_1) < kf(y|\theta_0)
\end{cases}$$

and $\alpha = E_{\theta_0}[\phi(Y)]$. Consider any level $\alpha$ test $\psi(y)$. Since $\psi(y)$ is a level $\alpha$ test,

$$E_{\theta_0}[\psi(Y)] \leq E_{\theta_0}[\phi(Y)] = \alpha. \quad (7.1)$$

Then the NP test is UMP if the power

$$\beta_{\psi}(\theta_1) = E_{\theta_1}[\psi(Y)] \leq \beta_{\phi}(\theta_1) = E_{\theta_1}[\phi(Y)].$$

Let $f_i(y) = f(y|\theta_i)$ for $i = 0, 1$. Notice that $\phi(y) = 1 \geq \psi(y)$ if $f_1(y) > kf_0(y)$ and $\phi(y) = 0 \leq \psi(y)$ if $f_1(y) < kf_0(y)$. Hence

$$\int [\phi(y) - \psi(y)][f_1(y) - kf_0(y)]dy \geq 0 \quad (7.2)$$

since the integrand is nonnegative. Hence the power

$$\beta_{\phi}(\theta_1) - \beta_{\psi}(\theta_1) = E_{\theta_1}[\phi(Y)] - E_{\theta_1}[\psi(Y)] \geq k(E_{\theta_0}[\phi(Y)] - E_{\theta_0}[\psi(Y)]) \geq 0$$

where the first inequality follows from (7.2) and the second inequality from Equation (7.1). QED
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

Theorem 7.2, One Sided UMP Tests via the Neyman Pearson Lemma. Suppose that the hypotheses are of the form $H_0 : \theta \leq \theta_o$ vs $H_1 : \theta > \theta_o$ or $H_0 : \theta \geq \theta_o$ vs $H_1 : \theta < \theta_o$, or that the inequality in $H_o$ is replaced by equality. Also assume that

$$\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_o).$$

Pick $\theta_1 \in \Theta_1$ and use the Neyman Pearson lemma to find the UMP test for $H_o^* : \theta = \theta_o$ vs $H_1^* : \theta = \theta_1$. Then the UMP test rejects $H_o^*$ if $f(y|\theta_1) > k f(y|\theta_o)$, and rejects $H_o^*$ with probability $\gamma$ if $f(y|\theta_1) = k f(y|\theta_o)$ for some $k \geq 0$ where $\alpha = \beta(\theta_o)$. This test is also the UMP level $\alpha$ test for $H_o : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$ if $k$ does not depend on the value of $\theta_1 \in \Theta_1$. If $R = f(Y|\theta_1)/f(Y|\theta_o)$, then $\alpha = P_{\theta_o}(R > k) + \gamma P_{\theta_o}(R = k)$.

Theorem 7.3, One Sided UMP Tests for Exponential Families. Let $Y_1, ..., Y_n$ be a sample with a joint pdf or pmf from a one parameter exponential family where $w(\theta)$ is increasing and $T(y)$ is the complete sufficient statistic. Alternatively, let $Y_1, ..., Y_n$ be iid with pdf or pmf

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)]$$

from a one parameter exponential family where $\theta$ is real and $w(\theta)$ is increasing. Here $T(y) = \sum_{i=1}^n t(y_i)$. I) Let $\theta_1 > \theta_o$. Consider the test that rejects $H_o$ if $T(y) > k$ and rejects $H_o$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) > k) + \gamma P_{\theta_o}(T(Y) = k)$. This test is the UMP level $\alpha$ test for

a) $H_o : \theta = \theta_o$ vs $H_1 : \theta = \theta_1$,

b) $H_o : \theta = \theta_o$ vs $H_1 : \theta > \theta_o$, and
c) $H_o : \theta < \theta_o$ vs $H_1 : \theta > \theta_o$.

II) Let $\theta_1 < \theta_o$. Consider the test that rejects $H_o$ if $T(y) < k$ and rejects $H_o$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) < k) + \gamma P_{\theta_o}(T(Y) = k)$. This test is the UMP level $\alpha$ test for
d) $H_o : \theta = \theta_o$ vs $H_1 : \theta = \theta_1$
e) $H_o : \theta = \theta_o$ vs $H_1 : \theta < \theta_o$, and

f) $H_o : \theta \geq \theta_o$ vs $H_1 : \theta < \theta_o$.

Proof. I) Let $\theta_1 > \theta_o$. a) Then

$$\frac{f(y|\theta_1)}{f(y|\theta_o)} = \left[\frac{c(\theta_1)}{c(\theta_o)}\right]^n \frac{\exp[w(\theta_1)\sum_{i=1}^n t(y_i)]}{\exp[w(\theta_o)\sum_{i=1}^n t(y_i)]} > c$$

Theorem 7.4, Two Sided UMP Tests for Exponential Families. Let $Y_1, ..., Y_n$ be a sample with a joint pdf or pmf from a one parameter exponential family where $w(\theta)$ is increasing and $T(y)$ is the complete sufficient statistic. Alternatively, let $Y_1, ..., Y_n$ be iid with pdf or pmf

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)]$$

from a one parameter exponential family where $\theta$ is real and $w(\theta)$ is increasing. Here $T(y) = \sum_{i=1}^n t(y_i)$. I) Let $\theta_1 > \theta_o$. Consider the test that rejects $H_o$ if $T(y) > k$ and rejects $H_o$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) > k) + \gamma P_{\theta_o}(T(Y) = k)$. This test is the UMP level $\alpha$ test for

a) $H_o : \theta = \theta_o$ vs $H_1 : \theta = \theta_1$,

b) $H_o : \theta = \theta_o$ vs $H_1 : \theta > \theta_o$, and
c) $H_o : \theta < \theta_o$ vs $H_1 : \theta > \theta_o$.

II) Let $\theta_1 < \theta_o$. Consider the test that rejects $H_o$ if $T(y) < k$ and rejects $H_o$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) < k) + \gamma P_{\theta_o}(T(Y) = k)$. This test is the UMP level $\alpha$ test for
d) $H_o : \theta = \theta_o$ vs $H_1 : \theta = \theta_1$
e) $H_o : \theta = \theta_o$ vs $H_1 : \theta < \theta_o$, and

f) $H_o : \theta \geq \theta_o$ vs $H_1 : \theta < \theta_o$.

Proof. I) Let $\theta_1 > \theta_o$. a) Then

$$\frac{f(y|\theta_1)}{f(y|\theta_o)} = \left[\frac{c(\theta_1)}{c(\theta_o)}\right]^n \frac{\exp[w(\theta_1)\sum_{i=1}^n t(y_i)]}{\exp[w(\theta_o)\sum_{i=1}^n t(y_i)]} > c$$
iff

\[ [w(\theta_1) - w(\theta_o)] \sum_{i=1}^{n} t(y_i) > d \]

iff \( \sum_{i=1}^{n} t(y_i) > k \) since \( w(\theta) \) is increasing. Hence the result holds by the NP lemma. b) The test in a) did not depend on \( \theta_1 > \theta_o \), so the test is UMP by Theorem 7.2. c) In a), \( \theta_o < \theta_1 \) were arbitrary, so \( \sup_{\theta \in \Theta_o} \beta(\theta) = \beta(\theta_o) \) where \( \Theta_o = \{ \theta \in \Theta | \theta \leq \theta_o \} \). So the test is UMP by Theorem 7.2. The proof of II) is similar, but \( \theta_1 < \theta_o \) so \( [w(\theta_1) - w(\theta_o)] < 0 \), and there is a sign change. QED

Remark 7.1. a) The UMP level \( \alpha \) tests in Theorems 7.1, 7.2 and 7.3 are also UMP size \( \alpha \) tests. b) As a mnemonic, note that the inequality used in the rejection region is the same as the inequality in the alternative hypothesis. Usually \( \gamma = 0 \) if \( f \) is a pdf. Suppose that the parameterization is

\[ f(y|\theta) = h(y)c(\theta)\exp[\tilde{w}(\theta)\tilde{t}(y)] \]

where \( \tilde{w}(\theta) \) is decreasing. Then set \( w(\theta) = -\tilde{w}(\theta) \) and \( t(y) = -\tilde{t}(y) \). In this text, \( w(\theta) \) is an increasing function if \( w(\theta_o) < w(\theta_1) \) for \( \theta_o < \theta_1 \) and nondecreasing if \( w(\theta_o) \leq w(\theta_1) \). Some texts use “strictly increasing” for “increasing” and use “increasing” for “nondecreasing.” c) A simple hypothesis consists of exactly one distribution for the sample. A composite hypothesis consists of more than one distribution for the sample.

If the data are iid from a one parameter exponential family, then Theorem 7.3 is simpler to use than the Neyman Pearson lemma since the test statistic \( T \) will have a distribution from an exponential family by Theorem 3.5. This result makes finding the cutoff value \( k \) easier. To find a UMP test via the Neyman Pearson lemma, you need to check that the cutoff value \( k \) does not depend on \( \theta_1 \in \Theta_1 \) and usually need to transform the NP test statistic to put the test in useful form. With exponential families, the transformed test statistic is often \( T \).

Example 7.2. Suppose that \( X_1, ..., X_{10} \) are iid Poisson with unknown mean \( \lambda \). Derive the most powerful level \( \alpha = 0.10 \) test for \( H_0 : \lambda = 0.30 \) versus \( H_1 : \lambda = 0.40 \).

Solution: Since

\[ f(x|\lambda) = \frac{1}{x!}e^{-\lambda}\exp[\log(\lambda)x] \]
and \( \log(\lambda) \) is an increasing function of \( \lambda \), by Theorem 7.3 the UMP test rejects \( H_0 \) if \( \sum x_i > k \) and rejects \( H_0 \) with probability \( \gamma \) if \( \sum x_i = k \) where \( \alpha = 0.1 = P_{\lambda=0.3}(\sum X_i > k) + \gamma P_{\lambda=0.3}(\sum X_i = k) \). Notice that
\[
\gamma = \alpha - P_{H_0}(\sum X_i > k) - P_{H_0}(\sum X_i = k).
\]

Alternatively use the Neyman Pearson lemma. Let
\[
r = f(x|0.4)/f(x|0.3) = \frac{e^{-n\lambda_1} \lambda_1^{\sum x_i} / \prod x_i!}{e^{-n\lambda_0} \lambda_0^{\sum x_i} / \prod x_i!} = e^{-n(\lambda_1-\lambda_0)} \left( \frac{\lambda_1}{\lambda_0} \right)^{\sum x_i}.
\]
Since \( \lambda_1 = 0.4 > 0.3 = \lambda_0 \), \( r > c \) is equivalent to \( \sum x_i > k \) and the NP UMP test has the same form as the UMP test found using the much simpler Theorem 7.3.

<table>
<thead>
<tr>
<th>( k )</th>
<th>( P(T = k) )</th>
<th>( F(k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0498</td>
<td>0.0498</td>
</tr>
<tr>
<td>1</td>
<td>0.1494</td>
<td>0.1992</td>
</tr>
<tr>
<td>2</td>
<td>0.2240</td>
<td>0.4232</td>
</tr>
<tr>
<td>3</td>
<td>0.2240</td>
<td>0.6472</td>
</tr>
<tr>
<td>4</td>
<td>0.1680</td>
<td>0.8152</td>
</tr>
<tr>
<td>5</td>
<td>0.1008</td>
<td>0.9160</td>
</tr>
</tbody>
</table>

If \( H_0 \) is true, then \( T = \sum_{i=1}^{10} X_i \sim \text{Pois}(3) \) since \( 3 = 10\lambda_0 = 10(0.3) \). The above table gives the probability that \( T = k \) and \( F(k) = P(T \leq k) \). First find the smallest integer \( k \) such that \( P_{\lambda=0.3}(\sum X_i > k) = P(T > k) < \alpha = 0.1 \). Since \( P(T > k) = 1 - F(k) \), find the smallest value of \( k \) such that \( F(k) > 0.9 \). This happens with \( k = 5 \). Next use (7.3) to find \( \gamma \).

\[
\gamma = \frac{0.1 - (1 - 0.9160)}{0.1008} = \frac{0.1 - 0.084}{0.1008} = \frac{0.016}{0.1008} \approx 0.1587.
\]

Hence the \( \alpha = 0.1 \) UMP test rejects \( H_0 \) if \( T = \sum_{i=1}^{10} X_i > 5 \) and rejects \( H_0 \) with probability 0.1587 if \( \sum_{i=1}^{10} X_i = 5 \). Equivalently, the test function \( \phi(T) \) gives the probability of rejecting \( H_0 \) for a given value of \( T \) where
\[
\phi(T) = \begin{cases} 
1, & T > 5 \\
0.1587, & T = 5 \\
0, & T < 5.
\end{cases}
\]

**Example 7.3.** Let \( X_1, \ldots, X_n \) be independent identically distributed random variables from a distribution with pdf
\[
f(x) = \frac{2}{\lambda \sqrt{2\pi x}} \exp \left[ -\frac{(\log(x))^2}{2\lambda^2} \right]
\]
where $\lambda > 0$ where and $0 \leq x \leq 1$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0: \lambda = 1$ vs. $H_1: \lambda = 2$?

b) If possible, find the UMP level $\alpha$ test for $H_0: \lambda = 1$ vs. $H_1: \lambda > 1$.

Solution. a) By the NP lemma reject $H_0$ if

$$f(x|\lambda = 2) > k'$$

The LHS =

$$\frac{1}{n} \exp \left[ \frac{-1}{8} \sum [\log(x_i)]^2 \right]$$

So reject $H_0$ if

$$\frac{1}{2n} \exp [\sum [\log(x_i)]^2 (\frac{1}{2} - \frac{1}{8})] > k'$$

or if $\sum [\log(X_i)]^2 > k$ where $P_{\lambda=1}(\sum [\log(X_i)]^2 > k) = \alpha$.

b) In the above argument, with any $\lambda_1 > 1$, get

$$\sum [\log(x_i)]^2 (\frac{1}{2} - \frac{1}{2\lambda_1^2})$$

and

$$\frac{1}{2} - \frac{1}{2\lambda_1^2} > 0$$

for any $\lambda_1^2 > 1$. Hence the UMP test is the same as in a).

Theorem 7.3 gives the same UMP test as a) for both a) and b) since the pdf is a 1P-REF and $w(\lambda^2) = -1/(2\lambda^2)$ is an increasing function of $\lambda^2$. Also, it can be shown that $\sum [\log(X_i)]^2 \sim \lambda^2 \chi_n^2$, so $k = \chi_{n,1-\alpha}^2$ where $P(W > \chi_{n,1-\alpha}^2) = \alpha$ if $W \sim \chi_n^2$.

Example 7.4. Let $X_1, ..., X_n$ be independent identically distributed (iid) random variables with probability density function

$$f(x) = \frac{2}{\lambda \sqrt{2\pi}} e^{x} \exp \left( \frac{-(e^x - 1)^2}{2\lambda^2} \right)$$

where $x > 0$ and $\lambda > 0$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0: \lambda = 1$ vs. $H_1: \lambda = 2$?
b) If possible, find the UMP level $\alpha$ test for $H_0 : \lambda = 1$ vs. $H_1 : \lambda > 1$.

a) By the NP lemma reject $H_0$ if

$$\frac{f(x | \lambda = 2)}{f(x | \lambda = 1)} > k'.$$

The LHS =

$$\frac{1}{2\pi} \exp\left[\frac{-1}{8} \sum (e^{x_i} - 1)^2\right] \exp\left[\frac{-1}{2} \sum (e^{x_i} - 1)^2\right].$$

So reject $H_0$ if

$$\frac{1}{2^n} \exp[\sum (e^{x_i} - 1)^2(\frac{1}{2} - \frac{1}{8})] > k'$$

or if $\sum (e^{x_i} - 1)^2 > k$ where $P_1(\sum (e^{X_i} - 1)^2 > k) = \alpha$.

b) In the above argument, with any $\lambda_1 > 1$, get

$$\sum (e^{x_i} - 1)^2(\frac{1}{2} - \frac{1}{2\lambda_1^2})$$

and

$$\frac{1}{2} - \frac{1}{2\lambda_1^2} > 0$$

for any $\lambda_1^2 > 1$. Hence the UMP test is the same as in a).

Alternatively, use the fact that this is an exponential family where $w(\lambda^2) = -1/(2\lambda^2)$ is an increasing function of $\lambda^2$ with $T(X_i) = (e^{X_i} - 1)^2$. Hence the same test in a) is UMP for both a) and b) by Theorem 7.3.

Example 7.5. Let $X_1, ..., X_n$ be independent identically distributed random variables from a half normal $HN(\mu, \sigma^2)$ distribution with pdf

$$f(x) = \frac{2}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > \mu$ and $\mu$ is real. Assume that $\mu$ is known.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 = 4$?

b) If possible, find the UMP level $\alpha$ test for $H_0 : \sigma^2 = 1$ vs. $H_1 : \sigma^2 > 1$.

Solution: a) By the NP lemma reject $H_0$ if

$$\frac{f(x | \sigma^2 = 4)}{f(x | \sigma^2 = 1)} > k'.$$
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

The LHS =
\[ \frac{1}{2^n} \exp\left[\frac{-\sum(x_i-\mu)^2}{2(4)}\right] \exp\left[\frac{\sum(x_i-\mu)^2}{2}\right]. \]

So reject Ho if
\[ \frac{1}{2^n} \exp\left[\sum(x_i-\mu)^2\right] > k' \]
or if \( \sum(x_i-\mu)^2 > k \) where \( P_{\lambda^2=1}(\sum(X_i-\mu)^2 > k) = \alpha. \)

Under Ho, \( \sum(X_i-\mu)^2 \sim \chi^2_n \) so \( k = \chi^2_n(1-\alpha) \) where \( P(\chi^2_n > \chi^2_n(1-\alpha)) = \alpha. \)

b) In the above argument,
\[ \frac{-1}{2(4)} + 0.5 = \frac{-1}{8} + 0.5 > 0 \]
but
\[ \frac{-1}{2\sigma^2} + 0.5 > 0 \]
for any \( \sigma^2 > 1. \) Hence the UMP test is the same as in a).

Alternatively, use the fact that this is an exponential family where \( w(\sigma^2) = -1/(2\sigma^2) \) is an increasing function of \( \sigma^2 \) with \( T(X_i) = (X_i-\mu)^2. \) Hence the test in a) is UMP for a) and b) by Theorem 7.3.

Example 7.6. Let \( Y_1, \ldots, Y_n \) be iid with pdf
\[ f(y) = \frac{1}{2\sqrt{2\pi} \theta y^{\nu}} \left( \frac{y^2 - \theta^2}{\sqrt{\theta} - \sqrt{\theta y}} \right) I(y > 0) \frac{1}{\nu} \exp\left[\frac{-1}{2\nu^2} t(y)\right] \]
where \( \nu > 0, \theta \) is known and \( t(y) \) is a function such that \( t(Y) \sim \nu^2 \chi^2_1. \)

a) Find the UMP level \( \alpha \) test for \( H_0 : \nu = 1 \) versus \( H_1 : \nu = 1.19. \)

b) Suppose \( n = 12 \) and \( \alpha = 0.05. \) Find the power \( \beta(1.19) \) when \( \nu = 1.19. \)

Solution: a) This is an exponential family. Note that \( 2\nu^2 \) is increasing, so \( 1/(2\nu^2) \) is decreasing and \( w(\lambda) = -1/(2\nu^2) \) is increasing. Thus the UMP test rejects \( H_0 \) if \( \sum_{i=1}^n t(y_i) > k \) where \( \alpha = P_l(\sum_{i=1}^n t(Y_i) > k). \)

b) Use \( \alpha \) to find \( k \) and then find the power. If \( H_0 \) is true so \( \nu = 1, \) then \( \sum_{i=1}^n t(Y_i) \sim \chi^2_{12}. \) Thus \( k = \chi^2_{12}(0.95) = 21.03 \) using a chi-square table. If \( \nu = 1.19, \) then \( \sum_{i=1}^n t(Y_i) \sim (1.19)^2 \chi^2_{12}. \) So \( \beta(1.19) = P_{1.19}(\sum_{i=1}^n t(Y_i) > 21.03) = P(X > 21.03/(1.19)^2) = P(X > 14.8506) = 0.25 \) using a chi-square table where \( X \sim \chi^2_{12}. \)
7.2 Likelihood Ratio Tests

Definition 7.6. Let \((Y_1, ..., Y_n)\) be the data with joint pdf or joint pmf \(f(y|\theta)\) where \(\theta\) is a vector of unknown parameters with parameter space \(\Theta\). Let \(\hat{\theta}\) be the MLE of \(\theta\) and let \(\hat{\theta}_o\) be the MLE of \(\theta\) if the parameter space is \(\Theta_0\) (where \(\Theta_0 \subset \Theta\)). A likelihood ratio test (LRT) statistic for testing \(H_0: \theta \in \Theta_0\) versus \(H_1: \theta \in \Theta_0^c\) is

\[
\lambda(y) = \frac{L(\hat{\theta}_o|y)}{L(\hat{\theta}|y)} = \frac{\sup_{\Theta_0} L(\theta|y)}{\sup_{\Theta} L(\theta|y)}.
\]

(7.4)

The likelihood ratio test (LRT) has a rejection region of the form

\[
R = \{y|\lambda(y) \leq c\}
\]

where \(0 \leq c \leq 1\), and \(\alpha = \sup_{\theta \in \Theta_0} P_{\theta}(\lambda(Y) \leq c)\). Suppose \(\theta_o \in \Theta_0\) and \(\sup_{\theta \in \Theta_0} P_{\theta}(\lambda(Y) \leq c) = P_{\theta_o}(\lambda(Y) \leq c)\). Then \(\alpha = P_{\theta_o}(\lambda(Y) \leq c)\).

**Rule of Thumb 7.1: Asymptotic Distribution of the LRT.** Let \(Y_1, ..., Y_n\) be iid. Then under strong regularity conditions, 

\[-2\log \lambda(Y) \approx \chi^2_j\]

for large \(n\) where \(j = r - q\), \(r\) is the number of free parameters specified by \(\theta \in \Theta\), and \(q\) is the number of free parameters specified by \(\theta \in \Theta_o\). Hence the approximate LRT rejects \(H_0\) if \(-2\log \lambda(y) > c\) where \(P(\chi^2_j > c) = \alpha\). Thus \(c = \chi^2_{j,1-\alpha}\) where \(P(\chi^2_j > \chi^2_{j,1-\alpha}) = \alpha\).

Often \(\hat{\theta}\) is called the **unrestricted MLE** of \(\theta\), and \(\hat{\theta}_o\) is called the **restricted MLE** of \(\theta\). Often \(\theta = \theta_o\) is a scalar parameter, \(\Theta_0 = (a, \theta_o]\) and \(\Theta_1 = \Theta_0^c = (\theta_o, b]\) or \(\Theta_0 = [\theta_o, b]\) and \(\Theta_1 = (a, \theta_o]\).

**Remark 7.2.** Suppose the problem wants the rejection region in useful form. Find the two MLEs and write \(L(\theta|y)\) in terms of a sufficient statistic. Then you should either I) simplify the LRT test statistic \(\lambda(y)\) and try to find an equivalent test that uses test statistic \(T(y)\) where the distribution of \(T(Y)\) is known (ie put the LRT in useful form). Often the LRT rejects \(H_0\) if \(T > k\) (or \(T < k\)). Getting the test into useful form can be very difficult. Monotone transformations such as log or power transformations can be useful. II) If you can not find a statistic \(T\) with a simple distribution, state that the Rule of Thumb 7.1 suggests that the LRT test rejects \(H_0\) if \(-2\log \lambda(y) > \chi^2_{j,1-\alpha}\) where \(\alpha = P(-2\log \lambda(Y) > \chi^2_{j,1-\alpha})\). Using II) is dangerous because for many data sets the asymptotic result will not be valid.
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

Example 7.7. Let $X_1, ..., X_n$ be independent identically distributed random variables from a $N(\mu, \sigma^2)$ distribution where the variance $\sigma^2$ is known. We want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

a) Derive the likelihood ratio test.

b) Let $\lambda$ be the likelihood ratio. Show that $-2 \log \lambda$ is a function of $(X - \mu_0)$.

c) Assuming that $H_0$ is true, find $P(-2 \log \lambda > 3.84)$.

Solution: a) The likelihood function

$$L(\mu) = (2\pi\sigma^2)^{-n/2} \exp[-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2]$$

and the MLE for $\mu$ is $\hat{\mu} = \bar{x}$. Thus the numerator of the likelihood ratio test statistic is $L(\mu_0)$ and the denominator is $L(\bar{x})$. So the test is reject $H_0$ if $\lambda(x) = L(\mu_0) / L(\bar{x}) \leq c$ where $\alpha = P_{\mu_0}(\lambda(X) \leq c)$.

b) As a statistic, $\log \lambda = \log L(\mu_0) - \log L(\bar{x}) = -\frac{1}{2\sigma^2} [\sum (x_i - \mu_0)^2 - \sum (x_i - \bar{x})^2] = \frac{n}{2\sigma^2} [\bar{x} - \mu_0]^2$ since $\sum (x_i - \mu_0)^2 = \sum (x_i - \bar{x} + \bar{x} - \mu_0)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2$. So $-2 \log \lambda = \frac{n}{\sigma^2} [\bar{x} - \mu_0]^2$.

c) $-2 \log \lambda \sim \chi^2_1$ and from a chi–square table, $P(-2 \log \lambda > 3.84) = 0.05$.

Example 7.8. Let $Y_1, ..., Y_n$ be iid $N(\mu, \sigma^2)$ random variables where $\mu$ and $\sigma^2$ are unknown. Set up the likelihood ratio test for $H_0 : \mu = \mu_o$ versus $H_A : \mu \neq \mu_o$.

Solution: Under $H_0$, $\mu = \mu_o$ is known and the MLE

$$(\hat{\mu}_o, \hat{\sigma}_o^2) = (\mu_o, \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_o)^2).$$

Recall that

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{Y}, \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2).$$

Now

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2} (y_i - \mu)^2].$$
Thus
\[ \lambda(y) = \frac{L(\hat{\mu}_o, \hat{\sigma}_o^2 | y)}{L(\mu, \hat{\sigma}_2 | y)} = \frac{1/(\hat{\sigma}_o^2)^n/2 \exp[1/(2\hat{\sigma}_o^2) \sum_{i=1}^n (y_i - \mu_o)^2]}{1/(\hat{\sigma}_2^2)^n/2 \exp[1/(2\hat{\sigma}_2^2) \sum_{i=1}^n (y_i - \bar{y})^2]} = \left( \frac{\hat{\sigma}_2^2}{\hat{\sigma}_o^2} \right)^{n/2} \exp(\frac{n}{2}) \right]. \]

The LRT rejects Ho iff \( \lambda(y) \leq c \) where \( \sup_{\sigma^2} P_{\mu_o, \sigma^2}(\lambda(Y) \leq c) = \alpha. \)

On an exam the above work may be sufficient, but to implement the LRT, more work is needed. Notice that the LRT rejects Ho iff \( \hat{\sigma}_2^2/\hat{\sigma}_o^2 \leq c' \) iff \( \hat{\sigma}_o^2/\hat{\sigma}_2^2 \geq k' \).

Using
\[ \sum_{i=1}^n (y_i - \mu_o)^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + n(y - \mu_o)^2, \]
the LRT rejects Ho iff
\[ \left[ 1 + \frac{n(\bar{y} - \mu_o)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right] \geq k'' \]
iff
\[ \frac{\sqrt{n} \left| \bar{y} - \mu_o \right|}{\left( \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{1/2}} \geq k \]
where \( s \) is the observed sample standard deviation. Hence the LRT is equivalent to the usual \( t \) test with test statistic
\[ T_o = \frac{\bar{y} - \mu_o}{S/\sqrt{n}} \]
that rejects Ho iff \( |T_o| \geq k \) with \( k = t_{n-1,1-\alpha/2} \) where \( P(T \leq t_{n-1,1-\alpha/2}) = 1 - \alpha/2 \) when \( T \sim t_{n-1} \).

**Example 7.9.** Suppose that \( X_1, ..., X_n \) are iid \( N(0, \sigma^2) \) where \( \sigma > 0 \) is the unknown parameter. With preassigned \( \alpha \in (0, 1) \), derive a level \( \alpha \) likelihood ratio test for the null hypothesis \( H_0 : \sigma^2 = \sigma_o^2 \) against an alternative hypothesis \( H_A : \sigma^2 \neq \sigma_o^2 \).

Solution: The likelihood function is given by
\[ L(\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp(- \frac{1}{2\sigma^2} \sum_{i=1}^n x_i^2) \]
for all \( \sigma^2 > 0 \), and \( \hat{\sigma}^2(x) = \frac{\sum_{i=1}^{n} x_i^2}{n} \) is the MLE for \( \sigma^2 \). Under Ho, \( \hat{\sigma}^2_o = \sigma^2_o \) since \( \sigma^2_o \) is the only value in the parameter space \( \Theta_o = \{\sigma^2_o\} \). Thus

\[
\lambda(x) = \frac{L(\hat{\sigma}^2_o|x)}{L(\sigma^2|x)} = \frac{\sup_{\Theta_o} L(\sigma^2|x)}{\sup_{\sigma^2} L(\sigma^2|x)} = \frac{(2\pi \sigma^2_o)^{-n/2} \exp(-\frac{1}{2\sigma^2_o} \sum_{i=1}^{n} x_i^2)}{(2\pi \hat{\sigma}^2)^{-n/2} \exp(-\frac{n}{2})}.
\]

So

\[
\lambda(x) = \left( \frac{\hat{\sigma}^2}{\sigma^2_o} \right)^{n/2} \exp\left( -\frac{n\hat{\sigma}^2}{2\sigma^2_o} \right) e^{n/2} = \left[ \frac{\hat{\sigma}^2}{\sigma^2_o} \exp\left( 1 - \frac{\hat{\sigma}^2}{\sigma^2_o} \right) \right]^{n/2}.
\]

The LRT rejects \( H_0 \) if \( \lambda(x) \leq c \) where \( P_{\sigma^2_o}(\lambda(X) \leq c) = \alpha \).

The function \( g(u) = ue^{1-u}I(u > 0) \) monotonically increases for \( 0 < u < d \), monotonically decreases for \( d < u < \infty \), and attains its maximum at \( u = d \), for some \( d > 0 \). So \( \lambda(x) \) will be small in the two tail areas.

Under \( H_0 \), \( T = \sum_{i=1}^{n} X_i^2 / \sigma^2_o \sim \chi^2_n \). Hence the LR test will reject Ho if \( T < a \) or \( T > b \) where \( 0 < a < b \). The \( a \) and \( b \) correspond to horizontal line drawn on the \( \chi^2_n \) pdf such that the tail area is \( \alpha \). Hence \( a \) and \( b \) need to be found numerically. An approximation that should be good for large \( n \) rejects Ho if \( T < \chi^2_n,\frac{1}{2} \) or \( T > \chi^2_n,1-\frac{1}{2} \) where \( P(\chi^2_n < \chi^2_{n,\alpha}) = \alpha \).

## 7.3 Summary

For hypothesis testing there is a null hypothesis \( H_0 \) and an alternative hypothesis \( H_1 \equiv H_A \). A hypothesis test is a rule for rejecting \( H_0 \). Either reject \( H_0 \) or fail to reject \( H_0 \). A simple hypothesis consists of exactly one distribution for the sample. A composite hypothesis consists of more than one distribution for the sample.

The power \( \beta(\theta) = P_{\theta}(\text{reject } H_o) \) is the probability of rejecting \( H_o \) when \( \theta \) is the true value of the parameter. Often the power function can not be calculated, but you should be prepared to calculate the power for a sample of size one for a test of the form \( H_o : f(y) = f_0(y) \) versus \( H_1 : f(y) = f_1(y) \) or if the test is of the form \( \sum t(Y_i) > k \) or \( \sum t(Y_i) < k \) when \( \sum t(Y_i) \) has an easily handled distribution under \( H_1 \), eg binomial, normal, Poisson or \( \chi^2_p \). To compute the power, you need to find \( k \) and \( \gamma \) for the given value of \( \alpha \).

Consider all level \( \alpha \) tests of \( H_o : \theta \in \Theta_o \) vs \( H_1 : \theta_1 \in \Theta_1 \). A uniformly most powerful (UMP) level \( \alpha \) test is a level \( \alpha \) test with power function
$\beta_{UMP}(\theta)$ such that $\beta_{UMP}(\theta) \geq \beta(\theta)$ for every $\theta \in \Theta_1$ where $\beta$ is the power function for any level $\alpha$ test of $H_0$ vs $H_1$.

One Sided UMP Tests for Exponential Families. Let $Y_1, \ldots, Y_n$ be iid with pdf or pmf

$$f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)]$$

from a one parameter exponential family where $\theta$ is real and $w(\theta)$ is increasing. Let $T(y) = \sum_{i=1}^n t(y_i)$. Then the UMP test for $H_0 : \theta \leq \theta_o$ vs $H_A : \theta > \theta_o$ rejects $H_0$ if $T(y) > k$ and rejects $H_0$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) > k) + \gamma P_{\theta_o}(T(Y) = k)$. The UMP test for $H_0 : \theta \geq \theta_o$ vs $H_A : \theta < \theta_o$ rejects $H_0$ if $T(y) < k$ and rejects $H_0$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_o}(T(Y) < k) + \gamma P_{\theta_o}(T(Y) = k)$.

Fact: if $f$ is a pdf, then usually $\gamma = 0$. For a pmf and $H_A : \theta > \theta_o$,

$$\gamma = \frac{\alpha - P_{\theta_o}[T(Y) > k]}{P_{\theta_o}[T(Y) = k]}.$$

For a pmf and $H_A : \theta < \theta_o$,

$$\gamma = \frac{\alpha - P_{\theta_o}[T(Y) < k]}{P_{\theta_o}[T(Y) = k]}.$$

As a mnemonic, note that the inequality used in the rejection region is the same as the inequality in the alternative hypothesis. Suppose that the parameterization is

$$f(y|\theta) = h(y)c(\theta)\exp[\tilde{w}(\theta)\tilde{t}(y)]$$

where $\tilde{w}(\theta)$ is decreasing. Then set $w(\theta) = -\tilde{w}(\theta)$ and $t(y) = -\tilde{t}(y)$.

Recall that $w(\theta)$ is increasing on $\Theta$ if $w'(\theta) > 0$ for $\theta \in \Theta$, and $w(\theta)$ is decreasing on $\Theta$ if $w'(\theta) < 0$ for $\theta \in \Theta$. Also $w(\theta)$ is nondecreasing on $\Theta$ if $w'(\theta) \geq 0$ for $\theta \in \Theta$, and $w(\theta)$ is nonincreasing on $\Theta$ if $w'(\theta) \leq 0$ for $\theta \in \Theta$.

The Neyman Pearson Lemma: Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta = \theta_1$ where the pdf or pmf corresponding to $\theta_i$ is $f(y|\theta_i)$ for $i =$
0.1. Suppose the test rejects $H_o$ if $f(y|\theta_1) > kf(y|\theta_0)$, and rejects $H_o$ with probability $\gamma$ if $f(y|\theta_1) = kf(y|\theta_0)$ for some $k \geq 0$. If

$$\alpha = \beta(\theta_0) = P_{\theta_0}[f(Y|\theta_1) > kf(Y|\theta_0)] + \gamma P_{\theta_0}[f(Y|\theta_1) = kf(Y|\theta_0)],$$

then this test is a UMP level $\alpha$ test and a UMP size $\alpha$ test.

**One Sided UMP Tests via the Neyman Pearson Lemma:** Suppose that the hypotheses are of the form $H_o : \theta \leq \theta_o$ vs $H_1 : \theta > \theta_o$ or $H_o : \theta \geq \theta_o$ vs $H_1 : \theta < \theta_o$, or that the inequality in $H_o$ is replaced by equality. Also assume that

$$\sup_{\theta \in \Theta_o} \beta(\theta) = \beta(\theta_o).$$

Pick $\theta_1 \in \Theta_1$ and use the Neyman Pearson lemma to find the UMP test for $H^*_o : \theta = \theta_o$ vs $H^*_1 : \theta = \theta_1$. Then the UMP test rejects $H^*_o$ if $f(y|\theta_1) > kf(y|\theta_0)$, and rejects $H^*_o$ with probability $\gamma$ if $f(y|\theta_1) = kf(y|\theta_0)$ for some $k \geq 0$ where $\alpha = \beta(\theta_o)$. This test is also the UMP level $\alpha$ test for $H_o : \theta \in \Theta_o$ vs $H_1 : \theta \in \Theta_1$ if $k$ does not depend on the value of $\theta_1 \in \Theta_1$. If $R = f(Y|\theta_1)/f(Y|\theta_0)$, then $\alpha = P_{\theta_o}(R > k) + \gamma P_{\theta_o}(R = k)$.

Fact: if $f$ is a pdf, then usually $\gamma = 0$ and $\alpha = P_{\theta_o}[f(Y|\theta_1) > kf(Y|\theta_0)]$. So $\gamma$ is important when $f$ is a pmf. For a pmf,

$$\gamma = \frac{\alpha - P_{\theta_0}[f(Y|\theta_1) > kf(Y|\theta_0)]}{P_{\theta_0}[f(Y|\theta_1) = kf(Y|\theta_0)].$$

Often it is too hard to give the UMP test in useful form. Then simply specify when the test rejects $H_o$ and specify $\alpha$ in terms of $k$ (eg $\alpha = P_{\theta_o}(T > k) + \gamma P_{\theta_o}(T = k)$).

The problem will be harder if you are asked to put the test in useful form. To find an UMP test with the NP lemma, often the ratio $\frac{f(y|\theta_1)}{f(y|\theta_0)}$ is computed. The test will certainly reject $H_o$ is the ratio is large, but usually the distribution of the ratio is not easy to use. Hence try to get an equivalent test by simplifying and transforming the ratio. Ideally, the ratio can be transformed into a statistic $T$ whose distribution is tabulated.

If the test rejects $H_o$ if $T > k$ (or if $T > k$ and with probability $\gamma$ if $T = k$, or if $T < k$, or if $T < k$ and with probability $\gamma$ if $T = k$) the test is
in useful form if for a given $\alpha$, you find $k$ and $\gamma$. If you are asked to find the power (perhaps with a table), put the test in useful form.

Let $(Y_1, \ldots, Y_n)$ be the data with joint pdf or pmf $f(y|\theta)$ where $\theta$ is a vector of unknown parameters with parameter space $\Theta$. Let $\hat{\theta}$ be the MLE of $\theta$ and let $\hat{\theta}_o$ be the MLE of $\theta$ if the parameter space is $\Theta_0$ (where $\Theta_0 \subset \Theta$). A LRT statistic for testing $H_o: \theta \in \Theta_0$ versus $H_1: \theta \in \Theta^c_0$ is

$$
\lambda(y) = \frac{L(\hat{\theta}_o|y)}{L(\hat{\theta}|y)}
$$

The LRT has a rejection region of the form

$$
R = \{y|\lambda(y) \leq c\}
$$

where $0 \leq c \leq 1$ and $\alpha = \sup_{\theta \in \Theta_0} P_\theta(\lambda(Y) \leq c)$.

Fact: Often $\Theta_0 = (a, \theta_o]$ and $\Theta_1 = [\theta_o, b)$ or $\Theta_0 = [\theta_o, b)$ and $\Theta_1 = (a, \theta_o)$.

If you are not asked to find the power or to put the LRT into useful form, it is often enough to find the two MLEs and write $L(\theta|y)$ in terms of a sufficient statistic. Simplify the statistic $\lambda(y)$ and state that the LRT test rejects $H_o$ if $\lambda(y) \leq c$ where $\alpha = \sup_{\theta \in \Theta_0} P_\theta(\lambda(Y) \leq c)$. If the sup is achieved at $\theta_o \in \Theta_0$, then $\alpha = P_{\theta_o}(\lambda(Y) \leq c)$.

Put the LRT into useful form if asked to find the power. Try to simplify $\lambda$ or transform $\lambda$ so that the test rejects $H_o$ if some statistic $T > k$ (or $T < k$). Getting the test into useful form can be very difficult. Monotone transformations such as log or power transformations can be useful. If you can not find a statistic $T$ with a simple distribution, use the large sample approximation to the LRT that rejects $H_o$ if $-2 \log \lambda(y) > \chi^2_{j,1-\alpha}$ where $P(\chi^2_j > \chi^2_{j,1-\alpha}) = \alpha$. Here $j = r - q$ where $r$ is the number of free parameters specified by $\theta \in \Theta$, and $q$ is the number of free parameters specified by $\theta \in \Theta_0$.

Some distribution facts useful for testing.

Memorize Theorems 2.17 and 4.1.

Suppose $Y_1, \ldots, Y_n$ are iid $N(\mu, \sigma^2)$. Then $Z = \frac{Y_i - \mu}{\sigma} \sim N(0, 1)$.

$$
Z = \frac{\sum Y_i - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad \text{while} \quad a + cY_i \sim N(a + c\mu, c^2\sigma^2).
$$
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

Suppose $Z, Z_1, ..., Z_n$ are iid $N(0,1)$. Then $Z^2 \sim \chi^2_1$.

Also $a + cZ_i \sim N(a, c^2)$ while $\sum_{i=1}^n Z_i^2 \sim \chi^2_n$.

If $X_i$ are independent $\chi^2_{k_i} \equiv \chi^2(k_i)$ for $i = 1, ..., n$, then $\sum_{i=1}^n X_i \sim \chi^2(\sum_{i=1}^n k_i)$.

Let $W \sim EXP(\lambda)$ and let $c > 0$. Then $cW \sim EXP(c\lambda)$.

Let $W \sim gamma(\nu, \lambda)$ and let $c > 0$. Then $cW \sim gamma(\nu, c\lambda)$.

If $W \sim EXP(\lambda) \sim gamma(1, \lambda)$, then $2W/\lambda \sim EXP(2) \sim gamma(1, 2) \sim \chi^2(2)$.

Let $k \geq 0.5$ and let $2k$ be an integer. If $W \sim gamma(k, \lambda)$, then $2W/\lambda \sim gamma(k, 2) \sim \chi^2(2k)$.

Let $W_1, ..., W_n$ be independent $gamma(\nu_i, \lambda)$. Then $\sum_{i=1}^n W_i \sim gamma(\sum_{i=1}^n \nu_i, \lambda)$.

7.4 Complements

Example 7.10. Let $Y_1, ..., Y_n$ be iid from a one parameter exponential family with pdf or pmf $f(y|\theta)$ with complete sufficient statistic $T(Y) = \sum_{i=1}^n t(Y_i)$ where $\theta > 0$, $t(Y_i) \sim \theta \chi^2_j$ and $j$ is some positive integer, often 1 or 2. Suppose the UMP test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$ rejects $H_0$ if $T(Y) > k$ where $\alpha = P_0(T(Y) > k)$. Since $\frac{T(Y)}{\theta} \sim \chi^2_{nj}$,

$$\alpha = P(\frac{T(Y)}{\theta_0} > \frac{k}{\theta_0}) = P(\chi^2_{nj} > \frac{k}{\theta_0}) = P(\chi^2_{nj,1-\alpha})$$

and $k = \theta_0 \chi^2_{nj,1-\alpha}$. Hence the power

$$\beta(\theta) = P_0(T(Y) > \theta_0 \chi^2_{nj,1-\alpha}) = P(\frac{T(Y)}{\theta} > \frac{\theta_0}{\theta} \chi^2_{nj,1-\alpha}) = P(\chi^2_{nj} > \frac{\theta_0}{\theta} \chi^2_{nj,1-\alpha})$$

which can be approximated with a chi-square table.

Definition 7.7. Let $Y_1, ..., Y_n$ have pdf or pmf $f(y|\theta)$ for $\theta \in \Theta$. Let $T(Y)$ be a statistic. Then $f(y|\theta)$ has a monotone likelihood ratio (MLR) in statistic $T$ if for any two values $\theta_0, \theta_1 \in \Theta$ with $\theta_0 < \theta_1$, the ratio...
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

$f(y|\theta_1)/f(y|\theta_0)$ depends on the vector $y$ only through $T(y)$, and this ratio is an increasing function of $T(y)$ over the possible values of $T(y)$.

**Remark 7.3.** Theorem 7.3 is a corollary of the following theorem, because under the conditions of Theorem 7.3, $f(y|\theta)$ has MLR in $T(y) = \sum_{i=1}^{n} t(y_i)$.

**Theorem 7.4, MLR UMP Tests.** Let $Y_1, \ldots, Y_n$ be a sample with a joint pdf or pmf $f(y|\theta)$ that has MLR in statistic $T(y)$. I) The UMP test for $H_0: \theta \leq \theta_0$ versus $H_1: \theta > \theta_0$ rejects $H_0$ if $T(y) > k$ and rejects $H_0$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_0}(T(Y) > k) + \gamma P_{\theta_0}(T(Y) = k)$. II) The UMP test for $H_0: \theta \geq \theta_0$ versus $H_1: \theta < \theta_0$ rejects $H_0$ if $T(y) < k$ and rejects $H_0$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_0}(T(Y) < k) + \gamma P_{\theta_0}(T(Y) = k)$.

**Proof.** I) Pick $\theta_1 > \theta_0$ and consider $H^*_0: \theta = \theta_0$ versus $H^*_1: \theta = \theta_1$. Let $h$ be the increasing function such that

$$\frac{f(y|\theta_1)}{f(y|\theta_0)} = h(T(y)) > c$$

iff $T(y) > k$. So the NP UMP test is equivalent to the test that rejects $H^*_0$ if $T(y) > k$ and rejects $H^*_0$ with probability $\gamma$ if $T(y) = k$ where $\alpha = P_{\theta_0}(T(Y) > k) + \gamma P_{\theta_0}(T(Y) = k)$. Since this test does not depend on $\theta_1 > \theta_0$, it is also the UMP test for $H^*_0: \theta = \theta_0$ versus $H_1: \theta > \theta_0$ by Theorem 7.2. Since $\theta_0 < \theta_1$ was arbitrary, $\sup_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0)$ if $\Theta_0 = \{\theta \in \Theta | \theta \leq \theta_0\}$, and the result follows. The proof of II) is similar, but $\theta_1 < \theta_0$. Thus $h$ is an increasing function with respect to $f(y|\theta_0)/f(y|\theta_1)$, but a decreasing function of $T(y)$ with respect to $f(y|\theta_0)/f(y|\theta_1)$. Thus

$$\frac{f(y|\theta_1)}{f(y|\theta_0)} = h(T(y)) > c$$

iff $T(y) < k$.

Lehmann and Romano (2005) is an authoritative PhD level text on testing statistical hypotheses. Many of the most used statistical tests of hypotheses are likelihood ratio tests, and several examples are given in DeGroot and Schervish (2001). Scott (2007) discusses the asymptotic distribution of the LRT test.
Birkes (1990) and Solomen (1975) compare the LRT and UMP tests. Rohatgi (1984, p. 725) claims that if the Neyman Pearson and likelihood ratio tests exist for a given size $\alpha$, then the two tests are equivalent, but this claim seems to contradict Solomen (1975). Exponential families have the MLR property, and Pfanzagl (1968) gives a partial converse.

### 7.5 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.**

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

**7.1.** Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$, $\sigma^2 > 0$. Let $\Theta_o = \{(\mu_o, \sigma^2) : \mu_o \text{ fixed}, \sigma^2 > 0\}$ and let $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$. Consider testing $H_0 : \theta \in \Theta_o$ vs $H_1 : \text{not } H_0$. The MLE $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2)$ while the restricted MLE is $\hat{\theta}_o = (\hat{\mu}_o, \hat{\sigma}_o^2) = (\mu_o, \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu_o)^2)$.

a) Show that the likelihood ratio statistic

$$\lambda(x) = (\frac{\hat{\sigma}^2}{\hat{\sigma}_o^2})^n/2 = [1 + \frac{n(\bar{X} - \mu_o)^2}{\sum_{i=1}^{n} (x_i - \bar{X})^2}]^{-n/2}.$$ 

b) Show that $H_o$ is rejected iff $|\sqrt{n}(\bar{X} - \mu_o)/S| \geq k$ and find $k$ if $n = 11$ and $\alpha = 0.05$. (Hint: show that $H_o$ is rejected iff $n(\bar{X} - \mu_o)^2/\sum_{i=1}^{n} (X_i - \bar{X})^2 \geq c$, then multiply both sides by a constant such that the left hand side has a $(t_{n-1})^2$ distribution. Use the fact that $\frac{\bar{X} - \mu_o}{S/\sqrt{n}} \sim t_{n-1}$ under $H_0$. Use a t-table to find $k$.)

**7.2.** Let $X_1, \ldots, X_n$ be a random sample from the distribution with pdf

$$f(x|\theta) = \frac{x^{\theta-1}e^{-x}}{\Gamma(\theta)}, \quad x > 0, \ \theta > 0.$$
For a) and b) do not put the rejection region into useful form.

a) Use the Neyman Pearson Lemma to find the UMP size $\alpha$ test for testing $H_0: \theta = 1$ vs $H_1: \theta = \theta_1$ where $\theta_1$ is a fixed number greater than 1.

b) Find the uniformly most powerful level $\alpha$ test of $H_0: \theta = 1$ versus $H_1: \theta > 1$.

Justify your steps. Hint: Use the statistic in part a).

7.3. Let $H_o : X_1, ..., X_n$ are iid $U(0,10)$ and $H_1 : X_1, ..., X_n$ are iid $U(4,7)$. Suppose you had a sample of size $n = 1000$. How would you decide which hypothesis is true?

Problems from old quizzes and exams.

7.4. Let $X_1, ..., X_{10}$ be iid Bernoulli$(p)$. The most powerful level $\alpha = 0.0547$ test of $H_o : p = 1/2$ vs $H_1 : p = 1/4$ rejects $H_o$ if $\sum_{i=1}^{10} x_i \leq 2$. $H_o$ is not rejected if $\sum_{i=1}^{10} x_i > 2$. Find the power of this test if $p = 1/4$.

7.5. Let $X_1, ..., X_n$ be iid exponential$(\beta)$. Hence the pdf is

$$f(x|\beta) = \frac{1}{\beta} \exp(-x/\beta)$$

where $0 \leq x$ and $0 < \beta$.

a) Find the MLE of $\beta$.

b) Find the level $\alpha$ likelihood ratio test for the hypotheses $H_o : \beta = \beta_o$ vs $H_1 : \beta \neq \beta_o$.

7.6. (Aug. 2002 Qual): Let $X_1, ..., X_n$ be independent, identically distributed random variables from a distribution with a beta$(\theta, \theta)$ pdf

$$f(x|\theta) = \frac{\Gamma(2\theta)}{\Gamma(\theta)\Gamma(\theta)} [x(1-x)]^{\theta-1}$$

where $0 < x < 1$ and $\theta > 0$.

a) Find the UMP (uniformly most powerful) level $\alpha$ test for $H_o : \theta = 1$ vs. $H_1 : \theta = 2$.

b) If possible, find the UMP level $\alpha$ test for $H_o : \theta = 1$ vs. $H_1 : \theta > 1$.

7.7. Let $X_1, ..., X_n$ be iid $N(\mu_1, 1)$ random variables and let $Y_1, ..., Y_n$ be iid $N(\mu_2, 1)$ random variables that are independent of the $X$’s.
a) Find the \( \alpha \) level likelihood ratio test for \( H_0: \mu_1 = \mu_2 \) vs \( H_1: \mu_1 \neq \mu_2 \). You may assume that \((X, Y)\) is the MLE of \((\mu_1, \mu_2)\) and that under the restriction \(\mu_1 = \mu_2 = \mu\), say, then the restricted MLE

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i}{2n}.
\]

b) If \( \lambda \) is the LRT test statistic of the above test, use the approximation

\[
-2 \log \lambda \approx \chi^2_d
\]

for the appropriate degrees of freedom \(d\) to find the rejection region of the test in useful form if \( \alpha = 0.05 \).

**7.8.** Let \( X_1, \ldots, X_n \) be independent identically distributed random variables from a distribution with pdf

\[
f(x) = \frac{2}{\sigma \sqrt{2\pi}} \exp \left( -\frac{[\log(x)]^2}{2\sigma^2} \right)
\]

where \( \sigma > 0 \) and \( x \geq 1 \).

If possible, find the UMP level \( \alpha \) test for \( H_0: \sigma = 1 \) vs. \( H_1: \sigma > 1 \).

**7.9.** Let \( X_1, \ldots, X_n \) be independent identically distributed random variables from a distribution with pdf

\[
f(x) = \frac{2}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right)
\]

where \( \sigma > 0 \) and \( x > \mu \) and \( \mu \) is real. Assume that \( \mu \) is known.

a) What is the UMP (uniformly most powerful) level \( \alpha \) test for \( H_0: \sigma^2 = 1 \) vs. \( H_1: \sigma^2 > 1 \)?

b) If possible, find the UMP level \( \alpha \) test for \( H_0: \sigma^2 = 1 \) vs. \( H_1: \sigma^2 > 1 \).

**7.10.** (Jan. 2001 SIU and 1990 Univ. MN Qual): Let \( X_1, \ldots, X_n \) be a random sample from the distribution with pdf

\[
f(x, \theta) = \frac{x^{\theta-1}e^{-x}}{\Gamma(\theta)}, \quad x > 0, \ \theta > 0.
\]

Find the uniformly most powerful level \( \alpha \) test of
CHAPTER 7. TESTING STATISTICAL HYPOTHESES

H: θ = 1 versus K: θ > 1.

7.11. (Jan 2001 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a $N(\mu, \sigma^2)$ distribution where the variance $\sigma^2$ is known. We want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$.

a) Derive the likelihood ratio test.

b) Let $\lambda$ be the likelihood ratio. Show that $-2 \log \lambda$ is a function of $(\bar{X} - \mu_0)$.

c) Assuming that $H_0$ is true, find $P(-2 \log \lambda > 3.84)$.

7.12. (Aug. 2001 Qual): Let $X_1, ..., X_n$ be iid from a distribution with pdf

$$f(x) = \frac{2x}{\lambda} \exp(-x^2/\lambda)$$

where $\lambda$ and $x$ are both positive. Find the level $\alpha$ UMP test for $H_o : \lambda = 1$ vs $H_1 : \lambda > 1$.

7.13. (Jan. 2003 Qual): Let $X_1, ..., X_n$ be iid from a distribution with pdf

$$f(x|\theta) = \frac{(\log \theta)\theta^x}{\theta - 1}$$

where $0 < x < 1$ and $\theta > 1$. Find the UMP (uniformly most powerful) level $\alpha$ test of $H_o : \theta = 2$ vs $H_1 : \theta = 4$.

7.14. (Aug. 2003 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a distribution with pdf

$$f(x) = \frac{x^2 \exp \left(\frac{-x^2}{2\sigma^2}\right)}{\sigma^3 \sqrt{2} \Gamma(3/2)}$$

where $\sigma > 0$ and $x \geq 0$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_o : \sigma = 1$ vs $H_1 : \sigma = 2$?

b) If possible, find the UMP level $\alpha$ test for $H_o : \sigma = 1$ vs $H_1 : \sigma > 1$. 

7.15. (Jan. 2004 Qual): Let $X_1, \ldots, X_n$ be independent identically distributed random variables from a distribution with pdf

$$f(x) = \frac{2}{\sigma \sqrt{2\pi} x} \exp\left(-\frac{[\log(x)]^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x \geq 1$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0 : \sigma = 1$ vs. $H_1 : \sigma = 2$?

b) If possible, find the UMP level $\alpha$ test for $H_0 : \sigma = 1$ vs. $H_1 : \sigma > 1$.

7.16. (Aug. 2004 Qual): Suppose $X$ is an observable random variable with its pdf given by $f(x)$, $x \in \mathbb{R}$. Consider two functions defined as follows:

$$f_0(x) = \begin{cases} \frac{3}{64} x^2 & 0 \leq x \leq 4 \\ 0 & \text{elsewhere} \end{cases}$$

$$f_1(x) = \begin{cases} \frac{3}{16} \sqrt{x} & 0 \leq x \leq 4 \\ 0 & \text{elsewhere} \end{cases}$$

Determine the most powerful level $\alpha$ test for $H_0 : f(x) = f_0(x)$ versus $H_a : f(x) = f_1(x)$ in the simplest implementable form. Also, find the power of the test when $\alpha = 0.01$.

7.17. (Sept. 2005 Qual): Let $X$ be one observation from the probability density function

$$f(x) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$$ 

a) Find the most powerful level $\alpha$ test of $H_0 : \theta = 1$ versus $H_1 : \theta = 2$.

b) For testing $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$, find the size and the power function of the test which rejects $H_0$ if $X > \frac{5}{8}$.

c) Is there a UMP test of $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$? If so, find it. If not, prove so.

7.18. Let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$ random variables where $\sigma^2 > 0$ is known.

a) Find the UMVUE of $\mu^2$.

(Hint: try estimators of the form $aT^2 + b$ where $T$ is the complete sufficient statistic for $\mu$. )
b) Suppose $\sigma^2 = 1$ and $n = 4$. Then the (uniformly) most powerful level $\alpha = 0.05$ test for $H_0: \mu = 0$ vs $H_1: \mu = 2$ rejects $H_0$ iff and only iff

$$\sum_{i=1}^{4} x_i > 3.29.$$  

Find the power $\beta(2)$ of this test if $\mu = 2$.

7.19. (Jan. 2009 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a half normal $HN(\mu, \sigma^2)$ distribution with pdf

$$f(x) = \frac{2}{\sigma \sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and $x > \mu$ and $\mu$ is real. **Assume that $\mu$ is known.**

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0: \sigma^2 = 1$ vs. $H_1: \sigma^2 = 4$ ?

b) If possible, find the UMP level $\alpha$ test for $H_0: \sigma^2 = 1$ vs. $H_1: \sigma^2 > 1$.

7.20. (Aug. 2009 Qual): Suppose that the test statistic $T(X)$ for testing $H_0: \lambda = 1$ versus $H_1: \lambda > 1$ has an exponential($1/\lambda_1$) distribution if $\lambda = \lambda_1$. The test rejects $H_0$ if $T(X) < \log(100/95)$.

a) Find the power of the test if $\lambda_1 = 1$.

b) Find the power of the test if $\lambda_1 = 50$.

c) Find the pvalue of this test.

7.21. (Aug. 2009 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a Burr type X distribution with pdf

$$f(x) = 2 \tau x e^{-x^2} (1 - e^{-x^2})^{\tau-1}$$

where $\tau > 0$ and $x > 0$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0: \tau = 2$ versus $H_1: \tau = 4$ ?

b) If possible, find the UMP level $\alpha$ test for $H_0: \tau = 2$ versus $H_1: \tau > 2$.

7.22. (Jan. 2010 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from an inverse exponential distribution with pdf

$$f(x) = \frac{\theta}{x^2} \exp\left(\frac{-\theta}{x}\right)$$
where $\theta > 0$ and $x > 0$.

a) What is the UMP (uniformly most powerful) level $\alpha$ test for $H_0 : \theta = 1$ versus $H_1 : \theta = 2$?

b) If possible, find the UMP level $\alpha$ test for $H_0 : \theta = 1$ versus $H_1 : \theta > 1$.

7.23. (Sept. 2010 Qual): Suppose that $X$ is an observable random variable with its pdf given by $f(x)$. Consider the two functions defined as follows: $f_0(x)$ is the probability density function of a Beta distribution with $\alpha = 1$ and $\beta = 2$ and and $f_1(x)$ is the pdf of a Beta distribution with $\alpha = 2$ and $\beta = 1$.

a) Determine the UMP level $\alpha = 0.10$ test for $H_0 : f(x) = f_0(x)$ versus $H_1 : f(x) = f_1(x)$. (Find the constant.)

b) Find the power of the test in a).

7.24. (Sept. 2010 Qual): The pdf of a bivariate normal distribution is

$$f(x, y) = \frac{1}{2\pi \sigma_1 \sigma_2 (1 - \rho^2)^{1/2}} \exp \left( \frac{-1}{2(1 - \rho^2)} \left[ \left( \frac{x - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x - \mu_1}{\sigma_1} \right) \left( \frac{y - \mu_2}{\sigma_2} \right) + \left( \frac{y - \mu_2}{\sigma_2} \right)^2 \right] \right)$$

where $-1 < \rho < 1$, $\sigma_1 > 0$, $\sigma_2 > 0$, while $x$, $y$, $\mu_1$, and $\mu_2$ are all real. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate normal distribution.

Let $\hat{\theta}(x, y)$ be the observed value of the MLE of $\theta$, and let $\hat{\theta}(X, Y)$ be the MLE as a random variable. Let the (unrestricted) MLEs be $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1$, $\hat{\sigma}_2$, and $\hat{\rho}$. Then

$$T_1 = \sum_{i=1}^{n} \left( \frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right)^2 = \frac{n\hat{\sigma}_1^2}{\hat{\sigma}_1^2} = n, \quad \text{and} \quad T_3 = \sum_{i=1}^{n} \left( \frac{y_i - \hat{\mu}_2}{\hat{\sigma}_2} \right)^2 = \frac{n\hat{\sigma}_2^2}{\hat{\sigma}_2^2} = n,$$

and

$$T_2 = \sum_{i=1}^{n} \left( \frac{x_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \left( \frac{y_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) = n\hat{\rho}.$$

Consider testing $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$. The (restricted) MLEs for $\mu_1, \mu_2, \sigma_1$ and $\sigma_2$ do not change under $H_0$, and hence are still equal to $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}_1$, and $\hat{\sigma}_2$.

a) Using the above information, find the likelihood ratio test for $H_0 : \rho = 0$ vs. $H_A : \rho \neq 0$. Denote the likelihood ratio test statistic by $\lambda(x, y)$.

b) Find the large sample (asymptotic) likelihood ratio test that uses test statistic $-2\log(\lambda(x, y))$. 
Chapter 8
Large Sample Theory

8.1 The CLT, Delta Method and an Exponential Family Limit Theorem

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size $n$ is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large $n$ must be for the approximation to be useful. Moreover, the (asymptotic) standard error (SE), an estimator of the asymptotic standard deviation, must be computable if the estimator is to be useful for inference.

**Theorem 8.1: the Central Limit Theorem (CLT).** Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \overset{D}{\to} N(0, \sigma^2).$$

Hence

$$\sqrt{n} \left( \frac{\overline{Y}_n - \mu}{\sigma} \right) = \sqrt{n} \left( \frac{\sum_{i=1}^{n} Y_i - n\mu}{n\sigma} \right) \overset{D}{\to} N(0, 1).$$

Note that the sample mean is estimating the population mean $\mu$ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal, and the SE $= S/\sqrt{n}$ where $S$ is the sample standard deviation. For many distributions
the central limit theorem provides a good approximation if the sample size \( n > 30 \). A special case of the CLT is proven at the end of Section 8.4.

**Notation.** The notation \( X \sim Y \) and \( X \overset{D}{=} Y \) both mean that the random variables \( X \) and \( Y \) have the same distribution. See Definition 1.24. The notation \( Y_n \overset{D}{\to} X \) means that for large \( n \) we can approximate the cdf of \( Y_n \) by the cdf of \( X \). The distribution of \( X \) is the limiting distribution or asymptotic distribution of \( Y_n \). For the CLT, notice that

\[
Z_n = \sqrt{n} \left( \frac{\bar{Y}_n - \mu}{\sigma} \right) = \left( \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}} \right)
\]

is the z-score of \( \bar{Y} \). If \( Z_n \overset{D}{\to} N(0,1) \), then the notation \( Z_n \approx N(0,1) \), also written as \( Z_n \sim AN(0,1) \), means approximate the cdf of \( Z_n \) by the standard normal cdf. Similarly, the notation

\[
\bar{Y}_n \approx N(\mu,\sigma^2/n),
\]

also written as \( \bar{Y}_n \sim AN(\mu,\sigma^2/n) \), means approximate the cdf of \( \bar{Y}_n \) as if \( \bar{Y}_n \sim N(\mu,\sigma^2/n) \).

The two main applications of the CLT are to give the limiting distribution of \( \sqrt{n}(\bar{Y}_n - \mu) \) and the limiting distribution of \( \sqrt{n}(Y_n/n - \mu_X) \) for a random variable \( Y_n \) such that \( Y_n = \sum_{i=1}^{n} X_i \) where the \( X_i \) are iid with \( E(X) = \mu_X \) and \( \text{VAR}(X) = \sigma_X^2 \). Several of the random variables in Theorems 2.17 and 2.18 can be approximated in this way.

**Example 8.1.** a) Let \( Y_1, \ldots, Y_n \) be iid \( \text{Ber}(\rho) \). Then \( E(Y) = \rho \) and \( \text{VAR}(Y) = \rho(1 - \rho) \). Hence

\[
\sqrt{n}(\bar{Y}_n - \rho) \overset{D}{\to} N(0, \rho(1 - \rho))
\]

by the CLT.

b) Now suppose that \( Y_n \sim BIN(n, \rho) \). Then \( Y_n \overset{D}{=} \sum_{i=1}^{n} X_i \) where \( X_1, \ldots, X_n \) are iid \( \text{Ber}(\rho) \). Hence

\[
\sqrt{n}(\frac{Y_n}{n} - \rho) \overset{D}{\to} N(0, \rho(1 - \rho))
\]

since

\[
\sqrt{n}(\frac{Y_n}{n} - \rho) \overset{D}{=} \sqrt{n}(\bar{X}_n - \rho) \overset{D}{\to} N(0, \rho(1 - \rho))
\]
by a).

(c) Now suppose that $Y_n \sim \text{BIN}(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}(Y_n/k_n - \rho) \approx N(0, \rho(1 - \rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1 - \rho)}{k_n}\right) \text{ or } Y_n \approx N(k_n\rho, k_n\rho(1 - \rho)).$$

**Theorem 8.2: the Delta Method.** If $g'(\theta) \neq 0$ and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

**Example 8.2.** Let $Y_1, \ldots, Y_n$ be iid with $E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2 \mu^2)$$

for $\mu \neq 0$ by the delta method.

**Example 8.3.** Let $X \sim \text{Binomial}(n, p)$ where the positive integer $n$ is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n} \left[ \left( \frac{X}{n} \right)^2 - p^2 \right]$.

Solution. Example 8.1b gives the limiting distribution of $\sqrt{n}(\overline{X}_n - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n} \left[ \left( \frac{X}{n} \right)^2 - p^2 \right] = \sqrt{n} \left( g\left( \frac{X}{n} \right) - g(p) \right) \xrightarrow{D} N(0, p(1 - p)(g'(p))^2) = N(0, 4p^3(1 - p)).$$

**Example 8.4.** Let $X_n \sim \text{Poisson}(n\lambda)$ where the positive integer $n$ is large and $0 < \lambda$. 


a) Find the limiting distribution of $\sqrt{n} \left( \frac{X_n}{n} - \lambda \right)$.

b) Find the limiting distribution of $\sqrt{n} \left[ \sqrt{\frac{X_n}{n}} - \sqrt{\lambda} \right]$.

Solution. a) $X_n \overset{D}{=} \sum_{i=1}^{n} Y_i$ where the $Y_i$ are iid Poisson($\lambda$). Hence $E(Y) = \lambda = Var(Y)$. Thus by the CLT,

$$\sqrt{n} \left( \frac{X_n}{n} - \lambda \right) \overset{D}{=} \sqrt{n} \left( \frac{\sum_{i=1}^{n} Y_i}{n} - \lambda \right) \overset{D}{\rightarrow} N(0, \lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n} \left[ \sqrt{\frac{X_n}{n}} - \sqrt{\lambda} \right] = \sqrt{n} \left( g\left( \frac{X_n}{n} \right) - g(\lambda) \right) \overset{D}{\rightarrow} \mathcal{N}(0, \lambda \left( g'(\lambda) \right)^2) = \mathcal{N}(0, \lambda \frac{1}{4\lambda}) = \mathcal{N}(0, \frac{1}{4}).$$

Example 8.5. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma($\alpha, \beta$) distribution.

a) Find the limiting distribution of $\sqrt{n} \left( \bar{Y} - \alpha \beta \right)$.

b) Find the limiting distribution of $\sqrt{n} \left( \bar{Y}^2 - c \right)$ for appropriate constant $c$.

Solution: a) Since $E(Y) = \alpha \beta$ and $V(Y) = \alpha \beta^2$, by the CLT

$$\sqrt{n} \left( \bar{Y} - \alpha \beta \right) \overset{D}{\rightarrow} N(0, \alpha \beta^2).$$

b) Let $\mu = \alpha \beta$ and $\sigma^2 = \alpha \beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2 \beta^2$. Then by the delta method, $\sqrt{n} \left( \bar{Y}^2 - c \right) \overset{D}{\rightarrow} \mathcal{N}(0, \sigma^2 [g'(\mu)]^2) = \mathcal{N}(0, 4\alpha^3 \beta^4)$ where $c = \mu^2 = \alpha^2 \beta^2$.

Barndorff-Nielsen (1982), Casella and Berger (2002, p. 472, 515), Cox and Hinckley (1974, p. 286), Lehmann and Casella (1998, Section 6.3), Schervish (1995, p. 418), and many others suggest that under regularity conditions if $Y_1, ..., Y_n$ are iid from a one parameter regular exponential family, and if $\hat{\theta}$ is the MLE of $\theta$, then

$$\sqrt{n}(\hat{\tau}(\hat{\theta}) - \tau(\theta)) \overset{D}{\rightarrow} \mathcal{N} \left( 0, \frac{[\tau'(\theta)]^2}{I_1(\theta)} \right) = \mathcal{N}[0, FCRLB_1(\tau(\theta))]$$ (8.1)
where the Fréchet Cramér Rao lower bound for $\tau(\theta)$ is

$$FCRLB_1(\tau(\theta)) = \frac{[\tau'(\theta)]^2}{I_1(\theta)}$$

and the Fisher information based on a sample of size one is

$$I_1(\theta) = -E_\theta[\frac{\partial^2}{\partial \theta^2} \log(f(X|\theta))].$$

Notice that if

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \frac{1}{I_1(\theta)}\right),$$

then (8.1) follows by the delta method. Also recall that $\tau(\hat{\theta})$ is the MLE of $\tau(\theta)$ by the invariance principle and that

$$I_1(\tau(\theta)) = \frac{I_1(\theta)}{[\tau'(\theta)]^2}$$

if $\tau'(\theta) \neq 0$ by Definition 6.3.

For a 1P–REF, $T_n = \frac{1}{n} \sum_{i=1}^{n} t(Y_i)$ is the UMVUE and generally the MLE of its expectation $\mu_t \equiv \mu_T = E_\theta(T_n) = E_\theta[t(Y)]$. Let $\sigma_t^2 = \text{VAR}_\theta[t(Y)]$. These values can be found by using the distribution of $t(Y)$ (see Theorems 3.6 and 3.7) or by the following result.

**Proposition 8.3.** Suppose $Y$ is a 1P–REF with pdf or pmf

$$f(y|\theta) = h(y)c(\theta) \exp[w(\theta)t(y)]$$

and natural parameterization

$$f(y|\eta) = h(y)b(\eta) \exp[\eta t(y)].$$

Then a)

$$\mu_t = E[t(Y)] = \frac{-c'(\theta)}{c(\theta)w'(\theta)} = \frac{-\partial}{\partial \eta} \log(b(\eta)), \quad \text{(8.2)}$$

and b)

$$\sigma_t^2 = V[t(Y)] = \frac{-\partial^2}{\partial \theta^2} \log(c(\theta)) - \frac{[w''(\theta)]\mu_t}{[w'(\theta)]^2} = \frac{-\partial^2}{\partial \eta^2} \log(b(\eta)). \quad \text{(8.3)}$$
Proof. The proof will be for pdfs. For pmfs replace the integrals by sums. By Theorem 3.3, only the middle equalities need to be shown. By Remark 3.2 the derivative and integral operators can be interchanged for a 1P–REF. a) Since $1 = \int f(y|\theta)dy$, 

$$0 = \frac{\partial}{\partial \theta} 1 = \frac{\partial}{\partial \theta} \int h(y) \exp[w(\theta)t(y) + \log(c(\theta))]dy$$

$$= \int h(y) \frac{\partial}{\partial \theta} \exp[w(\theta)t(y) + \log(c(\theta))]dy$$

$$= \int h(y) \exp[w(\theta)t(y) + \log(c(\theta))] \left( w'(\theta)t(y) + \frac{c'(\theta)}{c(\theta)} \right) dy$$

or

$$E[w'(\theta)t(Y)] = -\frac{c'(\theta)}{c(\theta)}$$

or

$$E[t(Y)] = -\frac{c'(\theta)}{c(\theta)w'(\theta)}.$$

b) Similarly,

$$0 = \int h(y) \frac{\partial^2}{\partial \theta^2} \exp[w(\theta)t(y) + \log(c(\theta))]dy.$$ 

From the proof of a) and since $\frac{\partial}{\partial \theta} \log(c(\theta)) = c'(\theta)/c(\theta)$,

$$0 = \int h(y) \frac{\partial}{\partial \theta} \left[ \exp[w(\theta)t(y) + \log(c(\theta))] \left( w'(\theta)t(y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right) \right] dy$$

$$= \int h(y) \exp[w(\theta)t(y) + \log(c(\theta))] \left( w'(\theta)t(y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right)^2 dy$$

$$+ \int h(y) \exp[w(\theta)t(y) + \log(c(\theta))] \left( w''(\theta)t(y) + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right) dy.$$ 

So

$$E \left( w'(\theta)t(Y) + \frac{\partial}{\partial \theta} \log(c(\theta)) \right)^2 = -E \left( w''(\theta)t(Y) + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right). \quad (8.4)$$
Using a) shows that the left hand side of (8.4) equals
\[ E \left( w'(\theta) \left( t(Y) + \frac{c'(\theta)}{c(\theta)w'(\theta)} \right) \right)^2 = [w'(\theta)]^2 \text{VAR}(t(Y)) \]
while the right hand side of (8.4) equals
\[ - \left( w''(\theta) \mu_t + \frac{\partial^2}{\partial \theta^2} \log(c(\theta)) \right) \]
and the result follows. QED

The simplicity of the following result is rather surprising. When (as is usually the case) \( \frac{1}{n} \sum_{i=1}^{n} t(Y_i) \) is the MLE of \( \mu_t \), \( \hat{\eta} = g^{-1}(\frac{1}{n} \sum_{i=1}^{n} t(Y_i)) \) is the MLE of \( \eta \) by the invariance principle.

**Theorem 8.4.** Let \( Y_1, \ldots, Y_n \) be iid from a 1P–REF with pdf or pmf
\[ f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)] \]
and natural parameterization
\[ f(y|\eta) = h(y)b(\eta)\exp[\eta t(y)]. \]
Let
\[ E(t(Y)) = \mu_t \equiv g(\eta) \]
and \( \text{VAR}(t(Y)) = \sigma_t^2 \).

a) Then
\[ \sqrt{n} \left[ \frac{1}{n} \sum_{i=1}^{n} t(Y_i) - \mu_t \right] \xrightarrow{D} N(0, I_1(\eta)) \]
where
\[ I_1(\eta) = \sigma_t^2 = g'(\eta) = \frac{[g'(\eta)]^2}{I_1(\eta)}. \]

b) If \( \eta = g^{-1}(\mu_t) \), \( \hat{\eta} = g^{-1}(\frac{1}{n} \sum_{i=1}^{n} t(Y_i)) \), and \( g^{-1}(\mu_t) \neq 0 \) exists, then
\[ \sqrt{n} [\hat{\eta} - \eta] \xrightarrow{D} N \left( 0, \frac{1}{I_1(\eta)} \right). \]
c) Suppose the conditions in b) hold. If \( \theta = w^{-1}(\eta), \hat{\theta} = w^{-1}(\hat{\eta}), w^{-1}' \) exists and is continuous, and \( w^{-1}(\eta) \neq 0 \), then

\[
\sqrt{n}[\hat{\theta} - \theta] \xrightarrow{D} N \left( 0, \frac{1}{I_1(\theta)} \right).
\]

\[
\sqrt{n}[\tau(\hat{\theta}) - \tau(\theta)] \xrightarrow{D} N \left( 0, \frac{[\tau'(\theta)]^2}{I_1(\theta)} \right).
\]

Proof: a) The result follows by the central limit theorem if \( \sigma_t^2 = I_1(\eta) = g'(\eta) \). Since \( \log(f(y|\eta)) = \log(h(y)) + \log(b(\eta)) + \eta t(y) \),

\[
\frac{\partial}{\partial \eta} \log(f(y|\eta)) = \frac{\partial}{\partial \eta} \log(b(\eta)) + t(y) = -\mu_t + t(y) = -g(\eta) + t(y)
\]

by Proposition 8.3 a). Hence

\[
\frac{\partial^2}{\partial \eta^2} \log(f(y|\eta)) = \frac{\partial^2}{\partial \eta^2} \log(b(\eta)) = -g'(\eta),
\]

and thus by Proposition 8.3 b)

\[
I_1(\eta) = -\frac{\partial^2}{\partial \eta^2} \log(b(\eta)) = \sigma_t^2 = g'(\eta).
\]

b) By the delta method,

\[
\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N(0, \sigma_t^2[g^{-1}'(\mu_t)]^2),
\]

but

\[
g^{-1}'(\mu_t) = \frac{1}{g'(g^{-1}(\mu_t))} = \frac{1}{g'(\eta)}.
\]

Since \( \sigma_t^2 = I_1(\eta) = g'(\eta) \), it follows that \( \sigma_t^2 = [g'(\eta)]^2/I_1(\eta) \), and

\[
\sigma_t^2[g^{-1}'(\mu_t)]^2 = \frac{[g'(\eta)]^2}{I_1(\eta)} \cdot \frac{1}{[g'(\eta)]^2} = \frac{1}{I_1(\eta)}.
\]

So

\[
\sqrt{n}(\hat{\eta} - \eta) \xrightarrow{D} N \left( 0, \frac{1}{I_1(\eta)} \right).
\]
CHAPTER 8. LARGE SAMPLE THEORY

236

c) By the delta method,

\[ \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N\left(0, \frac{[w^{-1'}(\eta)]^2}{I_1(\eta)}\right), \]

but

\[ \frac{[w^{-1'}(\eta)]^2}{I_1(\eta)} = \frac{1}{I_1(\theta)}. \]

The last equality holds since by Theorem 6.3c, if \( \theta = g(\eta) \), if \( g'(\theta) \) exists and is continuous, and if \( g'(\theta) \neq 0 \), then \( I_1(\theta) = I_1(\eta)/[g'(\eta)]^2 \). Use \( \eta = w(\theta) \) so \( \theta = g(\eta) = w^{-1}(\eta) \).

d) The result follows by the delta method. QED

Remark 8.1. Following DasGupta (2008, p. 241-242), let \( \psi(\eta) = -\log(b(\eta)) \). Then \( E_\eta[t(Y_1)] = \mu_t = \psi'(\eta) = g(\eta) \) by Proposition 8.3a, and the MLE \( \hat{\eta} \) is the solution of \( \frac{1}{n} \sum_{i=1}^{n} t(Y_i) \equiv E_\eta[t(Y_1)] = g(\eta) \) by Theorem 5.2, if the MLE exists. Now \( g(\eta) = E_\eta[t(Y_1)] \) is an increasing function of \( \eta \) since \( g'(\eta) = \psi''(\eta) = V_\eta(t(Y)) > 0 \) (1PREFs do not contain degenerate distributions). So for large \( n \), with probability tending to one, the MLE \( \hat{\eta} \) exists and \( \hat{\eta} = g^{-1}(\frac{1}{n} \sum_{i=1}^{n} t(Y_i)) \). Since \( g'(\eta) \) exists, \( g(\eta) \) and \( g^{-1}(\eta) \) are continuous and the delta method can be applied to \( \hat{\eta} \) as in Theorem 8.4b. By the proof of Theorem 8.4a, \( \psi''(\eta) = I_1(\eta) \). Notice that if \( \hat{\eta} \) is the MLE of \( \eta \), then \( \frac{1}{n} \sum_{i=1}^{n} t(Y_i) \) is the MLE of \( \mu_t = E[t(Y_1)] \) by invariance. Hence if \( n \) is large enough, Theorem 8.4ab is for the MLE of \( E[t(Y_1)] \) and the MLE of \( \eta \).

8.2 Asymptotically Efficient Estimators

Definition 8.1. Let \( Y_1, ..., Y_n \) be iid random variables. Let \( T_n \equiv T_n(Y_1, ..., Y_n) \) be an estimator of a parameter \( \mu_T \) such that

\[ \sqrt{n}(T_n - \mu_T) \xrightarrow{D} N(0, \sigma_A^2). \]

Then the asymptotic variance of \( \sqrt{n}(T_n - \mu_T) \) is \( \sigma_A^2 \) and the asymptotic variance (AV) of \( T_n \) is \( \sigma_A^2/n \). If \( S_A^2 \) is a consistent estimator of \( \sigma_A^2 \), then the (asymptotic) standard error (SE) of \( T_n \) is \( S_A/\sqrt{n} \). If \( Y_1, ..., Y_n \) are iid with cdf \( F \), then \( \sigma_A^2 \equiv \sigma_A^2(F) \) depends on \( F \).

Remark 8.2. Consistent estimators are defined in the following section. The parameter \( \sigma_A^2 \) is a function of both the estimator \( T_n \) and the underlying
distribution $F$ of $Y_1$. Frequently $n \text{VAR}(T_n)$ converges in distribution to $\sigma^2_A$, but not always. See Staudte and Sheather (1990, p. 51) and Lehmann (1999, p. 232).

**Example 8.6.** If $Y_1, \ldots, Y_n$ are iid from a distribution with mean $\mu$ and variance $\sigma^2$, then by the central limit theorem,

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Recall that $\text{VAR}(\bar{Y}_n) = \sigma^2/n = AV(\bar{Y}_n)$ and that the standard error $SE(\bar{Y}_n) = S_n/\sqrt{n}$ where $S^2_n$ is the sample variance. Note that $\sigma^2_A(F) = \sigma^2$. If $F$ is a $N(\mu, 1)$ cdf then $\sigma^2_A(F) = 1$, but if $F$ is the $G(\nu = 7, \lambda = 1)$ cdf then $\sigma^2_A(F) = 7$.

**Definition 8.2.** Let $T_{1,n}$ and $T_{2,n}$ be two estimators of a parameter $\theta$ such that

$$n^{\delta}(T_{1,n} - \theta) \xrightarrow{D} N(0, \sigma^2_A(F))$$

and

$$n^{\delta}(T_{2,n} - \theta) \xrightarrow{D} N(0, \sigma^2_B(F)),$$

then the **asymptotic relative efficiency** of $T_{1,n}$ with respect to $T_{2,n}$ is

$$\text{ARE}(T_{1,n}, T_{2,n}) = \frac{\sigma^2_B(F)}{\sigma^2_A(F)}.$$ 

This definition brings up several issues. First, both estimators must have the same convergence rate $n^{\delta}$. Usually $\delta = 0.5$. If $T_{i,n}$ has convergence rate $n^{\delta_i}$, then estimator $T_{1,n}$ is judged to be “better” than $T_{2,n}$ if $\delta_1 > \delta_2$. Secondly, the two estimators need to estimate the same parameter $\theta$. This condition will often not hold unless the distribution is symmetric about $\mu$. Then $\theta = \mu$ is a natural choice. Thirdly, estimators are often judged by their Gaussian efficiency with respect to the sample mean (thus $F$ is the normal distribution). Since the normal distribution is a location–scale family, it is often enough to compute the ARE for the standard normal distribution. If the data come from a distribution $F$ and the ARE can be computed, then $T_{1,n}$ is judged to be a “better” estimator (for the data distribution $F$) than $T_{2,n}$ if the $\text{ARE} > 1$. Similarly, $T_{1,n}$ is judged to be a “worse” estimator than $T_{2,n}$ if the $\text{ARE} < 1$. **Notice that the “better” estimator has the smaller asymptotic variance.**
The population median is any value \( \text{MED}(Y) \) such that
\[
P(Y \leq \text{MED}(Y)) \geq 0.5 \quad \text{and} \quad P(Y \geq \text{MED}(Y)) \geq 0.5.
\] (8.5)

In simulation studies, typically the underlying distribution \( F \) belongs to a symmetric location–scale family. There are at least two reasons for using such distributions. First, if the distribution is symmetric, then the population median \( \text{MED}(Y) \) is the point of symmetry and the natural parameter to estimate. Under the symmetry assumption, there are many estimators of \( \text{MED}(Y) \) that can be compared via their ARE with respect to the sample mean or the maximum likelihood estimator (MLE). Secondly, once the ARE is obtained for one member of the family, it is typically obtained for all members of the location–scale family. That is, suppose that \( Y_1, \ldots, Y_n \) are iid from a location–scale family with parameters \( \mu \) and \( \sigma \). Then \( Y_i = \mu + \sigma Z_i \) where the \( Z_i \) are iid from the same family with \( \mu = 0 \) and \( \sigma = 1 \). Typically
\[
\text{AV}[T_{i,n}(Y)] = \sigma^2 \text{AV}[T_{i,n}(Z)],
\]
so
\[
\text{ARE}[T_{1,n}(Y), T_{2,n}(Y)] = \text{ARE}[T_{1,n}(Z), T_{2,n}(Z)].
\]

**Theorem 8.5.** Let \( Y_1, \ldots, Y_n \) be iid with a pdf \( f \) that is positive at the population median: \( f(\text{MED}(Y)) > 0 \). Then
\[
\sqrt{n}(\text{MED}(n) - \text{MED}(Y)) \xrightarrow{D} N\left(0, \frac{1}{4[f(\text{MED}(Y))]^2}\right).
\]

**Example 8.7.** Let \( Y_1, \ldots, Y_n \) be iid \( N(\mu, \sigma^2) \), \( T_{1,n} = \overline{Y} \) and let \( T_{2,n} = \text{MED}(n) \) be the sample median. Let \( \theta = \mu = E(Y) = \text{MED}(Y) \). Find \( \text{ARE}(T_{1,n}, T_{2,n}) \).

Solution: By the CLT, \( \sigma_1^2(F) = \sigma^2 \) when \( F \) is the \( N(\mu, \sigma^2) \) distribution. By Theorem 8.5,
\[
\sigma_2^2(F) = \frac{1}{4[f(\text{MED}(Y))]^2} = \frac{1}{4[1/\sqrt{2\pi\sigma^2}]^2} = \frac{\pi\sigma^2}{2}.
\]
Hence
\[
\text{ARE}(T_{1,n}, T_{2,n}) = \frac{\pi\sigma^2/2}{\sigma^2} = \frac{\pi}{2} \approx 1.571.
\]
and the sample mean $\overline{Y}$ is a “better” estimator of $\mu$ than the sample median MED($n$) for the family of normal distributions.

Recall from Definition 6.3 that $I_1(\theta)$ is the information number for $\theta$ based on a sample of size 1. Also recall that $I_1(\tau(\theta)) = I_1(\theta)/[\tau'(\theta)]^2 = 1/FCRLB_1[\tau(\theta)]$. See Definition 6.4.

**Definition 8.3.** Assume $\tau'(\theta) \neq 0$. Then an estimator $T_n$ of $\tau(\theta)$ is asymptotically efficient if

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N \left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right) \sim N(0, FCRLB_1[\tau(\theta)]).$$

In particular, the estimator $T_n$ of $\theta$ is asymptotically efficient if

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N \left(0, \frac{1}{I_1(\theta)}\right) \sim N(0, FCRLB_1[\theta]).$$

Following Lehmann (1999, p. 486), if $T_{2,n}$ is an asymptotically efficient estimator of $\theta$, if $I_1(\theta)$ and $v(\theta)$ are continuous functions, and if $T_{1,n}$ is an estimator such that

$$\sqrt{n}(T_{1,n} - \theta) \xrightarrow{D} N(0, v(\theta)),$$

then under regularity conditions, $v(\theta) \geq 1/I_1(\theta)$ and

$$ARE(T_{1,n}, T_{2,n}) = \frac{I_1(\theta)}{v(\theta)} = \frac{1}{I_1(\theta) v(\theta)} \leq 1.$$

Hence asymptotically efficient estimators are “better” than estimators of the form $T_{1,n}$. When $T_{2,n}$ is asymptotically efficient,

$$AE(T_{1,n}) = ARE(T_{1,n}, T_{2,n}) = \frac{1}{I_1(\theta) v(\theta)}$$

is sometimes called the asymptotic efficiency of $T_{1,n}$.

Notice that for a 1P–REF, $\overline{T}_n = \frac{1}{n} \sum_{i=1}^{n} t(Y_i)$ is an asymptotically efficient estimator of $g(\eta) = E(t(Y))$ by Theorem 8.4. $\overline{T}_n$ is the UMVUE of $E(t(Y))$ by the LSU theorem.

The following rule of thumb suggests that MLEs and UMVUEs are often asymptotically efficient. The rule often holds for location families where the
support does not depend on $\theta$. The rule does not hold for the uniform $(0, \theta)$ family.

Rule of Thumb 8.1: a “Standard Limit Theorem”: Let $\hat{\theta}_n$ be the MLE or UMVUE of $\theta$. If $\tau'(\theta) \neq 0$, then under strong regularity conditions,
\[
\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{D} N\left(0, \frac{[\tau'(\theta)]^2}{I_1(\theta)}\right).
\]

8.3 Modes of Convergence and Consistency

Definition 8.4. Let $\{Z_n, n = 1, 2, \ldots\}$ be a sequence of random variables with cdfs $F_n$, and let $X$ be a random variable with cdf $F$. Then $Z_n$ converges in distribution to $X$, written
\[ Z_n \xrightarrow{D} X, \]
or $Z_n$ converges in law to $X$, written $Z_n \xrightarrow{L} X$, if
\[ \lim_{n \to \infty} F_n(t) = F(t) \]
at each continuity point $t$ of $F$. The distribution of $X$ is called the limiting distribution or the asymptotic distribution of $Z_n$.

An important fact is that the limiting distribution does not depend on the sample size $n$. Notice that the CLT, delta method and Theorem 8.4 give the limiting distributions of $Z_n = \sqrt{n}[\overline{Y}_n - \mu]$, $Z_n = \sqrt{n}(g(T_n) - g(\theta))$ and $Z_n = \sqrt{n}\left[\frac{1}{n} \sum_{i=1}^{n} t(Y_i) - E(t(Y))\right]$, respectively.

Convergence in distribution is useful because if the distribution of $X_n$ is unknown or complicated and the distribution of $X$ is easy to use, then for large $n$ we can approximate the probability that $X_n$ is in an interval by the probability that $X$ is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if $F$ is continuous at $a$ and $b$.

Warning: convergence in distribution says that the cdf $F_n(t)$ of $X_n$ gets close to the cdf of $F(t)$ of $X$ as $n \to \infty$ provided that $t$ is a continuity point of $F$. Hence for any $\epsilon > 0$ there exists $N_t$ such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that $N_t$ depends on the value of $t$. Convergence in
distribution does not imply that the random variables \( X_n \equiv X_n(\omega) \) converge to the random variable \( X \equiv X(\omega) \) for all \( \omega \).

**Example 8.8.** Suppose that \( X_n \sim U(-1/n, 1/n) \). Then the cdf \( F_n(x) \) of \( X_n \) is

\[
F_n(x) = \begin{cases} 
0, & x \leq -\frac{1}{n} \\
\frac{x}{2} + \frac{1}{2}, & -\frac{1}{n} \leq x \leq \frac{1}{n} \\
1, & x \geq \frac{1}{n}.
\end{cases}
\]

Sketching \( F_n(x) \) shows that it has a line segment rising from 0 at \( x = -1/n \) to 1 at \( x = 1/n \) and that \( F_n(0) = 0.5 \) for all \( n \geq 1 \). Examining the cases \( x < 0, x = 0 \) and \( x > 0 \) shows that as \( n \to \infty \),

\[
F_n(x) \to \begin{cases} 
0, & x < 0 \\
\frac{1}{2}, & x = 0 \\
1, & x > 0.
\end{cases}
\]

Notice that if \( X \) is a random variable such that \( P(X = 0) = 1 \), then \( X \) has cdf

\[
F_X(x) = \begin{cases} 
0, & x < 0 \\
1, & x \geq 0.
\end{cases}
\]

Since \( x = 0 \) is the only discontinuity point of \( F_X(x) \) and since \( F_n(x) \to F_X(x) \) for all continuity points of \( F_X(x) \) (ie for \( x \neq 0 \)),

\( X_n \overset{D}{\to} X \).

**Example 8.9.** Suppose \( Y_n \sim U(0, n) \). Then \( F_n(t) = t/n \) for \( 0 < t \leq n \) and \( F_n(t) = 0 \) for \( t \leq 0 \). Hence \( \lim_{n \to \infty} F_n(t) = 0 \) for \( t \leq 0 \). If \( t > 0 \) and \( n > t \), then \( F_n(t) = t/n \to 0 \) as \( n \to \infty \). Thus \( \lim_{n \to \infty} F_n(t) = 0 \) for all \( t \) and \( Y_n \) does not converge in distribution to any random variable \( Y \) since \( H(t) \equiv 0 \) is not a cdf.

**Definition 8.5.** A sequence of random variables \( X_n \) converges in distribution to a constant \( \tau(\theta) \), written

\[
X_n \overset{D}{\to} \tau(\theta), \quad \text{if} \quad X_n \overset{D}{\to} X
\]

where \( P(X = \tau(\theta)) = 1 \). The distribution of the random variable \( X \) is said to be degenerate at \( \tau(\theta) \).
Definition 8.6. A sequence of random variables $X_n$ converges in probability to a constant $\tau(\theta)$, written

$$X_n \overset{P}{\to} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{n \to \infty} P(|X_n - \tau(\theta)| \geq \epsilon) = 0.$$

The sequence $X_n$ converges in probability to $X$, written

$$X_n \overset{P}{\to} X,$$

if $X_n - X \overset{P}{\to} 0$.

Notice that $X_n \overset{P}{\to} X$ if for every $\epsilon > 0$,

$$\lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{n \to \infty} P(|X_n - X| \geq \epsilon) = 0.$$

Definition 8.7. A sequence of estimators $T_n$ of $\tau(\theta)$ is consistent for $\tau(\theta)$ if

$$T_n \overset{P}{\to} \tau(\theta)$$

for every $\theta \in \Theta$. If $T_n$ is consistent for $\tau(\theta)$, then $T_n$ is a consistent estimator of $\tau(\theta)$.

Consistency is a weak property that is usually satisfied by good estimators. $T_n$ is a consistent estimator for $\tau(\theta)$ if the probability that $T_n$ falls in any neighborhood of $\tau(\theta)$ goes to one, regardless of the value of $\theta \in \Theta$.

Definition 8.8. For a real number $r > 0$, $Y_n$ converges in $r$th mean to a random variable $Y$, written

$$Y_n \overset{r}{\to} Y,$$

if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, $Y_n$ converges in quadratic mean to $Y$, written

$$Y_n \overset{qm}{\to} Y \quad \text{or} \quad Y_n \overset{qm}{\to} Y,$$
if

\[ E[(Y_n - Y)^2] \to 0 \]

as \( n \to \infty \).

**Lemma 8.6: Generalized Chebyshev’s Inequality.** Let \( u : \mathbb{R} \to [0, \infty) \) be a nonnegative function. If \( E[u(Y)] \) exists then for any \( c > 0 \),

\[ P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}. \]

If \( \mu = E(Y) \) exists, then taking \( u(y) = |y - \mu|^r \) and \( \bar{c} = c^r \) gives

**Markov’s Inequality:** for \( r > 0 \) and any \( c > 0 \),

\[ P(|Y - \mu| \geq c] = P(|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}. \]

If \( r = 2 \) and \( \sigma^2 = \text{VAR}(Y) \) exists, then we obtain

**Chebyshev’s Inequality:**

\[ P(|Y - \mu| \geq c] \leq \frac{\text{VAR}(Y)}{c^2}. \]

**Proof.** The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

\[ E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y : u(y) \geq c\}} u(y)f(y)dy + \int_{\{y : u(y) < c\}} u(y)f(y)dy \geq \int_{\{y : u(y) \geq c\}} u(y)f(y)dy \]

since the integrand \( u(y)f(y) \geq 0 \). Hence

\[ E[u(Y)] \geq c \int_{\{y : u(y) \geq c\}} f(y)dy = cP[u(Y) \geq c]. \quad QED \]

The following proposition gives sufficient conditions for \( T_n \) to be a consistent estimator of \( \tau(\theta) \). Notice that \( MSE_{\tau(\theta)}(T_n) \to 0 \) for all \( \theta \in \Theta \) is equivalent to \( T_n \xrightarrow{qm} \tau(\theta) \) for all \( \theta \in \Theta \).
Proposition 8.7. a) If
\[ \lim_{n \to \infty} MSE_{\tau(\theta)}(T_n) = 0 \]
for all \( \theta \in \Theta \), then \( T_n \) is a consistent estimator of \( \tau(\theta) \).

b) If
\[ \lim_{n \to \infty} \text{VAR}_\theta(T_n) = 0 \quad \text{and} \quad \lim_{n \to \infty} E_\theta(T_n) = \tau(\theta) \]
for all \( \theta \in \Theta \), then \( T_n \) is a consistent estimator of \( \tau(\theta) \).

Proof. a) Using Lemma 8.6 with \( Y = T_n \), \( u(T_n) = (T_n - \tau(\theta))^2 \) and \( c = \epsilon^2 \) shows that for any \( \epsilon > 0 \),
\[ P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta((T_n - \tau(\theta))^2 \geq \epsilon^2) \leq \frac{E_\theta((T_n - \tau(\theta))^2)}{\epsilon^2}. \]
Hence
\[ \lim_{n \to \infty} E_\theta((T_n - \tau(\theta))^2) = \lim_{n \to \infty} MSE_{\tau(\theta)}(T_n) \to 0 \]
is a sufficient condition for \( T_n \) to be a consistent estimator of \( \tau(\theta) \).

b) Referring to Definition 6.1,
\[ MSE_{\tau(\theta)}(T_n) = \text{VAR}_\theta(T_n) + [\text{Bias}_{\tau(\theta)}(T_n)]^2 \]
where \( \text{Bias}_{\tau(\theta)}(T_n) = E_\theta(T_n) - \tau(\theta) \). Since \( MSE_{\tau(\theta)}(T_n) \to 0 \) if both \( \text{VAR}_\theta(T_n) \to 0 \) and \( \text{Bias}_{\tau(\theta)}(T_n) = E_\theta(T_n) - \tau(\theta) \to 0 \), the result follows from a). QED

The following result shows estimators that converge at a \( \sqrt{n} \) rate are consistent. Use this result and the delta method to show that \( g(T_n) \) is a consistent estimator of \( g(\theta) \). Note that b) follows from a) with \( X_\theta \sim N(0, v(\theta)) \). The WLLN shows that \( \bar{Y} \) is a consistent estimator of \( E(Y) = \mu \) if \( E(Y) \) exists.

Proposition 8.8. a) Let \( X \) be a random variable and \( 0 < \delta \leq 1 \). If
\[ n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X \]
then \( T_n \xrightarrow{P} \tau(\theta) \).

b) If
\[ \sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta)) \]
for all $\theta \in \Theta$, then $T_n$ is a consistent estimator of $\tau(\theta)$.

**Definition 8.9.** A sequence of random variables $X_n$ converges almost everywhere (or almost surely, or with probability 1) to $X$ if

$$P(\lim_{n \to \infty} X_n = X) = 1.$$ 

This type of convergence will be denoted by

$$X_n \xrightarrow{ae} X.$$ 

Notation such as “$X_n$ converges to $X$ ae” will also be used. Sometimes “ae” will be replaced with “as” or “wp1.” We say that $X_n$ converges almost everywhere to $\tau(\theta)$, written

$$X_n \xrightarrow{ae} \tau(\theta),$$

if $P(\lim_{n \to \infty} X_n = \tau(\theta)) = 1$.

**Theorem 8.9.** Let $Y_n$ be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\bar{Y}_n \xrightarrow{ae} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\bar{Y}_n \xrightarrow{p} \mu$.

**Proof of WLLN when $V(Y_i) = \sigma^2$:** By Chebyshev’s inequality, for every $\epsilon > 0$,

$$P(|\bar{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\bar{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. QED

### 8.4 Slutsky’s Theorem and Related Results

**Theorem 8.10: Slutsky’s Theorem.** Suppose $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{p} w$ for some constant $w$. Then

a) $Y_n + W_n \xrightarrow{D} Y + w$,

b) $Y_n W_n \xrightarrow{D} wY$, and

c) $Y_n / W_n \xrightarrow{D} Y / w$ if $w \neq 0$. 

Chapter 8. Large Sample Theory

Theorem 8.11. a) If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$.
b) If $X_n \xrightarrow{ae} X$ then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.
c) If $X_n \xrightarrow{r} X$ then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.
d) $X_n \xrightarrow{P} \tau(\theta)$ iff $X_n \xrightarrow{D} \tau(\theta)$.
e) If $X_n \xrightarrow{P} \theta$ and $\tau$ is continuous at $\theta$, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.
f) If $X_n \xrightarrow{D} \theta$ and $\tau$ is continuous at $\theta$, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Suppose that for all $\theta \in \Theta$, $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$ or $T_n \xrightarrow{ae} \tau(\theta)$. Then $T_n$ is a consistent estimator of $\tau(\theta)$ by Theorem 8.11.

Example 8.10. Let $Y_1, \ldots, Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n$ is a consistent estimator of $\mu$ since i) the SLLN holds (use Theorem 8.9 and 8.11), ii) the WLLN holds and iii) the CLT holds (use Proposition 8.8). Since

$$\lim_{n \to \infty} \text{VAR}_\mu(\overline{Y}_n) = \lim_{n \to \infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n \to \infty} E_\mu(\overline{Y}_n) = \mu,$$

$\overline{Y}_n$ is also a consistent estimator of $\mu$ by Proposition 8.7b. By the delta method and Proposition 8.8b, $T_n = g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 8.11e, $g(\overline{Y}_n)$ is a consistent estimator of $g(\mu)$ if $g$ is continuous at $\mu$ for all $\mu \in \Theta$.

Theorem 8.12. a) Generalized Continuous Mapping Theorem: If $X_n \xrightarrow{D} X$ and the function $g$ is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where $g$ is continuous, then $g(X_n) \xrightarrow{D} g(X)$.
b) Continuous Mapping Theorem: If $X_n \xrightarrow{D} X$ and the function $g$ is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Remark 8.3. For Theorem 8.11, a) follows from Slutsky’s Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \xrightarrow{D} Y = X$ and $W_n \xrightarrow{D} 0$. Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if $T_n$ is a consistent estimator of $\theta$ and $\tau$ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 8.12 says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.
Example 8.11. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$ then $1/X_n \xrightarrow{D} 1/X$ if $X$ is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 8.12. Show that if $Y_n \sim t_n$, a $t$ distribution with $n$ degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n = Z/\sqrt{V_n/n}$ where $Z \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky’s Theorem. But $V_n = \sum_{i=1}^{n} X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 8.11e.

Theorem 8.13: Continuity Theorem. Let $Y_n$ be sequence of random variables with characteristic functions $\phi_n(t)$. Let $Y$ be a random variable with cf $\phi(t)$.

a) $Y_n \xrightarrow{D} Y$ iff $\phi_n(t) \rightarrow \phi(t) \forall t \in \mathbb{R}$.

b) Also assume that $Y_n$ has mgf $m_n$ and $Y$ has mgf $m$. Assume that all of the mgfs $m_n$ and $m$ are defined on $|t| \leq d$ for some $d > 0$. Then if $m_n(t) \rightarrow m(t)$ as $n \rightarrow \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

Application: Proof of a Special Case of the CLT. Following Rohatgi (1984, p. 569-9), let $Y_1, ..., Y_n$ be iid with mean $\mu$, variance $\sigma^2$ and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1 and mgf $m_Z(t) = \exp(-t\mu/\sigma)m_Y(t/\sigma)$ for $|t| < \sigma t_o$. Want to show that

$$W_n = \sqrt{n} \left( \frac{\overline{Y}_n - \mu}{\sigma} \right) \xrightarrow{D} N(0, 1).$$

Notice that $W_n =$

$$n^{-1/2} \sum_{i=1}^{n} Z_i = n^{-1/2} \sum_{i=1}^{n} \left( \frac{Y_i - \mu}{\sigma} \right) = n^{-1/2} \sum_{i=1}^{n} \frac{Y_i - n\mu}{\sigma} = n^{-1/2} \frac{1}{n} \sum_{i=1}^{n} Y_i - n\mu.$$
Thus
\[ m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2} \sum_{i=1}^{n} Z_i)] = E[\exp(\sum_{i=1}^{n} tZ_i/\sqrt{n})] \]
\[ = \prod_{i=1}^{n} E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^{n} m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n. \]

Set \( \psi(x) = \log(m_Z(x)) \). Then
\[ \log[m_{W_n}(t)] = n \log[m_Z(t/\sqrt{n})] = n\psi(t/\sqrt{n}) = \frac{\psi(t/\sqrt{n})}{n}. \]

Now \( \psi(0) = \log[m_Z(0)] = \log(1) = 0 \). Thus by L'Hôpital's rule (where the derivative is with respect to \( n \)), \( \lim_{n \to \infty} \log[m_{W_n}(t)] = \)
\[ \lim_{n \to \infty} \frac{\psi(t/\sqrt{n})}{n} = \lim_{n \to \infty} \frac{\psi'(t/\sqrt{n}) \cdot \left(-\frac{t}{2n^{3/2}}\right)}{\frac{-1}{2n^{3/2}}} = \frac{t^2}{2} \lim_{n \to \infty} \frac{\psi''(t/\sqrt{n})}{n \sqrt{n}}. \]

Now
\[ \psi'(0) = \frac{m'_Z(0)}{m_Z(0)} = E(Z_i)/1 = 0, \]
so L'Hôpital's rule can be applied again, giving \( \lim_{n \to \infty} \log[m_{W_n}(t)] = \)
\[ \frac{t^2}{2} \lim_{n \to \infty} \psi''(t/\sqrt{n}) \cdot \left(-\frac{t}{2n^{3/2}}\right) = \frac{t^2}{2} \psi''(0). \]

Now
\[ \psi''(t) = \frac{d}{dt} \frac{m'_Z(t)}{m_Z(t)} = \frac{m''_Z(t)m_Z(t) - (m'_Z(t))^2}{[m_Z(t)]^2}. \]
So
\[ \psi''(0) = m''_Z(0) - [m'_Z(0)]^2 = E(Z_i^2) - [E(Z_i)]^2 = 1. \]
Hence \( \lim_{n \to \infty} \log[m_{W_n}(t)] = t^2/2 \) and
\[ \lim_{n \to \infty} m_{W_n}(t) = \exp(t^2/2) \]
which is the \( N(0,1) \) mgf. Thus by the continuity theorem,
\[ W_n = \sqrt{n} \left( \frac{\bar{Y}_n - \mu}{\sigma} \right) \xrightarrow{D} N(0,1). \]
8.5 Order Relations and Convergence Rates

Definition 8.10. Lehmann (1999, p. 53-54): a) A sequence of random variables \( W_n \) is tight or bounded in probability, written \( W_n = \mathcal{O}_P(1) \), if for every \( \epsilon > 0 \) there exist positive constants \( D_\epsilon \) and \( N_\epsilon \) such that

\[
P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon
\]

for all \( n \geq N_\epsilon \). Also \( W_n = \mathcal{O}_P(X_n) \) if \( |W_n/X_n| = \mathcal{O}_P(1) \) which means that

\[
n^\delta W_n \xrightarrow{P} 0.
\]

c) \( W_n \) has the same order as \( X_n \) in probability, written \( W_n \sim_P X_n \), if for every \( \epsilon > 0 \) there exist positive constants \( N_\epsilon \) and \( 0 < d_\epsilon < D_\epsilon \) such that

\[
P(d_\epsilon \leq |W_n/X_n| \leq D_\epsilon) \geq 1 - \epsilon
\]

for all \( n \geq N_\epsilon \).

d) Similar notation is used for a \( k \times r \) matrix \( A_n = [a_{i,j}(n)] \) if each element \( a_{i,j}(n) \) has the desired property. For example, \( A_n = \mathcal{O}_P(n^{-1/2}) \) if each \( a_{i,j}(n) = \mathcal{O}_P(n^{-1/2}) \).

Definition 8.11. Let \( \hat{\beta}_n \) be an estimator of a \( p \times 1 \) vector \( \beta \), and let \( W_n = \|\beta_n - \beta\| \).

a) If \( W_n \sim_P n^{-\delta} \) for some \( \delta > 0 \), then both \( W_n \) and \( \hat{\beta}_n \) have (tightness) rate \( n^\delta \).

b) If there exists a constant \( \kappa \) such that

\[
n^\delta(W_n - \kappa) \xrightarrow{D} X
\]

for some nondegenerate random variable \( X \), then both \( W_n \) and \( \hat{\beta}_n \) have convergence rate \( n^\delta \).

Proposition 8.14. Suppose there exists a constant \( \kappa \) such that

\[
n^\delta(W_n - \kappa) \xrightarrow{D} X.
\]

a) Then \( W_n = \mathcal{O}_P(n^{-\delta}) \).

b) If \( X \) is not degenerate, then \( W_n \sim_P n^{-\delta} \).
CHAPTER 8. LARGE SAMPLE THEORY

The above result implies that if \( W_n \) has convergence rate \( n^\delta \), then \( W_n \) has tightness rate \( n^\delta \), and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if \( W_n \sim P X_n \), then \( X_n \sim P W_n \). Notice that if \( W_n = O_P(n^{-\delta}) \), then \( n^\delta \) is a lower bound on the rate of \( W_n \). As an example, if the CLT holds then \( Y_n = O_P(n^{-1/3}) \), but \( Y_n \sim n^{-1/2} \).

**Proposition 8.15.** a) If \( W_n \sim P X_n \) then \( X_n \sim P W_n \).

b) If \( W_n \sim P X_n \) then \( W_n = O_P(X_n) \).

c) If \( W_n \sim P X_n \) then \( X_n = O_P(W_n) \).

d) \( W_n \sim P X_n \) iff \( W_n = O_P(X_n) \) and \( X_n = O_P(W_n) \).

**Proof.** a) Since \( W_n \sim P X_n \),

\[
P(d_\epsilon \leq \left| \frac{W_n}{X_n} \right| \leq D_\epsilon) = P\left( \frac{1}{D_\epsilon} \leq \left| \frac{X_n}{W_n} \right| \leq \frac{1}{d_\epsilon} \right) \geq 1 - \epsilon
\]

for all \( n \geq N_\epsilon \). Hence \( X_n \sim P W_n \).

b) Since \( W_n \sim P X_n \),

\[
P(|W_n| \leq |X_n D_\epsilon|) \geq P(d_\epsilon \leq \left| \frac{W_n}{X_n} \right| \leq D_\epsilon) \geq 1 - \epsilon
\]

for all \( n \geq N_\epsilon \). Hence \( W_n = O_P(X_n) \).

c) Follows by a) and b).

d) If \( W_n \sim P X_n \), then \( W_n = O_P(X_n) \) and \( X_n = O_P(W_n) \) by b) and c).

Now suppose \( W_n = O_P(X_n) \) and \( X_n = O_P(W_n) \). Then

\[
P(|W_n| \leq |X_n| D_{\epsilon/2}) \geq 1 - \epsilon/2
\]

for all \( n \geq N_1 \), and

\[
P(|X_n| \leq |W_n| 1/d_{\epsilon/2}) \geq 1 - \epsilon/2
\]

for all \( n \geq N_2 \). Hence

\[
P(A) \equiv P\left( \left| \frac{W_n}{X_n} \right| \leq D_{\epsilon/2} \right) \geq 1 - \epsilon/2
\]

and

\[
P(B) \equiv P(d_{\epsilon/2} \leq \left| \frac{W_n}{X_n} \right|) \geq 1 - \epsilon/2
\]
for all \( n \geq N = \max(N_1, N_2) \). Since \( P(A \cap B) = P(A) + P(B) - P(A \cup B) \geq P(A) + P(B) - 1 \),

\[
P(A \cap B) = P(d_{\epsilon/2} \leq \left| \frac{W_n}{X_n} \right| \leq D_{\epsilon/2}) \geq 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon
\]

for all \( n \geq N \). Hence \( W_n \sim P X_n \). QED

The following result is used to prove the following Theorem 8.17 which says that if there are \( K \) estimators \( T_{j,n} \) of a parameter \( \beta \), such that \( \|T_{j,n} - \beta\| = O_P(n^{-\delta}) \) where \( 0 < \delta \leq 1 \), and if \( T_n^* \) picks one of these estimators, then \( \|T_n^* - \beta\| = O_P(n^{-\delta}) \).

**Proposition 8.16: Pratt (1959).** Let \( X_{1,n}, \ldots, X_{K,n} \) each be \( O_P(1) \) where \( K \) is fixed. Suppose \( W_n = X_{i_n,n} \) for some \( i_n \in \{1, \ldots, K\} \). Then

\[
W_n = O_P(1).
\] (8.8)

**Proof.**

\[
P(\max\{X_{1,n}, \ldots, X_{K,n}\} \leq x) = P(X_{1,n} \leq x, \ldots, X_{K,n} \leq x) \leq \]

\[
F_{W_n}(x) \leq P(\min\{X_{1,n}, \ldots, X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, \ldots, X_{K,n} > x).\]

Since \( K \) is finite, there exists \( B > 0 \) and \( N \) such that \( P(X_{i,n} \leq B) > 1 - \epsilon/2K \) and \( P(X_{i,n} > B) > 1 - \epsilon/2K \) for all \( n > N \) and \( i = 1, \ldots, K \). Bonferroni’s inequality states that \( P(\cap_{i=1}^{K} A_i) \geq \sum_{i=1}^{K} P(A_i) - (K - 1) \). Thus

\[
F_{W_n}(B) \geq P(X_{1,n} \leq B, \ldots, X_{K,n} \leq B) \geq \\
K(1 - \epsilon/2K) - (K - 1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2
\]

and

\[
-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, \ldots, X_{K,n} > -B) \geq \\
-1 + K(1 - \epsilon/2K) - (K - 1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.
\]

Hence

\[
F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \text{ for } n > N. \text{ QED}
\]

**Theorem 8.17.** Suppose \( \|T_{j,n} - \beta\| = O_P(n^{-\delta}) \) for \( j = 1, \ldots, K \) where \( 0 < \delta \leq 1 \). Let \( T_{n}^* = T_{i_n,n} \) for some \( i_n \in \{1, \ldots, K\} \) where, for example, \( T_{i_n,n} \) is the \( T_{j,n} \) that minimized some criterion function. Then

\[
\|T_n^* - \beta\| = O_P(n^{-\delta}). \tag{8.9}
\]

**Proof.** Let \( X_{j,n} = n^\delta \|T_{j,n} - \beta\| \). Then \( X_{j,n} = O_P(1) \) so by Proposition 8.16, \( n^\delta \|T_n^* - \beta\| = O_P(1) \). Hence \( \|T_n^* - \beta\| = O_P(n^{-\delta}) \). QED
8.6 Multivariate Limit Theorems

Many of the univariate results of the previous 5 sections can be extended to random vectors. As stated in Section 2.7, the notation for random vectors is rather awkward. For the limit theorems, the vector $X$ is typically a $k \times 1$ column vector and $X^T$ is a row vector. Let $\|x\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $x$.

**Definition 8.12.** Let $X_n$ be a sequence of random vectors with joint cdfs $F_n(x)$ and let $X$ be a random vector with joint cdf $F(x)$.

a) $X_n$ converges in distribution to $X$, written $X_n \xrightarrow{D} X$, if $F_n(x) \rightarrow F(x)$ as $n \rightarrow \infty$ for all points $x$ at which $F(x)$ is continuous. The distribution of $X$ is the limiting distribution or asymptotic distribution of $X_n$.

b) $X_n$ converges in probability to $X$, written $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, $P(\|X_n - X\| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.

c) Let $r > 0$ be a real number. Then $X_n$ converges in $r$th mean to $X$, written $X_n \xrightarrow{L^r} X$, if $E(\|X_n - X\|^r) \rightarrow 0$ as $n \rightarrow \infty$.

d) $X_n$ converges almost everywhere to $X$, written $X_n \xrightarrow{ae} X$, if $P(\lim_{n \rightarrow \infty} X_n = X) = 1$.

Theorems 8.18, 8.19 and 8.21 below are the multivariate extensions of the limit theorems in Section 8.1. When the limiting distribution of $Z_n = \sqrt{n}(g(T_n) - g(\theta))$ is multivariate normal $N_k(0, \Sigma)$, approximate the joint cdf of $Z_n$ with the joint cdf of the $N_k(0, \Sigma)$ distribution. Thus to find probabilities, manipulate $Z_n$ as if $Z_n \approx N_k(0, \Sigma)$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$ and $V(X) = \Sigma = \sigma^2$.

**Theorem 8.18:** the Multivariate Central Limit Theorem (MCLT).
If $X_1, \ldots, X_n$ are iid $k \times 1$ random vectors with $E(X) = \mu$ and $\text{Cov}(X) = \Sigma$, then

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} N_k(0, \Sigma)$$

where the sample mean

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

To see that the delta method is a special case of the multivariate delta method, note that if $T_n$ and parameter $\theta$ are real valued, then $D g(T_n, \theta) = g'(\theta)$. 


CHAPTER 8. LARGE SAMPLE THEORY

Theorem 8.19: the Multivariate Delta Method. If

\[ \sqrt{n}(T_n - \theta) \xrightarrow{D} N_k(0, \Sigma), \]

then

\[ \sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N_d(0, Dg(\theta) \Sigma Dg(\theta)^T) \]

where the \(d \times k\) Jacobian matrix of partial derivatives

\[
Dg(\theta) = \begin{bmatrix}
\frac{\partial}{\partial \theta_1} g_1(\theta) & \cdots & \frac{\partial}{\partial \theta_k} g_1(\theta) \\
\vdots & \ddots & \vdots \\
\frac{\partial}{\partial \theta_1} g_d(\theta) & \cdots & \frac{\partial}{\partial \theta_k} g_d(\theta)
\end{bmatrix}.
\]

Here the mapping \(g : \mathbb{R}^k \to \mathbb{R}^d\) needs to be differentiable in a neighborhood of \(\theta \in \mathbb{R}^k\).

Example 8.13. If \(Y\) has a Weibull distribution, \(Y \sim W(\phi, \lambda)\), then the pdf of \(Y\) is

\[ f(y) = \frac{\phi}{\lambda} y^{\phi-1} e^{-\frac{y^\phi}{\lambda}} \]

where \(\lambda, y, \) and \(\phi\) are all positive. If \(\mu = \lambda^{1/\phi}\) so \(\mu^\phi = \lambda\), then the Weibull pdf

\[ f(y) = \frac{\phi}{\mu} \left( \frac{y}{\mu} \right)^{\phi-1} e^{-\left( \frac{y}{\mu} \right)^\phi} \cdot \]

Let \((\hat{\mu}, \hat{\phi})\) be the MLE of \((\mu, \phi)\). According to Bain (1978, p. 215),

\[ \sqrt{n} \left( \begin{bmatrix} \hat{\mu} \\ \hat{\phi} \end{bmatrix} - \begin{bmatrix} \mu \\ \phi \end{bmatrix} \right) \xrightarrow{D} N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.109\phi^2 & 0.257\mu \\ 0.257\mu & 0.608\phi^2 \end{bmatrix} \right) \]

\[ = N_2(0, I^{-1}(\theta)) \]

where \(I(\theta)\) is given in Definition 8.13.

Let column vectors \(\theta = (\mu, \phi)^T\) and \(\eta = (\lambda, \phi)^T\). Then

\[ \eta = g(\theta) = \begin{bmatrix} \lambda \\ \phi \end{bmatrix} = \begin{bmatrix} \mu^\phi \\ \phi \end{bmatrix} = \begin{bmatrix} g_1(\theta) \\ g_2(\theta) \end{bmatrix}. \]

So

\[
Dg(\theta) = \begin{bmatrix}
\frac{\partial}{\partial \theta_1} g_1(\theta) & \frac{\partial}{\partial \theta_2} g_1(\theta) \\
\frac{\partial}{\partial \theta_1} g_2(\theta) & \frac{\partial}{\partial \theta_2} g_2(\theta)
\end{bmatrix} = \begin{bmatrix}
\frac{\partial}{\partial \mu} \mu^\phi & \frac{\partial}{\partial \phi} \mu^\phi \\
\frac{\partial}{\partial \mu} \phi & \frac{\partial}{\partial \phi} \phi
\end{bmatrix} = \begin{bmatrix}
\phi \mu^{\phi-1} & \mu^\phi \log(\mu) \\
0 & 1
\end{bmatrix}. \]
Thus by the multivariate delta method,

$$\sqrt{n} \left( \begin{pmatrix} \hat{\lambda} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \lambda \\ \phi \end{pmatrix} \right) \xrightarrow{D} N_2(0, \Sigma)$$

where (see Definition 8.15 below)

$$\Sigma = I(\eta)^{-1} = [I(g(\theta))]^{-1} = Dg(\theta)I^{-1}(\theta)D^T g(\theta) =$$

$$\begin{bmatrix}
1.109\lambda^2(1 + 0.4635 \log(\lambda) + 0.5482(\log(\lambda))^2) & 0.257\phi\lambda + 0.608\lambda\phi \log(\lambda) \\
0.257\phi\lambda + 0.608\lambda\phi \log(\lambda) & 0.608\phi^2
\end{bmatrix}.$$

**Definition 8.13.** Let $X$ be a random variable with pdf or pmf $f(x|\theta)$. Then the information matrix

$$I(\theta) = [I_{i,j}]$$

where

$$I_{i,j} = E \left[ \frac{\partial}{\partial \theta_i} \log(f(X|\theta)) \frac{\partial}{\partial \theta_j} \log(f(X|\theta)) \right].$$

**Definition 8.14.** An estimator $T_n$ of $\theta$ is asymptotically efficient if

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N_k(0, I^{-1}(\theta)).$$

Following Lehmann (1999, p. 511), if $T_n$ is asymptotically efficient and if the estimator $W_n$ satisfies

$$\sqrt{n}(W_n - \theta) \xrightarrow{D} N_k(0, J(\theta))$$

where $J(\theta)$ and $I^{-1}(\theta)$ are continuous functions of $\theta$, then under regularity conditions, $J(\theta) - I^{-1}(\theta)$ is a positive semidefinite matrix, and $T_n$ is “better” than $W_n$.

**Definition 8.15.** Assume that $\eta = g(\theta)$. Then

$$I(\eta) = I(g(\theta)) = [Dg(\theta)I^{-1}(\theta)D^T g(\theta)]^{-1}. $$
CHAPTER 8. LARGE SAMPLE THEORY

255

Notice that this definition agrees with the multivariate delta method if

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N_k(0, \Sigma)$$

where \(\Sigma = I^{-1}(\theta)\).

Now suppose that \(X_1, \ldots, X_n\) are iid random variables from a \(k\)-parameter REF

$$f(x|\theta) = h(x)c(\theta)\exp\left[\sum_{i=1}^{k} w_i(\theta)t_i(x)\right]$$

(8.10)

with natural parameterization

$$f(x|\eta) = h(x)b(\eta)\exp\left[\sum_{i=1}^{k} \eta_i t_i(x)\right].$$

(8.11)

Then the complete minimal sufficient statistic is

$$T_n = \frac{1}{n} \left(\sum_{i=1}^{n} t_1(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)^T.$$

Let \(\mu_T = (E(t_1(X)), \ldots, E(t_k(X)))^T\). From Theorem 3.3, for \(\eta \in \Omega\),

$$E(t_i(X)) = -\frac{\partial}{\partial \eta_i} \log(b(\eta)),$$

and

$$\text{Cov}(t_i(X), t_j(X)) \equiv \sigma_{i,j} = -\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\eta)).$$

**Proposition 8.20.** If the random variable \(X\) is a kP–REF with pmf or pdf (8.11), then the information matrix

$$I(\eta) = [I_{i,j}]$$

where

$$I_{i,j} = E \left[\frac{\partial}{\partial \eta_i} \log(f(X|\eta)) \frac{\partial}{\partial \eta_j} \log(f(X|\eta))\right] = -E \left[\frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\eta))\right].$$

Several authors, including Barndorff–Nielsen (1982), have noted that the multivariate CLT can be used to show that \(\sqrt{n}(T_n - \mu_T) \xrightarrow{D} N_k(0, \Sigma)\). The fact that \(\Sigma = I(\eta)\) appears in Lehmann (1983, p. 127).
Theorem 8.21. If $X_1, ..., X_n$ are iid from a $k$-parameter regular exponential family, then

$$\sqrt{n}(T_n - \mu_T) \xrightarrow{D} N_k(0, I(\eta)).$$

Proof. By the multivariate central limit theorem,

$$\sqrt{n}(T_n - \mu_T) \xrightarrow{D} N_k(0, \Sigma)$$

where $\Sigma = [\sigma_{i,j}]$. Hence the result follows if $\sigma_{i,j} = I_{i,j}$. Since

$$\log(f(x|\eta)) = \log(h(x)) + \log(b(\eta)) + \sum_{l=1}^k \eta_l t_l(x),$$

$$\frac{\partial}{\partial \eta_i} \log(f(x|\eta)) = \frac{\partial}{\partial \eta_i} \log(b(\eta)) + t_i(X).$$

Hence

$$-I_{i,j} = E \left[ \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(f(X|\eta)) \right] = \frac{\partial^2}{\partial \eta_i \partial \eta_j} \log(b(\eta)) = -\sigma_{i,j}. \quad \text{QED}$$

To obtain standard results, use the multivariate delta method, assume that both $\theta$ and $\eta$ are $k \times 1$ vectors, and assume that $\eta = g(\theta)$ is a one to one mapping so that the inverse mapping is $\theta = g^{-1}(\eta)$. If $Dg(\theta)$ is nonsingular, then

$$D^{-1}g(\theta) = Dg^{-1}(\eta) \quad (8.12)$$

(see Searle 1982, p. 339), and

$$I(\eta) = [Dg(\theta)I^{-1}(\theta)D^T g(\theta)]^{-1} = [Dg(\theta)^{-1}]^T I(\theta)Dg(\theta)^{-1} = D^T \eta^{-1} I(\theta)Dg(\theta) \eta^{-1} \quad (8.13)$$


For example, suppose that $\mu_T$ and $\eta$ are $k \times 1$ vectors, and

$$\sqrt{n}(\eta - \eta) \xrightarrow{D} N_k(0, I^{-1}(\eta))$$

where $\mu_T = g(\eta)$ and $\eta = g^{-1}(\mu_T)$. Also assume that $T_n = g(\eta)$ and $\hat{\eta} = g^{-1}(T_n)$. Then by the multivariate delta method and Theorem 8.21,

$$\sqrt{n}(T_n - \mu_T) = \sqrt{n}(g(\eta) - g(\eta)) \xrightarrow{D} N_k[0, I(\eta)] = N_k[0, Dg(\eta)I^{-1}(\eta)D^T g(\eta)].$$
Hence
\[ I(\eta) = Dg(\eta)I^{-1}(\eta)D^Tg(\eta). \]

Similarly,
\[ \sqrt{n}(g^{-1}(T_n) - g^{-1}(\mu_T)) = \sqrt{n}(\eta - \eta) \xrightarrow{D} N_k[0, I^{-1}(\eta)] = N_k[0, Dg^{-1}(\mu_T)I(Dg^{-1}(\mu_T))]. \]

Thus
\[ I^{-1}(\eta) = Dg^{-1}(\mu_T)I(\eta)D^Tg^{-1}(\mu_T) = Dg^{-1}(\mu_T)Dg(\eta)I^{-1}(\eta)D^Tg(\eta)D^Tg^{-1}(\mu_T) \]
as expected by Equation (8.13). Typically \( \hat{\theta} \) is a function of the sufficient statistic \( T_n \) and is the unique MLE of \( \theta \). Replacing \( \eta \) by \( \theta \) in the above discussion shows that \( \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{D} N_k(0, I^{-1}(\theta)) \) is equivalent to \( \sqrt{n}(T_n - \mu_T) \xrightarrow{D} N_k(0, I(\theta)) \) provided that \( Dg(\theta) \) is nonsingular.

### 8.7 More Multivariate Results

**Definition 8.16.** If the estimator \( g(T_n) \xrightarrow{P} g(\theta) \) for all \( \theta \in \Theta \), then \( g(T_n) \) is a **consistent estimator** of \( g(\theta) \).

**Proposition 8.22.** If \( 0 < \delta \leq 1 \), \( X \) is a random vector, and
\[ n^{\delta}(g(T_n) - g(\theta)) \xrightarrow{D} X, \]
then \( g(T_n) \xrightarrow{P} g(\theta) \).

**Theorem 8.23.** If \( X_1, ..., X_n \) are iid, \( E(||X||) < \infty \) and \( E(X) = \mu \), then
a) WLLN: \( \bar{X}_n \xrightarrow{D} \mu \) and
b) SLLN: \( \bar{X}_n \xrightarrow{ae} \mu \).

**Theorem 8.24: Continuity Theorem.** Let \( X_n \) be a sequence of \( k \times 1 \) random vectors with characteristic function \( \phi_n(t) \) and let \( X \) be a \( k \times 1 \) random vector with cf \( \phi(t) \). Then
\[ X_n \xrightarrow{D} X \iff \phi_n(t) \to \phi(t) \]
CHAPTER 8. LARGE SAMPLE THEORY

258

for all $t \in \mathbb{R}^k$.

**Theorem 8.25: Cramér Wold Device.** Let $X_n$ be a sequence of $k \times 1$ random vectors and let $X$ be a $k \times 1$ random vector. Then

$$X_n \overset{D}{\to} X \quad \text{iff} \quad t^TX_n \overset{D}{\to} t^TX$$

for all $t \in \mathbb{R}^k$.

**Theorem 8.26.** a) If $X_n \overset{P}{\to} X$, then $X_n \overset{D}{\to} X$.

b) $X_n \overset{P}{\to} g(\theta)$ iff $X_n \overset{D}{\to} g(\theta)$.

Let $g(n) \geq 1$ be an increasing function of the sample size $n$: $g(n) \uparrow \infty$, eg $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $T_n - \mu$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $T_n$ has (tightness) rate $\sqrt{n}$.

**Definition 8.17.** Let $A_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.

a) $A_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

b) $A_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

c) $A_n \succ_p (1/(g(n)))$ if $a_{i,j}(n) \succ_p (1/(g(n)))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.

d) Let $A_{1,n} = T_n - \mu$ and $A_{2,n} = C_n - c\Sigma$ for some constant $c > 0$. If $A_{1,n} \succ_p (1/(g(n)))$ and $A_{2,n} \succ_p (1/(g(n)))$, then $(T_n, C_n)$ has (tightness) rate $g(n)$.

Recall that the smallest integer function $\lceil x \rceil$ rounds up, eg $\lceil 7.7 \rceil = 8$.

**Definition 8.18.** The sample $\rho$ quantile $\hat{\xi}_{n,\rho} = Y(\lceil n\rho \rceil)$. The population quantile $\xi_\rho = Q(\rho) = \inf \{ y : F(y) \geq \rho \}$.

**Theorem 8.27: Serfling (1980, p. 80).** Let $0 < \rho_1 < \rho_2 < \cdots < \rho_k < 1$. Suppose that $F$ has a density $f$ that is positive and continuous in neighborhoods of $\xi_{\rho_1}, \ldots, \xi_{\rho_k}$. Then

$$\sqrt{n}[(\hat{\xi}_{n,\rho_1}, \ldots, \hat{\xi}_{n,\rho_k})^T - (\xi_{\rho_1}, \ldots, \xi_{\rho_k})^T] \overset{D}{\to} N_k(0, \Sigma)$$

where $\Sigma = (\sigma_{ij})$ and

$$\sigma_{ij} = \frac{\rho_i(1 - \rho_j)}{f(\xi_{\rho_i})f(\xi_{\rho_j})}$$

for $i \leq j$ and $\sigma_{ij} = \sigma_{ji}$ for $i > j$. 

CHAPTER 8. LARGE SAMPLE THEORY

Theorem 8.28: Continuous Mapping Theorem. Let \( X_n \in \mathbb{R}^k \). If \( X_n \xrightarrow{D} X \) and if the function \( g : \mathbb{R}^k \to \mathbb{R}^j \) is continuous, then \( g(X_n) \xrightarrow{D} g(X) \).

8.8 Summary

1) CLT: Let \( Y_1, \ldots, Y_n \) be iid with \( E(Y) = \mu \) and \( V(Y) = \sigma^2 \). Then \( \sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2) \).

2) Delta Method: If \( g'(\theta) \neq 0 \) and \( \sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2) \), then \( \sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2) \).

3) 1P–REF Limit Theorem: Let \( Y_1, \ldots, Y_n \) be iid from a 1P–REF with pdf or pmf \( f(y|\theta) = h(y)c(\theta)\exp[w(\theta)t(y)] \) and natural parameterization \( f(y|\eta) = h(y)b(\eta)\exp[\eta t(y)] \). Let \( E(t(Y)) = \mu_t \equiv g(\eta) \) and \( V(t(Y)) = \sigma_t^2 \). Then \( \sqrt{n}\frac{1}{n} \sum_{i=1}^{n} t(Y_i) - \mu_t \xrightarrow{D} N(0, I(\eta)) \) where \( I(\eta) = \sigma_t^2 = g'(\eta) \).

4) Limit theorem for the Sample Median:
\[ \sqrt{n}(\text{MED}(n) - \text{MED}(Y)) \xrightarrow{D} N \left( 0, \frac{1}{4I^2(\text{MED}(Y))} \right) \]

5) If \( n^\delta(T_{1,n} - \theta) \xrightarrow{D} N(0, \sigma_1^2(F)) \) and \( n^\delta(T_{2,n} - \theta) \xrightarrow{D} N(0, \sigma_2^2(F)) \), then the asymptotic relative efficiency of \( T_{1,n} \) with respect to \( T_{2,n} \) is
\[ ARE(T_{1,n}, T_{2,n}) = \frac{\sigma_2^2(F)}{\sigma_1^2(F)} \]
The “better” estimator has the smaller asymptotic variance or \( \sigma_1^2(F) \).

6) An estimator \( T_n \) of \( \tau(\theta) \) is asymptotically efficient if
\[ \sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N \left( 0, \frac{[\tau'(\theta)]^2}{I(\theta)} \right) \]

7) For a 1P–REF, \( \frac{1}{n} \sum_{i=1}^{n} t(Y_i) \) is an asymptotically efficient estimator of \( g(\eta) = E(t(Y)) \).

8) Rule of thumb: If \( \hat{\theta}_n \) is the MLE or UMVUE of \( \theta \), then \( T_n = \tau(\hat{\theta}_n) \) is an asymptotically efficient estimator of \( \tau(\theta) \). Hence if \( \tau'(\theta) \neq 0 \), then
\[ \sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)] \xrightarrow{D} N \left( 0, \frac{[\tau'(\theta)]^2}{I(\theta)} \right) \].
9) \( X_n \overset{D}{\to} X \) if 
\[
\lim_{n \to \infty} F_n(t) = F(t)
\]
at each continuity point \( t \) of \( F \).

10) \( X_n \overset{P}{\to} \tau(\theta) \) if for every \( \epsilon > 0 \),
\[
\lim_{n \to \infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{n \to \infty} P(|X_n - \tau(\theta)| \geq \epsilon) = 0.
\]

11) \( T_n \) is a \textbf{consistent estimator} of \( \tau(\theta) \) if \( T_n \overset{P}{\to} \tau(\theta) \) for every \( \theta \in \Theta \).

12) \( T_n \) is a \textbf{consistent estimator} of \( \tau(\theta) \) if any of the following 3 conditions holds:
   i) \( \lim_{n \to \infty} \text{VAR}_\theta(T_n) = 0 \) and \( \lim_{n \to \infty} E_\theta(T_n) = \tau(\theta) \) for all \( \theta \in \Theta \).
   ii) \( \text{MSE}_{\tau(\theta)}(T_n) \to 0 \) for all \( \theta \in \Theta \).
   iii) \( E[(T_n - \tau(\theta))^2] \to 0 \) for all \( \theta \in \Theta \).

13) If 
\[
\sqrt{n}(T_n - \tau(\theta)) \overset{D}{\to} N(0, v(\theta))
\]
for all \( \theta \in \Theta \), then \( T_n \) is a consistent estimator of \( \tau(\theta) \).

14) \textbf{WLLN}: Let \( Y_1, \ldots, Y_n, \ldots \) be a sequence of iid random variables with \( E(Y_i) = \mu \). Then \( \bar{Y}_n \overset{P}{\to} \mu \). Hence \( \bar{Y}_n \) is a consistent estimator of \( \mu \).

15) i) If \( X_n \overset{P}{\to} X \) then \( X_n \overset{D}{\to} X \).
   ii) \( T_n \overset{P}{\to} \tau(\theta) \) iff \( T_n \overset{D}{\to} \tau(\theta) \).
   iii) If \( T_n \overset{P}{\to} \theta \) and \( \tau \) is continuous at \( \theta \), then \( \tau(T_n) \overset{P}{\to} \tau(\theta) \). Hence if \( T_n \) is a consistent estimator of \( \theta \), then \( \tau(T_n) \) is a consistent estimator of \( \tau(\theta) \) if \( \tau \) is a continuous function on \( \Theta \).

16) \textbf{Slutsky’s Theorem}: If \( Y_n \overset{D}{\to} Y \) and \( W_n \overset{P}{\to} w \) for some constant \( w \),
then \( Y_n W_n \overset{D}{\to} wY \), \( Y_n + W_n \overset{D}{\to} Y + w \) and \( Y_n/W_n \overset{D}{\to} Y/w \) for \( w \neq 0 \).

\section*{8.9 Complements}

Some authors state that if \( X_n \overset{P}{\to} X \) and \( g \) is continuous, then \( g(X_n) \overset{P}{\to} g(X) \), but Sen and Singer (1993, p. 59) says that \( g \) needs to be uniformly continuous.

The following extension of the delta method is sometimes useful.
Theorem 8.29. Suppose that $g'(\theta) = 0$, $g''(\theta) \neq 0$ and
\[ \sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \tau^2(\theta)). \]
Then
\[ n[g(T_n) - g(\theta)] \xrightarrow{D} \frac{1}{2} \tau^2(\theta) g''(\theta) \chi^2_1. \]

Example 8.14. Let $X_n \sim \text{Binomial}(n, p)$ where the positive integer $n$ is large and $0 < p < 1$. Let $g(\theta) = \theta^3 - \theta$. Find the limiting distribution of
\[ n \left[ g \left( \frac{X_n}{n} \right) - c \right] \]
for appropriate constant $c$ when $p = \frac{1}{\sqrt{3}}$.

Solution: Since $X_n \xrightarrow{D} \sum_{i=1}^{n} Y_i$ where $Y_i \sim BIN(1, p)$,
\[ \sqrt{n} \left( \frac{X_n}{n} - p \right) \xrightarrow{D} N(0, p(1-p)) \]
by the CLT. Let $\theta = p$. Then $g'(\theta) = 3\theta^2 - 1$ and $g''(\theta) = 6\theta$. Notice that
\[ g(1/\sqrt{3}) = (1/\sqrt{3})^3 - 1/\sqrt{3} = (1/\sqrt{3})(1/3 - 1) = \frac{-2}{3\sqrt{3}} = c. \]
Also $g'(1/\sqrt{3}) = 0$ and $g''(1/\sqrt{3}) = 6/\sqrt{3}$. Since $\tau^2(p) = p(1-p)$,
\[ \tau^2(1/\sqrt{3}) = \frac{1}{\sqrt{3}}(1 - \frac{1}{\sqrt{3}}). \]
Hence
\[ n \left[ g \left( \frac{X_n}{n} \right) - \left( -\frac{2}{3\sqrt{3}} \right) \right] \xrightarrow{D} \frac{1}{2} \frac{1}{\sqrt{3}} \left(1 - \frac{1}{\sqrt{3}}\right) \frac{6}{\sqrt{3}} \chi^2_1 = (1 - \frac{1}{\sqrt{3}}) \chi^2_1. \]


In analysis, convergence in probability is a special case of convergence in measure and convergence in distribution is a special case of weak convergence.
See Ash (1972, p. 322) and Sen and Singer (1993, p. 39). Almost sure convergence is also known as strong convergence. See Sen and Singer (1993, p. 34). Since $\overline{Y} \xrightarrow{P} \mu$ iff $\overline{Y} \xrightarrow{D} \mu$, the WLLN refers to weak convergence. Technically the $X_n$ and $X$ need to share a common probability space for convergence in probability and almost sure convergence.

Perlman (1972) and Wald (1949) give general results on the consistency of the MLE while Berk (1972), Lehmann (1980) and Schervish (1995, p. 418) discuss the asymptotic normality of the MLE in exponential families. Theorems 8.4 and 8.21 appear in Olive (2007). Also see Cox (1984) and McCulloch (1995, p. 418). A similar result to Theorem 8.21 for linear exponential families where $t_i(x) = x_i$, are given by Brown (1986, p. 172). Portnoy (1977) gives large sample theory for unbiased estimators in exponential families. Although $\overline{T}_n$ is the UMVUE of $E(t(Y)) = \mu_t$, asymptotic efficiency of UMVUEs is not simple in general. See Pfanzagl (1993).


Suppose $\theta = g^{-1}(\eta)$. In analysis, the fact that
\[
D^{-1}g(\theta) = Dg^{-1}(\eta)
\]

Casella and Berger (2002, p. 112, 133) give results similar to Proposition 8.3.

According to Rohatgi (1984, p. 616), if $Y_1, ..., Y_n$ are iid with pdf $f(y)$, if $Y_{r_n:n}$ is the $r_n$th order statistic, $r_n/n \to \rho$, $F(\xi_\rho) = \rho$ and if $f(\xi_\rho) > 0$, then
\[
\sqrt{n}(Y_{r_n:n} - \xi_\rho) \xrightarrow{D} N \left(0, \frac{\rho(1 - \rho)}{[f(\xi_\rho)]^2} \right).
\]
So there are many asymptotically equivalent ways of defining the sample $\rho$ quantile.

\section{8.10 Problems}

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.
Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

8.1*. a) Enter the following R/Splus function that is used to illustrate the central limit theorem when the data $Y_1, ..., Y_n$ are iid from an exponential distribution. The function generates a data set of size $n$ and computes $\bar{Y}_1$ from the data set. This step is repeated $n_{runs} = 100$ times. The output is a vector $(\bar{Y}_1, \bar{Y}_2, ..., \bar{Y}_{100})$. A histogram of these means should resemble a symmetric normal density once $n$ is large enough.

```r
cltsim <- function(n=100, nruns=100){
ybar <- 1:nruns
for(i in 1:nruns){
  ybar[i] <- mean(rexp(n))
}list(ybar=ybar)}
```

b) The following commands will plot 4 histograms with $n = 1, 5, 25$ and 100. Save the plot in Word.

```r
> z1 <- cltsim(n=1)
> z5 <- cltsim(n=5)
> z25 <- cltsim(n=25)
> z200 <- cltsim(n=200)
> par(mfrow=c(2,2))
> hist(z1$ybar)
> hist(z5$ybar)
> hist(z25$ybar)
> hist(z200$ybar)
```

c) Explain how your plot illustrates the central limit theorem.

d) Repeat parts a), b) and c), but in part a), change rexp(n) to rnorm(n). Then $Y_1, ..., Y_n$ are iid $N(0,1)$ and $\bar{Y} \sim N(0, 1/n)$.

8.2*. Let $X_1, ..., X_n$ be iid from a normal distribution with unknown mean $\mu$ and known variance $\sigma^2$. Let

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Find the limiting distribution of $\sqrt{n}((\bar{X})^3 - c)$ for an appropriate constant $c$. 

8.3*. (Aug. 03 Qual) Let $X_1, ..., X_n$ be a random sample from a population with pdf $f(x) = \begin{cases} \frac{\theta x^{\theta - 1}}{\theta^\theta} & 0 < x < 3 \\ 0 & \text{elsewhere} \end{cases}$

The method of moments estimator for $\theta$ is $T_n = \frac{\bar{X}}{3 - \bar{X}}$.

a) Find the limiting distribution of $\sqrt{n}(T_n - \theta)$ as $n \to \infty$.

b) Is $T_n$ asymptotically efficient? Why?

c) Find a consistent estimator for $\theta$ and show that it is consistent.

8.4*. From Theorems 2.17 and 2.18, if $Y_n = \sum_{i=1}^{n} X_i$ where the $X_i$ are iid from a nice distribution, then $Y_n$ also has a nice distribution. If $E(X) = \mu$ and $\text{VAR}(X) = \sigma^2$ then by the CLT

$$\sqrt{n}(X_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Hence

$$\sqrt{n}(\frac{Y_n}{n} - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Find $\mu$, $\sigma^2$ and the distribution of $X_i$ if

i) $Y_n \sim \text{BIN}(n, \rho)$ where BIN stands for binomial.

ii) $Y_n \sim \chi^2_n$.

iii) $Y_n \sim G(n\nu, \lambda)$ where G stands for gamma.

iv) $Y_n \sim \text{NB}(n, \rho)$ where NB stands for negative binomial.

v) $Y_n \sim \text{POIS}(n\theta)$ where POIS stands for Poisson.

vi) $Y_n \sim N(n\mu, n\sigma^2)$.

8.5*. Suppose that $X_n \sim U(-1/n, 1/n)$.

a) What is the cdf $F_n(x)$ of $X_n$?

b) What does $F_n(x)$ converge to?

(Hint: consider $x < 0$, $x = 0$ and $x > 0$.)

c) $X_n \xrightarrow{D} X$. What is $X$?

8.6. Continuity Theorem problem: Let $X_n$ be sequence of random variables with cdfs $F_n$ and mgfs $m_n$. Let $X$ be a random variable with cdf $F$ and mgf $m$. Assume that all of the mgfs $m_n$ and $m$ are defined if $|t| \leq d$ for
some $d > 0$. Thus if $m_n(t) \rightarrow m(t)$ as $n \rightarrow \infty$ for all $|t| < c$ where $0 < c < d$, then $X_n \rightarrow X$.

Let

$$m_n(t) = \frac{1}{[1 - (\lambda + \frac{1}{n})t]}$$

for $t < 1/(\lambda + 1/n)$. Then what is $m(t)$ and what is $X$?

8.7. Let $Y_1, \ldots, Y_n$ be iid, $T_{1,n} = \bar{Y}$ and let $T_{2,n} = MED(n)$ be the sample median. Let $\theta = \mu$.

Then

$$\sqrt{n}(MED(n) - MED(Y)) \rightarrow N\left(0, \frac{1}{4f^2(MED(Y))}\right)$$

where the population median is $MED(Y)$ (and $MED(Y) = \mu = \theta$ for a) and b) below).

a) Find $ARE(T_{1,n}, T_{2,n})$ if $F$ is the cdf of the normal $N(\mu, \sigma^2)$ distribution.

b) Find $ARE(T_{1,n}, T_{2,n})$ if $F$ is the cdf of the double exponential $DE(\theta, \lambda)$ distribution.

8.8. (Sept. 2005 Qual) Let $X_1, \ldots, X_n$ be independent identically distributed random variables with probability density function

$$f(x) = \theta x^{\theta - 1}, \ 0 < x < 1, \ \theta > 0.$$  

a) Find the MLE of $\frac{1}{\theta}$. Is it unbiased? Does it achieve the information inequality lower bound?

b) Find the asymptotic distribution of the MLE of $\frac{1}{\theta}$.

c) Show that $\bar{X}_n$ is unbiased for $\frac{\theta}{\theta + 1}$. Does $\bar{X}_n$ achieve the information inequality lower bound?

d) Find an estimator of $\frac{1}{\theta}$ from part (c) above using $\bar{X}_n$ which is different from the MLE in (a). Find the asymptotic distribution of your estimator using the delta method.

e) Find the asymptotic relative efficiency of your estimator in (d) with respect to the MLE in (b).
8.9. Many multiple linear regression estimators $\hat{\beta}$ satisfy

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N_p(0, V(\hat{\beta}, F) W)$$  \hspace{2cm} (8.14)

when

$$\frac{X^TX}{n} \xrightarrow{P} W^{-1},$$  \hspace{2cm} (8.15)

and when the errors $e_i$ are iid with a cdf $F$ and a unimodal pdf $f$ that is symmetric with a unique maximum at 0. When the variance $V(e_i)$ exists,

$$V(\text{OLS}, F) = V(e_i) = \sigma^2 \text{ while } V(L_1, F) = \frac{1}{4[f(0)]^2}. $$

In the multiple linear regression model,

$$Y_i = x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = x_i^T\beta + e_i$$  \hspace{2cm} (8.16)

for $i = 1, \ldots, n$. In matrix notation, these $n$ equations become

$$Y = X\beta + e,$$  \hspace{2cm} (8.17)

where $Y$ is an $n \times 1$ vector of dependent variables, $X$ is an $n \times p$ matrix of predictors, $\beta$ is a $p \times 1$ vector of unknown coefficients, and $e$ is an $n \times 1$ vector of unknown errors.

a) What is the $ij$th element of the matrix

$$\frac{X^TX}{n}?$$

b) Suppose $x_{k,1} = 1$ and that $x_{k,j} \sim X_j$ are iid with $E(X_j) = 0$ and $V(X_j) = 1$ for $k = 1, \ldots, n$ and $j = 2, \ldots, p$. Assume that $X_i$ and $X_j$ are independent for $i \neq j, i > 1$ and $j > 1$. (Often $x_{k,j} \sim N(0, 1)$ in simulations.) Then what is $W^{-1}$ for model (8.16)?

c) Suppose $p = 2$ and $Y_i = \alpha + \beta X_i + e_i$. Show

$$(X^TX)^{-1} = \begin{bmatrix} \frac{\sum x_i^2}{n\sum(x_i - \bar{x})^2} & \frac{-\sum x_i}{n\sum(x_i - \bar{x})^2} \\ \frac{-\sum x_i}{n\sum(x_i - \bar{x})^2} & \frac{n}{n\sum(x_i - \bar{x})^2} \end{bmatrix}.$$
d) Under the conditions of c), let $S^2_x = \sum(X_i - \bar{X})^2/n$. Show that

$$n(X^TX)^{-1} = \left(\frac{X^TX}{n}\right)^{-1} = \begin{bmatrix}
\frac{\frac{\sum X^2}{S^2_x}}{S^2_x} & \frac{-\bar{X}}{S^2_x} \\
\frac{-\bar{X}}{S^2_x} & \frac{1}{S^2_x}
\end{bmatrix}.$$ 


e) If the $X_i$ are iid with variance $V(X)$ then $n(X^TX)^{-1} \overset{D}{\rightarrow} W$. What is $W$?

f) Now suppose that $n$ is divisible by 5 and the $n/5$ of $X_i$ are at 0.1, $n/5$ at 0.3, $n/5$ at 0.7 and $n/5$ at 0.9. (Hence if $n = 100$, 20 of the $X_i$ are at 0.1, 0.3, 0.5, 0.7 and 0.9.)

Find $\sum X^2_i/n$, $\bar{X}$ and $S^2_x$. (Your answers should not depend on $n$.)

\text{g) Under the conditions of f), estimate } V(\hat{\alpha}) \text{ and } V(\hat{\beta}) \text{ if } L_1 \text{ is used and if the } e_i \text{ are iid } N(0,0.01). \\
\text{Hint: Estimate } W \text{ with } n(X^TX)^{-1} \text{ and } V(\hat{\beta},F) = V(L_1,F) = \frac{1}{4[f(0)]^2}. \text{ Hence}

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \approx N_2\left(\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \frac{1}{n} \frac{1}{4[f(0)]^2} \begin{pmatrix} \frac{\sum X^2}{S^2_x} & \frac{-\bar{X}}{S^2_x} \\
\frac{-\bar{X}}{S^2_x} & \frac{1}{S^2_x} \end{pmatrix}\right).$$

You should get an answer like $0.0648/n$.

**Problems from old quizzes and exams.**

8.10. Let $X_1,\ldots,X_n$ be iid Bernoulli($p$) random variables.

a) Find $I_1(p)$.

b) Find the FCRLB for estimating $p$.

c) Find the limiting distribution of $\sqrt{n} (\bar{X}_n - p)$.

d) Find the limiting distribution of $\sqrt{n} \left[ (\bar{X}_n)^2 - c \right]$ for an appropriate constant $c$.

8.11. Let $X_1,\ldots,X_n$ be iid Exponential($\beta$) random variables.

a) Find the FCRLB for estimating $\beta$.

b) Find the limiting distribution of $\sqrt{n} (\bar{X}_n - \beta)$.

c) Find the limiting distribution of $\sqrt{n} \left[ (\bar{X}_n)^2 - c \right]$ for an appropriate constant $c$. 

8.12. Let \( Y_1, \ldots, Y_n \) be iid Poisson (\( \lambda \)) random variables.
   a) Find the limiting distribution of \( \sqrt{n} \left( Y_n - \lambda \right) \).
   b) Find the limiting distribution of \( \sqrt{n} \left[ (Y_n)^2 - c \right] \) for an appropriate constant \( c \).

8.13. Let \( Y_n \sim \chi^2_n \).
   a) Find the limiting distribution of \( \sqrt{n} \left( \frac{Y_n}{n} - 1 \right) \).
   b) Find the limiting distribution of \( \sqrt{n} \left[ \left( \frac{Y_n}{n} \right)^3 - 1 \right] \).

8.14. Let \( X_1, \ldots, X_n \) be iid with cdf \( F(x) = P(X \leq x) \). Let \( Y_i = I(X_i \leq x) \) where the indicator equals 1 if \( X_i \leq x \) and 0, otherwise.
   a) Find \( E(Y_i) \).
   b) Find \( \text{VAR}(Y_i) \).
   c) Let \( \hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) \) for some fixed real number \( x \). Find the limiting distribution of \( \sqrt{n} \left( \hat{F}_n(x) - c_x \right) \) for an appropriate constant \( c_x \).

8.15. Suppose \( X_n \) has cdf
   \[
   F_n(x) = 1 - \left( 1 - \frac{x}{\theta_n} \right)^n
   \]
   for \( x \geq 0 \) and \( F_n(x) = 0 \) for \( x < 0 \). Show that \( X_n \xrightarrow{D} X \) by finding the cdf of \( X \).

8.16. Let \( X_n \) be a sequence of random variables such that
   \( P(X_n = 1/n) = 1 \). Does \( X_n \) converge in distribution? If yes, prove it by finding \( X \) and the cdf of \( X \). If no, prove it.

8.17. Suppose that \( Y_1, \ldots, Y_n \) are iid with \( E(Y) = (1-\rho)/\rho \) and \( \text{VAR}(Y) = (1-\rho)/\rho^2 \) where \( 0 < \rho < 1 \).
   a) Find the limiting distribution of
   \[
   \sqrt{n} \left( Y_n - \frac{1-\rho}{\rho} \right).
   \]
b) Find the limiting distribution of $\sqrt{n} \left[ g(Y_n) - \rho \right]$ for appropriate function $g$.

8.18. Let $X_n \sim \text{Binomial}(n, p)$ where the positive integer $n$ is large and $0 < p < 1$.

a) Find the limiting distribution of $\sqrt{n} \left( \frac{X_n}{n} - p \right)$.

b) Find the limiting distribution of $\sqrt{n} \left( \left( \frac{X_n}{n} \right)^2 - p^2 \right)$.

8.19. Let $Y_1, \ldots, Y_n$ be iid exponential ($\lambda$) so that $E(Y) = \lambda$ and $\text{MED}(Y) = \log(2)\lambda$.

a) Let $T_{1,n} = \log(2)Y$ and find the limiting distribution of $\sqrt{n}(T_{1,n} - \log(2)\lambda)$.

b) Let $T_{2,n} = \text{MED}(n)$ be the sample median and find the limiting distribution of $\sqrt{n}(T_{2,n} - \log(2)\lambda)$.

c) Find $\text{ARE}(T_{1,n}, T_{2,n})$.

8.20. Suppose that $\eta = g(\theta)$, $\theta = g^{-1}(\eta)$ and $g'(\theta) > 0$ exists. If $X$ has pdf or pmf $f(x|\theta)$, then in terms of $\eta$, the pdf or pmf is $f^*(x|\eta) = f(x|g^{-1}(\eta))$. Now

$$A = \frac{\partial}{\partial \eta} \log[f(x|g^{-1}(\eta))] = \frac{1}{f(x|g^{-1}(\eta))} \frac{\partial}{\partial \eta} f(x|g^{-1}(\eta)) = \left[ \frac{1}{f(x|g^{-1}(\eta))} \right] \left[ \frac{\partial}{\partial \theta} f(x|\theta) \right]_{\theta = g^{-1}(\eta)} \left[ \frac{\partial}{\partial \eta} g^{-1}(\eta) \right]$$

using the chain rule twice. Since $\theta = g^{-1}(\eta)$,

$$A = \left[ \frac{1}{f(x|\theta)} \right] \left[ \frac{\partial}{\partial \theta} f(x|\theta) \right] \left[ \frac{\partial}{\partial \eta} g^{-1}(\eta) \right].$$

Hence

$$A = \frac{\partial}{\partial \eta} \log[f(x|g^{-1}(\eta))] = \left[ \frac{\partial}{\partial \theta} \log[f(x|\theta)] \right] \left[ \frac{\partial}{\partial \eta} g^{-1}(\eta) \right].$$

Now show that

$$I_1^*(\eta) = \frac{I_1(\theta)}{[g'(\theta)]^2}.$$
8.21. Let $Y_1, \ldots, Y_n$ be iid exponential (1) so that $P(Y \leq y) = F(y) = 1 - e^{-y}$ for $y \geq 0$. Let $Y_{(n)} = \max(Y_1, \ldots, Y_n)$.

a) Show that $F_{Y_{(n)}}(t) = P(Y_{(n)} \leq t) = [1 - e^{-t}]^n$ for $t \geq 0$.

b) Show that $P(Y_{(n)} - \log(n) \leq t) \rightarrow \exp(-e^{-t})$ (for all $t \in (-\infty, \infty)$ since $t + \log(n) > 0$ implies $t \in \mathbb{R}$ as $n \rightarrow \infty$).

8.22. Let $Y_1, \ldots, Y_n$ be iid uniform (0, 2). 

a) Find the limiting distribution of $\sqrt{n}(\bar{Y} - \theta)$.

b) Let $T_{1,n} = \bar{Y}$ and find the limiting distribution of $\sqrt{n}(T_{1,n} - \theta)$.

c) Find $\text{ARE}(T_{1,n}, T_{2,n})$. Which estimator is better, asymptotically?

8.23. Suppose that $Y_1, \ldots, Y_n$ are iid from a distribution with pdf $f(y|\theta)$ and that the integral and differentiation operators of all orders can be interchanged (eg the data is from a one parameter exponential family).

a) Show that $0 = E\left[\frac{\partial}{\partial \theta} \log(f(Y|\theta))\right]$ by showing that

$$0 = \frac{\partial}{\partial \theta} \int f(y|\theta) dy = \int \left[\frac{\partial}{\partial \theta} \log(f(y|\theta))\right] f(y|\theta) dy.$$ 

b) Take 2nd derivatives of (*) to show that

$$I_1(\theta) = E_{\theta}[\left(\frac{\partial}{\partial \theta} \log(f(Y|\theta))\right)^2] = -E_{\theta}\left[\frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta))\right].$$

8.24. Suppose that $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$.

a) Find the limiting distribution of $\sqrt{n} \left(\bar{X}_n - \mu\right)$.

b) Let $g(\theta) = [\log(1 + \theta)]^2$. Find the limiting distribution of $\sqrt{n} \left(g(\bar{X}_n) - g(\mu)\right)$ for $\mu > 0$.

c) Let $g(\theta) = [\log(1 + \theta)]^2$. Find the limiting distribution of $\sqrt{n} \left(g(\bar{X}_n) - g(\mu)\right)$ for $\mu = 0$. Hint: Use Theorem 8.29.

8.25. Let $W_n = X_n - X$ and let $r > 0$. Notice that for any $\epsilon > 0$,

$$E|X_n - X|^r \geq E[|X_n - X|^r I(|X_n - X| \geq \epsilon)] \geq \epsilon^r P(|X_n - X| \geq \epsilon).$$

Show that $W_n \overset{P}{\rightarrow} 0$ if $E|X_n - X|^r \rightarrow 0$ as $n \rightarrow \infty$. 
8.26. Let $X_1, \ldots, X_n$ be iid with $E(X) = \mu$ and $V(X) = \sigma^2$. What is the limiting distribution of $n[(\bar{X})^2 - \mu^2]$ for the value or values of $\mu$ where the delta method does not apply? Hint: use Theorem 8.29.

8.27. (Sept. 05 Qual) Let $X \sim \text{Binomial}(n, p)$ where the positive integer $n$ is large and $0 < p < 1$.

a) Find the limiting distribution of $\sqrt{n} \left( \frac{X}{n} - p \right)$.

b) Find the limiting distribution of $\sqrt{n} \left[ \left( \frac{X}{n} \right)^2 - p^2 \right]$.

c) Show how to find the limiting distribution of $\left[ \left( \frac{X}{n} \right)^3 - \frac{X}{n} \right]$ when $p = \frac{1}{\sqrt{3}}$.

(Actually want the limiting distribution of $n \left( \left[ \left( \frac{X}{n} \right)^3 - \frac{X}{n} \right] - g(p) \right)$ where $g(\theta) = \theta^3 - \theta$.)

8.28. (Aug. 04 Qual) Let $X_1, \ldots, X_n$ be independent and identically distributed (iid) from a Poisson($\lambda$) distribution.

a) Find the limiting distribution of $\sqrt{n} \left( \bar{X} - \lambda \right)$.

b) Find the limiting distribution of $\sqrt{n} \left[ (\bar{X})^3 - (\lambda)^3 \right]$.

8.29. (Jan. 04 Qual) Let $X_1, \ldots, X_n$ be iid from a normal distribution with unknown mean $\mu$ and known variance $\sigma^2$. Let $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

a) Show that $\bar{X}$ and $S^2$ are independent.

b) Find the limiting distribution of $\sqrt{n}((\bar{X})^3 - c)$ for an appropriate constant $c$. 

8.30. Suppose that \( Y_1, \ldots, Y_n \) are iid logistic(\( \theta, 1 \)) with pdf
\[
f(y) = \frac{\exp(-(y - \theta))}{[1 + \exp(-(y - \theta))]^2}
\]
where and \( y \) and \( \theta \) are real.

a) \( I_1(\theta) = 1/3 \) and the family is regular so the “standard limit theorem” for the MLE \( \hat{\theta}_n \) holds. Using this standard theorem, what is the limiting distribution of \( \sqrt{n}(\hat{\theta}_n - \theta) \)?

b) Find the limiting distribution of \( \sqrt{n}(\bar{Y}_n - \theta) \).

c) Find the limiting distribution of \( \sqrt{n}(MED(n) - \theta) \).

d) Consider the estimators \( \hat{\theta}_n, \bar{Y}_n \) and MED(\( n \)). Which is the best estimator and which is the worst?

8.31. Let \( Y_n \sim \text{binomial}(n, p) \). Find the limiting distribution of
\[
\sqrt{n} \left( \arcsin \left( \frac{\bar{Y}_n}{n} \right) - \arcsin(\sqrt{p}) \right).
\]
(Hint: \( \frac{d}{dx} \arcsin(x) = \frac{1}{\sqrt{1 - x^2}} \).)

8.32. Suppose \( Y_n \sim \text{uniform}(-n,n) \). Let \( F_n(y) \) be the cdf of \( Y_n \).

a) Find \( F(y) \) such that \( F_n(y) \to F(y) \) for all \( y \) as \( n \to \infty \).

b) Does \( Y_n \xrightarrow{D} Y? \) Explain briefly.

8.33. Suppose \( Y_n \sim \text{uniform}(0,n) \). Let \( F_n(y) \) be the cdf of \( Y_n \).

a) Find \( F(y) \) such that \( F_n(y) \to F(y) \) for all \( y \) as \( n \to \infty \).

b) Does \( Y_n \xrightarrow{D} Y? \) Explain briefly.

8.34. (abc Jan. 2010 Qual): Let \( Y_1, \ldots, Y_n \) be independent and identically distributed (iid) from a distribution with probability mass function \( f(y) = \rho(1 - \rho)^y \) for \( y = 0, 1, 2, \ldots \) and \( 0 < \rho < 1 \). Then \( E(Y) = (1 - \rho)/\rho \) and \( \text{VAR}(Y) = (1 - \rho)/\rho^2 \).

a) Find the limiting distribution of \( \sqrt{n} \left( \bar{Y} - \frac{1 - \rho}{\rho} \right) \).

b) Show how to find the limiting distribution of \( g(\bar{Y}) = \frac{1}{1 + \bar{Y}} \). Deduc it completely. (This bad notation means find the limiting distribution of \( \sqrt{n}(g(\bar{Y}) - c) \) for some constant \( c \).)
c) Find the method of moments estimator of $\rho$.

d) Find the limiting distribution of $\sqrt{n} \left( (1 + \bar{Y}) - d \right)$ for appropriate constant $d$.

e) Note that $1 + E(Y) = 1/\rho$. Find the method of moments estimator of $1/\rho$.

8.35. (Sept. 2010 Qual): Let $X_1, ..., X_n$ be independent identically distributed random variables from a normal distribution with mean $\mu$ and variance $\sigma^2$.

a) Find the approximate distribution of $1/\bar{X}$. Is this valid for all values of $\mu$?

b) Show that $1/\bar{X}$ is asymptotically efficient for $1/\mu$, provided $\mu \neq \mu^*$. Identify $\mu^*$.

8.36. (Jan. 2011 Qual): Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a distribution with probability density function

$$f(y) = \frac{2y}{\theta^2}$$

for $0 < y \leq \theta$ and $f(y) = 0$, otherwise.

a) Find the limiting distribution of $\sqrt{n} \left( \bar{Y} - c \right)$ for appropriate constant $c$.

b) Find the limiting distribution of $\sqrt{n} \left( \log(\bar{Y}) - d \right)$ for appropriate constant $d$.

c) Find the method of moments estimator of $\theta^k$. 


Chapter 9
Confidence Intervals

9.1 Introduction

Definition 9.1. Let the data $Y_1, \ldots, Y_n$ have joint pdf or pmf $f(y|\theta)$ with parameter space $\Theta$ and support $\mathcal{Y}$. Let $L_n(Y)$ and $U_n(Y)$ be statistics such that $L_n(y) \leq U_n(y), \forall y \in \mathcal{Y}$. Then $(L_n(y), U_n(y))$ is a 100 $(1 - \alpha)$ % confidence interval (CI) for $\theta$ if

$$P_\theta(L_n(Y) < \theta < U_n(Y)) = 1 - \alpha$$

for all $\theta \in \Theta$. The interval $(L_n(y), U_n(y))$ is a large sample 100 $(1 - \alpha)$ % CI for $\theta$ if

$$P_\theta(L_n(Y) < \theta < U_n(Y)) \rightarrow 1 - \alpha$$

for all $\theta \in \Theta$ as $n \rightarrow \infty$.

Definition 9.2. Let the data $Y_1, \ldots, Y_n$ have joint pdf or pmf $f(y|\theta)$ with parameter space $\Theta$ and support $\mathcal{Y}$. The random variable $R(Y|\theta)$ is a pivot or pivotal quantity if the distribution of $R(Y|\theta)$ is independent $\theta$. The quantity $R(Y, \theta)$ is an asymptotic pivot or asymptotic pivotal quantity if the limiting distribution of $R(Y, \theta)$ is independent of $\theta$.

The first CI in Definition 9.1 is sometimes called an exact CI. The words “exact” and “large sample” are often omitted. In the following definition, the scaled asymptotic length is closely related to asymptotic relative efficiency of an estimator and high power of a test of hypotheses.
Definition 9.3. Let \((L_n, U_n)\) be a 100 \((1 - \alpha)\)% CI or large sample CI for \(\theta\). If

\[
n^\delta(U_n - L_n) \xrightarrow{P} A_\alpha,
\]

then \(A_\alpha\) is the \textit{scaled asymptotic length} of the CI. Typically \(\delta = 0.5\) but superefficient CIs have \(\delta = 1\). For fixed \(\delta\) and fixed coverage \(1 - \alpha\), a CI with smaller \(A_\alpha\) is "better" than a CI with larger \(A_\alpha\). If \(A_{1,\alpha}\) and \(A_{2,\alpha}\) are for two competing CIs with the same \(\delta\), then \((A_{2,\alpha}/A_{1,\alpha})^{1/\delta}\) is a measure of "asymptotic relative efficiency."

Definition 9.4. Suppose a nominal 100\((1 - \alpha)\)% CI for \(\theta\) has actual coverage \(1 - \delta\), so that

\[
P_\theta(L_n(Y) < \theta < U_n(Y)) = 1 - \delta \quad \text{for all } \theta \in \Theta.
\]

If \(1 - \delta > 1 - \alpha\), then the CI is \textit{conservative}. If \(1 - \delta < 1 - \alpha\), then the CI is \textit{liberal}. Conservative CIs are generally considered better than liberal CIs. Suppose a nominal 100\((1 - \alpha)\)% large sample CI for \(\theta\) has actual coverage \(1 - \delta_n\), where \(\delta_n \to \delta\) as \(n \to \infty\) for all \(\theta \in \Theta\). If \(1 - \delta > 1 - \alpha\), then the CI is \textit{asymptotically conservative}. If \(1 - \delta < 1 - \alpha\), then the CI is \textit{asymptotically liberal}. It is possible that \(\delta \equiv \delta(\theta)\) depends on \(\theta\), and that the CI is (asymptotically) conservative or liberal for different values of \(\theta\), in that the (asymptotic) coverage is higher or lower than the nominal coverage, depending on \(\theta\).

Example 9.1. a) Let \(Y_1, ..., Y_n\) be iid \(N(\mu, \sigma^2)\) where \(\sigma^2 > 0\). Then

\[
R(Y | \mu, \sigma^2) = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}
\]

is a pivot or pivotal quantity.

To use this pivot to find a CI for \(\mu\), let \(t_{p,\alpha}\) be the \(\alpha\) percentile of the \(t_p\) distribution. Hence \(P(T \leq t_{p,\alpha}) = \alpha\) if \(T \sim t_p\). Using \(t_{p,\alpha} = -t_{p,1-\alpha}\) for \(0 < \alpha < 0.5\), note that

\[
1 - \alpha = P(-t_{n-1,1-\alpha/2} \leq \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq t_{n-1,1-\alpha/2}) =
\]

\[
P(-t_{n-1,1-\alpha/2} S/\sqrt{n} \leq \bar{Y} - \mu \leq t_{n-1,1-\alpha/2} S/\sqrt{n}) =
\]

\[
P(-t_{n-1,1-\alpha/2} S/\sqrt{n} \leq \bar{Y} - \mu \leq t_{n-1,1-\alpha/2} S/\sqrt{n}) =
\]

\[
P(\bar{Y} - t_{n-1,1-\alpha/2} S/\sqrt{n} \leq \mu \leq \bar{Y} + t_{n-1,1-\alpha/2} S/\sqrt{n}).
\]
Thus
\[ \bar{Y} \pm t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}} \]
is a 100\((1 - \alpha)\)% CI for \(\mu\).

b) If \(Y_1, \ldots, Y_n\) are iid with \(E(Y) = \mu\) and \(\text{VAR}(Y) = \sigma^2 > 0\), then, by the CLT and Slutsky’s Theorem,
\[ R(Y|\mu, \sigma^2) = \frac{\bar{Y} - \mu}{S/\sqrt{n}} \overset{D}{\rightarrow} N(0, 1) \]
is an asymptotic pivot or asymptotic pivotal quantity.

To use this asymptotic pivot to find a large sample CI for \(\mu\), let \(z_{\alpha}\) be the \(\alpha\) percentile of the \(N(0, 1)\) distribution. Hence \(P(Z \leq z_{\alpha}) = \alpha\) if \(Z \sim N(0, 1)\). Using \(z_{\alpha} = -z_{1-\alpha}\) for \(0 < \alpha < 0.5\), note that for large \(n\),
\[ 1 - \alpha \approx P(-z_{1-\alpha/2} \leq \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}) = \]
\[ P(-z_{1-\alpha/2} \frac{S}{\sqrt{n}} \leq \bar{Y} - \mu \leq z_{1-\alpha/2} \frac{S}{\sqrt{n}}) = \]
\[ P(-\bar{Y} - z_{1-\alpha/2} \frac{S}{\sqrt{n}} \leq -\mu \leq -\bar{Y} + z_{1-\alpha/2} \frac{S}{\sqrt{n}}) = \]
\[ P(\bar{Y} - z_{1-\alpha/2} S/\sqrt{n} \leq \mu \leq \bar{Y} + z_{1-\alpha/2} S/\sqrt{n}). \]

Thus
\[ \bar{Y} \pm z_{1-\alpha/2} S/\sqrt{n} \]
is a large sample 100\((1 - \alpha)\)% CI for \(\mu\).

Since \(t_{n-1,1-\alpha/2} > z_{1-\alpha/2}\) but \(t_{n-1,1-\alpha/2} \to z_{1-\alpha/2}\) as \(n \to \infty\),
\[ \bar{Y} \pm t_{n-1,1-\alpha/2} S/\sqrt{n} \]
is also a large sample 100\((1 - \alpha)\)% CI for \(\mu\). This \(t\) interval is the same as that in a), and is the most widely used confidence interval in statistics. Replacing \(z_{1-\alpha/2}\) by \(t_{n-1,1-\alpha/2}\) makes the CI longer and hence less likely to be liberal.

Large sample theory can be used to find a CI from the asymptotic pivot. Suppose that \(Y = (Y_1, \ldots, Y_n)\) and that \(W_n = W_n(Y)\) is an estimator of some parameter \(\mu_W\) such that
\[ \sqrt{n}(W_n - \mu_W) \overset{D}{\rightarrow} N(0, \sigma_W^2) \]
where $\sigma^2_W/n$ is the asymptotic variance of the estimator $W_n$. The above notation means that if $n$ is large, then for probability calculations

$$W_n - \mu_W \approx N(0, \sigma^2_W/n).$$

Suppose that $S^2_W$ is a consistent estimator of $\sigma^2_W$ so that the (asymptotic) standard error of $W_n$ is $\text{SE}(W_n) = S_W/\sqrt{n}$. As in Example 9.1, let $P(Z \leq z_\alpha) = \alpha$ if $Z \sim N(0,1)$. Then for large $n$

$$1 - \alpha \approx P(-z_{1-\alpha/2} \leq \frac{W_n - \mu_W}{\text{SE}(W_n)} \leq z_{1-\alpha/2}),$$

and an approximate or large sample 100(1 - $\alpha$)% CI for $\mu_W$ is given by

$$(W_n - z_{1-\alpha/2}\text{SE}(W_n), W_n + z_{1-\alpha/2}\text{SE}(W_n)). \tag{9.1}$$

Since

$$\frac{t_{p,1-\alpha/2}}{z_{1-\alpha/2}} \to 1$$

if $p \equiv p_n \to \infty$ as $n \to \infty$, another large sample 100(1 - $\alpha$)% CI for $\mu_W$ is

$$(W_n - t_{p,1-\alpha/2}\text{SE}(W_n), W_n + t_{p,1-\alpha/2}\text{SE}(W_n)). \tag{9.2}$$

The CI (9.2) often performs better than the CI (9.1) in small samples. The quantity $t_{p,1-\alpha/2}/z_{1-\alpha/2}$ can be regarded as a small sample correction factor. The CI (9.2) is longer than the CI (9.1). Hence the CI (9.2) more conservative than the CI (9.1).

Suppose that there are two independent samples $Y_1, ..., Y_n$ and $X_1, ..., X_m$ and that

$$
\left( \frac{\sqrt{n}(W_n(Y) - \mu_W(Y))}{\sqrt{m}(W_m(X) - \mu_W(X))} \right) \to N_2\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y) & 0 \\ 0 & \sigma^2_W(X) \end{pmatrix} \right).
\tag{\text{9.3}}
$$

Then

$$
\begin{pmatrix} (W_n(Y) - \mu_W(Y)) \\ (W_m(X) - \mu_W(X)) \end{pmatrix} \approx N_2\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_W(Y)/n & 0 \\ 0 & \sigma^2_W(X)/m \end{pmatrix} \right),
\tag{\text{9.4}}
$$

and

$$W_n(Y) - W_m(X) - (\mu_W(Y) - \mu_W(X)) \approx N(0, \frac{\sigma^2_W(Y)}{n} + \frac{\sigma^2_W(X)}{m}).$$

(9.5)
Hence $SE(W_n(Y) - W_m(X)) = \sqrt{S_W(Y)/n + S_W(X)/m} = \sqrt{[SE(W_n(Y))]^2 + [SE(W_m(X))]^2}$,
and the large sample $100(1-\alpha)$% CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(Y) - W_m(X)) \pm z_{1-\alpha/2} SE(W_n(Y) - W_m(X)).$$

(9.3)

If $p_n$ is the degrees of freedom used for a single sample procedure when the sample size is $n$, let $p = \min(p_n, p_m)$. Then another large sample $100(1-\alpha)$% CI for $\mu_W(Y) - \mu_W(X)$ is given by

$$(W_n(Y) - W_m(X)) \pm t_{p,1-\alpha/2} SE(W_n(Y) - W_m(X)).$$

(9.4)

These CIs are known as Welch intervals. See Welch (1937) and Yuen (1974).

**Example 9.2.** Consider the single sample procedures where $W_n = \bar{Y}_n$. Then $\mu_W = E(Y)$, $\sigma^2_W = \text{VAR}(Y)$, $S_W = S_n$, and $p = n - 1$. Let $t_p$ denote a random variable with a $t$ distribution with $p$ degrees of freedom and let the $\alpha$ percentile $t_{p,\alpha}$ satisfy $P(t_p \leq t_{p,\alpha}) = \alpha$. Then the classical $t$-interval for $\mu \equiv E(Y)$ is

$$\bar{Y}_n \pm t_{n-1,1-\alpha/2} \frac{S_n}{\sqrt{n}}$$

and the $t$-test statistic for $H_o : \mu = \mu_o$ is

$$t_0 = \frac{\bar{Y} - \mu_o}{S_n/\sqrt{n}}.$$

The right tailed p-value is given by $P(t_{n-1} > t_0)$.

Now suppose that there are two samples where $W_n(Y) = \bar{Y}_n$ and $W_m(X) = \bar{X}_m$. Then $\mu_W(Y) = E(Y) \equiv \mu_Y$, $\mu_W(X) = E(X) \equiv \mu_X$, $\sigma^2_W(Y) = \text{VAR}(Y) \equiv \sigma^2_Y$, $\sigma^2_W(X) = \text{VAR}(X) \equiv \sigma^2_X$, and $p_n = n - 1$. Let $p = \min(n - 1, m - 1)$. Since

$$SE(W_n(Y) - W_m(X)) = \sqrt{\frac{S_n^2(Y)}{n} + \frac{S_m^2(X)}{m}},$$

the two sample $t$-interval for $\mu_Y - \mu_X$

$$(\bar{Y}_n - \bar{X}_m) \pm t_{p,1-\alpha/2} \sqrt{\frac{S_n^2(Y)}{n} + \frac{S_m^2(X)}{m}}$$
and two sample t-test statistic

\[ t_o = \frac{\overline{Y}_n - \overline{X}_m}{\sqrt{\frac{s^2(Y)}{n}} + \frac{s^2(X)}{m}}. \]

The right tailed p-value is given by \( P(t_p > t_o) \). For sample means, values of the degrees of freedom that are more accurate than \( p = \min(n - 1, m - 1) \) can be computed. See Moore (2007, p. 474).

The remainder of this section follows Olive (2008b, Section 2.4) closely. Let \( \lfloor x \rfloor \) denote the “greatest integer function” (so \( \lfloor 7.7 \rfloor = 7 \)). Let \( \lceil x \rceil \) denote the smallest integer greater than or equal to \( x \) (so \( \lceil 7.7 \rceil = 8 \)).

**Example 9.3: inference with the sample median.** Let \( U_n = n - L_n \) where \( L_n = \lfloor n/2 \rfloor - \lceil \sqrt{n}/4 \rceil \) and use

\[ SE(MED(n)) = 0.5(Y_{(U_n)} - Y_{(L_n+1)}). \]

Let \( p = U_n - L_n - 1 \). Then a large sample 100(1 - \( \alpha \))% confidence interval for the population median \( MED(Y) \) is

\[ MED(n) \pm t_{p,1-\alpha/2}SE(MED(n)). \] (9.5)

**Example 9.4: inference with the trimmed mean.** The symmetrically trimmed mean or the \( \delta \) trimmed mean

\[ T_n = T_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} Y_{(i)} \] (9.6)

where \( L_n = \lfloor n\delta \rfloor \) and \( U_n = n - L_n \). If \( \delta = 0.25 \), say, then the \( \delta \) trimmed mean is called the 25% trimmed mean.

The trimmed mean is estimating a truncated mean \( \mu_T \). Assume that \( Y \) has a probability density function \( f_Y(y) \) that is continuous and positive on its support. Let \( y_\delta \) be the number satisfying \( P(Y \leq y_\delta) = \delta \). Then

\[ \mu_T = \frac{1}{1 - 2\delta} \int_{y_\delta}^{y_\delta - \delta} yf_Y(y)dy. \] (9.7)

Notice that the 25% trimmed mean is estimating

\[ \mu_T = \int_{y_{0.25}}^{y_{0.75}} 2yf_Y(y)dy. \]
To perform inference, find \( d_1, \ldots, d_n \) where
\[
d_i = \begin{cases} 
Y_{(L_n+1)}, & i \leq L_n \\
Y_{(i)}, & L_n + 1 \leq i \leq U_n \\
Y_{(U_n)}, & i \geq U_n + 1.
\end{cases}
\]
Then the Winsorized variance is the sample variance \( S_n^2(d_1, \ldots, d_n) \) of \( d_1, \ldots, d_n \), and the scaled Winsorized variance
\[
V_{SW}(L_n, U_n) = \frac{S_n^2(d_1, \ldots, d_n)}{(U_n - L_n)^2/n}.
\] (9.8)
The standard error of \( T_n \) is \( SE(T_n) = \sqrt{V_{SW}(L_n, U_n)/n} \).

A large sample 100 \((1 - \alpha)\)% confidence interval (CI) for \( \mu_T \) is
\[
T_n \pm t_{p,1-\frac{\alpha}{2}} SE(T_n)
\] (9.9)
where \( P(t_p \leq t_{p,1-\frac{\alpha}{2}}) = 1 - \alpha/2 \) if \( t_p \) is from a \( t \) distribution with \( p = U_n - L_n - 1 \) degrees of freedom. This interval is the classical \( t \)-interval when \( \delta = 0 \), but \( \delta = 0.25 \) gives a robust CI.

Example 9.5. Suppose the data below is from a symmetric distribution with mean \( \mu \). Find a 95% CI for \( \mu \).

6, 9, 9, 7, 8, 9, 9, 7

Solution. When computing small examples by hand, the steps are to sort the data from smallest to largest value, find \( n, L_n, U_n, Y_{(L_n)}, Y_{(U_n)}, p, MED(n) \) and \( SE(MED(n)) \). After finding \( t_{p,1-\alpha/2} \), plug the relevant quantities into the formula for the CI. The sorted data are 6, 7, 7, 8, 9, 9, 9, 9. Thus \( MED(n) = (8 + 9)/2 = 8.5 \). Since \( n = 8 \), \( L_n = \lfloor 4 \rfloor - \lceil \sqrt{2} \rceil = 4 - 2 = 2 \) and \( U_n = n - L_n = 8 - 2 = 6 \). Hence \( SE(MED(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 \ast (9 - 7) = 1 \). The degrees of freedom \( p = U_n - L_n - 1 = 6 - 2 - 1 = 3 \). The cutoff \( t_{3,0.975} = 3.182 \). Thus the 95% CI for MED(Y) is
\[
MED(n) \pm t_{3,0.975} SE(MED(n))
\]
\[= 8.5 \pm 3.182(1) = (5.318, 11.682) \]. The classical \( t \)-interval uses \( \bar{Y} = (6 + 7 + \ldots + 9 + 9 + 9 + 9 + 9)/8 \) and \( S_n^2 = (1/7)[(\sum_{i=1}^{n} Y_i^2) - 8(8^2)] = (1/7)[522 - 8(64)] = 10/7 \approx 1.4286 \), and \( t_{7,0.975} \approx 2.365 \). Hence the 95% CI for \( \mu \) is \( 8 \pm 2.365(\sqrt{1.4286}/8) = (7.001, 8.999) \). Notice that the \( t \)-cutoff = 2.365 for the classical interval is less than the \( t \)-cutoff = 3.182 for the median interval and that \( SE(\bar{Y}) < SE(MED(n)) \).
Example 9.6. In the last example, what happens if the 6 becomes 66 and a 9 becomes 99?

Solution. Then the ordered data are 7, 7, 8, 9, 9, 9, 66, 99. Hence \( \text{MED}(n) = 9 \). Since \( L_n \) and \( U_n \) only depend on the sample size, they take the same values as in the previous example and \( SE(\text{MED}(n)) = 0.5(Y_{(6)} - Y_{(3)}) = 0.5 \times (9 - 8) = 0.5 \). Hence the 95% CI for \( \text{MED}(Y) \) is \( \text{MED}(n) \pm t_{3.0.975}SE(\text{MED}(n)) = 9 \pm 3.182(0.5) = (7.409, 10.591) \). Notice that with discrete data, it is possible to drive \( SE(\text{MED}(n)) \) to 0 with a few outliers if \( n \) is small. The classical confidence interval \( \bar{Y} \pm t_{7.0.975}S/\sqrt{n} \) blows up and is equal to \((-2.955, 56.455)\).

Example 9.7. The Buxton (1920) data contains 87 heights of men, but five of the men were recorded to be about 0.75 inches tall! The mean height is \( \bar{Y} = 1598.862 \) and the classical 95% CI is \((1514.206, 1683.518)\). \( \text{MED}(n) = 1693.0 \) and the resistant 95% CI based on the median is \((1678.517, 1707.483)\). The 25% trimmed mean \( T_n = 1689.689 \) with 95% CI \((1672.096, 1707.282)\).

The heights for the five men were recorded under their head lengths, so the outliers can be corrected. Then \( \bar{Y} = 1692.356 \) and the classical 95% CI is \((1678.595, 1706.118)\). Now \( \text{MED}(n) = 1694.0 \) and the 95% CI based on the median is \((1678.403, 1709.597)\). The 25% trimmed mean \( T_n = 1693.200 \) with 95% CI \((1676.259, 1710.141)\). Notice that when the outliers are corrected, the three intervals are very similar although the classical interval length is slightly shorter. Also notice that the outliers roughly shifted the median confidence interval by about 1 mm while the outliers greatly increased the length of the classical t–interval.

9.2 Some Examples

Example 9.8. Suppose that \( Y_1, ..., Y_n \) are iid from a one parameter exponential family with parameter \( \tau \). Assume that \( T_n = \sum_{i=1}^{n} t(Y_i) \) is a complete sufficient statistic. Then from Theorems 3.6 and 3.7, often \( T_n \sim G(na, 2b \tau) \) where \( a \) and \( b \) are known positive constants. Then
\[
\hat{\tau} = \frac{T_n}{2nab}
\]
is the UMVUE and often the MLE of \( \tau \). Since \( T_n/(b \tau) \sim G(na, 2) \), a
100(1 − α)% confidence interval for τ is

\[
\left( \frac{T_n/b}{G(na, 2, 1 - \alpha/2)}, \frac{T_n/b}{G(na, 2, \alpha/2)} \right) \approx \left( \frac{T_n/b}{\chi_d^2(1 - \alpha/2)}, \frac{T_n/b}{\chi_d^2(\alpha/2)} \right)
\] (9.10)

where \( d = \lfloor 2na \rfloor \), \( \lfloor x \rfloor \) is the greatest integer function (e.g. \( \lfloor 7.7 \rfloor = 7 \)), \( P[G \leq G(\nu, \lambda, \alpha)] = \alpha \) if \( G \sim G(\nu, \lambda) \), and \( P[X \leq \chi_d^2(\alpha)] = \alpha \) if \( X \) has a chi-square \( \chi_d^2 \) distribution with \( d \) degrees of freedom.

This confidence interval can be inverted to perform two tail tests of hypotheses. By Theorem 7.3, if \( w(\theta) \) is increasing, then the uniformly most powerful (UMP) test of \( H_o: \tau \leq \tau_o \) versus \( H_A: \tau > \tau_o \) rejects \( H_o \) if and only if \( T_n > k \) where \( P[G > k] = \alpha \) when \( G \sim G(na, 2b \tau_o) \). Hence

\[
k = G(na, 2b \tau_o, 1 - \alpha). \tag{9.11}
\]

A good approximation to this test rejects \( H_o \) if and only if

\[
T_n > b \tau_o \chi_d^2(1 - \alpha)
\]

where \( d = \lfloor 2na \rfloor \).

**Example 9.9.** If \( Y \) is half normal \( \text{HN}(\mu, \sigma) \) then the pdf of \( Y \) is

\[
f(y) = \frac{2}{\sqrt{2\pi} \sigma} \exp\left( -\frac{(y - \mu)^2}{2\sigma^2} \right)
\]

where \( \sigma > 0 \) and \( y > \mu \) and \( \mu \) is real. Since

\[
f(y) = \frac{2}{\sqrt{2\pi} \sigma} \mathbb{I}[y > \mu] \exp\left( \frac{-1}{2\sigma^2}(y - \mu)^2 \right),
\]

\( Y \) is a 1P–REF if \( \mu \) is known.

Since \( T_n = \sum(Y_i - \mu)^2 \sim G(n/2, 2\sigma^2) \), in Example 9.8 take \( a = 1/2 \), \( b = 1 \), \( d = n \) and \( \tau = \sigma^2 \). Then a 100(1 − α)% confidence interval for \( \sigma^2 \) is

\[
\left( \frac{T_n}{\chi_n^2(1 - \alpha/2)}, \frac{T_n}{\chi_n^2(\alpha/2)} \right) \tag{9.12}
\]

The UMP test of \( H_o: \sigma^2 \leq \sigma_o^2 \) versus \( H_A: \sigma^2 > \sigma_o^2 \) rejects \( H_o \) if and only if

\[
T_n/\sigma_o^2 > \chi_n^2(1 - \alpha).
\]
CHAPTER 9. CONFIDENCE INTERVALS

Now consider inference when both \( \mu \) and \( \sigma \) are unknown. Then the family is no longer an exponential family since the support depends on \( \mu \). Let

\[
D_n = \sum_{i=1}^{n} (Y_i - Y_{1:n})^2. \tag{9.13}
\]

Pewsey (2002) showed that \((\hat{\mu}, \hat{\sigma}^2) = (Y_{1:n}, 1/n D_n)\) is the MLE of \((\mu, \sigma^2)\), and that

\[
\frac{Y_{1:n} - \mu}{\sigma \Phi^{-1}(\frac{1}{2} + \frac{1}{2n})} \xrightarrow{D} EXP(1).
\]

Since \((\sqrt{\pi/2})/n\) is an approximation to \(\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})\) based on a first order Taylor series expansion such that

\[
\frac{\Phi^{-1}(\frac{1}{2} + \frac{1}{2n})}{(\sqrt{\pi/2})/n} \to 1,
\]

it follows that

\[
\frac{n(Y_{1:n} - \mu)}{\sigma \sqrt{\frac{\pi}{2}}} \xrightarrow{D} EXP(1). \tag{9.14}
\]

Using this fact, it can be shown that a large sample 100(1 - \( \alpha \))% CI for \( \mu \) is

\[
(\hat{\mu} + \hat{\sigma} \log(\alpha) \Phi^{-1}(\frac{1}{2} + \frac{1}{2n}) \left(1 + \frac{13}{n^2}\right), \ \hat{\mu}) \tag{9.15}
\]

where the term \((1+13/n^2)\) is a small sample correction factor. See Abuhassan and Olive (2008).

Note that

\[
D_n = \sum_{i=1}^{n} (Y_i - Y_{1:n})^2 = \sum_{i=1}^{n} (Y_i - \mu + \mu - Y_{1:n})^2 = \sum_{i=1}^{n} (Y_i - \mu)^2 + n(\mu - Y_{1:n})^2 + 2(\mu - Y_{1:n}) \sum_{i=1}^{n} (Y_i - \mu).
\]

Hence

\[
D_n = T_n + \frac{1}{n}[n(Y_{1:n} - \mu)]^2 - 2[n(Y_{1:n} - \mu)] \sum_{i=1}^{n} (Y_i - \mu) / n,
\]
or
\[
\frac{D_n}{\sigma^2} = \frac{T_n}{\sigma^2} + \frac{1}{n\sigma^2} \left[n(Y_{1:n} - \mu)\right]^2 - \frac{2n(Y_{1:n} - \mu)}{\sigma} \sum_{i=1}^{n} (Y_i - \mu) \frac{n}{n\sigma}.
\]

(9.16)

Consider the three terms on the right hand side of (9.16). The middle term converges to 0 in distribution while the third term converges in distribution to a \(-2\text{EXP}(1)\) or \(-\chi^2_2\) distribution since \(\sum_{i=1}^{n} (Y_i - \mu)/(\sigma n)\) is the sample mean of \(\text{HN}(0,1)\) random variables and \(E(X) = \sqrt{2/\pi}\) when \(X \sim HN(0,1)\).

Let \(T_{n-p} = \sum_{i=1}^{n-p} (Y_i - \mu)^2\). Then
\[D_n = T_{n-p} + \sum_{i=n-p+1}^{n} (Y_i - \mu)^2 - V_n\]

(9.17)

where
\[\frac{V_n}{\sigma^2} \xrightarrow{D} \chi^2_2.\]

Hence
\[\frac{D_n}{T_{n-p}} \xrightarrow{D} 1\]

and \(D_n/\sigma^2\) is asymptotically equivalent to a \(\chi^2_{n-p}\) random variable where \(p\) is an arbitrary nonnegative integer. Pewsey (2002) used \(p = 1\).

Thus when both \(\mu\) and \(\sigma^2\) are unknown, a large sample 100(1 - \(\alpha\))\% confidence interval for \(\sigma^2\) is
\[
\left(\frac{D_n}{\chi^2_{n-1}(1 - \alpha/2)}, \frac{D_n}{\chi^2_{n-1}(\alpha/2)}\right).
\]

(9.18)

It can be shown that \(\sqrt{n}\) CI length converges to \(\sigma^2\sqrt{2(z_{1-\alpha/2} - z_{\alpha/2})}\) for CIs (9.12) and (9.18) while \(n\) length CI (9.15) converges to \(-\sigma \log(\alpha)\sqrt{\pi/2}\).

When \(\mu\) and \(\sigma^2\) are unknown, an approximate \(\alpha\) level test of \(H_0 : \sigma^2 \leq \sigma_o^2\) versus \(H_A : \sigma^2 > \sigma_o^2\) that rejects \(H_0\) if and only if
\[D_n/\sigma_o^2 > \chi^2_{n-1}(1 - \alpha)\]

(9.19)

has nearly as much power as the \(\alpha\) level UMP test when \(\mu\) is known if \(n\) is large.
**Example 9.10.** Following Mann, Schafer, and Singpurwalla (1974, p. 176), let $W_1, \ldots, W_n$ be iid $\text{EXP}(\theta, \lambda)$ random variables. Let

$$W_{1:n} = \min(W_1, \ldots, W_n).$$

Then the MLE

$$(\hat{\theta}, \hat{\lambda}) = \left( W_{1:n}, \frac{1}{n} \sum_{i=1}^{n} (W_i - W_{1:n}) \right) = (W_{1:n}, \bar{W} - W_{1:n}).$$

Let $D_n = n\lambda$. For $n > 1$, a 100$(1 - \alpha)$% confidence interval (CI) for $\theta$ is

$$(W_{1:n} - \hat{\lambda}[\alpha^{-1/(n-1)} - 1], W_{1:n}) \quad (9.20)$$

while a 100$(1 - \alpha)$% CI for $\lambda$ is

$$\left( \frac{2D_n}{\chi_{2(n-1),1-\alpha/2}^2}, \frac{2D_n}{\chi_{2(n-1),\alpha/2}^2} \right). \quad (9.21)$$

Let $T_n = \sum_{i=1}^{n} (W_i - \theta) = n(\bar{W} - \theta)$. If $\theta$ is known, then

$$\hat{\lambda}_\theta = \frac{\sum_{i=1}^{n} (W_i - \theta)}{n} = \bar{W} - \theta$$

is the UMVUE and MLE of $\lambda$, and a 100$(1 - \alpha)$% CI for $\lambda$ is

$$\left( \frac{2T_n}{\chi_{2n,1-\alpha/2}^2}, \frac{2T_n}{\chi_{2n,\alpha/2}^2} \right). \quad (9.22)$$

Using $\chi_{n,\alpha}^2 / \sqrt{n} \approx \sqrt{2}z_{\alpha} + \sqrt{n}$, it can be shown that $\sqrt{n}$ CI length converges to $\lambda(z_{1-\alpha/2} - z_{\alpha/2})$ for CIs (9.21) and (9.22) (in probability). It can be shown that $n$ length CI (9.20) converges to $-\lambda\log(\alpha)$.

When a random variable is a simple transformation of a distribution that has an easily computed CI, the transformed random variable will often have an easily computed CI. Similarly the MLEs of the two distributions are often closely related. See the discussion above Example 5.11. The first 3 of the following 4 examples are from Abuhassan and Olive (2008).

**Example 9.11.** If $Y$ has a Pareto distribution, $Y \sim \text{PAR}(\sigma, \lambda)$, then $W = \log(Y) \sim \text{EXP}(\theta = \log(\sigma), \lambda)$. If $\theta = \log(\sigma)$ so $\sigma = e^{\theta}$, then a
100 (1 − α)% CI for θ is (9.20). A 100 (1 − α)% CI for σ is obtained by exponentiating the endpoints of (9.20), and a 100 (1 − α)% CI for λ is (9.21). The fact that the Pareto distribution is a log-location-scale family (W = log(Y) is from a location-scale family) and hence has simple inference does not seem to be well known.

**Example 9.12.** If Y has a power distribution, \( Y \sim POW(\lambda) \), then

\[
W = -\log(Y) \sim EXP(-\log(\tau), \lambda).
\]

A 100 (1 − α)% CI for λ is (9.22).

If Y has a two parameter power distribution, \( Y \sim \text{power}(\tau, \lambda) \), then

\[
F(y) = \left(\frac{y}{\tau}\right)^{\frac{1}{\lambda}} I(0 < y \leq \tau).
\]

Then \( W = -\log(Y) = EXP(-\log(\tau), \lambda) \). Thus (9.21) is an exact 100(1 − α)% CI for λ, and (9.20) = (\( L_n, U_n \)) is an exact 100(1 − α)% CI for \(-\log(\tau)\). Hence (\( e^{L_n}, e^{U_n} \)) is a 100(1 − α)% CI for 1/τ, and (\( e^{-U_n}, e^{-L_n} \)) is a 100(1 − α)% CI for τ.

**Example 9.13.** If Y has a truncated extreme value distribution, \( Y \sim TEV(\lambda) \), then \( W = e^Y - 1 \) is EXP(0, λ). A 100 (1 − α)% CI for λ is (9.22).

**Example 9.14.** If Y has a lognormal distribution, \( Y \sim LN(\mu, \sigma^2) \), then \( W_i = \log(Y_i) \sim N(\mu, \sigma^2) \). Thus a (1 − α)100% CI for \( \mu \) when \( \sigma \) is unknown is

\[
(W_n - t_{n-1,\frac{1}{2}} \frac{S_W}{\sqrt{n}}, W_n + t_{n-1,\frac{1}{2}} \frac{S_W}{\sqrt{n}}),
\]

where

\[
S_W = \frac{n}{n-1} \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2},
\]

and \( P(t \leq t_{n-1,\frac{1}{2}}) = 1 - \alpha/2 \) when \( t \sim t_{n-1} \).

**Example 9.15.** Let \( X_1, ..., X_n \) be iid Poisson(\( \theta \)) random variables. The classical large sample 100 (1 − α)% CI for \( \theta \) is

\[
\bar{X} \pm z_{1-\alpha/2} \sqrt{\frac{\bar{X}}{n}}.
\]
where $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$ if $Z \sim N(0, 1)$.

Following Byrne and Kabaila (2005), a modified large sample $100(1-\alpha)$% CI for $\theta$ is $(L_n, U_n)$ where

$$L_n = \frac{1}{n} \left( \sum_{i=1}^{n} X_i - 0.5 + 0.5 z_{1-\alpha/2} - z_{1-\alpha/2} \sqrt{\sum_{i=1}^{n} X_i - 0.5 + 0.25 z_{1-\alpha/2}^2} \right)$$

and

$$U_n = \frac{1}{n} \left( \sum_{i=1}^{n} X_i + 0.5 + 0.5 z_{1-\alpha/2} + z_{1-\alpha/2} \sqrt{\sum_{i=1}^{n} X_i + 0.5 + 0.25 z_{1-\alpha/2}^2} \right).$$

Following Grosh (1989, p. 59, 197–200), let $W = \sum_{i=1}^{n} X_i$ and suppose that $W = w$ is observed. Let $P(T < \chi^2_d(\alpha)) = \alpha$ if $T \sim \chi^2_d$. Then an “exact” $100(1-\alpha)$% CI for $\theta$ is

$$\left( \frac{\chi^2_w(\alpha)}{2n}, \frac{\chi^2_{w+2}(1-\alpha/2)}{2n} \right)$$

for $w \neq 0$ and

$$\left( 0, \frac{\chi^2_w(1-\alpha)}{2n} \right)$$

for $w = 0$.

The “exact” CI is conservative: the actual coverage $(1 - \delta_n) \geq 1 - \alpha$ = the nominal coverage. This interval performs well if $\theta$ is very close to 0. See Problem 9.3.

**Example 9.16.** Let $Y_1, \ldots, Y_n$ be iid bin(1, $\rho$). Let $\hat{\rho} = \sum_{i=1}^{n} Y_i/n = \text{number of "successes" } / n$. The classical large sample 100 $(1-\alpha)$% CI for $\rho$ is

$$\hat{\rho} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n}}$$

where $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$ if $Z \sim N(0, 1)$.

The Agresti Cott CI takes $\tilde{n} = n + z_{1-\alpha/2}^2$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5 z_{1-\alpha/2}^2}{\tilde{n}}.$$
(The method “adds” $0.5z_{1-\alpha/2}^2$ “0’s” and $0.5z_{1-\alpha/2}^2$ “1’s” to the sample, so the “sample size” increases by $z_{1-\alpha/2}^2$. Then the large sample 100 $(1 - \alpha)$% Agresti Coull CI for $\rho$ is

$$\hat{\rho} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n}}.$$ 

Now let $Y_1, \ldots, Y_n$ be independent bin$(m_i, \rho)$ random variables, let $W = \sum_{i=1}^{n} Y_i \sim \text{bin}(\sum_{i=1}^{n} m_i, \rho)$ and let $n_w = \sum_{i=1}^{n} m_i$. Often $m_i \equiv 1$ and then $n_w = n$. Let $P(F_{d_1, d_2} \leq F_{d_1, d_2}(\alpha)) = \alpha$ where $F_{d_1, d_2}$ has an $F$ distribution with $d_1$ and $d_2$ degrees of freedom. Assume $W = w$ is observed. Then the Clopper Pearson “exact” 100 $(1 - \alpha)$% CI for $\rho$ is

$$\left(0, 1 + \frac{1}{n_w F_{2n_w, 2}(\alpha)}\right) \quad \text{for } w = 0,$$

$$\left(\frac{n_w}{n_w + F_{2n_w, 2}(1-\alpha), 1}, 1\right) \quad \text{for } w = n_w,$$

and $(\rho_L, \rho_U)$ for $0 < w < n_w$ with

$$\rho_L = \frac{w}{w + (n_w - w + 1)F_{2(n_w - w + 1), 2w}(1 - \alpha/2)},$$

and

$$\rho_U = \frac{w + 1}{w + 1 + (n_w - w)F_{2(n_w), 2w+1}(\alpha/2)}.$$ 

The “exact” CI is conservative: the actual coverage $(1 - \delta_n) \geq 1 - \alpha = \text{the nominal coverage}$. This interval performs well if $\rho$ is very close to 0 or 1. The classical interval should only be used if it agrees with the Agresti Coull interval. See Problem 9.2.

**Example 9.17.** Let $\hat{\rho} = \text{number of “successes”}/n$. Consider a taking a simple random sample of size $n$ from a finite population of known size $N$. Then the classical finite population large sample 100 $(1 - \alpha)$% CI for $\rho$ is

$$\hat{\rho} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\rho}(1-\hat{\rho})}{n - 1} \left(\frac{N - n}{N}\right)} = \hat{\rho} \pm z_{1-\alpha/2}SE(\hat{\rho}) \quad (9.23)$$

where $P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$ if $Z \sim N(0, 1)$.
CHAPTER 9. CONFIDENCE INTERVALS

Following DasGupta (2008, p. 121), suppose the number of successes $Y$ has a hypergeometric $(C, N - C, n)$ where $p = C/N$. If $n/N \approx \lambda \in (0, 1)$ where $n$ and $N$ are both large, then

$$\hat{\rho} \approx N \left(\rho, \frac{\rho(1 - \rho)(1 - \lambda)}{n}\right).$$

Hence CI (9.23) should be good if the above normal approximation is good.

Let $\tilde{n} = n + z_{1-\alpha/2}^2$ and

$$\tilde{\rho} = \frac{n\hat{\rho} + 0.5z_{1-\alpha/2}^2}{n + z_{1-\alpha/2}^2}.$$

(Heuristically, the method adds $0.5z_{1-\alpha/2}^2$ “0's” and $0.5z_{1-\alpha/2}^2$ “1's” to the sample, so the “sample size” increases by $z_{1-\alpha/2}^2$.) Then a large sample 100(1 - $\alpha$)% Agresti Coull type (ACT) finite population CI for $\rho$ is

$$\tilde{\rho} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\rho}(1 - \hat{\rho})}{\tilde{n}} \left(\frac{N - n}{N}\right)} = \hat{\rho} \pm z_{1-\alpha/2} SE(\hat{\rho}).$$

(9.24)

Notice that a 95% CI uses $z_{1-\alpha/2} = 1.96 \approx 2$.

For data from a finite population, large sample theory gives useful approximations as $N$ and $n \to \infty$ and $n/N \to 0$. Hence theory suggests that the ACT CI should have better coverage than the classical CI if the $p$ is near 0 or 1, if the sample size $n$ is moderate, and if $n$ is small compared to the population size $N$. The coverage of the classical and ACT CIs should be very similar if $n$ is large enough but small compared to $N$ (which may only be possible if $N$ is enormous). As $n$ increases to $N$, $\hat{\rho}$ goes to $p$, $SE(\hat{\rho})$ goes to 0, and the classical CI may perform well. $SE(\hat{\rho})$ also goes to 0, but $\hat{\rho}$ is a biased estimator of $\rho$ and the ACT CI will not perform well if $n/N$ is too large.

Want an interval that gives good coverage even if $\rho$ is near 0 or 1 or if $n/N$ is large. A simple method is to combine the two intervals. Let $(L_C, U_C)$ and $(L_A, U_A)$ be the classical and ACT 100(1 - $\alpha$)% intervals. Let the modified 100(1 - $\alpha$)% interval be

$$(\max[0, \min(L_C, L_U)], \min(1, \max(U_C, U_A))).$$

(9.25)

The modified interval seems to perform well. See Problem 9.4.
Example 9.18. If $Y_1, \ldots, Y_n$ are iid Weibull ($\phi, \lambda$), then the MLE ($\hat{\phi}, \hat{\lambda}$) must be found before obtaining CIs. The likelihood

$$L(\phi, \lambda) = \frac{\phi^n}{\lambda^n} \prod_{i=1}^{n} y_i^{\phi-1} \exp \left[ \frac{-1}{\lambda} \sum_{i=1}^{n} y_i^{\phi} \right],$$

and the log likelihood

$$\log(L(\phi, \lambda)) = n \log(\phi) - n \log(\lambda) + (\phi - 1) \sum_{i=1}^{n} \log(y_i) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi}.$$

Hence

$$\frac{\partial}{\partial \lambda} \log(L(\phi, \lambda)) = -\frac{n}{\lambda} + \frac{\sum y_i^{\phi}}{\lambda^2} = 0,$$

or $\sum y_i^{\phi} = n\lambda$, or

$$\hat{\lambda} = \frac{\sum y_i^{\phi}}{n}.$$

Now

$$\frac{\partial}{\partial \phi} \log(L(\phi, \lambda)) = \frac{n}{\phi} + \sum_{i=1}^{n} \log(y_i) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi} \log(y_i) = 0,$$

so

$$n + \phi [\sum_{i=1}^{n} \log(y_i) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi} \log(y_i)] = 0,$$

or

$$\hat{\phi} = \frac{n}{\frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi} \log(y_i) - \sum_{i=1}^{n} \log(y_i)}.$$

One way to find the MLE is to use iteration

$$\hat{\lambda}_k = \frac{\sum y_i^{\hat{\phi}_{k-1}}}{n},$$

and

$$\hat{\phi}_k = \frac{n}{\frac{1}{\hat{\lambda}_k} \sum_{i=1}^{n} y_i^{\hat{\phi}_{k-1}} \log(y_i) - \sum_{i=1}^{n} \log(y_i)}.$$

Since $W = \log(Y) \sim SEV(\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)$, let

$$\hat{\sigma}_R = MAD(W_1, \ldots, W_n)/0.767049.$$
and
\[ \hat{\theta}_R = MED(W_1, ..., W_n) - \log(\log(2))\hat{\sigma}_R. \]
Then \( \hat{\phi}_0 = 1/\hat{\sigma}_R \) and \( \hat{\lambda}_0 = \exp(\hat{\theta}_R/\hat{\sigma}_R) \). The iteration might be run until both \( |\hat{\phi}_k - \hat{\phi}_{k-1}| < 10^{-6} \) and \( |\hat{\lambda}_k - \hat{\lambda}_{k-1}| < 10^{-6} \). Then take \((\hat{\phi}, \hat{\lambda}) = (\hat{\phi}_k, \hat{\lambda}_k)\).

By Example 8.13,
\[ \sqrt{n}\left( \begin{pmatrix} \hat{\lambda} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \lambda \\ \phi \end{pmatrix} \right) \overset{D}{\to} N_2(0, \Sigma) \]
where
\[ \Sigma = \begin{bmatrix} 1.109\lambda^2(1 + 0.4635 \log(\lambda) + 0.5482(\log(\lambda))^2) & 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda) \\ 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda) & 0.608\phi^2 \end{bmatrix}. \]

Thus \( 1 - \alpha \approx P\left(-z_{1-\alpha/2}\sqrt{0.608}\hat{\phi} < \sqrt{n}(\hat{\phi} - \phi) < z_{1-\alpha/2}\sqrt{0.608}\hat{\phi}\right) \) and a large sample 100\((1 - \alpha)\)% CI for \( \phi \) is
\[ \hat{\phi} \pm z_{1-\alpha/2} \hat{\phi} \sqrt{0.608/n}. \quad (9.26) \]
Similarly, a large sample 100\((1 - \alpha)\)% CI for \( \lambda \) is
\[ \hat{\lambda} \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{1.109\hat{\lambda}^2[1 + 0.4635 \log(\hat{\lambda}) + 0.5824(\log(\hat{\lambda}))^2]} . \quad (9.27) \]

In simulations, for small \( n \) the number of iterations for the MLE to converge could be in the thousands, and the coverage of the large sample CIs is poor for \( n < 50 \). See Problem 9.7.

Iterating the likelihood equations until “convergence” to a point \( \hat{\theta} \) is called a fixed point algorithm. Such algorithms may not converge, so check that \( \hat{\theta} \) satisfies the likelihood equations. Other methods such as Newton’s method may perform better.

Newton’s method is used to solve \( g(\theta) = 0 \) for \( \theta \), where the solution is called \( \theta \), and uses
\[ \theta_{k+1} = \theta_k - [Dg(\theta_k)]^{-1}g(\theta_k) \quad (9.28) \]
where
\[ Dg(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\theta) & \cdots & \frac{\partial}{\partial \theta_p} g_1(\theta) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1} g_p(\theta) & \cdots & \frac{\partial}{\partial \theta_p} g_p(\theta) \end{bmatrix}. \]
CHAPTER 9. CONFIDENCE INTERVALS

If the MLE is the solution of the likelihood equations, then use \( g(\theta) = (g_1(\theta), \ldots, g_p(\theta))^T \) where

\[
g_i(\theta) = \frac{\partial}{\partial \theta_i} \log(L(\theta)).
\]

Let \( \theta_0 \) be an initial estimator, such as the method of moments estimator of \( \theta \). Let \( D = Dg(\theta) \). Then

\[
D_{ij} = \frac{\partial g_i(\theta)}{\partial \theta_j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(L(\theta)) = \sum_{k=1}^{n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(x_k|\theta)),
\]

and

\[
\frac{1}{n} D_{ij} = \frac{1}{n} \sum_{k=1}^{n} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(x_k|\theta)) \overset{D}{\rightarrow} E \left[ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(X|\theta)) \right].
\]

Newton’s method converges if the initial estimator is sufficiently close to \( \theta \), but may diverge otherwise. Hence \( \sqrt{n} \) consistent initial estimators are recommended. Newton’s method is also popular because if the partial derivative and integration operations can be interchanged, then

\[
\frac{1}{n} Dg(\hat{\theta}) \overset{D}{\rightarrow} -I(\theta).
\]  

(9.29)

For example, the regularity conditions hold for a kP-REF by Proposition 8.20. Then a 100 \((1-\alpha)\)% large sample CI for \( \theta_i \) is

\[
\hat{\theta}_i \pm z_{1-\alpha/2} \sqrt{-D^{-1}_{ii}}
\]

(9.30)

where

\[
D^{-1} = \left[ Dg(\hat{\theta}) \right]^{-1}.
\]

This result follows because

\[
\sqrt{-D^{-1}_{ii}} \approx \sqrt{[I^{-1}(\hat{\theta})]_{ii}/n}.
\]

Example 9.19. Problem 9.8 simulates CIs for the Rayleigh \((\mu, \sigma)\) distribution of the form (9.30) although no check has been made on whether (9.29) holds for the Rayleigh distribution (which is not a 2P-REF).

\[
L(\mu, \sigma) = \left( \prod \frac{y_i - \mu}{\sigma^2} \right) \exp \left[ -\frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \right].
\]
CHAPTER 9. CONFIDENCE INTERVALS

Notice that for fixed \( \sigma, L(Y(1), \sigma) = 0 \). Hence the MLE \( \hat{\mu} < Y(1) \). Now the log
likelihood

\[
\log(L(\mu, \sigma)) = \sum_{i=1}^{n} \log(y_i - \mu) - 2n \log(\sigma) - \frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \mu)^2}{\sigma^2}.
\]

Hence \( g_1(\mu, \sigma) = \)

\[
\frac{\partial}{\partial \mu} \log(L(\mu, \sigma)) = -\sum_{i=1}^{n} \frac{1}{y_i - \mu} + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 = 0,
\]

and \( g_2(\mu, \sigma) = \)

\[
\frac{\partial}{\partial \sigma} \log(L(\mu, \sigma)) = -\frac{2n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu)^2 = 0,
\]

which has solution

\[
\hat{\sigma}^2 = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2.
\] (9.31)

To obtain initial estimators, let \( \hat{\sigma}_M = \sqrt{S^2/0.429204} \) and \( \hat{\mu}_M = \bar{Y} - 1.253314\hat{\sigma}_M \). These would be the method of moments estimators if \( S^2_M \) was
used instead of the sample variance \( S^2 \). Then use \( \mu_0 = \min(\hat{\mu}_M, 2Y(1) - \hat{\mu}_M) \)
and \( \sigma_0 = \sqrt{\sum(Y_i - \mu_0)^2/(2n)} \). Now \( \theta = (\mu, \sigma)^T \) and

\[
D \equiv Dg(\theta) = \begin{bmatrix}
\frac{\partial}{\partial \mu} g_1(\theta) & \frac{\partial}{\partial \sigma} g_1(\theta) \\
\frac{\partial}{\partial \mu} g_2(\theta) & \frac{\partial}{\partial \sigma} g_2(\theta)
\end{bmatrix} =
\begin{bmatrix}
-\sum_{i=1}^{n} \frac{1}{(y_i - \mu)^2} - \frac{n}{\sigma^2} & -\frac{2}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu) \\
-\frac{2}{\sigma^3} \sum_{i=1}^{n} (y_i - \mu) & \frac{2n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} (y_i - \mu)^2
\end{bmatrix}.
\]

So

\[
\theta_{k+1} = \theta_k - D^{-1} g(\theta_k)
\]
where
\[ g(\theta_k) = \left( -\sum_{i=1}^{n} \frac{1}{y_i - \mu_k} + \frac{1}{\sigma_k^2} \sum_{i=1}^{n} (y_i - \mu_k) \right) \]
\[ + \frac{-2n}{\sigma_k^2} + \frac{1}{\sigma_k^2} \sum_{i=1}^{n} (y_i - \mu_k)^2 \]
This formula could be iterated for 100 steps resulting in \( \theta_{101} = (\mu_{101}, \sigma_{101}) \).
Then take \( \hat{\mu} = \min(\mu_{101}, 2Y(1) - \mu_{101}) \) and
\[ \hat{\sigma} = \sqrt{\frac{1}{2n} \sum_{i=1}^{n} (Y_i - \hat{\mu})^2}. \]
Then \( \hat{\theta} = (\hat{\mu}, \hat{\sigma})^T \) and compute \( D \equiv D_{g(\hat{\theta})} \). Then (assuming (9.29) holds) a 100 \( (1 - \alpha) \)% large sample CI for \( \mu \) is
\[ \hat{\mu} \pm z_{1-\alpha/2} \sqrt{-D_{11}^{-1}} \]
and a 100 \( (1 - \alpha) \)% large sample CI for \( \sigma \) is
\[ \hat{\sigma} \pm z_{1-\alpha/2} \sqrt{-D_{22}^{-1}}. \]

**Example 9.20.** Assume that \( Y_1, \ldots, Y_n \) are iid discrete uniform \( (1, \eta) \) where \( \eta \) is an integer. For example, each \( Y_i \) could be drawn with replacement from a population of \( \eta \) tanks with serial numbers 1, 2, ..., \( \eta \). The \( Y_i \) would be the serial number observed, and the goal would be to estimate the population size \( \eta = \text{number of tanks} \). Then \( P(Y_i = i) = 1/\eta \) for \( i = 1, \ldots, \eta \). Then the cdf of \( Y \) is
\[ F(y) = \sum_{i=1}^{\lfloor y \rfloor} \frac{1}{\eta} = \frac{\lfloor y \rfloor}{\eta} \]
for \( 1 \leq y \leq \eta \). Here \( \lfloor y \rfloor \) is the greatest integer function, eg, \( \lfloor 7.7 \rfloor = 7 \).
Now let \( Z_i = Y_i/\eta \) which has cdf
\[ F_Z(t) = P(Z \leq t) = P(Y \leq t\eta) = \frac{\lfloor t\eta \rfloor}{\eta} \approx t \]
for \( 0 < t < 1 \). Let \( Z_{(n)} = Y_{(n)}/\eta = \max(Z_1, \ldots, Z_n) \). Then
\[ F_{Z_{(n)}}(t) = P(\frac{Y_{(n)}}{\eta} \leq t) = \left( \frac{\lfloor t\eta \rfloor}{\eta} \right)^n \]
for $1/\eta < t < 1$.

Want $c_n$ so that

$$P(c_n \leq \frac{Y_n}{\eta} \leq 1) = 1 - \alpha$$

for $0 < \alpha < 1$. So

$$1 - F_{Z(n)}(c_n) = 1 - \alpha \quad \text{or} \quad 1 - \left(\frac{|c_n\eta|}{\eta}\right)^n = 1 - \alpha$$

or

$$\frac{|c_n\eta|}{\eta} = \alpha^{1/n}.$$

The solution may not exist, but $c_n - 1/\eta \leq \alpha^{1/n} \leq c_n$. Take $c_n = \alpha^{1/n}$ then

$$[Y_n, \frac{Y_n}{\alpha^{1/n}})$$

is a CI for $\eta$ that has coverage slightly less than $100(1 - \alpha)%$ for small $n$, but the coverage converges in probability to 1 as $n \to \infty$.

For small $n$ the midpoint of the 95% CI might be a better estimator of $\eta$ than $Y_n$. The left endpoint is closed since $Y_n$ is a consistent estimator of $\eta$. If the endpoint was open, coverage would go to 0 instead of 1. It can be shown that $n$ (length CI) converges to $-\eta \log(\alpha)$ in probability. Hence $n$ (length 95% CI) $\approx 3\eta$. Problem 9.9 provides simulations that suggest that the 95% CI coverage and length is close to the asymptotic values for $n \geq 10$.

**Example 9.21.** Assume that $Y_1, ..., Y_n$ are iid uniform $(0, \theta)$. Let $Z_i = Y_i/\theta \sim U(0, 1)$ which has cdf $F_Z(t) = t$ for $0 < t < 1$. Let $Z(n) = Y_n/\theta = \max(Z_1, ..., Z_n)$. Then

$$F_{Z(n)}(t) = P\left(\frac{Y_n}{\theta} \leq t\right) = t^n$$

for $0 < t < 1$.

Want $c_n$ so that

$$P\left(c_n \leq \frac{Y_n}{\theta} \leq 1\right) = 1 - \alpha$$

for $0 < \alpha < 1$. So

$$1 - F_{Z(n)}(c_n) = 1 - \alpha \quad \text{or} \quad 1 - c_n^n = 1 - \alpha$$
or
\[ c_n = \alpha^{1/n}. \]

Then
\[ \left( Y_{(n)}, \frac{Y_{(n)}}{\alpha^{1/n}} \right) \]
is an exact 100(1 - \alpha)\% CI for \( \theta \). It can be shown that \( n \) (length CI) converges to \( -\theta \log(\alpha) \) in probability.

If \( Y_1, \ldots, Y_n \) are iid \( U(\theta_1, \theta_2) \) where \( \theta_1 \) is known, then \( Y_i - \theta_1 \) are iid \( U(0, \theta_2 - \theta_1) \) and
\[ \left( Y_{(n)} - \theta_1, \frac{Y_{(n)} - \theta_1}{\alpha^{1/n}} \right) \]
is a 100(1 - \alpha)\% CI for \( \theta_2 - \theta_1 \). Thus if \( \theta_1 \) is known, then
\[ \left( Y_{(n)}, \theta_1(1 - \frac{1}{\alpha^{1/n}}) + \frac{Y_{(n)}}{\alpha^{1/n}} \right) \]
is a 100(1 - \alpha)\% CI for \( \theta_2 \).

**Example 9.22.** Assume \( Y_1, \ldots, Y_n \) are iid with mean \( \mu \) and variance \( \sigma^2 \). Bickel and Doksum (2007, p. 279) suggest that
\[ W_n = n^{-1/2} \left[ \frac{(n - 1)S^2}{\sigma^2} - n \right] \]
can be used as an asymptotic pivot for \( \sigma^2 \) if \( E(Y^4) < \infty \). Notice that \( W_n = \)
\[ n^{-1/2} \left[ \frac{\sum (Y_i - \mu)^2}{\sigma^2} - \frac{n(\bar{Y} - \mu)^2}{\sigma^2} - n \right] = \]
\[ \sqrt{n} \left[ \frac{\sum (Y_i - \mu)^2}{n} - 1 \right] - \frac{1}{\sqrt{n}} n \left( \frac{\bar{Y} - \mu}{\sigma} \right)^2 = X_n - Z_n. \]

Since \( \sqrt{n}Z_n \xrightarrow{D} \chi^2_1 \), the term \( Z_n \xrightarrow{D} 0 \). Now \( X_n = \sqrt{n}(\bar{Y} - 1) \xrightarrow{D} N(0, \tau) \) by the CLT since \( U_i = [(Y_i - \mu)/\sigma]^2 \) has mean \( E(U_i) = 1 \) and variance
\[ V(U_i) = \tau = E(U_i^2) - (E(U_i))^2 = \frac{E[(Y_i - \mu)^4]}{\sigma^4} - 1 = \kappa + 2 \]
where \( \kappa \) is the kurtosis of \( Y_i \). Thus \( W_n \xrightarrow{D} N(0, \tau). \)
Hence

\[ 1 - \alpha \approx P\left(-z_{1-\alpha/2} < \frac{W_n}{\sqrt{\hat{\tau}}} < z_{1-\alpha/2}\right) = P\left(-z_{1-\alpha/2}\sqrt{n\hat{\tau}} < W_n < z_{1-\alpha/2}\sqrt{n\hat{\tau}}\right) = P\left(-\frac{(n-1)S^2}{\sigma^2} - n < z_{1-\alpha/2}\sqrt{n\hat{\tau}}\right) = P\left(n - z_{1-\alpha/2}\sqrt{n\hat{\tau}} < \frac{(n-1)S^2}{\sigma^2} < n + z_{1-\alpha/2}\sqrt{n\hat{\tau}}\right). \]

Hence a large sample 100(1 - \alpha)\% CI for \( \sigma^2 \) is

\[
\left( \frac{(n-1)S^2}{n + z_{1-\alpha/2}\sqrt{n\hat{\tau}}}, \frac{(n-1)S^2}{n - z_{1-\alpha/2}\sqrt{n\hat{\tau}}} \right)
\]

where

\[ \hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^4 \frac{S^4}{n}. \]

Notice that this CI needs \( n > z_{1-\alpha/2}\sqrt{n\hat{\tau}} \) for the right endpoint to be positive. It can be shown that \( \sqrt{n} \) (length CI) converges to \( 2\sigma^2 z_{1-\alpha/2}\sqrt{\tau} \) in probability.

Problem 9.10 uses an asymptotically equivalent 100(1 - \alpha)\% CI of the form

\[
\left( \frac{(n-a)S^2}{n + t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}}, \frac{(n+b)S^2}{n - t_{n-1,1-\alpha/2}\sqrt{n\hat{\tau}}} \right)
\]

where \( a \) and \( b \) depend on \( \hat{\tau} \). The goal was to make a 95\% CI with good coverage for a wide variety of distributions (with 4th moments) for \( n \geq 100 \). The price is that the CI is too long for some of the distributions with small kurtosis. The \( N(\mu, \sigma^2) \) distribution has \( \tau = 2 \), while the \( \text{EXP}(\lambda) \) distribution has \( \sigma^2 = \lambda^2 \) and \( \tau = 8 \). The quantity \( \tau \) is small for the uniform distribution but large for the lognormal \( \text{LN}(0,1) \) distribution.

By the binomial theorem, if \( E(Y^4) \) exists and \( E(Y) = \mu \) then

\[
E(Y - \mu)^4 = \sum_{j=0}^{4} \binom{4}{j} E[Y^j](-\mu)^{4-j} = \mu^4 - 4\mu^3 E(Y) + 6\mu^2 (V(Y) + [E(Y)]^2) - 4\mu E(Y^3) + E(Y^4).
\]

This fact can be useful for computing

\[
\tau = \frac{E[(Y_i - \mu)^4]}{\sigma^4} - 1 = \kappa + 2.
\]
Example 9.23. Following DasGupta (2008, p. 402-404), consider the pooled t CI for $\mu_1 - \mu_2$. Let $X_1, ..., X_{n_1}$ be iid with mean $\mu_1$ and variance $\sigma_1^2$. Let $Y_1, ..., Y_{n_2}$ be iid with mean $\mu_2$ and variance $\sigma_2^2$. Assume that the two samples are independent and that $n_i \to \infty$ for $i = 1, 2$ in such a way that $\hat{\rho} = \frac{n_1}{n_1 + n_2} \to \rho \in (0, 1)$. Let $\theta = \sigma_2^2/\sigma_1^2$, and let the pooled sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}.$$  

Then

$$\left( \frac{1}{\sqrt{n_1}} (\bar{X} - \mu_1) \right) \left( \frac{1}{\sqrt{n_2}} (\bar{Y} - \mu_2) \right) \xrightarrow{D} N_2(0, \Sigma)$$

where $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2)$. Hence

$$\left( \frac{1}{\sqrt{n_1}} \bar{X} - \mu_1 \right) \left( \frac{1}{\sqrt{n_2}} \bar{Y} - \mu_2 \right) = \bar{X} - \bar{Y} - (\mu_1 - \mu_2) \xrightarrow{D} N(0, \frac{\sigma_2^2}{n_1} + \frac{\sigma_2^2}{n_2}).$$

So

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \xrightarrow{D} N(0, 1).$$

Thus

$$\frac{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{D} N(0, \tau^2)$$

where

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \frac{1/\sigma_1^2}{1/\sigma_1^2} \frac{n_1n_2}{n_1 + n_2} \frac{1/\sigma_1^2}{1/\sigma_1^2} \frac{n_1n_2}{n_1 + n_2} = \frac{1/\sigma_1^2}{\rho + (1 - \rho)\theta} \frac{n_1n_2}{\rho + (1 - \rho)\theta} \xrightarrow{D} \frac{1 - \rho + \rho \theta}{\rho + (1 - \rho)\theta} = \tau^2.$$  

Now let $\hat{\theta} = S_2^2/S_1^2$ and

$$\hat{\tau}^2 = \frac{1 - \hat{\rho} + \hat{\rho} \hat{\theta}}{\hat{\rho} + (1 - \hat{\rho})\hat{\theta}}.$$  

Notice that $\hat{\tau} = 1$ if $\hat{\rho} = 1/2$, and $\hat{\tau} = 1$ if $\hat{\theta} = 1$. 
The usual large sample \((1 - \alpha)100\%\) pooled t CI for \((\mu_1 - \mu_2)\) is

\[
\bar{X} - \bar{Y} \pm t_{n_1+n_2-2,1-\alpha/2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{9.32}
\]

The large sample theory says that this CI is valid if \(\tau = 1\), and that

\[
\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{D} N(0, 1).
\]

Hence a large sample \((1 - \alpha)100\%\) CI for \((\mu_1 - \mu_2)\) is

\[
\bar{X} - \bar{Y} \pm z_{1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.
\]

Then the large sample \((1 - \alpha)100\%\) modified pooled t CI for \((\mu_1 - \mu_2)\) is

\[
\bar{X} - \bar{Y} \pm t_{n_1+n_2-4,1-\alpha/2} \hat{\sigma} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}. \tag{9.33}
\]

The large sample \((1 - \alpha)100\%\) Welch CI for \((\mu_1 - \mu_2)\) is

\[
\bar{X} - \bar{Y} \pm t_{d,1-\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{9.34}
\]

where \(d = max(1, [d_0])\), and

\[
d_0 = \frac{(S_1^2/n_1 + S_2^2/n_2)^2}{\frac{1}{n_1-1}(S_1^2/n_1)^2 + \frac{1}{n_2-1}(S_2^2/n_2)^2}.
\]

Suppose \(n_1/(n_1 + n_2) \to \rho\). It can be shown that if the CI length is multiplied by \(\sqrt{n_1}\), then the scaled length of the pooled t CI converges in probability to \(2z_{1-\alpha/2} \sqrt{\frac{\rho}{1-\rho}} \sigma_1^2 + \sigma_2^2\) while the scaled lengths of the modified pooled t CI and Welch CI both converge in probability to \(2z_{1-\alpha/2} \sqrt{\sigma_1^2 + \frac{\rho}{1-\rho} \sigma_2^2}\).
CHAPTER 9. CONFIDENCE INTERVALS

9.3 Bootstrap and Randomization Tests

Randomization tests and bootstrap tests and confidence intervals are resampling algorithms used to provide information about the sampling distribution of a statistic $T_n \equiv T_n(F) \equiv T_n(Y_n)$ where $Y_n = (Y_1, \ldots, Y_n)$ and the $Y_i$ are iid from a distribution with cdf $F(y) = P(Y \leq y)$. Then $T_n$ has a cdf $H_n(y) = P(T_n \leq y)$. If $F(y)$ is known, then $m$ independent samples $Y_{j,n}^* = (Y_{j,1}^*, \ldots, Y_{j,n}^*)$ of size $n$ could be generated, where the $Y_{j,n}^*$ are iid from a distribution with cdf $F$ and $j = 1, \ldots, m$. Then the statistic $\bar{T}_n$ is computed for each sample, resulting in $m$ statistics $T_{1,n}(F), \ldots, T_{m,n}(F)$ which are iid from a distribution with cdf $H_n(y)$. Equivalent notation $T_{i,n}(F) \equiv T_{i,n}(Y_{i,n}^*)$ is often used, where $i = 1, \ldots, m$.

If $W_1, \ldots, W_m$ are iid from a distribution with cdf $F_W$, then the empirical cdf $F_m$ corresponding to $F_W$ is given by

$$F_m(y) = \frac{1}{m} \sum_{i=1}^{m} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix $m$ and $y$. Then $mF_m(y) \sim \text{binomial} (m, F_W(y))$. Thus $E[F_m(y)] = F_W(y)$ and $V[F_m(y)] = F_W(y)[1 - F_W(y)]/m$. By the central limit theorem,

$$\sqrt{m}(F_m(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_m(y) - F_W(y) = O_P(m^{-1/2})$, and $F_m$ is a reasonable estimator of $F_W$ if the number of samples $m$ is large.

Let $W_i = T_{i,n}(F)$. Then $F_m \equiv \bar{H}_{m,n}$ is an empirical cdf corresponding to $H_n$. Let $W_i = Y_i$ and $m = n$. Then $F_n$ is the empirical cdf corresponding to $F$. Let $y_n = (y_1, \ldots, y_n)^T$ be the observed data. Now $F_n$ is the cdf of the population that consists of $y_1, \ldots, y_n$ where the probability of selecting $y_i$ is $1/n$. Hence an iid sample of size $d$ from $F_n$ is obtained by drawing a sample of size $d$ with replacement from $y_1, \ldots, y_n$. If $d = n$, let $Y_{j,n}^* = (Y_{j,1}^*, \ldots, Y_{j,n}^*)$ be an iid sample of size $n$ from the empirical cdf $F_n$. Hence each $Y_{j,k}^*$ is one of the $y_1, \ldots, y_n$ where repetition is allowed. Take $m$ independent samples from $F_n$ and compute the statistic $T_n$ for each sample, resulting in $m$ statistics $T_{1,n}(F_n), \ldots, T_{m,n}(F_n)$ where $T_{i,n}(F_n) \equiv T_{i,n}(Y_{i,n}^*)$ for $i = 1, \ldots, m$. This type of sampling can be done even if $F$ is unknown, and if $T_n(F_n) \approx T_n(F)$, then the empirical cdf based on the $T_{i,n}(F_n)$ may be a useful approximation for $H_n$. 


For general resampling algorithms let $T^*_{i,n}(Y^*_{i,n})$ be the statistic based on a randomly chosen sample $Y^*_{i,n}$ used by the resampling algorithm. Let $H_{A,n}$ be the cdf of the $T^*_{i,n}$ based on all $J_n$ possible samples, and let $H_{m,n}$ be the cdf of the $T^*_{i,n}$ based on $m$ randomly chosen samples. Often theoretical results are given for $H_{A,n}$ but are not known for $H_{m,n}$. Let $G_{N,n}$ be a cdf based on a normal approximation for $H_{n}$. Central limit type theorems are used and $G_{N,n}$ is often first order accurate: $H_{n}(y) - G_{N,n}(y) = O_P(n^{-1/2})$. Approximations $G_{E,n}$ based on the Edgeworth expansion (which is not a cdf) and $H_{A,n}$ are sometimes second order accurate: $H_{n}(y) - H_{A,n}(y) = O_P(n^{-1})$.

The following two examples follow DasGupta (2008, p. 462, 469, 513).

**Example 9.24.** Let $Y_1,...,Y_n$ be iid with cdf $F$. Then the ordinary bootstrap distribution of $T_n$ is $H_{A,n}(y) = P_{F_{n}}(T_n(Y^*_{i,n}) \leq y)$ where $Y^*_{i,n} = (Y^*_{i,1},...,Y^*_{i,n})$ is an iid sample of size $n$ from the empirical cdf $F_{n}$ obtained by selecting with replacement from $Y_1,...,Y_n$. Here $T^*_{i,n}(Y^*_{i,n}) = T_{n}(Y^*_{i,n})$. Note that there are $J_n = n^n$ ordered samples and $n^n/n!$ unordered samples from $F_{n}$. The bootstrap distribution $H_{m,n}$ typically used in practice is based on $m$ samples randomly selected with replacement. Both $H_{A,n}$ and $H_{m,n}$ are estimators of $H_{n}$, the cdf of $T_n$.

For example, suppose the data is 1, 2, 3, 4, 5, 6, 7. Then $n = 7$ and the sample median $T_n$ is 4. Using R, we drew $m = 2$ bootstrap samples (samples of size $n$ drawn with replacement from the original data) and computed the sample median $T^*_{1,n} = 3$ and $T^*_{2,n} = 4$.

```r
b1 <- sample(1:7,replace=T)
b1
[1] 3 2 3 2 5 2 6
median(b1)
[1] 3
b2 <- sample(1:7,replace=T)
b2
[1] 3 5 3 4 3 5 7
median(b2)
[1] 4
```

Heuristically, suppose $T_{n}(F_{n})$ is an unbiased estimator of $\theta$. Let $T^*_{i,n} = T^*_{i,n}(Y^*_{i,n})$. Then $T^*_{1,n},...,T^*_{m,n}$ each gives an unbiased estimator of $\theta$. If $m$ is large, then typical values of $T^*_{i,n}$ should provide information about $\theta$.
example, the middle 95% of the $T_i^*$ should be an approximate 95% percentile method CI for $\theta$. Then reject $H_0 : \theta = \theta_0$ if $\theta_0$ is not in the CI. This bootstrap inference has two sources of error. First, $n$ needs to be large enough so that $T_n(F_n) \approx T_n(F)$. Second, the $T_{i,n}^*$ are used to form an empirical cdf $H_{m,n}$ corresponding to $H_{A,n}$, so $m$ needs to be large enough so that the empirical cdf $H_{m,n}$ is a good estimator of $H_{A,n}$.

**Example 9.25.** Let $X_1, ..., X_{k_1}$ be iid with pdf $f(y)$ while $Y_1, ..., Y_{k_2}$ are iid with pdf $f(y - \mu)$. Let $n = k_1 + k_2$ and consider testing $H_0 : \mu = 0$. Let $T_n \equiv T_{k_1,k_2}$ be the two sample $t$-statistic. Under $H_0$, the random variables in the combined sample $X_1, ..., X_{k_1}, Y_1, ..., Y_{k_2}$ are iid with pdf $f(y)$. Let $Z_n$ be any permutation of $(X_1, ..., X_{k_1}, Y_1, ..., Y_{k_2})$ and compute $T_n(Z_n)$ for each permutation. Then $H_{A,n}$ is the cdf based on all of the $T_n(Z_n)$. $H_0$ is rejected if $T_n$ is in the extreme tails of $H_{A,n}$. The number of ordered samples is $J_n = n!$ while the number of unordered samples is $(\binom{n}{k_1})$. Such numbers get enormous quickly. Usually $m$ randomly drawn permutations are selected with replacement, resulting in a cdf $H_{m,n}$ used to choose the appropriate cutoffs $c_L$ and $c_U$.

For randomization tests that used a fixed number $m = B$ of permutations, calculations using binomial approximations suggest that $B = 999$ to 5000 will give a test similar to those based on using all permutations. See Efron and Tibshirani (1993, pp. 208-210). Jöckel (1986) shows, under regularity conditions, that the power of a randomization test is increasing and converges as $m \to \infty$. It is suggested that the tests have good power if $m = 999$, but the pvalue of such a test is bounded below by 0.001 since the pvalue = $(1 + \text{the number of the } m \text{ test statistics at least as extreme as the observed statistic})/(m + 1)$. Buckland (1984) shows that the expected coverage of the nominal $100(1 - \alpha)\%$ percentile method confidence interval is approximately correct, but the standard deviation of the coverage is proportional to $1/\sqrt{m}$. Hence the percentile method is a large sample confidence interval, in that the true coverage converges in probability to the nominal coverage, only if $m \to \infty$ as $n \to \infty$. These results are good reasons for using $m = \text{max}(B, \lfloor n \log(n) \rfloor)$ samples.

The key observation for theory is that $H_{m,n}$ is an empirical cdf. To see this claim, recall that $H_{A,n}(y) \equiv H_{A,n}(y|Y_n)$ is a random cdf: it depends on the data $Y_n$. Hence $H_{A,n}(y) \equiv H_{A,n}(y|y_n)$ is the observed cdf based on the observed data. $H_{A,n}(y|y_n)$ can be computed by finding $T_{i,n}^*(Y_{i,n})$ for
all \( J_n \) possible samples \( Y_{i,n}^* \). If \( m \) samples are selected with replacement from all possible samples, then the samples are iid and \( T_{1,n}^*, \ldots, T_{m,n}^* \) are iid with cdf \( H_{A,n}(y|y_n) \). Hence \( F_m \equiv H_{m,n} \) is an empirical cdf corresponding to \( F \equiv H_{A,n}(y|y_n) \).

Thus empirical cdf theory can be applied to \( H_{m,n} \). Fix \( n \) and \( y_n \). Then \( mH_{m,n}(y_n) \sim \text{binomial} (m, H_{A,n}(y|y_n)) \). Thus \( E[H_{m,n}(y_n)] = H_{A,n}(y|y_n) \) and \( \text{Var}[H_{m,n}(y_n)] = H_{A,n}(y|y_n)[1 - H_{A,n}(y|y_n)]/m \). Also

\[
\sqrt{m}(H_{m,n}(y_n) - H_{A,n}(y|y_n)) \xrightarrow{D} N(0, H_{A,n}(y|y_n)[1 - H_{A,n}(y|y_n)]).
\]

Thus \( H_{m,n}(y_n) - H_{A,n}(y|y_n) = O_p(m^{-1/2}) \). Note that the probabilities and expectations depend on \( m \) and on the observed data \( y_n \).

This result suggests that if \( H_{A,n} \) is a first order accurate estimator of \( H_n \), then \( H_{m,n} \) can not be a first order accurate estimator of \( H_n \) unless \( m \) is proportional to \( n \). If \( m = \max(1000, \lfloor n \log(n) \rfloor) \), then \( H_{m,n} \) is asymptotically equivalent to \( H_{A,n} \) up to terms of order \( n^{-1/2} \). If \( m = \max(1000, [0.1n^2 \log(n)]) \), then \( H_{m,n} \) asymptotically equivalent to \( H_{A,n} \) up to terms of order \( n^{-1} \).

As an application, Efron and Tibshirani (1993, p. 187, 275) state that percentile method for bootstrap confidence intervals is first order accurate and that the coefficient of variation of a bootstrap percentile is proportional to \( \sqrt{\frac{1}{n} + \frac{1}{m}} \). If \( m = 1000 \), then the percentile bootstrap is not first order accurate. If \( m = \max(1000, \lfloor n \log(n) \rfloor) \), then the percentile bootstrap is first order accurate. Similarly, claims that a bootstrap method is second order accurate are false unless \( m \) is proportional to \( n^2 \). See a similar result in Robinson (1988).

Practical resampling algorithms often use \( m = B = 1000, 5000 \) or 10000. The choice of \( m = 10000 \) works well for small \( n \) and for simulation studies since the cutoffs based on \( H_{m,n} \) will be close to those based on \( H_{A,n} \) with high probability since \( V[H_{10000,n}(y)] \leq 1/40000 \). For the following theorem, also see Serfling (1981, p. 59-61).

**Theorem 9.1.** Let \( Y_1, \ldots, Y_n \) be iid \( k \times 1 \) random vectors from a distribution with cdf \( F(y) = P(Y_1 \leq y_1, \ldots, Y_k \leq y_k) \). Let

\[
D_n = \sup_{y \in \mathbb{R}^k} |F_n(y) - F(y)|.
\]

a) Massart (1990) \( k = 1 \): \( P(D_n > d) \leq 2 \exp(-2nd^2) \) if \( nd^2 \geq 0.5 \log(2) \).
b) Kiefer (1961) \( k \geq 2 \): \( P(D_n > d) \leq C \exp(-(2-\epsilon)nd^2) \) where \( \epsilon > 0 \) is fixed and the positive constant \( C \) depends on \( \epsilon \) and \( k \) but not on \( F \).

To use Theorem 9.1a, fix \( n \) (and suppressing the dependence on \( y_n \)), take \( F = H_{A,n} \) computed from the observed data and take \( F_m = H_{m,n} \). Then

\[
D_m = \sup_{y \in \mathbb{R}} |H_{m,n}(y) - H_{A,n}(y)|.
\]

Recalling that the probability is with respect to the observed data, consider the following choices of \( m \).

i) If \( m = 10000 \), then \( P(D_m > 0.01) \leq 2e^{-2} \approx 0.271 \).

ii) If \( m = \max(10000, [0.25n \log(n)]) \), then for \( n > 5000 \)

\[
P \left( D_m > \frac{1}{\sqrt{n}} \right) \leq 2 \exp(-2[0.25n \log(n)]/n) \approx 2/\sqrt{n}.
\]

iii) If \( m = \max(10000, [0.5n^2 \log(n)]) \), then for \( n > 70 \)

\[
P \left( D_m > \frac{1}{n} \right) \leq 2 \exp(-2[0.5n^2 \log(n)]/n^2) \approx 2/n.
\]

Two tail tests with nominal level \( \alpha \) and confidence intervals with nominal coverage \( 1 - \alpha \) tend to use the lower and upper \( \alpha/2 \) percentiles from \( H_{m,n} \). This procedure corresponds to an interval covering \( 100(1 - \alpha)\% \) of the mass. The interval is short if the distribution corresponding to \( H_{m,n} \) is approximately symmetric. Skewness or approximate symmetry can be checked by plotting the \( T^*_{i,n} \). Shorter intervals can be found if the distribution is skewed by using the shorth\((c)\) estimator where \( c = \lceil m(1 - \alpha) \rceil \) and \( \lceil x \rceil \) is the smallest integer \( \geq x \), e.g., \( \lceil 7.7 \rceil = 8 \). See Grubel (1988).

That is, let \( T^*_1, \ldots, T^*_m \) be the order statistics of the \( T^*_{1,n}, \ldots, T^*_{m,n} \) computed by the resampling algorithm. Compute \( T^*_c - T^*_1, T^*_c - T^*_{(c+1)}, \ldots, T^*_m - T^*_{(m-c+1)} \). Let \( \lceil T^*_s \rceil \) correspond to the closed interval with the smallest distance. Then reject \( H_0: \theta = \theta_0 \) if \( \theta_0 \) is not in the interval.

Resampling methods can be used in courses on resampling methods, non-parametric statistics, and experimental design. In such courses it can be stated that it is well known that \( H_{m,n} \) has good statistical properties (under regularity conditions) if \( m \to \infty \) as \( n \to \infty \), but algorithms tend to use \( m = B \) between 999 and 10000. Such algorithms may perform well in simulations, but lead to tests with p-value bounded away from 0, confidence intervals with coverage that fails to converge to the nominal coverage, and fail to take advantage of the theory derived for the impractical all subset algorithms. Since \( H_{m,n} \) is the empirical cdf corresponding to the all subset
algorithm cdf $H_{A,n}$, taking $m = \max(B, \lfloor n \log(n) \rfloor)$ leads to a practical algorithm with good theoretical properties (under regularity conditions) that performs well in simulations.

Although theory for resampling algorithms given in Lehmann (1999, p. 425) and Sen and Singer (1993, p. 365) has the number of samples $m \to \infty$ as the sample size $n \to \infty$, much of the literature suggests using $m = B$ between 999 and 10000. This choice is often justified using simulations and binomial approximations. An exception is Shao (1989) where $n/m \to 0$ as $n \to \infty$. Let $[x]$ be the integer part of $x$, so $[7.7] = 7$. Then $m = \lfloor n^{1.01} \rfloor$ may give poor results for $n < 900$. To combine theory with empirical results, we suggest using $m = \max(B, \lfloor n \log(n) \rfloor)$.

Theory for resampling algorithms such as first order accuracy of the bootstrap and the power of randomization tests is usually for the impractical algorithm that uses all $J_n$ samples. Practical algorithms use $B$ randomly drawn samples where $B$ is chosen to give good performance when $n$ is small. We suggest using $m = \max(B, \lfloor n \log(n) \rfloor)$ randomly drawn samples results in a practical algorithm that is asymptotically equivalent to the impractical algorithm up to terms of order $n^{-1/2}$ while also having good small sample performance.

**Example 9.26.** Suppose $F$ is the cdf of the $N(\mu, \sigma^2)$ distribution and $T_n(F) = \bar{Y}_n \sim N(\mu, \sigma^2/n)$. Suppose $m$ independent samples $(Y_{j,1}^*, ..., Y_{j,n}^*) = Y_{j,n}^*$ of size $n$ are generated, where the $Y_{j,k}^*$ are iid $N(\mu, \sigma^2)$ and $j = 1, ..., m$.

Then let the sample mean $T_{j,n}^* = \bar{Y}_{j,n}^* \sim N(\mu, \sigma^2/n)$ for $j = 1, ..., m$.

We want to examine, for a given $m$ and $n$, how well do the sample quantiles $T_{(\lfloor m \rho \rfloor)}^* = \bar{Y}_{(\lfloor m \rho \rfloor),n}^*$ of the $\bar{Y}_{j,n}^*$ estimate the quantiles $\xi_{\rho,n}$ of the $N(\mu, \sigma^2/n)$ distribution and how does $(T_{(\lfloor m 0.025 \rfloor)}^*, T_{(\lfloor m 0.975 \rfloor)}^*)$ perform as a 95% CI for $\mu$. Here $P(X \leq \xi_{\rho,n}) = \rho$ if $X \sim N(\mu, \sigma^2/n)$. Note that $\xi_{\rho,n} = \mu + z_\rho \sigma / \sqrt{n}$ where $P(Z \leq z_\rho) = \rho$ if $Z \sim N(0, 1)$.

Fix $n$ and let $f_n$ be the pdf of the $N(\mu, \sigma^2/n)$ distribution. By Theorem 8.27, as $m \to \infty$

$$\sqrt{m}(\bar{Y}_{(\lfloor m \rho \rfloor),n} - \xi_{\rho,n}) \overset{D}{\to} N(0, \tau_n^2)$$

where

$$\tau_n^2 = \frac{\rho(1-\rho)}{[f_n(\xi_{\rho})]^2} = \frac{\rho(1-\rho)2\pi\sigma^2}{n \exp(-z_\rho^2)}$$

Since the quantile $\xi_{\rho,n} = \mu + z_\rho \sigma / \sqrt{n}$, need $m$ fairly large for the estimated
quantile to be good. To see this claim, suppose we want

\[ P(\xi_{0.975,n} - 0.04\sigma/\sqrt{n} < Y^*_{(\lceil m \cdot 0.975 \rceil),n} < \xi_{0.975,n} - 0.04\sigma/\sqrt{n}) > 0.9. \]

(For \( N(0,1) \) data, this would be similar to wanting the estimated 0.975 quantile to be between 1.92 and 2.00 with high probability.) Then 0.9 \( \approx \)

\[ P\left( \frac{-0.04\sigma\sqrt{m}}{\tau_n\sqrt{n}} < Z < \frac{0.04\sigma\sqrt{m}}{\tau_n\sqrt{n}} \right) \approx P\left( -0.01497\sqrt{m} < Z < 0.01497\sqrt{m} \right) \]

or

\[ m \approx \left( \frac{z_{0.05}}{-0.01497} \right)^2 \approx 12076. \]

With \( m = 1000 \), the above probability is only about 0.36. To have the probability go to one, need \( m \to \infty \) as \( n \to \infty \).

Note that if \( m = B = 1000 \), say, then the sample quantile is not a consistent estimator of the population quantile \( \xi_{\rho,n} \). Also, \( (\bar{Y}_{(\lceil m \cdot \rho \rceil),n} - \xi_{\rho,n}) = O_P(n^{-\delta}) \) needs \( m \propto n^{2\delta} \) where \( \delta = 1/2 \) or 1 are the most interesting cases. For good simulation results, typically need \( m \) larger than a few hundred, eg \( B = 1000 \), for small \( n \). Hence \( m = \max(B, n\log(n)) \) combines theory with good simulation results.

The CI length behaves fairly well for large \( n \). For example, the 95% CI length will be close to \( 3.92/\sqrt{n} \) since roughly 95% of the \( \bar{Y}_{j,n} \) are between \( \mu - 1.96\sigma/\sqrt{n} \) and \( \mu + 1.96\sigma/\sqrt{n} \). The coverage is conservative (higher than 95%) for moderate \( m \). To see this, note that the 95% CI contains \( \mu \) if \( T_{(\lceil m \cdot 0.025 \rceil)} > \mu \) and \( T_{(\lceil m \cdot 0.975 \rceil)} > \mu \). Let \( W \sim \text{binomial} (m, 0.5) \). Then

\[ P(T_{(\lceil m \cdot 0.025 \rceil)} > \mu) \approx P(W > 0.025m) \approx P(Z > \frac{0.025m - 0.5m}{0.5\sqrt{m}}) = P(Z > -0.95\sqrt{m}) \to 1 \text{ as } m \to \infty. \]

(Note that if \( m = 1000 \), then \( T_{(\lceil m \cdot 0.975 \rceil)} > \mu \) if 225 or more \( \bar{Y}_{j,n} > \mu \) or if fewer than 975 \( \bar{Y}_{j,n} < \mu \).)

Since \( F \) is not known, we can not sample from \( T_n(F) \), but sampling from \( T_n(F_n) \) can at least be roughly approximated using computer generated random numbers. The bootstrap replaces \( m \) samples from \( T_n(F) \) by \( m \) samples from \( T_n(F_n) \), that is, there is a single sample \( Y_1, \ldots, Y_n \) of data. Take a sample of size \( n \) with replacement from \( Y_1, \ldots, Y_n \) and compute the sample mean \( \bar{Y}_{1,n} \). Repeat to obtain the bootstrap sample \( \bar{Y}_{1,n}^*, \ldots, \bar{Y}_{m,n}^* \). Expect the bootstrap estimator of the quantile to perform less well than that based on samples
from $T_n(F)$. So still need $m$ large so that the estimated quantiles are near the population quantiles.

Simulated coverage for the bootstrap percentile 95% CI tends to be near 0.95 for moderate $m$, and we expect the length of the 95% CI to again be near $3.92/\sqrt{n}$. The bootstrap sample tends to be centered about the observed value of $\bar{Y}$. If there is a “bad sample” so that $\bar{Y}$ is in the left or right tail of the sampling distribution, say $\bar{Y} > \mu + 1.96\sigma / \sqrt{n}$ or $\bar{Y} < \mu - 1.96\sigma / \sqrt{n}$, then the coverage may be much less that 95%. But the probability of a “bad sample” is 0.05 for this example.

### 9.4 Complements


A comparison of CIs with other intervals (such as prediction intervals) is given in Vardeman (1992). Also see Hahn and Meeker (1991).

Newton’s method is described, for example, in Peressini, Sullivan and Uhl (1988, p. 85).

### 9.5 Problems

**PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.**

Refer to Chapter 10 for the pdf or pmf of the distributions in the problems below.

**9.1.** (Aug. 2003 Qual): Suppose that $X_1, ..., X_n$ are iid with the Weibull
distribution, that is the common pdf is

\[ f(x) = \begin{cases} \frac{b}{a} x^{b-1} e^{-\frac{x}{a}} & 0 < x \\ 0 & \text{elsewhere} \end{cases} \]

where \( a \) is the unknown parameter, but \( b(>0) \) is assumed known.

a) Find a minimal sufficient statistic for \( a \).

b) Assume \( n = 10 \). Use the Chi-Square Table and the minimal sufficient statistic to find a 95\% two sided confidence interval for \( a \).

R/Splus Problems

Use the command `source("G:/sipack.txt")` to download the functions. See Section 11.1. Typing the name of the `sipack` function, eg `accisimf`, will display the code for the function. Use the `args` command, eg `args(accisimf)`, to display the needed arguments for the function.

9.2. Let \( Y_1, \ldots, Y_n \) be iid binomial(1, \( \rho \)) random variables.

From the website (www.math.siu.edu/olive/sipack.txt), enter the R/Splus function `bcisim` into R/Splus. This function simulates the 3 CIs (classical, Agresti Coull and exact) from Example 9.16, but changes the CI (L,U) to (max(0,L),min(1,U)) to get shorter lengths.

To run the function for \( n = 10 \) and \( \rho \equiv p = 0.001 \), enter the R/Splus command `bcisim(n=10,p=0.001)`. Make a table with header “n p ccov clen accov aclen ecov elen.” Fill the table for \( n = 10 \) and \( p = 0.001, 0.01, 0.5, 0.99, 0.999 \) and then repeat for \( n = 100 \). The “cov” is the proportion of 500 runs where the CI contained \( p \) and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal. Typically want the true coverage \( \geq \) to the nominal coverage, so conservative intervals are better than liberal CIs. The “len” is the average scaled length of the CI and for large \( n \) should be near \( 2(1.96)\sqrt{p(1-p)} \).

From your table, is the classical estimator or the Agresti Coull CI better? When is the exact interval good? Explain briefly.

9.3. Let \( X_1, \ldots, X_n \) be iid Poisson(\( \theta \)) random variables.

From the website (www.math.siu.edu/olive/sipack.txt), enter the R/Splus function `poiscisim` into R/Splus. This function simulates the 3 CIs (classical, modified and exact) from Example 9.15. To run the function for \( n = 100 \)
and $\theta = 5$, enter the $R$/Splus command `poiscisim(theta=5)`. Make a table with header “theta ccov clen mcov mlen ecov elen.” Fill the table for $\theta =$ 0.001, 0.1, 1.0, and 5.

The “cov” is the proportion of 500 runs where the CI contained $\theta$ and the nominal coverage is 0.95. A coverage between 0.92 and 0.98 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.98 suggests that the CI is conservative while a coverage less than 0.92 suggests that the CI is liberal (too short). Typically want the true coverage $\geq$ to the nominal coverage, so conservative intervals are better than liberal CIs. The “len” is the average scaled length of the CI and for large $n\theta$ should be near $2(1.96)\sqrt{\theta}$ for the classical and modified CIs.

From your table, is the classical CI or the modified CI or the exact CI better? Explain briefly. (Warning: in a 1999 version of $R$, there was a bug for the Poisson random number generator for $\theta \geq 10$. The 2011 version of $R$ seems to work.)

9.4. This problem simulates the CIs from Example 9.17.

a) Download the function `accisimf` into $R$/Splus.

b) The function will be used to compare the classical, ACT and modified 95% CIs when the population size $N =$ 500 and $p$ is close to 0.01. The function generates such a population, then selects 5000 independent simple random samples from the population. The 5000 CIs are made for both types of intervals, and the number of times the true population $p$ is in the $i$th CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for $n =$ 50 and $p \approx 0.01$, enter the command `accisimf(n=50,p=0.01)`. Make a table with header “n p ccov clen accov aclen mcov mlen.” Fill the table for $n =$ 50 and then repeat for $n =$ 100, 150, 200, 250, 300, 350, 400 and 450. The “len” is $\sqrt{n}$ times the mean length from the 5000 runs. The “cov” is the proportion of 5000 runs where the CI contained $p$ and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. Typically want the true coverage $\geq$ to the nominal coverage, so conservative intervals are better than liberal CIs. The “ccov” is for the classical CI, “accov” is for the Agresti Cough type (ACT) CI and “mcov” is for the modified interval. Given good coverage $> 0.94$, want short length.
c) First compare the classical and ACT intervals. From your table, for what values of \( n \) is the ACT CI better, for what values of \( n \) are the 3 intervals about the same, and for what values of \( n \) is the classical CI better?

d) Was the modified CI ever good?

9.5. This problem simulates the CIs from Example 9.10.

a) Download the function \texttt{expsim} into \texttt{R/Splus}.

The output from this function are the coverages \( \text{scov, lcov and ccov} \) of the CI for \( \lambda, \theta \) and of \( \lambda \) if \( \theta \) is known. The scaled average lengths of the CIs are also given. The lengths of the CIs for \( \lambda \) are multiplied by \( \sqrt{n} \) while the length of the CI for \( \theta \) is multiplied by \( n \).

b) The 5000 CIs are made for 3 intervals, and the number of times the true population parameter \( \lambda \) or \( \theta \) is in the \( i \)th CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The nominal coverage is 0.95. To run the function for \( n = 5, \theta = 0 \) and \( \lambda = 1 \) enter the command \texttt{expsim(n=5)}. Make a table with header “CI for \( \lambda \) CI for \( \theta \) CI for \( \lambda, \theta \) known.” Then make a second header “n cov slen cov slen cov slen” where “cov slen” is below each of the three CI headers. Fill the table for \( n = 5 \) and then repeat for \( n = 10, 20, 50, 100 \) and 1000. The “cov” is the proportion of 5000 runs where the CI contained \( \lambda \) or \( \theta \) and the nominal coverage is 0.95. For 5000 runs, an observed coverage between 0.94 and 0.96 gives little evidence that the true coverage differs from the nominal coverage of 0.95. A coverage greater that 0.96 suggests that the CI is conservative while a coverage less than 0.94 suggests that the CI is liberal. As \( n \) gets large, the values of slen should get closer to 3.92, 2.9957 and 3.92.

9.6. This problem simulates the CIs from Example 9.9.

a) Download the function \texttt{hnsim} into \texttt{R/Splus}.

The output from this function are the coverages \( \text{scov, lcov and ccov} \) of the CI for \( \sigma^2, \mu \) and of \( \sigma^2 \) if \( \mu \) is known. The scaled average lengths of the CIs are also given. The lengths of the CIs for \( \sigma^2 \) are multiplied by \( \sqrt{n} \) while the length of the CI for \( \mu \) is multiplied by \( n \).

b) The 5000 CIs are made for 3 intervals, and the number of times the true population parameter \( \theta = \mu \) or \( \sigma^2 \) is in the \( i \)th CI is counted. The simulated coverage is this count divided by 5000 (the number of CIs). The
CHAPTER 9. CONFIDENCE INTERVALS

nominal coverage is 0.95. To run the function for \( n = 5, \mu = 0 \) and \( \sigma^2 = 1 \)
enter the command \texttt{hnsim(n=5)}. Make a table with header
“CI for \( \sigma^2 \)  CI for \( \mu \)  CI for \( \sigma^2, \mu \) known.”
Then make a second header “n cov slen cov slen cov slen” where “cov slen” is
below each of the three CI headers. Fill the table for \( n = 5 \) and then repeat
for \( n = 10, 20, 50, 100 \) and 1000. The “cov” is the proportion of 5000 runs
where the CI contained \( \theta \) and the nominal coverage is 0.95. For 5000 runs,
an observed coverage between 0.94 and 0.96 gives little evidence that the
true coverage differs from the nominal coverage of 0.95. A coverage greater
that 0.96 suggests that the CI is conservative while a coverage less than 0.94
suggests that the CI is liberal. As \( n \) gets large, the values of slen should get
closer to 5.5437, 3.7546 and 5.5437.

9.7. a) Download the function \texttt{wcisim} into \texttt{R/Splus}.
The output from this function includes the coverages \( \text{pcov} \) and \( \text{lco} \)
of the CIs for \( \phi \) and \( \lambda \) if the simulated data \( Y_1, ..., Y_n \) are iid Weibull (\( \phi, \lambda \)). The
scaled average lengths of the CIs are also given. The values \( \text{pconv} \) and \( \text{lconv} \)
should be less than \( 10^{-5} \). If this is not the case, increase \texttt{iter}. 100 samples
of size \( n = 100 \) are used to create the 95\% large sample CIs for \( \phi \) and \( \lambda \)
given in Example 9.18. If the sample size is large, then \( \text{sdphihat} \), the sample
standard deviation of the 100 values of the MLE \( \hat{\phi} \), should be close to \( \text{phiasd} = \phi \sqrt{.608} \). Similarly, \( \text{sdlamhat} \) should be close to the asymptotic standard
deviation \( \text{lamasd} = \sqrt{1.109\lambda^2(1 + 0.4635 \log(\lambda) + 0.5282(\log(\lambda))^2}). \)

b) Type the command
\texttt{wcisim(n = 100, phi = 1, lam = 1, iter = 100)}
and record the coverages for the CIs for \( \phi \) and \( \lambda \).

c) Type the command
\texttt{wcisim(n = 100, phi = 20, lam = 20, iter = 100)}
and record the coverages for the CIs for \( \phi \) and \( \lambda \).

9.8. a) Download the function \texttt{raysim} into \texttt{R/Splus} to simulate the CI
of Example 9.19.

b) Type the command
\texttt{raysim(n = 100, mu = 20, sigma = 20, iter = 100)}
and record the coverages for the CIs for \( \mu \) and \( \sigma \).

9.9. a) Download the function \texttt{ducisim} into \texttt{R/Splus} to simulate the CI
of Example 9.20.
b) Type the command  
`ducisim(n=10,nruns=1000,eta=1000)`.
Repeat for \( n = 50, 100, 500 \) and make a table with header 
“n coverage 95% CI length.”
Fill in the table for \( n = 10, 50, 100 \) and 500.

c) Are the coverages close to or higher than 0.95 and is the scaled length close to \( 3\eta = 3000 \)?

9.10. a) Download the function `varcisim` into `R/Splus` to simulate a modified version of the CI of Example 9.22.

b) Type the command `varcisim(n = 100, nruns = 1000, type = 1)` to simulate the 95% CI for the variance for iid N(0,1) data. Is the coverage `vcov` close to or higher than 0.95? Is the scaled length `vlen = \sqrt{n}` (CI length) 
\[ = 2(1.96)\sigma^2\sqrt{\tau} = 5.554\sigma^2 \] close to 5.554?

c) Type the command `varcisim(n = 100, nruns = 1000, type = 2)` to simulate the 95% CI for the variance for iid EXP(1) data. Is the coverage `vcov` close to or higher than 0.95? Is the scaled length `vlen = \sqrt{n}` (CI length) 
\[ = 2(1.96)\sigma^2\sqrt{\tau} = 2(1.96)\lambda^2\sqrt{\tau} = 11.087\lambda^2 \] close to 11.087?

d) Type the command `varcisim(n = 100, nruns = 1000, type = 3)` to simulate the 95% CI for the variance for iid LN(0,1) data. Is the coverage `vcov` close to or higher than 0.95? Is the scaled length `vlen` long?

9.11. a) Download the function `pcisim` into `R/Splus` to simulate the three CIs of Example 9.23. The modified pooled t CI is almost the same as the Welch CI, but uses degrees of freedom \( = n_1 + n_2 - 4 \) instead of the more complicated formula for the Welch CI. The pooled t CI should have coverage that is too low if 
\[ \frac{\rho}{1 - \rho} \sigma_1^2 + \sigma_2^2 < \sigma_1^2 + \frac{\rho}{1 - \rho} \sigma_2^2. \]

b) Type the command `pcisim(n1=100,n2=200,var1=10,var2=1)` to simulate the CIs for \( N(\mu_i, \sigma_i^2) \) data for \( i = 1, 2 \). The terms `pcov`, `mpcov` and `wcov` are the simulated coverages for the pooled, modified pooled and Welch 95% CIs. Record these quantities. Are they near 0.95?
9.12. (Aug. 2009 qual): Let \( X_1, \ldots, X_n \) be a random sample from a \( \text{uniform}(0, \theta) \) distribution. Let \( Y = \max(X_1, X_2, \ldots, X_n) \).

a) Find the pdf of \( Y/\theta \).

b) To find a confidence interval for \( \theta \), can \( Y/\theta \) be used as a pivot?

c) Find the shortest \( (1 - \alpha) \)% confidence interval for \( \theta \).
Chapter 10

Some Useful Distributions

**Definition 10.1.** The population median is any value \( \text{MED}(Y) \) such that

\[
P(Y \leq \text{MED}(Y)) \geq 0.5 \quad \text{and} \quad P(Y \geq \text{MED}(Y)) \geq 0.5.
\]

(10.1)

**Definition 10.2.** The population median absolute deviation is

\[
\text{MAD}(Y) = \text{MED}(|Y - \text{MED}(Y)|).
\]

(10.2)

Finding \( \text{MED}(Y) \) and \( \text{MAD}(Y) \) for symmetric distributions and location–scale families is made easier by the following lemma. Let \( F(y_\alpha) = P(Y \leq y_\alpha) = \alpha \) for \( 0 < \alpha < 1 \) where the cdf \( F(y) = P(Y \leq y) \). Let \( D = \text{MAD}(Y) \), \( M = \text{MED}(Y) = y_{0.5} \) and \( U = y_{0.75} \).

**Lemma 10.1.** a) If \( W = a + bY \), then \( \text{MED}(W) = a + b\text{MED}(Y) \) and \( \text{MAD}(W) = |b|\text{MAD}(Y) \).

b) If \( Y \) has a pdf that is continuous and positive on its support and symmetric about \( \mu \), then \( \text{MED}(Y) = \mu \) and \( \text{MAD}(Y) = y_{0.75} - \text{MED}(Y) \). Find \( M = \text{MED}(Y) \) by solving the equation \( F(M) = 0.5 \) for \( M \), and find \( U \) by solving \( F(U) = 0.75 \) for \( U \). Then \( D = \text{MAD}(Y) = U - M \).

c) Suppose that \( W \) is from a location–scale family with standard pdf \( f_Y(y) \) that is continuous and positive on its support. Then \( W = \mu + \sigma Y \) where \( \sigma > 0 \). First find \( M \) by solving \( F_Y(M) = 0.5 \). After finding \( M \), find \( D \) by solving \( F_Y(M + D) - F_Y(M - D) = 0.5 \). Then \( \text{MED}(W) = \mu + \sigma M \) and \( \text{MAD}(W) = \sigma D \).
Definition 10.3. The *gamma function* $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$ for $x > 0$.

Some properties of the gamma function follow.

i) $\Gamma(k) = (k - 1)!$ for integer $k \geq 1$.

ii) $\Gamma(x + 1) = x \Gamma(x)$ for $x > 0$.

iii) $\Gamma(x) = (x - 1) \Gamma(x - 1)$ for $x > 1$.

iv) $\Gamma(0.5) = \sqrt{\pi}$.

Some lower case Greek letters are alpha: $\alpha$, beta: $\beta$, gamma: $\gamma$, delta: $\delta$, epsilon: $\epsilon$, zeta: $\zeta$, eta: $\eta$, theta: $\theta$, iota: $\iota$, kappa: $\kappa$, lambda: $\lambda$, mu: $\mu$, nu: $\nu$, xi: $\xi$, omicron: o, pi: $\pi$, rho: $\rho$, sigma: $\sigma$, upsilon: $\upsilon$, phi: $\phi$, chi: $\chi$, psi: $\psi$ and omega: $\omega$.

Some capital Greek letters are gamma: $\Gamma$, theta: $\Theta$, sigma: $\Sigma$ and phi: $\Phi$.

For the discrete uniform and geometric distributions, the following facts on series are useful.

Lemma 10.2. Let $n$, $n_1$ and $n_2$ be integers with $n_1 \leq n_2$, and let $a$ be a constant. Notice that $\sum_{i=n_1}^{n_2} a^i = n_2 - n_1 + 1$ if $a = 1$.

- $a) \sum_{i=n_1}^{n_2} a^i = \frac{a^{n_1} - a^{n_2+1}}{1-a}, \ a \neq 1$.
- $b) \sum_{i=0}^{\infty} a^i = \frac{1}{1-a}, \ |a| < 1$.
- $c) \sum_{i=1}^{\infty} a^i = \frac{a}{1-a}, \ |a| < 1$.
- $d) \sum_{i=n_1}^{\infty} a^i = \frac{a^{n_1}}{1-a}, \ |a| < 1$.
- $e) \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.
- $f) \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$. 
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

For the special case of \(0 \leq n_1 \leq n_2\), notice that
\[
\sum_{i=0}^{n_2} a^i = \frac{1 - a^{n_2+1}}{1 - a}, \quad a \neq 1.
\]
To see this, multiply both sides by \((1 - a)\). Then
\[
(1 - a) \sum_{i=0}^{n_2} a^i = (1 - a)(1 + a + a^2 + \cdots + a^{n_2-1} + a^{n_2}) =
1 + a + a^2 + \cdots + a^{n_2-1} + a^{n_2} - a - a^2 - \cdots - a^{n_2} - a^{n_2+1} = 1 - a^{n_2+1}
\]
and the result follows. Hence for \(a \neq 1\),
\[
\sum_{i=n_1}^{n_2} a^i = \sum_{i=0}^{n_2} a^i - \sum_{i=0}^{n_1-1} a^i = \frac{1 - a^{n_2+1}}{1 - a} - \frac{1 - a^{n_1}}{1 - a} = \frac{a^{n_1} - a^{n_2+1}}{1 - a}.
\]
The binomial theorem below is sometimes useful.

**Theorem 10.3, the Binomial Theorem.** For any real numbers \(x\) and \(y\) and for any integer \(n \geq 0\),
\[
(x + y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i} = (y + x)^n = \sum_{i=0}^{n} \binom{n}{i} y^i x^{n-i}.
\]
For the following theorem, see Marshall and Olkin (2007, p. 15-16). Note that part b) follows from part a).

**Theorem 10.4.** If \(E(X)\) exists, then
a) \(E(X) = \int_0^\infty (1 - F(x)) \, dx - \int_{-\infty}^0 F(x) \, dx\).

b) If \(F(x) = 0\) for \(x < 0\), then \(E(X) = \int_0^\infty (1 - F(x)) \, dx\).

### 10.1 The Beta Distribution

If \(Y\) has a beta distribution, \(Y \sim \text{beta}(\delta, \nu)\), then the probability density function (pdf) of \(Y\) is
\[
f(y) = \frac{\Gamma(\delta + \nu)}{\Gamma(\delta)\Gamma(\nu)} y^{\delta-1}(1 - y)^{\nu-1}
\]
where $\delta > 0$, $\nu > 0$ and $0 \leq y \leq 1$.

$$E(Y) = \frac{\delta}{\delta + \nu}.$$  

$$\text{VAR}(Y) = \frac{\delta \nu}{(\delta + \nu)^2(\delta + \nu + 1)}. $$

Notice that

$$f(y) = \frac{\Gamma(\delta + \nu)}{\Gamma(\delta)\Gamma(\nu)} I_{[0,1]}(y) \exp[(\delta - 1)\log(y) + (\nu - 1)\log(1-y)]$$

is a 2P–REF (2-parameter regular exponential family). Hence $\Theta = (0, \infty) \times (0, \infty)$, $\eta_1 = 1 - \delta$, $\eta_2 = 1 - \nu$ and $\Omega = (-\infty, 1) \times (-\infty, 1)$.

If $\delta = 1$, then $W = -\log(1-Y) \sim \text{EXP}(1/\nu).$ Hence $T_n = -\sum \log(1-Y_i) \sim G(n, 1/\nu)$ and if $r > -n$ then $T_n^r$ is the UMVUE of

$$E(T_n^r) = \frac{1}{\nu^r} \frac{\Gamma(r+n)}{\Gamma(n)}. $$

If $\nu = 1$, then $W = -\log(Y) \sim \text{EXP}(1/\delta).$ Hence $T_n = -\sum \log(Y_i) \sim G(n, 1/\delta)$ and if $r > -n$ then $[T_n]^r$ is the UMVUE of

$$E([T_n]^r) = \frac{1}{\delta^r} \frac{\Gamma(r+n)}{\Gamma(n)}. $$

### 10.2 The Beta–Binomial Distribution

If $Y$ has a beta–binomial distribution, $Y \sim \text{BB}(m, \rho, \theta)$, then the probability mass function (pmf) of $Y$ is

$$f(y) = P(Y = y) = \binom{m}{y} \frac{B(\delta + y, \nu + m - y)}{B(\delta, \nu)}$$

for $y = 0, 1, 2, ..., m$ where $0 < \rho < 1$ and $\theta > 0$. Here $\delta = \rho/\theta$ and $\nu = (1-\rho)/\theta$, so $\rho = \delta/(\delta + \nu)$ and $\theta = 1/(\delta + \nu)$. Also

$$B(\delta, \nu) = \frac{\Gamma(\delta)\Gamma(\nu)}{\Gamma(\delta + \nu)}. $$
Hence \( \delta > 0 \) and \( \nu > 0 \). Then \( E(Y) = m\delta/(\delta + \nu) = m\rho \) and \( V(Y) = m\rho(1 - \rho)[1 + (m - 1)\theta/(1 + \theta)] \). If \( Y | \pi \sim \text{binomial}(m, \pi) \) and \( \pi \sim \text{beta}(\delta, \nu) \), then \( Y \sim \text{BB}(m, \rho, \theta) \).

As \( \theta \to 0 \), the beta-binomial \((m, \rho, \theta)\) distribution converges to a binomial\((m, \rho)\) distribution.

### 10.3 The Binomial and Bernoulli Distributions

If \( Y \) has a binomial distribution, \( Y \sim \text{BIN}(k, \rho) \), then the pmf of \( Y \) is

\[
f(y) = P(Y = y) = \binom{k}{y} \rho^y (1 - \rho)^{k-y}
\]

for \( y = 0, 1, \ldots, k \) where \( 0 < \rho < 1 \).

If \( \rho = 0 \), \( P(Y = 0) = 1 = (1 - \rho)^k \) while if \( \rho = 1 \), \( P(Y = k) = 1 = \rho^k \).

The moment generating function

\[
m(t) = [(1 - \rho) + \rho e^t]^k,
\]

and the characteristic function \( c(t) = [(1 - \rho) + \rho e^{it}]^k \).

\[
E(Y) = kp.
\]

\[
\text{VAR}(Y) = kp(1 - \rho).
\]

The Bernoulli \((\rho)\) distribution is the binomial \((k = 1, \rho)\) distribution.

Pourahmadi (1995) showed that the moments of a binomial \((k, \rho)\) random variable can be found recursively. If \( r \geq 1 \) is an integer, \( \binom{0}{0} = 1 \) and the last term below is 0 for \( r = 1 \), then

\[
E(Y^r) = kp \sum_{i=0}^{r-1} \binom{r - 1}{i} E(Y^i) - \rho \sum_{i=0}^{r-2} \binom{r - 1}{i} E(Y^{i+1}).
\]

The following normal approximation is often used.

\[
Y \approx N(kp, kp(1 - \rho))
\]
when $k\rho(1 - \rho) > 9$. Hence

$$ P(Y \leq y) \approx \Phi \left( \frac{y + 0.5 - k\rho}{\sqrt{k\rho(1 - \rho)}} \right). $$

Also

$$ P(Y = y) \approx \frac{1}{\sqrt{k\rho(1 - \rho)}} \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} \frac{(y - k\rho)^2}{k\rho(1 - \rho)} \right). $$

See Johnson, Kotz and Kemp (1992, p. 115). This approximation suggests that $\text{MED}(Y) \approx k\rho$, and $\text{MAD}(Y) \approx 0.674\sqrt{k\rho(1 - \rho)}$. Hamza (1995) states that $|E(Y) - \text{MED}(Y)| \leq \max(\rho, 1 - \rho)$ and shows that

$$ |E(Y) - \text{MED}(Y)| \leq \log(2). $$

If $k$ is large and $k\rho$ small, then $Y \approx \text{Poisson}(k\rho)$.

If $Y_1, \ldots, Y_n$ are independent $\text{BIN}(k_i, \rho)$ then $\sum_{i=1}^n Y_i \sim \text{BIN}(\sum_{i=1}^n k_i, \rho)$.

Notice that

$$ f(y) = \binom{k}{y} I[y \in \{0, \ldots, k\}] (1 - \rho)^k \exp \left[ \log(\frac{\rho}{1 - \rho}) y \right] $$

is a $\text{1P–REF}$ (1-parameter regular exponential family) in $\rho$ if $k$ is known. Thus $\Theta = (0, 1)$,

$$ \eta = \log \left( \frac{\rho}{1 - \rho} \right) $$

and $\Omega = (-\infty, \infty)$.

Assume that $Y_1, \ldots, Y_n$ are iid $\text{BIN}(k, \rho)$, then

$$ T_n = \sum_{i=1}^n Y_i \sim \text{BIN}(nk, \rho). $$

If $k$ is known, then the likelihood

$$ L(\rho) = c \rho^\sum_{i=1}^n y_i (1 - \rho)^{nk - \sum_{i=1}^n y_i}, $$

and the log likelihood

$$ \log(L(\rho)) = d + \log(\rho) \sum_{i=1}^n y_i + (nk - \sum_{i=1}^n y_i) \log(1 - \rho). $$
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

Hence

\[
\frac{d}{d\rho} \log(L(\rho)) = \frac{\sum_{i=1}^{n} y_i}{\rho} + \frac{nk - \sum_{i=1}^{n} y_i}{1 - \rho} (-1) \overset{set}{=} 0,
\]

or \((1 - \rho) \sum_{i=1}^{n} y_i = \rho(nk - \sum_{i=1}^{n} y_i)\), or \(\sum_{i=1}^{n} y_i = \rho nk\) or

\[\hat{\rho} = \sum_{i=1}^{n} y_i / (nk).\]

This solution is unique and

\[
\frac{d^2}{d\rho^2} \log(L(\rho)) = -\frac{\sum_{i=1}^{n} y_i}{\rho^2} - \frac{nk - \sum_{i=1}^{n} y_i}{(1 - \rho)^2} < 0
\]

if \(0 < \sum_{i=1}^{n} y_i < nk\). Hence \(k\hat{\rho} = \bar{Y}\) is the UMVUE, MLE and MME (method of moments estimator) of \(k\rho\) if \(k\) is known.

Let \(\hat{\rho} = \text{number of “successes”} / n\) and let \(P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2\) if \(Z \sim N(0,1)\). Let \(\hat{n} = n + z_{1-\alpha/2}^2\) and

\[\tilde{\rho} = \frac{n\hat{\rho} + 0.5z_{1-\alpha/2}^2}{n + z_{1-\alpha/2}^2}.
\]

Then the large sample 100 \((1 - \alpha)\)% Agresti Coull confidence interval (CI) for \(\rho\) is

\[\hat{\rho} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{\rho}(1 - \hat{\rho})}{\hat{n}}}.
\]

Let \(W = \sum_{i=1}^{n} Y_i \sim \text{bin}(\sum_{i=1}^{n} k_i, \rho)\) and let \(n_w = \sum_{i=1}^{n} k_i\). Often \(k_i \equiv 1\) and then \(n_w = n\). Let \(P(F_{d_1,d_2} \leq F_{d_1,d_2}(\alpha)) = \alpha\) where \(F_{d_1,d_2}\) has an \(F\) distribution with \(d_1\) and \(d_2\) degrees of freedom. Then the Clopper Pearson “exact” 100 \((1 - \alpha)\)% CI for \(\rho\) is

\[
\left(0, \frac{1}{1 + \frac{n_w}{F_{2n_w,2}(\alpha)}}\right) \text{ for } W = 0,
\]

\[
\left(\frac{n_w}{n_w + F_{2,2n_w}(1 - \alpha)}, 1\right) \text{ for } W = n_w,
\]

and \((\rho_L, \rho_U)\) for \(0 < W < n_w\) with

\[
\rho_L = \frac{W}{W + (n_w - W + 1)F_{2(n_w-W+1),2W}(1 - \alpha/2)}
\]

and

\[
\rho_U = \frac{W + 1}{W + 1 + (n_w - W)F_{2(n_w-W),2(W+1)}(\alpha/2)}.
\]
10.4 The Birnbaum Saunders Distribution

If $Y$ has a Birnbaum Saunders distribution, $Y \sim BS(\nu, \theta)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{2\sqrt{2\pi}} \left( \frac{1}{\theta} \sqrt{\frac{\theta}{y} + \frac{\theta}{y^2}} \right) \frac{1}{\nu} \exp \left[ \frac{-1}{2\nu^2} \left( \frac{y}{\theta} + \frac{\theta}{y} - 2 \right) \right] =$$

$$\frac{1}{2\sqrt{2\pi}} \frac{y^2 - \theta^2}{\sqrt{\frac{\theta}{y}} - \sqrt{\frac{\theta}{y}}} \frac{1}{\nu} \exp \left[ \frac{-1}{2\nu^2} \left( \frac{y}{\theta} + \frac{\theta}{y} - 2 \right) \right]$$

$y > 0, \theta > 0$ and $\nu > 0$.

The cdf of $Y$ is

$$F(y) = \Phi \left[ \frac{1}{\nu} \left( \sqrt{\frac{y}{\theta}} - \sqrt{\frac{\theta}{y}} \right) \right]$$

where $\Phi(x)$ is the N(0,1) cdf and $y > 0$. Hence MED($Y$) = $\theta$. Let $W = t(Y) = \frac{Y}{\theta} + \frac{\theta}{y} - 2$, then $W \sim \chi^2_1$.

$E(Y) = \theta (1 + \nu^2/2)$ and $V(Y) = (\nu\theta)^2 (1 + 5\nu^2/6)$.

Suppose $\theta$ is known and $Y_1, ..., Y_i$ are iid. Let $W_i = t(Y_i)$. Then the likelihood

$$L(\nu) = d \frac{1}{\nu^n} \exp \left( \frac{-1}{2\nu^2} \sum_{i=1}^{n} w_i \right),$$

and the log likelihood

$$\log(L(\nu)) = c - n \log(\nu) - \frac{1}{2\nu^2} \sum_{i=1}^{n} w_i.$$ 

Hence

$$\frac{d}{d\nu} \log(L(\nu)) = -\frac{n}{\nu} + \frac{1}{\nu^3} \sum_{i=1}^{n} w_i^\text{set} = 0,$$

or

$$\hat{\nu} = \sqrt{\frac{\sum_{i=1}^{n} w_i}{n}}.$$ 

This solution is unique and

$$\frac{d^2}{d\nu^2} \log(L(\nu)) = \frac{n}{\nu^2} - \frac{3\sum_{i=1}^{n} w_i}{\nu^4} \bigg|_{\nu = \hat{\nu}} = \frac{n}{\hat{\nu}^2} - \frac{3n\hat{\nu}^2}{\hat{\nu}^4} = \frac{-2n}{\hat{\nu}^2} < 0.$$
Thus
\[ \hat{\nu} = \sqrt{\frac{\sum_{i=1}^{n} W_i}{n}} \]
is the MLE of \( \nu \) if \( \hat{\nu} > 0 \).

If \( \nu \) is fixed, this distribution is a scale family with scale parameter \( \theta \).
If \( \theta \) is known this distribution is a 1P-REF. This family of distributions is similar to the lognormal family.

### 10.5 The Burr Type III Distribution

If \( Y \) has a Burr Type III distribution, \( Y \sim BTIII(\phi, \lambda) \), then the pdf of \( Y \) is

\[
f(y) = I(y > 0) \frac{\phi}{\lambda} y^{-(\phi+1)} (1 + y^{-\phi})^{-\left(\frac{1}{\lambda} + 1\right)}
\]

\[
= I(y > 0) \phi y^{-(\phi+1)} \frac{1}{1 + y^{-\phi}} \frac{1}{\lambda} \exp\left[\frac{-1}{\lambda} \log(1 + y^{-\phi})\right]
\]

where \( \phi > 0 \) and \( \lambda > 0 \). This family is a 1P-REF if \( \phi \) is known.

The cdf
\[
F(y) = (1 + y^{-\phi})^{-1/\lambda}
\]
for \( y > 0 \).

\( X = 1/Y \sim BTXII(\lambda, \phi) \).

\( W = \log(1 + Y^{-\phi}) \sim EXP(\lambda) \).

If \( Y_1, \ldots, Y_n \) are iid \( BTIII(\lambda, \phi) \), then

\[
T_n = \sum_{i=1}^{n} \log(1 + Y_i^{-\phi}) \sim G(n, \lambda).
\]

If \( \phi \) is known, then the likelihood

\[
L(\lambda) = c \frac{1}{\lambda^n} \exp\left[\frac{-1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^{-\phi})\right],
\]

and the log likelihood \( \log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^{-\phi}) \). Hence

\[
\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{\sum_{i=1}^{n} \log(1 + y_i^{-\phi})}{\lambda^2} = 0,
\]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

or \( \sum_{i=1}^{n} \log(1 + y_i^{-\phi}) = n\lambda \) or

\[
\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + y_i^{-\phi})}{n}.
\]

This solution is unique and

\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2n\lambda}{\lambda^3} = n\frac{\lambda^2 - 2\lambda^3}{\lambda^3} = -n < 0.
\]

Thus

\[
\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + Y_i^{-\phi})}{n}
\]

is the UMVUE and MLE of \( \lambda \) if \( \phi \) is known.

If \( \phi \) is known and \( r > -n \), then \( T_{nr} \) is the UMVUE of

\[
E(T_{nr}) = \lambda \frac{\Gamma(r+n)}{\Gamma(n)}.
\]

10.6 The Burr Type X Distribution

If \( Y \) has a Burr Type X distribution, \( Y \sim BTX(\tau) \), then the pdf of \( Y \) is

\[
f(y) = I(y > 0) \frac{2}{\tau} y e^{-y^2} (1 - e^{-y^2})^{\tau-1} =
\]

\[
I(y > 0) \frac{2y e^{-y^2}}{1 - e^{-y^2}} \tau \exp[\tau \log(1 - e^{-y^2})]
\]

where \( \tau > 0 \). This family is a 1P-REF. \( W = -\log(1 - e^{-Y^2}) \sim EXP(1/\tau) \) and \( MED(W) = \log(2)/\tau \).

Given data \( Y_1, ..., Y_n \), a robust estimator of \( \tau \) is \( \hat{\tau} = \log(2)/MED(n) \) where \( MED(n) \) is the sample median of \( W_1, ..., W_n \) and \( W_i = -\log(1 - e^{-Y_i^2}) \).

If \( Y_1, ..., Y_n \) are iid BTX(\( \tau \)), then

\[
T_n = -\sum_{i=1}^{n} \log(1 - e^{-Y_i^2}) \sim G(n, 1/\tau).
\]

The likelihood \( L(\tau) = c \tau^n \exp[\tau \sum_{i=1}^{n} \log(1 - e^{-y_i^2})] \), and the log likelihood \( \log(L(\tau)) = d + n\log(\tau) + \tau \sum_{i=1}^{n} \log(1 - e^{-y_i^2}) \).
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

Hence
\[ \frac{d}{d\tau} \log(L(\tau)) = \frac{n}{\tau} + \sum_{i=1}^{n} \log(1 - e^{-y_i^2}) \leq 0, \]
or
\[ \hat{\tau} = \frac{n}{-\sum_{i=1}^{n} \log(1 - e^{-y_i^2})}. \]
This solution is unique and
\[ \frac{d^2}{d\tau^2} \log(L(\tau)) = \frac{-n}{\tau^2} < 0. \]
Thus
\[ \hat{\tau} = \frac{n}{-\sum_{i=1}^{n} \log(1 - e^{-Y_i^2})} \]
is the MLE of \( \tau \).
Now \((n - 1)\hat{\tau}/n\) is the UMVUE of \( \tau \), and if \( r > -n \), then \([T_n]^r\) is the UMVUE of
\[ E([T_n]^r) = \frac{\Gamma(r + n)}{\tau^r \Gamma(n)}. \]

10.7 The Burr Type XII Distribution

If \( Y \) has a Burr Type XII distribution, \( Y \sim \text{BTXII}(\phi, \lambda) \), then the pdf of \( Y \) is
\[ f(y) = \frac{1}{\lambda} \frac{\phi y^{\phi-1}}{(1 + y^{\phi})^{1+1}} \]
where \( y, \phi, \) and \( \lambda \) are all positive. The cdf of \( Y \) is
\[ F(y) = 1 - \exp\left[ -\frac{\log(1 + y^\phi)}{\lambda} \right] = 1 - (1 + y^\phi)^{-1/\lambda} \text{ for } y > 0. \]
MED(\( Y \)) = \( [e^{\lambda \log(2)} - 1]^{1/\phi} \). See Patel, Kapadia and Owen (1976, p. 195).
\( W = \log(1 + Y^\phi) \) is EXP(\( \lambda \)).

Notice that
\[ f(y) = \phi y^{\phi-1} \frac{1}{1 + y^\phi} \frac{1}{\lambda} \exp\left[ -\frac{1}{\lambda} \log(1 + y^\phi) \right] I(y > 0) \]
is a 1P-REF if \( \phi \) is known.
If $Y_1, \ldots, Y_n$ are iid BTXII($\lambda, \phi$), then

$$T_n = \sum_{i=1}^{n} \log(1 + Y_i^\phi) \sim G(n, \lambda).$$

If $\phi$ is known, then the likelihood

$$L(\lambda) = c \frac{1}{\lambda^d} \exp \left[ -\frac{1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^\phi) \right],$$

and the log likelihood $\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^\phi)$. Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{\sum_{i=1}^{n} \log(1 + y_i^\phi)}{\lambda^2} \text{ set } = 0,$$

or $\sum_{i=1}^{n} \log(1 + y_i^\phi) = n\lambda$ or

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + y_i^\phi)}{n}.$$

This solution is unique and

$$\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2 \sum_{i=1}^{n} \log(1 + y_i^\phi)}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}} = \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = \frac{-n}{\lambda^2} < 0.$$

Thus

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + Y_i^\phi)}{n}$$

is the UMVUE and MLE of $\lambda$ if $\phi$ is known.

If $\phi$ is known and $r > -n$, then $T_r^n$ is the UMVE of

$$E(T_r^n) = \lambda r \frac{\Gamma(r + n)}{\Gamma(n)}.$$

### 10.8 The Cauchy Distribution

If $Y$ has a Cauchy distribution, $Y \sim C(\mu, \sigma)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\pi \sigma^2} \frac{1}{\pi \sigma^2 + (y - \mu)^2} = \frac{1}{\pi \sigma [1 + (\frac{y - \mu}{\sigma})^2]}$$
where \( y \) and \( \mu \) are real numbers and \( \sigma > 0 \). The cumulative distribution function (cdf) of \( Y \) is

\[
F(y) = \frac{1}{\pi} \arctan\left(\frac{y - \mu}{\sigma}\right) + \frac{\pi}{2}.
\]

See Ferguson (1967, p. 102). This family is a location–scale family that is symmetric about \( \mu \).

The moments of \( Y \) do not exist, but the characteristic function of \( Y \) is

\[
c(t) = \exp(it\mu - |t|\sigma).
\]

MED(\( Y \)) = \( \mu \), the upper quartile = \( \mu + \sigma \), and the lower quartile = \( \mu - \sigma \). MAD(\( Y \)) = \( F^{-1}(3/4) - \text{MED}(Y) = \sigma \).

If \( Y_1, \ldots, Y_n \) are independent \( C(\mu_i, \sigma_i) \), then

\[
\sum_{i=1}^{n} a_i Y_i \sim C\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} |a_i| \sigma_i\right).
\]

In particular, if \( Y_1, \ldots, Y_n \) are iid \( C(\mu, \sigma) \), then \( \overline{Y} \sim C(\mu, \sigma) \).

If \( W \sim U(-\pi/2, \pi/2) \), then \( Y = \tan(W) \sim C(0, 1) \).

**10.9 The Chi Distribution**

If \( Y \) has a chi distribution (also called a \( p \)-dimensional Rayleigh distribution), \( Y \sim \text{chi}(p, \sigma) \), then the pdf of \( Y \) is

\[
f(y) = \frac{y^{p-1} e^{-\frac{y^2}{2\sigma^2}}}{\sigma^p 2^{\frac{p}{2}-1} \Gamma(p/2)}
\]

where \( y > 0 \) and \( \sigma, p > 0 \). This is a scale family if \( p \) is known.

\[
E(Y) = \sigma \sqrt{2} \frac{\Gamma(\frac{1+p}{2})}{\Gamma(p/2)}.
\]

\[
\text{VAR}(Y) = 2\sigma^2 \left[ \frac{\Gamma(\frac{2+p}{2})}{\Gamma(p/2)} - \left( \frac{\Gamma(\frac{1+p}{2})}{\Gamma(p/2)} \right)^2 \right],
\]
CHAPTER 10.  SOME USEFUL DISTRIBUTIONS

and

\[ E(Y^r) = 2^{r/2} \sigma^r \frac{\Gamma\left(\frac{r+p}{2}\right)}{\Gamma(p/2)} \]

for \( r > -p \).

The mode is at \( \sigma \sqrt{p-1} \) for \( p \geq 1 \). See Cohen and Whitten (1988, ch. 10). Note that \( W = Y^2 \sim G(p/2, 2\sigma^2) \).

\[ Y \sim \text{generalized gamma} \ (\nu = p/2, \lambda = \sigma \sqrt{2}, \phi = 2). \]

If \( p = 1 \), then \( Y \) has a half normal distribution, \( Y \sim \text{HN}(0, \sigma^2) \).

If \( p = 2 \), then \( Y \) has a Rayleigh distribution, \( Y \sim \text{R}(0, \sigma) \).

If \( p = 3 \), then \( Y \) has a Maxwell–Boltzmann distribution, \( Y \sim \text{MB} (0, \sigma) \).

If \( p \) is an integer and \( Y \sim \chi(p, 1) \), then \( Y^2 \sim \chi_p^2 \).

Since

\[ f(y) = I(y > 0) \frac{1}{2^{p-1} \Gamma(p/2) \sigma^p} \exp[(p-1) \log(y) - \frac{1}{2\sigma^2} y^2], \]

this family is a \textbf{2P–REF}. Notice that \( \Theta = (0, \infty) \times (0, \infty) \), \( \eta_1 = p - 1 \), \( \eta_2 = -1/(2\sigma^2) \), and \( \Omega = (-1, \infty) \times (-\infty, 0) \).

If \( p \) is known then

\[ f(y) = \frac{y^{p-1}}{2^{p-1} \Gamma(p/2) \sigma^p} I(y > 0) \frac{1}{\sigma^p} \exp\left[\frac{-1}{2\sigma^2} y^2\right] \]

is a \textbf{1P–REF}.

If \( Y_1, ..., Y_n \) are iid \( \chi(p, \sigma) \), then

\[ T_n = \sum_{i=1}^{n} Y_i^2 \sim G(np/2, 2\sigma^2). \]

If \( p \) is known, then the likelihood

\[ L(\sigma^2) = c \frac{1}{\sigma^{np}} \exp\left[\frac{-1}{2\sigma^2} \sum_{i=1}^{n} y_i^2\right], \]

and the log likelihood

\[ \log(L(\sigma^2)) = d - \frac{np}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2. \]
Hence
\[
\frac{d}{d(\sigma^2)} \log(\sigma^2) = \frac{-np}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} y_i^2 \gtrless 0,
\]
or \[
\sum_{i=1}^{n} y_i^2 = np\sigma^2 \text{ or } \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} y_i^2}{np}.
\]
This solution is unique and
\[
\frac{d^2}{d(\sigma^2)^2} \log(L(\sigma^2)) = \frac{np}{2(\sigma^2)^2} - \left. \frac{\sum_{i=1}^{n} y_i^2}{(\sigma^2)^3} \right|_{\sigma^2 = \hat{\sigma}^2} = \frac{np}{2(\hat{\sigma}^2)^2} - \frac{np\hat{\sigma}^2}{(\hat{\sigma}^2)^3} \cdot \frac{2}{2} = -\frac{np}{2(\hat{\sigma}^2)^2} < 0.
\]
Thus \( \hat{\sigma}^2 \)
\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} Y_i^2}{np}
\]
is the UMVUE and MLE of \( \sigma^2 \) when \( p \) is known.

If \( p \) is known and \( r > -np/2 \), then \( T_n^r \) is the UMVUE of
\[
E(T_n^r) = \frac{2^r \sigma^{2r} \Gamma(r + np/2)}{\Gamma(np/2)}.
\]

### 10.10 The Chi–square Distribution

If \( Y \) has a chi–square distribution, \( Y \sim \chi^2_p \), then the pdf of \( Y \) is
\[
f(y) = \frac{y^{p-1} e^{-y/2}}{2^p \Gamma(p/2)}
\]
where \( y \geq 0 \) and \( p \) is a positive integer. The mgf of \( Y \) is
\[
m(t) = \left( \frac{1}{1-2t} \right)^{p/2} = (1 - 2t)^{-p/2}
\]
for \( t < 1/2 \). The characteristic function
\[
c(t) = \left( \frac{1}{1 - i2t} \right)^{p/2}.
\]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

\( E(Y) = p. \)
\( \text{VAR}(Y) = 2p. \)

Since \( Y \) is gamma \( G(\nu = p/2, \lambda = 2) \),
\[
E(Y^r) = \frac{2^r \Gamma(r + p/2)}{\Gamma(p/2)}, \quad r > -p/2.
\]

\( \text{MED}(Y) \approx p - 2/3. \) See Pratt (1968, p. 1470) for more terms in the expansion of \( \text{MED}(Y) \). Empirically,
\[
\text{MAD}(Y) \approx \sqrt{2p} \left( 1 - \frac{2}{9p} \right)^2 \approx 0.9536 \sqrt{p}.
\]

There are several normal approximations for this distribution. The Wilson–Hilferty approximation is
\[
\left( \frac{Y}{p} \right)^{1/3} \approx N(1 - \frac{2}{9p}, \frac{2}{9p}).
\]

See Bowman and Shenton (1992, p. 6). This approximation gives
\[
P(Y \leq x) \approx \Phi(\sqrt{\frac{2}{9p}} \left( 1 + \frac{2}{9p} \right) \sqrt{\frac{1}{9p/2}}),
\]
and
\[
\chi^2_{p, \alpha} \approx p(z_{\alpha} \sqrt{\frac{2}{9p}} + 1 - \frac{2}{9p})^3
\]
where \( z_{\alpha} \) is the standard normal percentile, \( \alpha = \Phi(z_{\alpha}) \). The last approximation is good if \( p > -1.24 \log(\alpha) \). See Kennedy and Gentle (1980, p. 118).

This family is a one parameter exponential family, but is not a REF since the set of integers does not contain an open interval.

10.11 The Discrete Uniform Distribution

If \( Y \) has a discrete uniform distribution, \( Y \sim DU(\theta_1, \theta_2) \), then the pmf of \( Y \) is
\[
f(y) = P(Y = y) = \frac{1}{\theta_2 - \theta_1 + 1}
\]
for \( \theta_1 \leq y \leq \theta_2 \) where \( y \) and the \( \theta_i \) are integers. Let \( \theta_2 = \theta_1 + \tau - 1 \) where \( \tau = \theta_2 - \theta_1 + 1 \). The cdf for \( Y \) is

\[
F(y) = \frac{\lfloor y \rfloor - \theta_1 + 1}{\theta_2 - \theta_1 + 1}
\]

for \( \theta_1 \leq y \leq \theta_2 \). Here \( \lfloor y \rfloor \) is the greatest integer function, eg, \( \lfloor 7.7 \rfloor = 7 \). This result holds since for \( \theta_1 \leq y \leq \theta_2 \),

\[
E(Y) = \frac{\theta_1 + \theta_2}{2} = \theta_1 + \frac{(\tau - 1)}{2} \text{ while } V(Y) = \frac{(\tau^2 - 1)}{12}. \text{ The result for } E(Y) \text{ follows by symmetry, or because}
\]

\[
E(Y) = \sum_{y=\theta_1}^{\theta_2} \frac{y}{\theta_2 - \theta_1 + 1} = \frac{\theta_1(\theta_2 - \theta_1 + 1) + [0 + 1 + 2 + \cdots + (\theta_2 - \theta_1)]}{\theta_2 - \theta_1 + 1}
\]

where last equality follows by adding and subtracting \( \theta_1 \) to \( y \) for each of the \( \theta_2 - \theta_1 + 1 \) terms in the middle sum. Thus

\[
E(Y) = \theta_1 + \frac{(\theta_2 - \theta_1)(\theta_2 - \theta_1 + 1)}{2(\theta_2 - \theta_1 + 1)} = \frac{2\theta_1}{2} + \frac{\theta_2 - \theta_1}{2} = \frac{\theta_1 + \theta_2}{2}
\]

since \( \sum_{i=1}^{n} i = n(n + 1)/2 \) by Lemma 10.2 e) with \( n = \theta_2 - \theta_1 \).

To see the result for \( V(Y) \), let \( W = Y - \theta_1 + 1 \). Then \( V(Y) = V(W) \) and \( f(w) = 1/\tau \) for \( w = 1, ..., \tau \). Hence \( W \sim DU(1, \tau) \),

\[
E(W) = \frac{1}{\tau} \sum_{w=1}^{\tau} w = \frac{\tau(\tau + 1)}{2\tau} = \frac{1 + \tau}{2},
\]

and

\[
E(W^2) = \frac{1}{\tau} \sum_{w=1}^{\tau} w^2 = \frac{\tau(\tau + 1)(2\tau + 1)}{6\tau} = \frac{(\tau + 1)(2\tau + 1)}{6}
\]

by Lemma 10.2 e) and f). So

\[
V(Y) = V(W) = E(W^2) - (E(W))^2 = \frac{(\tau + 1)(2\tau + 1)}{6} - \left(\frac{1 + \tau}{2}\right)^2
\]
2(\tau + 1)(2\tau + 1) - 3(\tau + 1)^2 = 2(\tau + 1)[2(\tau + 1) - 1] - 3(\tau + 1)^2 = \frac{4(\tau + 1)^2 - 2(\tau + 1) - 3(\tau + 1)^2}{12} = \frac{(\tau + 1)^2 - 2\tau - 2}{12} = \frac{\tau^2 + 2\tau + 1 - 2\tau - 2}{12} = \frac{\tau^2 - 1}{12}.

Let \mathcal{Z} be the set of integers and let \( Y_1, \ldots, Y_n \) be iid \( DU(\theta_1, \theta_2) \). Then the likelihood function \( L(\theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1 + 1)^n} I(\theta_1 \leq Y_1) I(\theta_2 \geq Y_n) I(\theta_1 \leq \theta_2) I(\theta_1 \in \mathcal{Z}) I(\theta_2 \in \mathcal{Z}) \) is maximized by making \( \theta_2 - \theta_1 - 1 \) as small as possible where integers \( \theta_2 \geq \theta_1 \).

So need \( \theta_2 \) as small as possible and \( \theta_1 \) as large as possible, and the MLE of \( (\theta_1, \theta_2) \) is \((Y_1, Y_n)\).

### 10.12 The Double Exponential Distribution

If \( Y \) has a double exponential distribution (or Laplace distribution), \( Y \sim DE(\theta, \lambda) \), then the pdf of \( Y \) is

\[
f(y) = \frac{1}{2\lambda} \exp \left( -\frac{|y - \theta|}{\lambda} \right)
\]

where \( y \) is real and \( \lambda > 0 \). The cdf of \( Y \) is

\[
F(y) = 0.5 \exp \left( \frac{y - \theta}{\lambda} \right) \text{ if } y \leq \theta,
\]

and

\[
F(y) = 1 - 0.5 \exp \left( -\frac{(y - \theta)}{\lambda} \right) \text{ if } y \geq \theta.
\]

This family is a location–scale family which is symmetric about \( \theta \).

The mgf

\[
m(t) = \exp(\theta t)/(1 - \lambda^2 t^2)
\]

for \(|t| < 1/\lambda\), and the characteristic function \( c(t) = \exp(\theta it)/(1 + \lambda^2 t^2) \).

\[E(Y) = \theta,\text{ and}
\]

\[\text{MED}(Y) = \theta.\]
VAR(Y) = 2\lambda^2, and
MAD(Y) = \log(2)\lambda \approx 0.693\lambda.
Hence \lambda = \text{MAD}(Y) / \log(2) \approx 1.443\text{MAD}(Y).
To see that MAD(Y) = \lambda \log(2), note that F(\theta + \lambda \log(2)) = 1 - 0.25 = 0.75.
The maximum likelihood estimators are \hat{\theta}_{MLE} = \text{MED}(n) and
\hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \text{MED}(n)|.
A 100(1 - \alpha)\% CI for \lambda is
\left( \frac{2\sum_{i=1}^{n} |Y_i - \text{MED}(n)|}{\chi_{2n-1,1-\frac{\alpha}{2}}^2}, \frac{2\sum_{i=1}^{n} |Y_i - \text{MED}(n)|}{\chi_{2n-1,\frac{\alpha}{2}}^2} \right),
and a 100(1 - \alpha)\% CI for \theta is
\left( \text{MED}(n) \pm \frac{z_{1-\alpha/2} \sum_{i=1}^{n} |Y_i - \text{MED}(n)|}{n \sqrt{n - z_{1-\alpha/2}^2}} \right)
where \chi_{p,\alpha}^2 and z_\alpha are the \alpha percentiles of the \chi_p^2 and standard normal distributions, respectively. See Patel, Kapadia and Owen (1976, p. 194).
W = |Y - \theta| \sim \text{EXP(\lambda)}.
Notice that
f(y) = \frac{1}{2\lambda} \exp \left[ \frac{-1}{\lambda} |y - \theta| \right]
is a one parameter exponential family in \lambda if \theta is known.
If Y_1, ..., Y_n are iid \text{DE}(\theta, \lambda) then
T_n = \sum_{i=1}^{n} |Y_i - \theta| \sim G(n, \lambda).
If \theta is known, then the likelihood
L(\lambda) = c \frac{1}{\lambda^n} \exp \left[ \frac{-1}{\lambda} \sum_{i=1}^{n} |y_i - \theta| \right],
and the log likelihood
\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} |y_i - \theta|. 
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

Hence
\[ \frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} |y_i - \theta| \]
or \[ \sum_{i=1}^{n} |y_i - \theta| = n\lambda \] or
\[ \hat{\lambda} = \frac{\sum_{i=1}^{n} |y_i - \theta|}{n}. \]
This solution is unique and
\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2 \sum_{i=1}^{n} |y_i - \theta|}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}} = \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = -\frac{n}{\lambda^2} < 0. \]
Thus
\[ \hat{\lambda} = \frac{\sum_{i=1}^{n} |Y_i - \theta|}{n} \]
is the UMVUE and MLE of \( \lambda \) if \( \theta \) is known.

10.13 The Exponential Distribution

If \( Y \) has an exponential distribution, \( Y \sim \text{EXP}(\lambda) \), then the pdf of \( Y \) is
\[ f(y) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right) I(y \geq 0) \]
where \( \lambda > 0 \). The cdf of \( Y \) is
\[ F(y) = 1 - \exp(-y/\lambda), \ y \geq 0. \]
This distribution is a scale family with scale parameter \( \lambda \).

The mgf
\[ m(t) = 1/(1 - \lambda t) \]
for \( t < 1/\lambda \), and the characteristic function \( c(t) = 1/(1 - i\lambda t) \).

\( E(Y) = \lambda \),
and \( \text{VAR}(Y) = \lambda^2 \).
\( W = 2Y/\lambda \sim \chi^2_2 \).

Since \( Y \) is gamma \( G(\nu = 1, \lambda) \), \( E(Y^r) = \lambda^r \Gamma(r + 1) \) for \( r > -1 \).

\( \text{MED}(Y) = \log(2)\lambda \) and
\( \text{MAD}(Y) \approx \lambda/2.0781 \) since it can be shown that
\[ \exp(\text{MAD}(Y)/\lambda) = 1 + \exp(-\text{MAD}(Y)/\lambda). \]
Hence $\text{MAD}(Y) \approx \lambda$.

The classical estimator is $\hat{\lambda} = \overline{Y}_n$ and the $100(1 - \alpha)\%$ CI for $E(Y) = \lambda$ is

\[
\left( \frac{2 \sum_{i=1}^n Y_i}{\chi^2_{2n,1-\frac{\alpha}{2}}}, \frac{2 \sum_{i=1}^n Y_i}{\chi^2_{2n,\frac{\alpha}{2}}} \right)
\]

where $P(Y \leq \chi^2_{2n,\frac{\alpha}{2}}) = \alpha/2$ if $Y$ is $\chi^2_{2n}$. See Patel, Kapadia and Owen (1976, p. 188).

Notice that $f(y) = I(y \geq 0) \frac{1}{\lambda} \exp\left[\frac{-1}{\lambda} y\right]$ is a $1P$–$\text{REF}$. Hence $\Theta = (0, \infty)$, $\eta = -1/\lambda$ and $\Omega = (-\infty, 0)$.

Suppose that $Y_1, ..., Y_n$ are iid $\text{EXP}(\lambda)$, then

\[ T_n = \sum_{i=1}^n Y_i \sim G(n, \lambda). \]

The likelihood

\[ L(\lambda) = \frac{1}{\lambda^n} \exp\left[\frac{-1}{\lambda} \sum_{i=1}^n y_i\right], \]

and the log likelihood

\[ \log(L(\lambda)) = -n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^n y_i. \]

Hence

\[ \frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^n y_i \overset{\text{set}}{=} 0, \]

or $\sum_{i=1}^n y_i = n\lambda$ or $\hat{\lambda} = \overline{y}$.

Since this solution is unique and

\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) \bigg|_{\lambda = \hat{\lambda}} = \frac{n}{\lambda^2} - \frac{2 \sum_{i=1}^n y_i}{\lambda^3} = \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = \frac{-n}{\lambda^2} < 0, \]

the $\hat{\lambda} = \overline{Y}$ is the UMVUE, MLE and MME of $\lambda$.

If $r > -n$, then $T_n^r$ is the UMVUE of

\[ E(T_n^r) = \frac{\lambda^r \Gamma(r + n)}{\Gamma(n)}. \]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

10.14 The Two Parameter Exponential Distribution

If $Y$ has a 2 parameter exponential distribution, $Y \sim \text{EXP}(\theta, \lambda)$ then the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda} \exp \left( -\frac{(y - \theta)}{\lambda} \right) I(y \geq \theta)$$

where $\lambda > 0$ and $\theta$ is real. The cdf of $Y$ is

$$F(y) = 1 - \exp[-(y - \theta)/\lambda], \; y \geq \theta.$$  

This family is an asymmetric location-scale family. The mgf

$$m(t) = \exp(t\theta)/(1 - \lambda t)$$

for $t < 1/\lambda$, and the characteristic function $c(t) = \exp(it\theta)/(1 - i\lambda t)$. $E(Y) = \theta + \lambda$, and $\text{VAR}(Y) = \lambda^2$.

\[\text{MED}(Y) = \theta + \lambda \log(2)\]

and

\[\text{MAD}(Y) \approx \lambda/2.0781.\]

Hence $\theta \approx \text{MED}(Y) - 2.0781 \log(2)\text{MAD}(Y)$. See Rousseeuw and Croux (1993) for similar results. Note that $2.0781 \log(2) \approx 1.44$.

To see that $2.0781 \text{MAD}(Y) \approx \lambda$, note that

\[0.5 = \int_{\theta + \lambda \log(2) - \text{MAD}}^{\theta + \lambda \log(2) + \text{MAD}} \frac{1}{\lambda} \exp\left(-\frac{(y - \theta)}{\lambda}\right) dy\]

\[= 0.5\left[ -e^{-\text{MAD}/\lambda} + e^{\text{MAD}/\lambda}\right]\]

assuming $\lambda \log(2) > \text{MAD}$. Plug in $\text{MAD} = \lambda/2.0781$ to get the result.

If $\theta$ is known, then

$$f(y) = I(y \geq \theta) \frac{1}{\lambda} \exp \left[ -\frac{1}{\lambda} (y - \theta) \right]$$
is a 1P–REF in $\lambda$. Notice that $Y - \theta \sim EXP(\lambda)$. Let

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} (Y_i - \theta)}{n}. $$

Then $\hat{\lambda}$ is the UMVUE and MLE of $\lambda$ if $\theta$ is known.

If $Y_1, ..., Y_n$ are iid $EXP(\theta, \lambda)$, then the likelihood

$$L(\theta, \lambda) = \frac{1}{\lambda^n} \exp \left[ -\frac{1}{\lambda} \sum_{i=1}^{n} (y_i - \theta) \right] I(y_{(1)} \geq \theta).$$

Need $y_{(1)} \geq \hat{\theta}$, and for $y_{(1)} \geq \theta$, the log likelihood

$$\log(L(\theta, \lambda)) = [-n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} (y_i - \theta)] I(y_{(1)} \geq \theta).$$

For any fixed $\lambda > 0$, the log likelihood is maximized by maximizing $\theta$. Hence $\hat{\theta} = Y_{(1)}$, and the profile log likelihood is

$$\log(L(\lambda|y_{(1)})) = -n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} (y_i - y_{(1)})$$

is maximized by $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} (y_i - y_{(1)})$. Hence the MLE

$$(\hat{\theta}, \hat{\lambda}) = \left( Y_{(1)}, \frac{1}{n} \sum_{i=1}^{n} (Y_i - Y_{(1)}) \right) = (Y_{(1)}, \overline{Y} - Y_{(1)}).$$

Let $D_n = \sum_{i=1}^{n} (Y_i - Y_{(1)}) = n \hat{\lambda}$. Then for $n \geq 2$,

$$\left( \frac{2D_n}{\chi^2_{2(n-1),1-\alpha/2}}, \frac{2D_n}{\chi^2_{2(n-1),\alpha/2}} \right)$$

(10.3)

is a $100(1 - \alpha)\%$ CI for $\lambda$, while

$$\left( Y_{(1)} - \hat{\lambda}[(\alpha)^{-1/(n-1)} - 1], Y_{(1)} \right)$$

(10.4)

is a $100(1 - \alpha)\%$ CI for $\theta$. See Mann, Schafer, and Singpurwalla (1974, p. 176).

If $\theta$ is known and $T_n = \sum_{i=1}^{n} (Y_i - \theta)$, then a $100(1 - \alpha)\%$ CI for $\lambda$ is

$$\left( \frac{2T_n}{\chi^2_{2(n-1),1-\alpha/2}}, \frac{2T_n}{\chi^2_{2(n-1),\alpha/2}} \right).$$

(10.5)
10.15 The F Distribution

If \( Y \) has an F distribution, \( Y \sim F(\nu_1, \nu_2) \), then the pdf of \( Y \) is

\[
f(y) = \frac{\Gamma\left(\frac{\nu_1 + \nu_2}{2}\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{y^{(\nu_1-2)/2}}{\left(1 + \left(\frac{\nu_1}{\nu_2}\right)y\right)^{(\nu_1+\nu_2)/2}}
\]

where \( y > 0 \) and \( \nu_1 \) and \( \nu_2 \) are positive integers.

\[
E(Y) = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2
\]

and

\[
\text{VAR}(Y) = 2 \left(\frac{\nu_2}{\nu_2 - 2}\right)^2 \frac{(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}, \quad \nu_2 > 4.
\]

\[
E(Y^r) = \frac{\Gamma\left(\frac{\nu_1 + 2r}{2}\right)\Gamma\left(\frac{\nu_2 - 2r}{2}\right)}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_2}{\nu_1}\right)^r, \quad r < \nu_2/2.
\]

Suppose that \( X_1 \) and \( X_2 \) are independent where \( X_1 \sim \chi^2_{\nu_1} \) and \( X_2 \sim \chi^2_{\nu_2} \). Then

\[
W = \left(\frac{X_1/\nu_1}{X_2/\nu_2}\right) \sim F(\nu_1, \nu_2).
\]

Notice that \( E(Y^r) = E(W^r) = \left(\frac{\nu_2}{\nu_1}\right)^r E(X_1^r) E(X_2^{-r}) \).

If \( W \sim t_v \), then \( Y = W^2 \sim F(1, \nu) \).

10.16 The Gamma Distribution

If \( Y \) has a gamma distribution, \( Y \sim G(\nu, \lambda) \), then the pdf of \( Y \) is

\[
f(y) = \frac{y^{\nu-1}e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)}
\]

where \( \nu, \lambda, \) and \( y \) are positive. The mgf of \( Y \) is

\[
m(t) = \left(\frac{1/\lambda}{1 - t}\right)^\nu = \left(\frac{1}{1 - \lambda t}\right)^\nu
\]
for $t < 1/\lambda$. The characteristic function

$$c(t) = \left( \frac{1}{1 - i\lambda t} \right)^\nu.$$ 

$E(Y) = \nu\lambda$.

$\text{VAR}(Y) = \nu\lambda^2$.

$$E(Y^r) = \frac{\lambda^r \Gamma(r + \nu)}{\Gamma(\nu)} \quad \text{if} \quad r > -\nu. \quad (10.6)$$

Chen and Rubin (1986) show that $\lambda(\nu - 1/3) < \text{MED}(Y) < \lambda\nu = E(Y)$.

Empirically, for $\nu > 3/2$,

$$\text{MED}(Y) \approx \lambda(\nu - 1/3),$$

and

$$\text{MAD}(Y) \approx \frac{\lambda\sqrt{\nu}}{1.483}.$$ 

This family is a scale family for fixed $\nu$, so if $Y$ is $G(\nu, \lambda)$ then $cY$ is $G(\nu, c\lambda)$ for $c > 0$. If $W$ is $\text{EXP}(\lambda)$ then $W$ is $G(1, \lambda)$. If $W$ is $\chi^2_p$, then $W$ is $G(p/2, 2)$.

Some classical estimators are given next. Let

$$w = \log \left[ \frac{\bar{y}}{\text{geometric mean}(n)} \right]$$

where geometric mean$(n) = (y_1 y_2 \ldots y_n)^{1/n} = \exp\left[ \frac{1}{n} \sum_{i=1}^n \log(y_i) \right]$. Then

Thom’s estimator (Johnson and Kotz 1970a, p. 188) is

$$\hat{\nu} \approx 0.25(1 + \sqrt{1 + 4w/3}) \frac{1}{w}.$$ 

Also

$$\hat{\nu}_{MLE} \approx \frac{0.5000876 + 0.1648852w - 0.0544274w^2}{w}$$

for $0 < w \leq 0.5772$, and

$$\hat{\nu}_{MLE} \approx \frac{8.898919 + 9.059950w + 0.9775374w^2}{w(17.79728 + 11.968477w + w^2)}$$

for $0.5772 < w \leq 17$. If $W > 17$ then estimation is much more difficult, but a rough approximation is $\hat{\nu} \approx 1/w$ for $w > 17$. See Bowman and Shenton
(1988, p. 46) and Greenwood and Durand (1960). Finally, \( \hat{\lambda} = \bar{Y}_n/\hat{\nu} \). Notice that \( \hat{\lambda} \) may not be very good if \( \hat{\nu} < 1/17 \).

Several normal approximations are available. The Wilson–Hilferty approximation says that for \( \nu > 0.5 \),

\[
Y^{1/3} \approx N \left( (\nu \lambda)^{1/3}(1 - \frac{1}{9\nu}), (\nu \lambda)^{2/3} \frac{1}{9\nu} \right).
\]

Hence if \( Y \) is \( G(\nu, \lambda) \) and

\[
\alpha = P[Y \leq G_\alpha],
\]

then

\[
G_\alpha \approx \nu \lambda \left[ z_\alpha \sqrt{\frac{1}{9\nu} + 1 - \frac{1}{9\nu}} \right]^3
\]

where \( z_\alpha \) is the standard normal percentile, \( \alpha = \Phi(z_\alpha) \). Bowman and Shenton (1988, p. 101) include higher order terms.

Notice that

\[
f(y) = I(y > 0) \frac{1}{\lambda^{\nu} \Gamma(\nu)} \exp\left[ \frac{-1}{\nu} y + (\nu - 1) \log(y) \right]
\]

is a 2P–REF. Hence \( \Theta = (0, \infty) \times (0, \infty) \), \( \eta_1 = -1/\lambda \), \( \eta_2 = \nu - 1 \) and \( \Omega = (-\infty, 0) \times (-1, \infty) \).

If \( Y_1, ..., Y_n \) are independent \( G(\nu_i, \lambda) \) then \( \sum_{i=1}^n Y_i \sim G(\sum_{i=1}^n \nu_i, \lambda) \).

If \( Y_1, ..., Y_n \) are iid \( G(\nu, \lambda) \), then

\[
T_n = \sum_{i=1}^n Y_i \sim G(n\nu, \lambda).
\]

Since

\[
f(y) = \frac{1}{\Gamma(\nu)} \exp((\nu - 1) \log(y)) I(y > 0) \frac{1}{\lambda^{\nu}} \exp\left[ \frac{-1}{\nu} y \right],
\]

\( Y \) is a 1P–REF when \( \nu \) is known.

If \( \nu \) is known, then the likelihood

\[
L(\beta) = c \frac{1}{\lambda^{nn}} \exp\left[ \frac{-1}{\lambda} \sum_{i=1}^n y_i \right].
\]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

The log likelihood
\[ \log(L(\lambda)) = d - n\nu \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i. \]

Hence
\[ \frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n\nu}{\lambda} + \frac{\sum_{i=1}^{n} y_i}{\lambda^2} \equiv 0, \]
or \[ \sum_{i=1}^{n} y_i = n\nu \lambda \]
or \[ \hat{\lambda} = \bar{y}/\nu. \]

This solution is unique and
\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n\nu}{\lambda^2} - 2 \frac{\sum_{i=1}^{n} y_i}{\lambda^3} \bigg|_{\lambda = \hat{\lambda}} = \frac{n\nu}{\hat{\lambda}^2} - \frac{2n\nu\hat{\lambda}}{\hat{\lambda}^3} = -\frac{n\nu}{\hat{\lambda}^2} < 0. \]

Thus \( \bar{y} \) is the UMVUE, MLE and MME of \( \nu \lambda \) if \( \nu \) is known.

The (lower) incomplete gamma function is
\[ \gamma_{\nu}(x) = \frac{1}{\Gamma(\nu)} \int_{0}^{x} t^{\nu-1} e^{-t} dt. \]

If \( \nu \) is a positive integer, then
\[ \gamma_{\nu}(x) = 1 - e^{-x} \sum_{k=0}^{\nu-1} \frac{x^k}{k!}. \]

If \( Y \sim G(\nu, \lambda) \), the \( F(y) = \gamma_{\nu}(y/\lambda) \). Hence if \( \nu > 0 \) is an integer,
\[ F(y) = 1 - \sum_{k=0}^{\nu-1} e^{-y/\lambda} \frac{(y/\lambda)^k}{k!}. \]

For example, if \( Y \sim G(\nu = 4, \lambda = 2) \), then \( P(Y \leq 14) = F(14) = \gamma_4(7) \) and \( P(Y \leq 10) = F(10) = \gamma_4(5) \).

10.17 The Generalized Gamma Distribution

If \( Y \) has a generalized gamma distribution, \( Y \sim GG(\nu, \lambda, \phi) \), then the pdf of \( Y \) is
\[ f(y) = \frac{\phi y^{\phi\nu-1}}{\lambda^{\phi\nu} \Gamma(\nu)} \exp(-y^{\phi}/\lambda^\phi). \]
where \( \nu, \lambda, \phi \) and \( y \) are positive. This family is a scale family with scale parameter \( \lambda \) if \( \phi \) and \( \nu \) are known.

\[
E(Y^k) = \frac{\lambda^k \Gamma(\nu + \frac{k}{\phi})}{\Gamma(\nu)} \quad \text{if} \quad k > -\phi \nu. 
\] (10.7)

If \( \phi \) and \( \nu \) are known, then

\[
f(y) = \frac{\phi y^{\phi y - 1}}{\Gamma(\nu)}I(y > 0) \frac{1}{\lambda^{\phi y}} \exp \left[ \frac{-1}{\lambda^{\phi y}} \right],
\]

which is a one parameter exponential family.

Notice that \( W = Y^\phi \sim G(\nu, \lambda^\phi) \). If \( Y_1, \ldots, Y_n \) are iid \( G(\nu, \lambda, \phi) \) where \( \phi \) and \( \nu \) are known, then \( T_n = \sum_{i=1}^n Y_i^\phi \sim G(n\nu, \lambda^\phi) \), and \( T_r^n \) is the UMVUE of

\[
E(T_r^n) = \lambda^{\phi r} \frac{\Gamma(r + n\nu)}{\Gamma(n\nu)}
\]

for \( r > -n\nu \).

### 10.18 The Generalized Negative Binomial Distribution

If \( Y \) has a generalized negative binomial distribution, \( Y \sim GNB(\mu, \kappa) \), then the pmf of \( Y \) is

\[
f(y) = P(Y = y) = \frac{\Gamma(y + \kappa)}{\Gamma(\kappa)\Gamma(y + 1)} \left( \frac{\kappa}{\mu + \kappa} \right)^\kappa \left( 1 - \frac{\kappa}{\mu + \kappa} \right)^y
\]

for \( y = 0, 1, 2, \ldots \) where \( \mu > 0 \) and \( \kappa > 0 \). This distribution is a generalization of the negative binomial \((\kappa, \rho)\) distribution with \( \rho = \kappa/(\mu + \kappa) \) and \( \kappa > 0 \) is an unknown real parameter rather than a known integer.

The mgf is

\[
m(t) = \left[ \frac{\kappa}{\kappa + \mu(1 - e^t)} \right]^\kappa
\]

for \( t < -\log(\mu/(\mu + \kappa)) \).

\( E(Y) = \mu \) and

\( \text{VAR}(Y) = \mu + \mu^2/\kappa. \)

If \( Y_1, \ldots, Y_n \) are iid \( GNB(\mu, \kappa) \), then \( \sum_{i=1}^n Y_i \sim GNB(n\mu, n\kappa). \)
When $\kappa$ is known, this distribution is a \textbf{1P–REF}. If $Y_1, \ldots, Y_n$ are iid $\text{GNB}(\mu, \kappa)$ where $\kappa$ is known, then $\hat{\mu} = \overline{Y}$ is the MLE, UMVUE and MME of $\mu$.

Let $\tau = 1/\kappa$. As $\tau \to 0$, the $\text{GNB}(\mu, \kappa)$ distribution converges to the $\text{Poisson}(\lambda = \mu)$ distribution.

### 10.19 The Geometric Distribution

If $Y$ has a geometric distribution, $Y \sim \text{geom}(\rho)$ then the pmf of $Y$ is

$$f(y) = P(Y = y) = \rho (1 - \rho)^y$$

for $y = 0, 1, 2, \ldots$ and $0 < \rho < 1$.

The cdf for $Y$ is $F(y) = 1 - (1 - \rho)^{\lfloor y \rfloor + 1}$ for $y \geq 0$ and $F(y) = 0$ for $y < 0$. Here $\lfloor y \rfloor$ is the greatest integer function, eg, $\lfloor 7.7 \rfloor = 7$. To see this, note that for $y \geq 0$,

$$F(y) = \rho \sum_{i=0}^{\lfloor y \rfloor} (1 - \rho)^i = \frac{1 - (1 - \rho)^{\lfloor y \rfloor + 1}}{1 - (1 - \rho)}$$

by Lemma 10.2a with $n_1 = 0$, $n_2 = \lfloor y \rfloor$ and $a = 1 - \rho$.

$E(Y) = (1 - \rho)/\rho$.
$\text{VAR}(Y) = (1 - \rho)/\rho^2$.
$Y \sim \text{NB}(1, \rho)$.
Hence the mgf of $Y$ is

$$m(t) = \frac{\rho}{1 - (1 - \rho)e^t}$$

for $t < - \log(1 - \rho)$.

Notice that

$$f(y) = 1[y \in \{0, 1, \ldots\}] \rho \exp[\log(1 - \rho)y]$$

is a \textbf{1P–REF}. Hence $\Theta = (0, 1)$, $\eta = \log(1 - \rho)$ and $\Omega = (-\infty, 0)$.

If $Y_1, \ldots, Y_n$ are iid $\text{geom}(\rho)$, then

$$T_n = \sum_{i=1}^{n} Y_i \sim \text{NB}(n, \rho).$$
The likelihood

\[ L(\rho) = \rho^n \exp\left[ \log(1 - \rho) \sum_{i=1}^{n} y_i \right], \]

and the log likelihood

\[ \log(L(\rho)) = n \log(\rho) + \log(1 - \rho) \sum_{i=1}^{n} y_i. \]

Hence

\[ \frac{d}{d\rho} \log(L(\rho)) = \frac{n}{\rho} - \frac{1}{1 - \rho} \sum_{i=1}^{n} y_i \overset{set}{=} 0 \]

or \( n(1 - \rho)/\rho = \sum_{i=1}^{n} y_i \) or \( n - n\rho - \rho \sum_{i=1}^{n} y_i = 0 \) or

\[ \hat{\rho} = \frac{n}{n + \sum_{i=1}^{n} y_i}. \]

This solution is unique and

\[ \frac{d^2}{d\rho^2} \log(L(\rho)) = \frac{-n}{\rho^2} - \frac{\sum_{i=1}^{n} y_i}{(1 - \rho)^2} < 0. \]

Thus

\[ \hat{\rho} = \frac{n}{n + \sum_{i=1}^{n} Y_i} \]

is the MLE of \( \rho \).

The UMVUE, MLE and MME of \((1 - \rho)/\rho\) is \( \bar{Y} \).

**10.20 The Gompertz Distribution**

If \( Y \) has a Gompertz distribution, \( Y \sim Gomp(\theta, \nu) \), then the pdf of \( Y \) is

\[ f(y) = \nu e^{\theta y} \exp \left[ \frac{-\nu}{\theta} (e^{\theta y} - 1) \right] \]

for \( \theta > 0 \) where \( \nu > 0 \) and \( y > 0 \). If \( \theta = 0 \), the \( Gomp(\theta = 0, \nu) \) distribution is the exponential \((1/\nu)\) distribution. The cdf is

\[ F(y) = 1 - \exp \left[ \frac{-\nu}{\theta} (e^{\theta y} - 1) \right] \]
for \( \theta > 0 \) and \( y > 0 \). For fixed \( \theta \) this distribution is a scale family with scale parameter \( 1/\nu \).

This family is a 1P–REF if \( \theta \in (0, \infty) \) is known, and \( W = e^{\theta Y} - 1 \sim \text{EXP}(\theta/\nu) \). Thus \( e^{\theta Y} \sim \text{EXP}(1, t/\nu) \) for \( t > 0 \) and \( E(e^{\theta Y}) = 1 + t/\nu \) for \( t \geq 0 \). But the mgf for \( Y \) does not exist since the \( k \)th derivative of \( (1 + t/\nu) \) is 0 for integer \( k \geq 2 \).

### 10.21 The Half Cauchy Distribution

If \( Y \) has a half Cauchy distribution, \( Y \sim \text{HC}(\mu, \sigma) \), then the pdf of \( Y \) is

\[
f(y) = \frac{2}{\pi \sigma [1 + (y - \mu)/\sigma]^2}
\]

where \( y \geq \mu \), \( \mu \) is a real number and \( \sigma > 0 \). The cdf of \( Y \) is

\[
F(y) = \frac{2}{\pi} \arctan\left(\frac{y - \mu}{\sigma}\right)
\]

for \( y \geq \mu \) and is 0, otherwise. This distribution is a right skewed location-scale family.

MED(\( Y \)) = \( \mu + \sigma \).

MAD(\( Y \)) = 0.73205\( \sigma \).

### 10.22 The Half Logistic Distribution

If \( Y \) has a half logistic distribution, \( Y \sim \text{HL}(\mu, \sigma) \), then the pdf of \( Y \) is

\[
f(y) = \frac{2 \exp\left(-\frac{(y - \mu)}{\sigma}\right)}{\sigma [1 + \exp\left(-\frac{(y - \mu)}{\sigma}\right)]^2}
\]

where \( \sigma > 0 \), \( y \geq \mu \) and \( \mu \) are real. The cdf of \( Y \) is

\[
F(y) = \frac{\exp\left[\frac{(y - \mu)}{\sigma}\right] - 1}{1 + \exp\left[\frac{(y - \mu)}{\sigma}\right]}
\]

for \( y \geq \mu \) and 0 otherwise. This family is a right skewed location–scale family.

MED(\( Y \)) = \( \mu + \log(3)\sigma \).

MAD(\( Y \)) = 0.67346\( \sigma \).
10.23 The Half Normal Distribution

If $Y$ has a half normal distribution, $Y \sim \text{HN}(\mu, \sigma^2)$, then the pdf of $Y$ is

$$f(y) = \frac{2}{\sqrt{2\pi} \sigma} \exp \left( \frac{-(y - \mu)^2}{2\sigma^2} \right)$$

where $\sigma > 0$ and $y \geq \mu$ and $\mu$ is real. Let $\Phi(y)$ denote the standard normal cdf. Then the cdf of $Y$ is

$$F(y) = 2\Phi\left(\frac{y - \mu}{\sigma}\right) - 1$$

for $y > \mu$ and $F(y) = 0$, otherwise.

$$E(Y) = \mu + \sigma \sqrt{\frac{2}{\pi}} \approx \mu + 0.797885\sigma.$$  

$$\text{VAR}(Y) = \frac{\sigma^2(\pi - 2)}{\pi} \approx 0.363380\sigma^2.$$

This is an asymmetric location-scale family that has the same distribution as $\mu + \sigma|Z|$ where $Z \sim N(0, 1)$. Note that $Z^2 \sim \chi_1^2$. Hence the formula for the $r$th moment of the $\chi_1^2$ random variable can be used to find the moments of $Y$.

$$\text{MED}(Y) = \mu + 0.6745\sigma.$$  

$$\text{MAD}(Y) = 0.3990916\sigma.$$  

Notice that

$$f(y) = I(y \geq \mu) \frac{2}{\sqrt{2\pi} \sigma} \exp \left[ \frac{-1}{2\sigma^2}(y - \mu)^2 \right]$$

is a 1P–REF if $\mu$ is known. Hence $\Theta = (0, \infty)$, $\eta = -1/(2\sigma^2)$ and $\Omega = (-\infty, 0)$.

$W = (Y - \mu)^2 \sim G(1/2, 2\sigma^2)$.

If $Y_1, \ldots, Y_n$ are iid HN($\mu, \sigma^2$), then

$$T_n = \sum_{i=1}^{n} (Y_i - \mu)^2 \sim G(n/2, 2\sigma^2).$$

If $\mu$ is known, then the likelihood

$$L(\sigma^2) = c \frac{1}{\sigma^n} \exp \left[ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right],$$
and the log likelihood
\[
\log(L(\sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(y_i - \mu)^2.
\]
Hence
\[
\frac{d}{d(\sigma^2)} \log(L(\sigma^2)) = \frac{-n}{2(\sigma^2)} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n}(y_i - \mu)^2 \overset{se}{=} 0,
\]
or \[
\sum_{i=1}^{n}(y_i - \mu)^2 = n\sigma^2 \text{ or } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n}(y_i - \mu)^2.
\]
This solution is unique and
\[
\frac{d^2}{d(\sigma^2)^2} \log(L(\sigma^2)) = \left. \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{(\sigma^2)^3} \right|_{\sigma^2 = \hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{2(\hat{\sigma}^2)^3} 2 = \frac{-n}{2\hat{\sigma}^2} < 0.
\]
Thus
\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n}(Y_i - \mu)^2
\]
is the UMVUE and MLE of \(\sigma^2\) if \(\mu\) is known.

If \(r > -n/2\) and if \(\mu\) is known, then \(T_n^r\) is the UMVUE of \(E(T_n^r) = 2^r \sigma^2 r \Gamma(r + n/2)/\Gamma(n/2)\).

Example 5.3 shows that \((\hat{\mu}, \hat{\sigma}^2) = (\bar{Y}_1, \frac{1}{n} \sum_{i=1}^{n}(Y_i - Y_1)^2)\) is the MLE of \((\mu, \sigma^2)\). Following Abuhassan and Olive (2008), a large sample 100(1 - \(\alpha\))% confidence interval for \(\sigma^2\) is
\[
\left( \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(1-\alpha/2)}, \frac{n\hat{\sigma}^2}{\chi^2_{n-1}(\alpha/2)} \right) \quad (10.8)
\]
while a large sample 100(1 - \(\alpha\))% CI for \(\mu\) is
\[
(\hat{\mu} + \hat{\sigma} \log(\alpha) \Phi^{-1}(\frac{1}{2} + \frac{1}{2n}) (1 + \frac{13}{n^2}), \; \hat{\mu} ). \quad (10.9)
\]
If \(\mu\) is known, then a 100(1 - \(\alpha\))% CI for \(\sigma^2\) is
\[
\left( \frac{T_n}{\chi^2_n(1-\alpha/2)}, \frac{T_n}{\chi^2_n(\alpha/2)} \right). \quad (10.10)
\]
10.24 The Hypergeometric Distribution

If $Y$ has a hypergeometric distribution, $Y \sim \text{HG}(C, N - C, n)$, then the data set contains $N$ objects of two types. There are $C$ objects of the first type (that you wish to count) and $N - C$ objects of the second type. Suppose that $n$ objects are selected at random without replacement from the $N$ objects. Then $Y$ counts the number of the $n$ selected objects that were of the first type. The pmf of $Y$ is

$$
f(y) = P(Y = y) = \frac{\binom{C}{y} \binom{N-C}{n-y}}{\binom{N}{n}}$

where the integer $y$ satisfies $\max(0, n - N + C) \leq y \leq \min(n, C)$. The right inequality is true since if $n$ objects are selected, then the number of objects $y$ of the first type must be less than or equal to both $n$ and $C$. The first inequality holds since $n - y$ counts the number of objects of second type.

Hence $n - y \leq N - C$.

Let $p = C/N$. Then

$$
E(Y) = \frac{nC}{N} = np
$$

and

$$
\text{VAR}(Y) = \frac{nC(N - C)}{N^2} \frac{N - n}{N - 1} = np(1 - p) \frac{N - n}{N - 1}.
$$

If $n$ is small compared to both $C$ and $N - C$ then $Y \approx \text{BIN}(n, p)$. If $n$ is large but $n$ is small compared to both $C$ and $N - C$ then $Y \approx N(np, np(1-p))$.

10.25 The Inverse Exponential Distribution

If $Y$ has an inverse exponential distribution, $Y \sim \text{IEXP}(\theta)$, then the pdf of $Y$ is

$$
f(y) = \frac{\theta}{y^2} \exp\left(-\frac{-\theta}{y}\right)
$$

where $y > 0$ and $\theta > 0$. The cdf $F(y) = \exp(-\theta/y)$ for $y > 0$. $E(Y)$ and $V(Y)$ do not exist. $\text{MED}(Y) = \theta/\log(2)$. This distribution is a 1P-REF and a scale family with scale parameter $\theta$. This distribution is the inverted gamma($\nu = 1, \lambda = 1/\theta$) distribution.

$W = 1/Y \sim \text{EXP}(1/\theta)$. 

Suppose that $Y_1, \ldots, Y_n$ are iid $\text{IEXP}(\theta)$, then

$$T_n = \sum_{i=1}^{n} \frac{1}{Y_i} \sim G(n, 1/\theta).$$

The likelihood

$$L(\theta) = c \theta^n \exp \left[-\theta \sum_{i=1}^{n} \frac{1}{y_i}\right],$$

and the log likelihood

$$\log(L(\theta)) = d + n \log(\theta) - \theta \sum_{i=1}^{n} \frac{1}{y_i}.$$ 

Hence

$$\frac{d}{d\theta} \log(L(\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} \frac{1}{y_i} \text{set} = 0,$$

or

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}.$$ 

Since this solution is unique and

$$\frac{d^2}{d\theta^2} \log(L(\theta)) = \frac{n}{\theta^2} < 0,$$

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \frac{1}{y_i}}$$

is the MLE of $\theta$.

If $r > -n$, then $[T_n]^r$ is the UMVUE of

$$E([T_n]^r) = \frac{\Gamma(r + n)}{\theta^r \Gamma(n)}.$$ 

### 10.26 The Inverse Gaussian Distribution

If $Y$ has an inverse Gaussian distribution (also called a Wald distribution), $Y \sim \text{IG}(\theta, \lambda)$, then the pdf of $Y$ is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp \left[-\frac{\lambda(y - \theta)^2}{2\theta^2y}\right]$$
where $y, \theta, \lambda > 0$. The mgf is

$$m(t) = \exp \left[ \frac{\lambda}{\theta} \left( 1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}} \right) \right]$$

for $t < \lambda/(2\theta^2)$. The characteristic function is

$$\phi(t) = \exp \left[ \frac{\lambda}{\theta} \left( 1 - \sqrt{1 - \frac{2\theta^2 it}{\lambda}} \right) \right].$$

$E(Y) = \theta$ and

$$\text{VAR}(Y) = \frac{\theta^3}{\lambda}.$$

See Datta (2005) and Schwarz and Samanta (1991) for additional properties.

Notice that

$$f(y) = \sqrt{\frac{1}{y^3}} I(y > 0) \sqrt{\frac{\lambda}{2\pi}} e^{\lambda/y} \exp \left[ -\frac{\lambda}{2\theta^2} y - \frac{\lambda 1}{2} \right]$$

is a two parameter exponential family.

If $Y_1, \ldots, Y_n$ are iid $IG(\theta, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim IG(n\theta, n^2\lambda) \quad \text{and} \quad \mathbf{\bar{Y}} \sim IG(\theta, n\lambda).$$

If $\lambda$ is known, then the likelihood

$$L(\theta) = c e^{n\lambda/\theta} \exp \left[ -\frac{\lambda}{2\theta^2} \sum_{i=1}^{n} y_i \right],$$

and the log likelihood

$$\log(L(\theta)) = d + \frac{n\lambda}{\theta} - \frac{\lambda}{2\theta^2} \sum_{i=1}^{n} y_i.$$

Hence

$$\frac{d}{d\theta} \log(L(\theta)) = -\frac{n\lambda}{\theta^2} + \frac{\lambda}{\theta^3} \sum_{i=1}^{n} y_i \overset{\text{set}}{=} 0.$$
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

or \( \sum_{i=1}^{n} y_i = n\theta \) or

\[ \hat{\theta} = \overline{y}. \]

This solution is unique and

\[
\frac{d^2}{d\theta^2} \log(L(\theta)) = \left. \frac{2n\lambda}{\theta^3} - 3\lambda \sum_{i=1}^{n} \frac{y_i}{\theta^4} \right|_{\theta=\hat{\theta}} = \frac{2n\lambda}{\hat{\theta}^3} - 3n\lambda \frac{\hat{\theta}}{\hat{\theta}^4} = -\frac{n\lambda}{\hat{\theta}^3} < 0.
\]

Thus \( \overline{Y} \) is the UMVUE, MLE and MME of \( \theta \) if \( \lambda \) is known.

If \( \theta \) is known, then the likelihood

\[ L(\lambda) = c \lambda^{n/2} \exp \left( \frac{-\lambda}{2\theta^2} \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i} \right), \]

and the log likelihood

\[ \log(L(\lambda)) = d + \frac{n}{2} \log(\lambda) - \frac{\lambda}{2\theta^2} \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i}. \]

Hence

\[ \frac{d}{d\lambda} \log(L(\lambda)) = \frac{n}{2\lambda} - \frac{1}{2\theta^2} \sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i} \text{ set } 0 \]

or

\[ \hat{\lambda} = \frac{n\theta^2}{\sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i}}. \]

This solution is unique and

\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) = -\frac{n}{2\lambda^2} < 0. \]

Thus

\[ \hat{\lambda} = \frac{n\theta^2}{\sum_{i=1}^{n} \frac{(y_i - \theta)^2}{y_i}} \]

is the MLE of \( \lambda \) if \( \theta \) is known.

Another parameterization of the inverse Gaussian distribution takes \( \theta = \sqrt{\lambda/\psi} \) so that

\[
f(y) = \sqrt{\frac{1}{y^3}} f[y > 0] \sqrt{\frac{\lambda}{2\pi}} e^{\lambda y^2} \exp \left[ -\frac{\psi}{2} y - \frac{\lambda}{2} \right].
\]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

where $\lambda > 0$ and $\psi \geq 0$. Here $\Theta = (0, \infty) \times [0, \infty)$, $\eta_1 = -\psi/2$, $\eta_2 = -\lambda/2$ and $\Omega = (-\infty, 0] \times (-\infty, 0)$. Since $\Omega$ is not an open set, this is a 2 parameter full exponential family that is not regular. If $\psi$ is known then $Y$ is a 1P–REF, but if $\lambda$ is known the $Y$ is a one parameter full exponential family. When $\psi = 0$, $Y$ has a one sided stable distribution with index $1/2$. See Barndorff–Nielsen (1978, p. 117).

10.27 The Inverse Weibull Distribution

$Y$ has an inverse Weibull distribution, $Y \sim IW(\phi, \lambda)$, if the pdf of $Y$ is

$$f(y) = \frac{\phi}{\lambda y^{\phi+1}} \exp\left(-\frac{1}{\lambda} \frac{1}{y^\phi}\right)$$

where $y > 0$, $\lambda > 0$ and $\phi > 0$ is known. The cdf is

$$F(y) = \exp\left(-\frac{1}{\lambda} \frac{1}{y^\phi}\right)$$

for $y > 0$. This family is a 1P-REF if $\phi$ is known.

$1/Y \sim W(\phi, \lambda)$ and $1/Y^\phi \sim EXP(\lambda)$. See Mahmoud, Sultan and Amer (2003).

If $Y_1, \ldots, Y_n$ are iid $IW(\phi, \lambda)$, then

$$T_n = \sum_{i=1}^{n} \frac{1}{Y_i^\phi} \sim G(n, \lambda).$$

If $\phi$ is known, then the likelihood

$$L(\lambda) = c \lambda^n \exp\left[\frac{-1}{\lambda} \sum_{i=1}^{n} \frac{1}{y_i^\phi}\right],$$

and the log likelihood

$$\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} \frac{1}{y_i^\phi}.$$ 

Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} \frac{1}{y_i^\phi} \neq 0,$$
or

\[ \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{Y_i^\phi}. \]

This solution is unique and

\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} \frac{1}{y_i^\phi} \bigg|_{\lambda=\hat{\lambda}} = \frac{n}{\hat{\lambda}^2} - \frac{2n\hat{\lambda}}{\hat{\lambda}^3} = -\frac{n}{\hat{\lambda}^2} < 0. \]

Thus

\[ \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{Y_i^\phi} \]

is the MLE of \( \lambda \) if \( \phi \) is known.

If \( r > -n \) and \( \phi \) is known, then \( [T_n]^r \) is the UMVUE of

\[ E([T_n]^r) = \frac{\lambda^r \Gamma(r + n)}{\Gamma(n)}. \]

### 10.28 The Inverted Gamma Distribution

If \( Y \) has an inverted gamma distribution, \( Y \sim INVG(\nu, \lambda) \), then the pdf of \( Y \) is

\[ f(y) = \frac{1}{y^{\nu+1} \Gamma(\nu)} I(y > 0) \frac{1}{\lambda^\nu} \exp \left( \frac{-1}{\lambda} \frac{1}{y} \right) \]

where \( \lambda, \nu \) and \( y \) are all positive. It can be shown that \( W = 1/Y \sim G(\nu, \lambda) \).

This family is a scale family with scale parameter \( \tau = 1/\lambda \) if \( \nu \) is known.

\[ E(Y^r) = \frac{\Gamma(\nu - r)}{\lambda^r \Gamma(\nu)} \]

for \( \nu > r \).

If \( \nu \) is known, this family is a 1 parameter exponential family. If \( Y_1, ..., Y_n \) are iid \( INVG(\nu, \lambda) \) and \( \nu \) is known, then \( T_n = \sum_{i=1}^{n} \frac{1}{Y_i} \sim G(n\nu, \lambda) \) and \( T_n^r \) is the UMVUE of

\[ \lambda^r \frac{\Gamma(r + n\nu)}{\Gamma(n\nu)} \]

for \( r > -n\nu \).

If \( W \) has an inverse exponential distribution, then \( W \sim INVG(\nu = 1, \lambda) \).
If $Y$ has a largest extreme value distribution (or Gumbel distribution), $Y \sim \text{LEV}(\theta, \sigma)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\sigma} \exp\left(-\left(\frac{y - \theta}{\sigma}\right)\right) \exp\left[-\exp\left(-\left(\frac{y - \theta}{\sigma}\right)\right)\right]$$

where $y$ and $\theta$ are real and $\sigma > 0$. The cdf of $Y$ is

$$F(y) = \exp\left[-\exp\left(-\left(\frac{y - \theta}{\sigma}\right)\right)\right].$$

This family is an asymmetric location–scale family with a mode at $\theta$. The mgf

$$m(t) = \exp(t\theta)\Gamma(1 - \sigma t)$$

for $|t| < 1/\sigma$.

$E(Y) \approx \theta + 0.57721\sigma$, and

$\text{VAR}(Y) = \sigma^2\pi^2/6 \approx 1.64493\sigma^2$.

$\text{MED}(Y) = \theta - \sigma \log(\log(2)) \approx \theta + 0.36651\sigma$

and

$\text{MAD}(Y) \approx 0.767049\sigma$.

$W = \exp\left(-\left(Y - \theta\right)/\sigma\right) \sim \text{EXP}(1)$.

Notice that

$$f(y) = e^{-y/\sigma} \frac{1}{\sigma} e^{\theta/\sigma} \exp\left[-e^{\theta/\sigma}e^{-y/\sigma}\right]$$

is a one parameter exponential family in $\theta$ if $\sigma$ is known.

If $Y_1, ..., Y_n$ are iid $\text{LEV}(\theta, \sigma)$ where $\sigma$ is known, then the likelihood

$$L(\sigma) = c e^{n\theta/\sigma} \exp\left[-\frac{\theta}{\sigma} \sum_{i=1}^{n} e^{-y_i/\sigma}\right],$$

and the log likelihood

$$\log(L(\theta)) = d + \frac{n\theta}{\sigma} - e^{\theta/\sigma} \sum_{i=1}^{n} e^{-y_i/\sigma}.$$
Hence
\[ \frac{d}{d\theta} \log(L(\theta)) = \frac{n}{\sigma} - e^{\theta/\sigma} \sum_{i=1}^{n} e^{-y_i/\sigma} \neq 0, \]
or
\[ e^{\theta/\sigma} \sum_{i=1}^{n} e^{-y_i/\sigma} = n, \]
or
\[ e^{\theta/\sigma} = \frac{n}{\sum_{i=1}^{n} e^{-y_i/\sigma}}, \]
or
\[ \hat{\theta} = \sigma \log \left( \frac{n}{\sum_{i=1}^{n} e^{-y_i/\sigma}} \right). \]

Since this solution is unique and
\[ \frac{d^2}{d\theta^2} \log(L(\theta)) = -\frac{1}{\sigma^2} e^{\theta/\sigma} \sum_{i=1}^{n} e^{-y_i/\sigma} < 0, \]
\[ \hat{\theta} = \sigma \log \left( \frac{n}{\sum_{i=1}^{n} e^{-y_i/\sigma}} \right) \]
is the MLE of \( \theta \).

**10.30 The Logarithmic Distribution**

If \( Y \) has a logarithmic distribution, then the pmf of \( Y \) is
\[ f(y) = P(Y = y) = \frac{-1}{\log(1 - \theta)} \frac{\theta^y}{y} \]
for \( y = 1, 2, \ldots \) and \( 0 < \theta < 1 \). This distribution is sometimes called the logarithmic series distribution or the log-series distribution. The mgf
\[ m(t) = \frac{\log(1 - \theta e^t)}{\log(1 - \theta)} \]
for \( t < -\log(\theta) \).
\[ E(Y) = \frac{-1}{\log(1 - \theta)} \frac{\theta}{1 - \theta}. \]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

\[ V(Y) = \frac{-\theta}{(1 - \theta)^2 \log(\theta)} \left( 1 + \frac{\theta}{\log(1 - \theta)} \right). \]

Notice that
\[ f(y) = \frac{-1}{\log(1 - \theta)} \frac{1}{y} \exp(\log(\theta) y) \]
is a 1P–REF. Hence \( \Theta = (0, 1), \eta = \log(\theta) \) and \( \Omega = (-\infty, 0). \)

If \( Y_1, \ldots, Y_n \) are iid logarithmic \( (\theta) \), then \( \bar{Y} \) is the UMVUE of \( E(Y) \).

10.31 The Logistic Distribution

If \( Y \) has a logistic distribution, \( Y \sim L(\mu, \sigma) \), then the pdf of \( Y \) is
\[ f(y) = \frac{\exp \left( -(y - \mu)/\sigma \right)}{\sigma [1 + \exp \left( -(y - \mu)/\sigma \right)]^2} \]
where \( \sigma > 0 \) and \( y \) and \( \mu \) are real. The cdf of \( Y \) is
\[ F(y) = \frac{1}{1 + \exp \left( -(y - \mu)/\sigma \right)} = \frac{\exp \left( (y - \mu)/\sigma \right)}{1 + \exp \left( (y - \mu)/\sigma \right)}. \]

This family is a symmetric location–scale family.

The mgf of \( Y \) is \( m(t) = \pi \sigma te^{\pi t} \csc(\pi \sigma t) \) for \( |t| < 1/\sigma \), and the chf is \( c(t) = \pi \sigma te^{\pi t} \csc(\pi \sigma t) \) where \( \csc(t) \) is the cosecant of \( t \).

\( E(Y) = \mu \), and
\( \text{MED}(Y) = \mu \).
\( \text{VAR}(Y) = \sigma^2 \pi^2/3 \), and
\( \text{MAD}(Y) = \log(3)\sigma \approx 1.0986 \sigma. \)

Hence \( \sigma = \text{MAD}(Y)/\log(3). \)

The estimators \( \hat{\mu} = \bar{Y} \) and \( S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \) are sometimes used.

Note that if
\[ q = F_{L(0,1)}(c) = \frac{e^c}{1 + e^c} \quad \text{then} \quad c = \log \left( \frac{q}{1 - q} \right). \]

Taking \( q = .9995 \) gives \( c = \log(1999) \approx 7.6. \)
To see that \( \text{MAD}(Y) = \log(3)\sigma \), note that \( F(\mu + \log(3)\sigma) = .75 \), \( F(\mu - \log(3)\sigma) = .25 \), and \( .75 = \exp (\log(3))/(1 + \exp(\log(3))). \)
10.32 The Log-Cauchy Distribution

If $Y$ has a log–Cauchy distribution, $Y \sim LC(\mu, \sigma)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\pi \sigma y [1 + \left( \frac{\log(y) - \mu}{\sigma} \right)^2]}$$

where $y > 0$, $\sigma > 0$ and $\mu$ is a real number. This family is a scale family with scale parameter $\tau = e^\mu$ if $\sigma$ is known. It can be shown that $W = \log(Y)$ has a Cauchy($\mu, \sigma$) distribution.

10.33 The Log-Gamma Distribution

$Y$ has a log-gamma distribution, $Y \sim LG(\nu, \lambda)$, if the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda^\nu \Gamma(\nu)} \exp(\nu y + \left( -\frac{1}{\lambda} \right)e^y)$$

where $y$ is real, $\nu > 0$ and $\lambda > 0$. The mgf

$$m(t) = \frac{\Gamma(t + \nu)\lambda^t}{\Gamma(\nu)}$$

for $t > -\nu$. This family is a 2P-REF.

$W = e^Y \sim$ gamma $(\nu, \lambda)$.

10.34 The Log-Logistic Distribution

If $Y$ has a log–logistic distribution, $Y \sim LL(\phi, \tau)$, then the pdf of $Y$ is

$$f(y) = \frac{\phi \tau (\phi y)^{\tau - 1}}{[1 + (\phi y)^{\tau}]^2}$$

where $y > 0$, $\phi > 0$ and $\tau > 0$. The cdf of $Y$ is

$$F(y) = 1 - \frac{1}{1 + (\phi y)^\tau}$$

for $y > 0$. This family is a scale family with scale parameter $\phi^{-1}$ if $\tau$ is known.
MED(Y) = 1/φ.

It can be shown that \( W = \log(Y) \) has a logistic(\( \mu = -\log(\phi), \tau = 1/\sigma \)) distribution. Hence \( \phi = e^{-\mu} \) and \( \tau = 1/\sigma \). Kalbfleisch and Prentice (1980, p. 27-28) suggest that the log-logistic distribution is a competitor of the lognormal distribution.

### 10.35 The Lognormal Distribution

If \( Y \) has a lognormal distribution, \( Y \sim LN(\mu, \sigma^2) \), then the pdf of \( Y \) is

\[
f(y) = \frac{1}{y \sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log(y) - \mu)^2}{2\sigma^2}\right)
\]

where \( y > 0 \) and \( \sigma > 0 \) and \( \mu \) is real. The cdf of \( Y \) is

\[
F(y) = \Phi\left(\frac{\log(y) - \mu}{\sigma}\right) \quad \text{for} \quad y > 0
\]

where \( \Phi(y) \) is the standard normal \( N(0,1) \) cdf. This family is a scale family with scale parameter \( \tau = e^\mu \) if \( \sigma^2 \) is known.

\[
E(Y) = \exp(\mu + \sigma^2/2)
\]

and

\[
\text{VAR}(Y) = \exp(\sigma^2)(\exp(\sigma^2) - 1) \exp(2\mu).
\]

For any \( r \),

\[
E(Y^r) = \exp(r\mu + r^2\sigma^2/2).
\]

MED(Y) = \( \exp(\mu) \) and

\[
\exp(\mu)[1 - \exp(-0.6744\sigma)] \leq \text{MAD}(Y) \leq \exp(\mu)[1 + \exp(0.6744\sigma)].
\]

Notice that

\[
f(y) = \frac{1}{y} I(y \geq 0) \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left[-\frac{1}{2\sigma^2}(\log(y))^2 + \frac{\mu}{\sigma^2} \log(y)\right]
\]

is a 2P–REF. Hence \( \Theta = (-\infty, \infty) \times (0, \infty) \), \( \eta_1 = -1/(2\sigma^2) \), \( \eta_2 = \mu/\sigma^2 \) and \( \Omega = (-\infty, 0) \times (-\infty, \infty) \).

Note that \( W = \log(Y) \sim N(\mu, \sigma^2) \).
Notice that
\[ f(y) = \frac{1}{y} I(y \geq 0) \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2\sigma^2} (\log(y) - \mu)^2 \right] \]
is a 1P–REF if \( \mu \) is known.

If \( Y_1, \ldots, Y_n \) are iid LN(\( \mu, \sigma^2 \)) where \( \mu \) is known, then the likelihood
\[ L(\sigma^2) = c \left( \frac{1}{\sigma^2} \sum_{i=1}^{n} (\log(y_i) - \mu)^2 \right), \]
and the log likelihood
\[ \log(L(\sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log(y_i) - \mu)^2. \]

Hence
\[ \frac{d}{d(\sigma^2)} \log(L(\sigma^2)) = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (\log(y_i) - \mu)^2 \]
set \( = 0 \),
or \( \sum_{i=1}^{n} (\log(y_i) - \mu)^2 = n\sigma^2 \) or
\[ \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log(y_i) - \mu)^2}{n}. \]

Since this solution is unique and
\[ \frac{d^2}{d(\sigma^2)^2} \log(L(\sigma^2)) = \]
\[ \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n} (\log(y_i) - \mu)^2}{(\sigma^2)^3} \bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{(\hat{\sigma}^2)^3} \frac{2}{2} = \frac{-n}{2(\hat{\sigma}^2)^2} < 0, \]

\( \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\log(Y_i) - \mu)^2}{n} \)
is the UMVUE and MLE of \( \sigma^2 \) if \( \mu \) is known.

Since \( T_n = \sum_{i=1}^{n} [\log(Y_i) - \mu]^2 \sim G(n/2, 2\sigma^2) \), if \( \mu \) is known and \( r > -n/2 \) then \( T_n^r \) is UMVUE of
\[ E(T_n^r) = 2^r \sigma^2 \frac{\Gamma(r + n/2)}{\Gamma(n/2)}. \]
If \( \sigma^2 \) is known,

\[
f(y) = \frac{1}{\sqrt{2\pi} \sigma} I(y \geq 0) \exp\left(\frac{-1}{2\sigma^2} (\log(y))^2\right) \exp\left(\frac{-\mu^2}{2\sigma^2}\right) \exp\left[\frac{\mu}{\sigma^2} \log(y)\right]
\]
is a 1P–REF.

If \( Y_1, \ldots, Y_n \) are iid LN(\( \mu, \sigma^2 \)), where \( \sigma^2 \) is known, then the likelihood

\[
L(\mu) = c \exp\left(\frac{-n\mu^2}{2\sigma^2}\right) \exp\left[\frac{\mu}{\sigma^2} \sum_{i=1}^{n} \log(y_i)\right],
\]

and the log likelihood

\[
\log(L(\mu)) = d - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} \log(y_i).
\]

Hence

\[
\frac{d}{d\mu} \log(L(\mu)) = \frac{-2n\mu}{2\sigma^2} + \frac{\sum_{i=1}^{n} \log(y_i)}{\sigma^2} \equiv 0,
\]
or \( \sum_{i=1}^{n} \log(y_i) = n\mu \) or

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} \log(y_i)}{n}.
\]

This solution is unique and

\[
\frac{d^2}{d\mu^2} \log(L(\mu)) = -\frac{n}{\sigma^2} < 0.
\]

Since \( T_n = \sum_{i=1}^{n} \log(Y_i) \sim N(n\mu, n\sigma^2) \),

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} \log(Y_i)}{n}
\]
is the UMVUE and MLE of \( \mu \) if \( \sigma^2 \) is known.

When neither \( \mu \) nor \( \sigma \) are known, the log likelihood

\[
\log(L(\mu, \sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\log(y_i) - \mu)^2.
\]

Let \( w_i = \log(y_i) \). Then the log likelihood is

\[
\log(L(\mu, \sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (w_i - \mu)^2,
\]
which has the same form as the normal \( N(\mu, \sigma^2) \) log likelihood. Hence the MLE

\[
(\hat{\mu}, \hat{\sigma}) = \left( \frac{1}{n} \sum_{i=1}^{n} W_i, \sqrt{\frac{1}{n} \sum_{i=1}^{n} (W_i - \bar{W})^2} \right).
\]

Hence inference for \( \mu \) and \( \sigma \) is simple. Use the fact that \( W_i = \log(Y_i) \sim N(\mu, \sigma^2) \) and then perform the corresponding normal based inference on the \( W_i \). For example, a the classical \((1 - \alpha)100\% \) CI for \( \mu \) when \( \sigma \) is unknown is

\[
(W_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S_W}{\sqrt{n}}, W_n + t_{n-1,1-\frac{\alpha}{2}} \frac{S_W}{\sqrt{n}})
\]

where

\[
S_W = \frac{n}{n-1} \hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2},
\]

and \( P(t \leq t_{n-1,1-\frac{\alpha}{2}}) = 1 - \alpha/2 \) when \( t \) is from a \( t \) distribution with \( n-1 \) degrees of freedom. Compare Meeker and Escobar (1998, p. 175).

### 10.36 The Maxwell-Boltzmann Distribution

If \( Y \) has a Maxwell–Boltzmann distribution (also known as a Boltzmann distribution or a Maxwell distribution), \( Y \sim MB(\mu, \sigma) \), then the pdf of \( Y \) is

\[
f(y) = \frac{\sqrt{2}(y-\mu)^2e^{-\frac{(y-\mu)^2}{2\sigma^2}}}{\sigma^3\sqrt{\pi}}
\]

where \( \mu \) is real, \( y \geq \mu \) and \( \sigma > 0 \). This is a location–scale family.

\[
E(Y) = \mu + \sigma \sqrt{2} \frac{1}{\Gamma(3/2)} = \mu + \sigma \frac{2\sqrt{2}}{\sqrt{\pi}}.
\]

\[
\text{VAR}(Y) = 2\sigma^2 \left[ \frac{\Gamma(\frac{5}{2})}{\Gamma(3/2)} - \left( \frac{1}{\Gamma(3/2)} \right)^2 \right] = \sigma^2 \left( 3 - \frac{8}{\pi} \right).
\]

\[
\text{MED}(Y) = \mu + 1.5381722\sigma \quad \text{and} \quad \text{MAD}(Y) = 0.460244\sigma.
\]

This distribution a one parameter exponential family when \( \mu \) is known. Note that \( W = (Y - \mu)^2 \sim G(3/2, 2\sigma^2) \).
If $Z \sim MB(0, \sigma)$, then $Z \sim \text{chi}(p = 3, \sigma)$, and

$$E(Z^r) = 2^{r/2} \sigma^r \frac{\Gamma(r+3/2)}{\Gamma(3/2)}$$

for $r > -3$. The mode of $Z$ is at $\sigma \sqrt{2}$.

### 10.37 The Modified DeMoivre’s Law Distribution

If $Y$ has a modified DeMoivre’s law distribution, $Y \sim MDL(\theta, \phi)$, then the pdf of $Y$ is

$$f(y) = \frac{\phi}{\theta} \left(\frac{\theta - y}{\theta}\right)^{\phi - 1}$$

for $0 < y < \theta$ where $\phi > 0$ and $\theta > 0$. The cdf of $Y$ is

$$F(y) = 1 - \left(1 - \frac{y}{\theta}\right)^\phi$$

for $0 < y < \theta$.

$$E(Y) = \frac{\theta}{\phi + 1}.$$

Notice that

$$f(y) = \frac{\phi}{\theta} I(0 < y < \theta) \frac{\theta}{\theta - y} \exp \left[ -\phi \log\left(\frac{\theta}{\theta - y}\right) \right]$$

is a 1P–REF if $\theta$ is known. Thus $\Theta = (0, \infty)$, $\eta = -\phi$ and $\Omega = (-\infty, 0)$.

$$W = t(Y) = -\log\left(\frac{\theta - Y}{\theta}\right) = \log\left(\frac{\theta}{\theta - Y}\right) \sim EXP(1/\phi).$$

If $Y_1, \ldots, Y_n$ are iid $MDL(\theta, \phi)$, then

$$T_n = -\sum_{i=1}^n \log\left(\frac{\theta - Y_i}{\theta}\right) = \sum_{i=1}^n \log\left(\frac{\theta}{\theta - Y_i}\right) \sim G(n, 1/\phi).$$
If $\theta$ is known, then the likelihood
\[ L(\phi) = c \phi^n \exp \left[ -\phi \sum_{i=1}^{n} \log \left( \frac{\theta}{\theta - y_i} \right) \right], \]
and the log likelihood
\[ \log(L(\phi)) = d + n \log(\phi) - \phi \sum_{i=1}^{n} \log \left( \frac{\theta}{\theta - y_i} \right). \]

Hence
\[ \frac{d}{d\phi} \log(L(\phi)) = \frac{n}{\phi} - \sum_{i=1}^{n} \log \left( \frac{\theta}{\theta - y_i} \right) \Rightarrow 0, \]

or
\[ \hat{\phi} = \frac{n}{\sum_{i=1}^{n} \log \left( \frac{\theta}{\theta - y_i} \right)}. \]

This solution is unique and
\[ \frac{d^2}{d\phi^2} \log(L(\phi)) = -\frac{n}{\phi^2} < 0. \]

Hence if $\theta$ is known, then
\[ \hat{\phi} = \frac{n}{\sum_{i=1}^{n} \log \left( \frac{\theta}{\theta - y_i} \right)} \]
is the MLE of $\phi$. If $r > -n$ and $\theta$ is known, then $T_n^r$ is the UMVUE of
\[ E(T_n^r) = \frac{1}{\phi^r} \frac{\Gamma(r + n)}{\Gamma(n)}. \]

### 10.38 The Negative Binomial Distribution

If $Y$ has a negative binomial distribution (also called the Pascal distribution), $Y \sim NB(r, \rho)$, then the pmf of $Y$ is
\[ f(y) = P(Y = y) = \binom{r + y - 1}{y} \rho^r (1 - \rho)^y. \]
for \( y = 0, 1, \ldots \) where \( 0 < \rho < 1 \). The moment generating function

\[
m(t) = \left[ \frac{\rho}{1 - (1 - \rho) e^t} \right]^r
\]

for \( t < -\log(1 - \rho) \).

\( E(Y) = r(1 - \rho)/\rho \), and

\[
\text{VAR}(Y) = \frac{r(1 - \rho)}{\rho^2}.
\]

Notice that

\[
f(y) = \mathbf{1}[y \in \{0, 1, \ldots\}] \left( r + \frac{y - 1}{\rho} \right)^r \exp[\log(1 - \rho) y]
\]
is a 1P–REF in \( \rho \) for known \( r \). Thus \( \Theta = (0, 1) \), \( \eta = \log(1 - \rho) \) and \( \Omega = (-\infty, 0) \).

If \( Y_1, \ldots, Y_n \) are independent NB\((r, \rho)\), then

\[
\sum_{i=1}^{n} Y_i \sim \text{NB}(\sum_{i=1}^{n} r_i, \rho).
\]

If \( Y_1, \ldots, Y_n \) are iid NB\((r, \rho)\), then

\[
T_n = \sum_{i=1}^{n} Y_i \sim \text{NB}(nr, \rho).
\]

If \( r \) is known, then the likelihood

\[
L(p) = c \rho^{nr} \exp[\log(1 - \rho) \sum_{i=1}^{n} y_i],
\]

and the log likelihood

\[
\log(L(\rho)) = d + nr \log(\rho) + \log(1 - \rho) \sum_{i=1}^{n} y_i.
\]

Hence

\[
\frac{d}{d\rho} \log(L(\rho)) = \frac{nr}{\rho} - \frac{1}{1 - \rho} \sum_{i=1}^{n} y_i \geq 0,
\]
or
\[ \frac{1 - \rho}{\rho} nr = \sum_{i=1}^{n} y_i, \]
or \[ nr - \rho nr - \rho \sum_{i=1}^{n} y_i = 0 \] or
\[ \hat{\rho} = \frac{nr}{nr + \sum_{i=1}^{n} y_i}. \]

This solution is unique and
\[ \frac{d^2}{d\rho^2} \log(L(\rho)) = \frac{-nr}{\rho^2} - \frac{1}{(1 - \rho)^2} \sum_{i=1}^{n} y_i < 0. \]
Thus
\[ \hat{\rho} = \frac{nr}{nr + \sum_{i=1}^{n} Y_i} \]
is the MLE of \( \rho \) if \( r \) is known.

Notice that \( \bar{Y} \) is the UMVUE, MLE and MME of \( r(1 - \rho)/\rho \) if \( r \) is known.

### 10.39 The Normal Distribution

If \( Y \) has a normal distribution (or Gaussian distribution), \( Y \sim N(\mu, \sigma^2) \), then the pdf of \( Y \) is
\[ f(y) = \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(\frac{-(y - \mu)^2}{2\sigma^2}\right) \]
where \( \sigma > 0 \) and \( \mu \) and \( y \) are real. Let \( \Phi(y) \) denote the standard normal cdf. Recall that \( \Phi(y) = 1 - \Phi(-y) \). The cdf \( F(y) \) of \( Y \) does not have a closed form, but
\[ F(y) = \Phi\left(\frac{y - \mu}{\sigma}\right), \]
and
\[ \Phi(y) \approx 0.5(1 + \sqrt{1 - \exp(-2y^2/\pi)}) \]
for \( y \geq 0 \). See Johnson and Kotz (1970a, p. 57).

The moment generating function is
\[ m(t) = \exp(t\mu + t^2\sigma^2/2). \]
The characteristic function is $c(t) = \exp(it\mu - t^2\sigma^2/2)$.

$E(Y) = \mu$ and $\text{VAR}(Y) = \sigma^2$.

$$E[|Y - \mu|^r] = \sigma^r \frac{2^r/2\Gamma((r + 1)/2)}{\sqrt{\pi}} \quad \text{for} \quad r > -1.$$  

If $k \geq 2$ is an integer, then $E(Y^k) = (k - 1)\sigma^2 E(Y^{k-2}) + \mu E(Y^{k-1})$. See Stein (1981) and Casella and Berger (2002, p. 125).

$\text{MED}(Y) = \mu$ and $\text{MAD}(Y) = \Phi^{-1}(0.75) \approx 0.6745\sigma$.

Hence $\sigma = [\Phi^{-1}(0.75)]^{-1}\text{MAD}(Y) \approx 1.483\text{MAD}(Y)$.

This family is a location–scale family which is symmetric about $\mu$.

Suggested estimators are

$$\overline{Y}_n = \hat{\mu} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \text{and} \quad S^2 = S^2_Y = \frac{1}{n - 1} \sum_{i=1}^n (Y_i - \overline{Y}_n)^2.$$  

The classical $(1 - \alpha)100\%$ CI for $\mu$ when $\sigma$ is unknown is

$$\left( \overline{Y}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{S_Y}{\sqrt{n}}, \overline{Y}_n + t_{n-1,1-\frac{\alpha}{2}} \frac{S_Y}{\sqrt{n}} \right)$$

where $P(t \leq t_{n-1,1-\frac{\alpha}{2}}) = 1 - \alpha/2$ when $t$ is from a $t$ distribution with $n - 1$ degrees of freedom.

If $\alpha = \Phi(z_\alpha)$, then

$$z_\alpha \approx m - \frac{c_0 + c_1m + c_2m^2}{1 + d_1m + d_2m^2 + d_3m^3}$$

where

$$m = [-2\log(1 - \alpha)]^{1/2},$$

$c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$, $d_3 = 0.001308$, and $0.5 \leq \alpha$. For $0 < \alpha < 0.5$,

$$z_\alpha = -z_{1-\alpha}.$$  

See Kennedy and Gentle (1980, p. 95).
To see that $\text{MAD}(Y) = \Phi^{-1}(0.75)\sigma$, note that $3/4 = F(\mu + \text{MAD})$ since $Y$ is symmetric about $\mu$. However,

$$F(y) = \Phi \left( \frac{y - \mu}{\sigma} \right)$$

and

$$\frac{3}{4} = \Phi \left( \frac{\mu + \Phi^{-1}(3/4)\sigma - \mu}{\sigma} \right).$$

So $\mu + \text{MAD} = \mu + \Phi^{-1}(3/4)\sigma$. Cancel $\mu$ from both sides to get the result.

Notice that

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{-\mu^2}{2\sigma^2} \right) \exp \left[ -\frac{1}{2\sigma^2} y^2 + \frac{\mu}{\sigma^2} y \right]$$

is a 2P–REF. Hence $\Theta = (0, \infty) \times (-\infty, \infty)$, $\eta_1 = -1/(2\sigma^2)$, $\eta_2 = \mu/\sigma^2$ and $\Omega = (-\infty, 0) \times (-\infty, \infty)$.

If $\sigma^2$ is known,

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} y^2 \right] \exp \left( \frac{-\mu^2}{2\sigma^2} \right) \exp \left( \frac{\mu}{\sigma^2} y \right)$$

is a 1P–REF. Also the likelihood

$$L(\mu) = c \exp \left( \frac{-n\mu^2}{2\sigma^2} \right) \exp \left( \frac{\mu}{\sigma^2} \sum_{i=1}^{n} y_i \right)$$

and the log likelihood

$$\log(L(\mu)) = d - \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} y_i.$$

Hence

$$\frac{d}{d\mu} \log(L(\mu)) = -\frac{2n\mu}{2\sigma^2} + \sum_{i=1}^{n} y_i \quad \text{set} \quad 0,$$

or $n\mu = \sum_{i=1}^{n} y_i$, or

$$\hat{\mu} = \overline{y}.$$

This solution is unique and

$$\frac{d^2}{d\mu^2} \log(L(\mu)) = -\frac{n}{\sigma^2} < 0.$$
Since \( T_n = \sum_{i=1}^{n} Y_i \sim N(n \mu, n \sigma^2) \), \( \bar{Y} \) is the UMVUE, MLE and MME of \( \mu \) if \( \sigma^2 \) is known.

If \( \mu \) is known,

\[
f(y) = \frac{1}{\sqrt{2\pi \sigma^2}} \exp \left[ -\frac{1}{2\sigma^2} (y - \mu)^2 \right]
\]

is a 1P–REF. Also the likelihood

\[
L(\sigma^2) = c \frac{1}{\sigma^n} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]
\]

and the log likelihood

\[
\log(L(\sigma^2)) = d - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.
\]

Hence

\[
\frac{d}{d\sigma^2} \log(L(\sigma^2)) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \mu)^2 \text{ set } = 0,
\]

or \( n \sigma^2 = \sum_{i=1}^{n} (y_i - \mu)^2 \), or

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{n}.
\]

This solution is unique and

\[
\frac{d^2}{d(\sigma^2)^2} \log(L(\sigma^2)) = \frac{n}{2(\sigma^2)^2} - \left. \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{(\sigma^2)^3} \right|_{\sigma^2=\hat{\sigma}^2} = \frac{n}{2(\hat{\sigma}^2)^2} - \frac{n\hat{\sigma}^2}{(\hat{\sigma}^2)^3} \cdot \frac{2}{2}
\]

\[
= -\frac{n}{2(\hat{\sigma}^2)^2} < 0.
\]

Since \( T_n = \sum_{i=1}^{n} (Y_i - \mu)^2 \sim G(n/2, 2\sigma^2) \),

\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{n}
\]

is the UMVUE and MLE of \( \sigma^2 \) if \( \mu \) is known.

Note that if \( \mu \) is known and \( r > -n/2 \), then \( T_n^r \) is the UMVUE of

\[
E(T_n^r) = 2^r \sigma^{2r} \frac{\Gamma(r + n/2)}{\Gamma(n/2)}.
\]
10.40 The One Sided Stable Distribution

If $Y$ has a one sided stable distribution (with index $1/2$, also called a Lévy distribution), $Y \sim OSS(\sigma)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\sqrt{2\pi y^3}} \sqrt{\sigma} \exp \left( \frac{-\sigma}{2} \frac{1}{y} \right)$$

for $y > 0$ and $\sigma > 0$. The cdf

$$F(y) = 2 \left[ 1 - \Phi \left( \frac{\sqrt{\sigma}}{\sqrt{y}} \right) \right]$$

for $y > 0$ where $\Phi(x)$ is the cdf of a $N(0, 1)$ random variable.

This distribution is a scale family with scale parameter $\sigma$ and a 1P–REF. When $\sigma = 1$, $Y \sim INVG(\nu = 1/2, \lambda = 2)$ where INVG stands for inverted gamma. This family is a special case of the inverse Gaussian IG distribution. It can be shown that $W = 1/Y \sim G(1/2, 2/\sigma)$. This distribution is even more outlier prone than the Cauchy distribution. See Feller (1971, p. 52) and Lehmann (1999, p. 76). For applications see Besbeas and Morgan (2004).

If $Y_1, ..., Y_n$ are iid OSS($\sigma$) then $T_n = \sum_{i=1}^{n} \frac{1}{Y_i} \sim G(n/2, 2/\sigma)$. The likelihood

$$L(\sigma) = \prod_{i=1}^{n} f(y_i) = \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi y_i^3}} \right) \sigma^{n/2} \exp \left( \frac{-\sigma}{2} \sum_{i=1}^{n} \frac{1}{y_i} \right),$$

and the log likelihood

$$\log(L(\sigma)) = \log \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi y_i^3}} \right) + \frac{n}{2} \log(\sigma) - \frac{\sigma}{2} \sum_{i=1}^{n} \frac{1}{y_i}. $$

Hence

$$\frac{d}{d\sigma} \log(L(\sigma)) = \frac{n}{2} \frac{1}{\sigma} - \frac{1}{2} \sum_{i=1}^{n} \frac{1}{y_i} = 0,$$

or

$$\frac{n}{2} = \sigma \frac{1}{2} \sum_{i=1}^{n} \frac{1}{y_i}.$$
or
\[ \hat{\sigma} = \frac{n}{\sum_{i=1}^{n} y_i}. \]

This solution is unique and
\[ \frac{d^2}{d\sigma^2} \log(L(\sigma)) = -\frac{n}{2} \frac{1}{\sigma^2} < 0. \]
Hence the MLE
\[ \hat{\sigma} = \frac{n}{\sum_{i=1}^{n} y_i}. \]

Notice that \( T_n/n \) is the UMVUE and MLE of \( 1/\sigma \) and \( T_n^r \) is the UMVUE of
\[ \frac{1}{\sigma^r} \frac{2^r \Gamma(r + n/2)}{\Gamma(n/2)} \]
for \( r > -n/2 \).

### 10.41 The Pareto Distribution

If \( Y \) has a Pareto distribution, \( Y \sim \text{PAR}(\sigma, \lambda) \), then the pdf of \( Y \) is
\[ f(y) = \frac{\frac{1}{\lambda} \sigma^{1/\lambda}}{y^{1+1/\lambda}} \]
where \( y \geq \sigma, \sigma > 0, \) and \( \lambda > 0. \) The mode is at \( Y = \sigma. \) The cdf of \( Y \) is
\[ F(y) = 1 - (\sigma/y)^{1/\lambda} \text{ for } y > \sigma. \]

This family is a scale family with scale parameter \( \sigma \) when \( \lambda \) is fixed.
\[ E(Y) = \frac{\sigma}{1 - \lambda} \]
for \( \lambda < 1. \)
\[ E(Y^r) = \frac{\sigma^r}{1 - r\lambda} \text{ for } r < 1/\lambda. \]

\( \text{MED}(Y) = \sigma 2^\lambda. \)
\( X = \log(Y/\sigma) \) is EXP(\( \lambda \)) and \( W = \log(Y) \) is EXP(\( \theta = \log(\sigma), \lambda \)).
Notice that
\[ f(y) = \frac{1}{y} I[y \geq \sigma] \frac{1}{\lambda} \exp \left[ -\frac{1}{\lambda} \log(y/\sigma) \right] \]
is a one parameter exponential family if $\sigma$ is known.

If $Y_1, \ldots, Y_n$ are iid $\text{PAR}(\sigma, \lambda)$ then

$$T_n = \sum_{i=1}^{n} \log(Y_i/\sigma) \sim G(n, \lambda).$$

If $\sigma$ is known, then the likelihood

$$L(\lambda) = c \frac{1}{\lambda^n} \exp\left[-(1 + \frac{1}{\lambda}) \sum_{i=1}^{n} \log(y_i/\sigma)\right],$$

and the log likelihood

$$\log(L(\lambda)) = d - n \log(\lambda) - \left(1 + \frac{1}{\lambda}\right) \sum_{i=1}^{n} \log(y_i/\sigma).$$

Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} \log(y_i/\sigma) \bigg|_{\lambda=\hat{\lambda}} = 0,$$

or $\sum_{i=1}^{n} \log(y_i/\sigma) = n\lambda$ or

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(y_i/\sigma)}{n}.$$

This solution is unique and

$$\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} \log(y_i/\sigma) \bigg|_{\lambda=\hat{\lambda}} =$$

$$\frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = -\frac{n}{\lambda^2} < 0.$$

Hence

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(Y_i/\sigma)}{n}$$

is the UMVUE and MLE of $\lambda$ if $\sigma$ is known.

If $\sigma$ is known and $r > -n$, then $T_n^r$ is the UMVUE of

$$E(T_n^r) = \lambda^r \frac{\Gamma(r + n)}{\Gamma(n)}.$$
If neither $\sigma$ nor $\lambda$ are known, notice that
\[ f(y) = \frac{1}{y \lambda} \exp \left[ - \left( \frac{\log(y) - \log(\sigma)}{\lambda} \right) \right] I(y \geq \sigma). \]

Hence the likelihood
\[ L(\lambda, \sigma) = c \frac{1}{\lambda^n} \exp \left[ - \sum_{i=1}^{n} \left( \frac{\log(y_i) - \log(\sigma)}{\lambda} \right) \right] I(y_{(1)} \geq \sigma), \]
and the log likelihood is
\[ \log L(\lambda, \sigma) = \left[ d - n \log(\lambda) - \sum_{i=1}^{n} \left( \frac{\log(y_i) - \log(\sigma)}{\lambda} \right) \right] I(y_{(1)} \geq \sigma). \]

Let $w_i = \log(y_i)$ and $\theta = \log(\sigma)$, so $\sigma = e^{\theta}$. Then the log likelihood is
\[ \log L(\lambda, \theta) = \left[ d - n \log(\lambda) - \sum_{i=1}^{n} \left( \frac{w_i - \theta}{\lambda} \right) \right] I(w_{(1)} \geq \theta), \]
which has the same form as the log likelihood of the $\text{EXP}(\theta, \lambda)$ distribution. Hence $(\hat{\lambda}, \hat{\theta}) = (\mathbf{W} - W_{(1)}, W_{(1)})$, and by invariance, the MLE
\[ (\hat{\lambda}, \hat{\sigma}) = (\mathbf{W} - W_{(1)}, Y_{(1)}). \]

Following Abuhassan and Olive (2008), let $D_n = \sum_{i=1}^{n}(W_i - W_{1:n}) = n\hat{\lambda}$ where $W_{(1)} = W_{1:n}$. For $n > 1$, a 100(1 - $\alpha$)% CI for $\theta$ is
\[ (W_{1:n} - \hat{\lambda}[(\alpha)^{-1/(n-1)} - 1], W_{1:n}). \]  \hspace{1cm} (10.11)

Exponentiate the endpoints for a 100(1 - $\alpha$)% CI for $\sigma$. A 100(1 - $\alpha$)% CI for $\lambda$ is
\[ \left( \frac{2D_n}{\chi^2_{2(n-1),1-\alpha/2}}, \frac{2D_n}{\chi^2_{2(n-1),\alpha/2}} \right). \]  \hspace{1cm} (10.12)

This distribution is used to model economic data such as national yearly income data, size of loans made by a bank, et cetera.
10.42 The Poisson Distribution

If \( Y \) has a Poisson distribution, \( Y \sim \text{POIS}(\theta) \), then the pmf of \( Y \) is

\[
f(y) = P(Y = y) = \frac{e^{-\theta} \theta^y}{y!}
\]

for \( y = 0, 1, \ldots \), where \( \theta > 0 \). The mgf of \( Y \) is

\[
m(t) = \exp(\theta(e^t - 1)),
\]

and the characteristic function of \( Y \) is \( c(t) = \exp(\theta(e^{it} - 1)) \).

\[E(Y) = \theta, \quad \text{and} \quad \text{VAR}(Y) = \theta.\]

Chen and Rubin (1986) and Adell and Jodrá (2005) show that

\[-1 < MED(Y) - E(Y) < 1/3.\]

Pourahmadi (1995) showed that the moments of a Poisson \((\theta)\) random variable can be found recursively. If \( k \geq 1 \) is an integer and \( \binom{k}{0} = 1 \), then

\[
E(Y^k) = \theta \sum_{i=0}^{k-1} \binom{k-1}{i} E(Y^i).
\]

The classical estimator of \( \theta \) is \( \hat{\theta} = \overline{Y}_n \).

The approximations \( Y \approx N(\theta, \theta) \) and \( 2\sqrt{Y} \approx N(2\sqrt{\theta}, 1) \) are sometimes used.

Notice that

\[
f(y) = \frac{1}{y!} I[y \in \{0, 1, \ldots\}] e^{-\theta} \exp[\log(\theta)y]
\]
is a 1P–REF. Thus \( \Theta = (0, \infty), \eta = \log(\theta) \) and \( \Omega = (-\infty, \infty) \).

If \( Y_1, \ldots, Y_n \) are independent \( \text{POIS}(\theta_i) \) then \( \sum_{i=1}^{n} Y_i \sim \text{POIS}(\sum_{i=1}^{n} \theta_i) \).

If \( Y_1, \ldots, Y_n \) are iid \( \text{POIS}(\theta) \) then

\[
T_n = \sum_{i=1}^{n} Y_i \sim \text{POIS}(n\theta).
\]

The likelihood

\[
L(\theta) = c \ e^{-n\theta} \exp[\log(\theta) \sum_{i=1}^{n} y_i],
\]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

and the log likelihood

$$\log(L(\theta)) = d - n\theta + \log(\theta) \sum_{i=1}^{n} y_i.$$  

Hence

$$\frac{d}{d\theta} \log(L(\theta)) = -n + \frac{1}{\theta} \sum_{i=1}^{n} y_i = 0,$$

or $\sum_{i=1}^{n} y_i = n\theta$, or

$$\hat{\theta} = \bar{y}.$$  

This solution is unique and

$$\frac{d^2}{d\theta^2} \log(L(\theta)) = \frac{-\sum_{i=1}^{n} y_i}{\theta^2} < 0$$

unless $\sum_{i=1}^{n} y_i = 0$.

Hence $\bar{Y}$ is the UMVUE and MLE of $\theta$.

Let $W = \sum_{i=1}^{n} Y_i$ and suppose that $W = w$ is observed. Let $P(T < \chi^2_d(\alpha)) = \alpha$ if $T \sim \chi^2_d$. Then an “exact” 100 $(1 - \alpha)$% CI for $\theta$ is

$$\left( \frac{\chi^2_{2w}(\frac{\alpha}{2})}{2n}, \frac{\chi^2_{2w+2}(1 - \frac{\alpha}{2})}{2n} \right)$$

for $w \neq 0$ and

$$\left( 0, \frac{\chi^2_{2}(1 - \alpha)}{2n} \right)$$

for $w = 0$.

10.43 The Power Distribution

If $Y$ has a power distribution, $Y \sim \text{POW}(\lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda} y^{\lambda - 1},$$

where $\lambda > 0$ and $0 < y \leq 1$. The cdf of $Y$ is $F(y) = y^{1/\lambda}$ for $0 < y \leq 1$.

MED$(Y) = (1/2)^{\lambda}$.

$W = -\log(Y)$ is EXP$(\lambda)$. Notice that $Y \sim \beta(\delta = 1/\lambda, \nu = 1)$.  

CHAPTER 10. SOME USEFUL DISTRIBUTIONS

Notice that

\[ f(y) = \frac{1}{\lambda} I_{(0,1]}(y) \exp \left[ \left( \frac{1}{\lambda} - 1 \right) \log(y) \right] \]

\[ = \frac{1}{y} I_{(0,1]}(y) \frac{1}{\lambda} \exp \left[ \frac{-1}{\lambda} (-\log(y)) \right] \]

is a \textbf{1P–REF}. Thus \( \Theta = (0, \infty), \eta = -1/\lambda \) and \( \Omega = (-\infty, 0) \).

If \( Y_1, ..., Y_n \) are iid \( \text{POW}(\lambda) \), then

\[ T_n = -\sum_{i=1}^n \log(Y_i) \sim G(n, \lambda). \]

The likelihood

\[ L(\lambda) = \frac{1}{\lambda^n} \exp \left[ \left( \frac{1}{\lambda} - 1 \right) \sum_{i=1}^n \log(y_i) \right], \]

and the log likelihood

\[ \log(L(\lambda)) = -n \log(\lambda) + \left( \frac{1}{\lambda} - 1 \right) \sum_{i=1}^n \log(y_i). \]

Hence

\[ \frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} - \frac{\sum_{i=1}^n \log(y_i)}{\lambda^2} \text{ set } = 0, \]

or \( -\sum_{i=1}^n \log(y_i) = n \lambda \), or

\[ \hat{\lambda} = \frac{-\sum_{i=1}^n \log(y_i)}{n}. \]

This solution is unique and

\[ \frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} + \frac{2 \sum_{i=1}^n \log(y_i)}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}} \]

\[ = \frac{n}{\lambda^2} - \frac{2 \hat{\lambda}}{\lambda^3} = -\frac{n}{\lambda^2} < 0. \]

Hence

\[ \hat{\lambda} = \frac{-\sum_{i=1}^n \log(Y_i)}{n}. \]
is the UMVUE and MLE of $\lambda$.

If $r > -n$, then $T_n^r$ is the UMVUE of

$$E(T_n^r) = \lambda^r \frac{\Gamma(r + n)}{\Gamma(n)}.$$

A 100(1 - $\alpha$)% CI for $\lambda$ is

$$\left( \frac{2T_n}{\chi^2_{2n,1-\alpha/2}}, \frac{2T_n}{\chi^2_{2n,\alpha/2}} \right).$$

(10.13)

10.44 The Two Parameter Power Distribution

$Y$ has a two parameter power distribution, $Y \sim \text{power}(\tau, \lambda)$, if the pdf of $Y$ is

$$f(y) = \frac{1}{\tau \lambda} \left( \frac{y}{\tau} \right)^{\tau-1} I(0 < y \leq \tau)$$

where $\tau > 0$ and $\lambda > 0$. The cdf

$$F(y) = \left( \frac{y}{\tau} \right)^{1/\lambda}$$

for $0 < y \leq \tau$. This is a scale family for fixed $\lambda$.

$W = -\log(Y) \sim EXP(-\log(\tau), \lambda)$.

If $Y_1, ..., Y_n$ are iid $\text{power}(\tau, \lambda)$, then the likelihood is

$$L(\tau, \lambda) = \frac{1}{\tau^{n/\lambda}} I(0 < y(n) \leq \tau) \frac{1}{\lambda^n} \exp \left[ \frac{1}{\lambda} - 1 \right] \sum_{i=1}^{n} \log(y_i) \right].$$

For fixed $\lambda$, the likelihood $L$ is maximized by minimizing $\tau$. So $\hat{\tau} = Y(n)$.

Then the profile likelihood

$$L_p(\lambda) = L(y(n), \lambda) = \frac{1}{y(n)^{n/\lambda}} \frac{1}{\lambda^n} \exp \left[ \frac{1}{\lambda} - 1 \right] \sum_{i=1}^{n} \log(y_i) \right],$$

and the log profile likelihood

$$\log(L_p(\lambda)) = \log(L(y(n), \lambda) = -\frac{n}{\lambda} \log(y(n)) - n \log(\lambda) + \frac{1}{\lambda} - 1 \sum_{i=1}^{n} \log(y_i).$$
Hence
\[ \frac{d}{d\lambda} \log(L_p(\lambda)) = \frac{n}{\lambda^2} \log(y(n)) - \frac{n}{\lambda} \frac{\sum_{i=1}^{n} \log(y_i)}{\lambda^2} = 0, \]
or
\[ -n\lambda + n \log(y(n)) - \sum_{i=1}^{n} \log(y_i) = 0 \]
or
\[ - \sum_{i=1}^{n} \frac{\log(y_i)}{y_i} = n\lambda, \]
or
\[ \hat{\lambda} = \frac{-1}{n} \sum_{i=1}^{n} \log\left(\frac{y_i}{y(n)}\right). \]

This solution is unique and
\[ \frac{d^2}{d\lambda^2} \log(L_p(\lambda)) = \frac{n}{\lambda^2} + \frac{2 \sum_{i=1}^{n} \log\left(\frac{y_i}{y(n)}\right)}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}} \]
\[ = \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = \frac{-n}{\lambda^2} < 0. \]

Hence
\[ (\hat{\tau}, \hat{\lambda}) = \left( Y(n), \frac{-1}{n} \sum_{i=1}^{n} \log\left(\frac{Y_i}{Y(n)}\right) \right) \]
is the MLE of \((\tau, \lambda)\).

Confidence intervals for \(\tau\) and \(\lambda\) are given in Example 9.12.

### 10.45 The Rayleigh Distribution

If \(Y\) has a Rayleigh distribution, \(Y \sim R(\mu, \sigma)\), then the pdf of \(Y\) is
\[ f(y) = \frac{y - \mu}{\sigma^2} \exp\left[ -\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right] \]
where \(\sigma > 0\), \(\mu\) is real, and \(y \geq \mu\). See Cohen and Whitten (1988, Ch. 10).

This is an asymmetric location–scale family. The cdf of \(Y\) is
\[ F(y) = 1 - \exp\left[ -\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2 \right] \]
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

for \( y \geq \mu \), and \( F(y) = 0 \), otherwise.

\[
E(Y) = \mu + \sigma \sqrt{\pi/2} \approx \mu + 1.253314\sigma. \\

\text{VAR}(Y) = \sigma^2 (4 - \pi) / 2 \approx 0.429204\sigma^2.
\]

\( \text{MED}(Y) = \mu + \sigma \sqrt{\log(4) / 2} \approx \mu + 1.17741\sigma. \)

Hence \( \mu \approx \text{MED}(Y) - 2.6255\text{MAD}(Y) \) and \( \sigma \approx 2.230\text{MAD}(Y). \)

Let \( \sigma D = \text{MAD}(Y). \) If \( \mu = 0 \), and \( \sigma = 1 \), then

\[
0.5 = \exp[-0.5(\sqrt{\log(4) - D})^2] - \exp[-0.5(\sqrt{\log(4) + D})^2].
\]

Hence \( D \approx 0.448453 \) and \( \text{MAD}(Y) \approx 0.448453\sigma. \)

It can be shown that \( W = (Y - \mu)^2 \sim \exp(2\sigma^2). \)

Other parameterizations for the Rayleigh distribution are possible.

Note that

\[
f(y) = (y - \mu)I(y \geq \mu) \frac{1}{\sigma^2} \exp\left[-\frac{1}{2\sigma^2}(y - \mu)^2\right]
\]

is a 1P–REF if \( \mu \) is known.

If \( Y_1, \ldots, Y_n \) are iid \( R(\mu, \sigma) \), then

\[
T_n = \sum_{i=1}^{n} (Y_i - \mu)^2 \sim G(n, 2\sigma^2).
\]

If \( \mu \) is known, then the likelihood

\[
L(\sigma^2) = c \frac{1}{\sigma^{2n}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right],
\]

and the log likelihood

\[
\log(L(\sigma^2)) = d - n \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.
\]

Hence

\[
\frac{d}{d(\sigma^2)} \log(L(\sigma^2)) = -\frac{n}{\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (y_i - \mu)^2 \equiv 0,
\]
or \( \sum_{i=1}^{n} (y_i - \mu)^2 = 2n \sigma^2 \), or
\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2n}.
\]
This solution is unique and
\[
\frac{d^2}{d(\sigma^2)^2} \log(L(\sigma^2)) = \frac{n}{(\sigma^2)^2} - \frac{\sum_{i=1}^{n} (y_i - \mu)^2}{(\sigma^2)^3} \bigg|_{\sigma^2 = \hat{\sigma}^2} = \frac{n}{(\hat{\sigma}^2)^2} - \frac{2n \hat{\sigma}^2}{(\hat{\sigma}^2)^3} = -\frac{n}{(\hat{\sigma}^2)^2} < 0.
\]
Hence
\[
\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{2n}
\]
is the UMVUE and MLE of \( \sigma^2 \) if \( \mu \) is known.

If \( \mu \) is known and \( r > -n \), then \( T^r_n \) is the UMVUE of
\[
E(T^r_n) = 2^r \sigma^2 r \frac{\Gamma(r + n)}{\Gamma(n)}.
\]

### 10.46 The Smallest Extreme Value Distribution

If \( Y \) has a smallest extreme value distribution (or log-Weibull distribution), \( Y \sim SEV(\theta, \sigma) \), then the pdf of \( Y \) is
\[
f(y) = \frac{1}{\sigma} \exp\left(\frac{y - \theta}{\sigma}\right) \exp[-\exp\left(\frac{y - \theta}{\sigma}\right)]
\]
where \( y \) and \( \theta \) are real and \( \sigma > 0 \). The cdf of \( Y \) is
\[
F(y) = 1 - \exp[-\exp\left(\frac{y - \theta}{\sigma}\right)].
\]
This family is an asymmetric location-scale family with a longer left tail than right.
\[
E(Y) \approx \theta - 0.57721 \sigma, \text{ and}
\]
\[
\text{VAR}(Y) = \sigma^2 \pi^2 / 6 \approx 1.64493 \sigma^2.
\]
\[
\text{MED}(Y) = \theta - \sigma \log(\log(2)).
\]
\[
\text{MAD}(Y) \approx 0.767049 \sigma.
\]
\( Y \) is a one parameter exponential family in \( \theta \) if \( \sigma \) is known.

If \( Y \) has a \( SEV(\theta, \sigma) \) distribution, then \( W = -Y \) has an \( LEV(-\theta, \sigma) \) distribution.
10.47 The Student’s t Distribution

If $Y$ has a Student’s $t$ distribution, $Y \sim t_p$, then the pdf of $Y$ is

$$f(y) = \frac{\Gamma\left(\frac{p+1}{2}\right)}{(p\pi)^{1/2}\Gamma(p/2)} \left(1 + \frac{y^2}{p}\right)^{-\frac{p+1}{2}}$$

where $p$ is a positive integer and $y$ is real. This family is symmetric about 0. The $t_1$ distribution is the Cauchy$(0, 1)$ distribution. If $Z \sim N(0, 1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(W/p)^{1/2}}$$

is $t_p$.

$E(Y) = 0$ for $p \geq 2$.

MED($Y$) = 0.

$\text{VAR}(Y) = p/(p - 2)$ for $p \geq 3$, and

$\text{MAD}(Y) = t_{p,0.75}$ where $P(t_p \leq t_{p,0.75}) = 0.75$.

If $\alpha = P(t_p \leq t_{p,\alpha})$, then Cooke, Craven, and Clarke (1982, p. 84) suggest the approximation

$$t_{p,\alpha} \approx \sqrt{p[\exp\left(\frac{w_{\alpha}^2}{p}\right) - 1]}$$

where

$$w_{\alpha} = \frac{z_\alpha(8p + 3)}{8p + 1}$$

$z_\alpha$ is the standard normal cutoff: $\alpha = \Phi(z_\alpha)$, and $0.5 \leq \alpha$. If $0 < \alpha < 0.5$, then

$$t_{p,\alpha} = -t_{p,1-\alpha}.$$ 

This approximation seems to get better as the degrees of freedom increase.

10.48 The Topp-Leone Distribution

If $Y$ has a Topp–Leone distribution, $Y \sim TL(\nu)$, then pdf of $Y$ is

$$f(y) = \nu(2 - 2y)(2y - y^2)^{\nu-1}$$
for \( \nu > 0 \) and \( 0 < y < 1 \). The cdf of \( Y \) is \( F(y) = (2y - y^2)^\nu \) for \( 0 < y < 1 \).

This distribution is a 1P–REF since

\[
f(y) = (2 - 2y)I_{(0,1)}(y) \nu \exp[(1 - \nu)(-\log(2y - y^2))].
\]

\[
\text{MED}(Y) = 1 - \sqrt{1 - (1/2)^{1/\nu}}, \quad \text{and Example 2.17 showed that} \quad W = -\log(2Y - Y^2) \sim EXP(1/\nu).
\]

The likelihood

\[
L(\nu) = c \nu^n \prod_{i=1}^{n} (2y_i - y_i^2)^{\nu - 1},
\]

and the log likelihood

\[
\log(L(\nu)) = d + n \log(\nu) + (\nu - 1) \sum_{i=1}^{n} \log(2y_i - y_i^2).
\]

Hence

\[
\frac{d}{d\nu} \log(L(\nu)) = \frac{n}{\nu} + \sum_{i=1}^{n} \log(2y_i - y_i^2) \overset{se}{=} 0,
\]

or \( n + \nu \sum_{i=1}^{n} \log(2y_i - y_i^2) = 0 \), or

\[
\hat{\nu} = \frac{-n}{\sum_{i=1}^{n} \log(2y_i - y_i^2)}.
\]

This solution is unique and

\[
\frac{d^2}{d\nu^2} \log(L(\nu)) = -\frac{n}{\nu^2} < 0.
\]

Hence

\[
\hat{\nu} = \frac{-n}{\sum_{i=1}^{n} \log(2Y_i - Y_i^2)} = \frac{n}{-\sum_{i=1}^{n} \log(2Y_i - Y_i^2)}
\]

is the MLE of \( \nu \).

If \( T_n = -\sum_{i=1}^{n} \log(2Y_i - Y_i^2) \sim G(n, 1/\nu) \), then \( T_n^r \) is the UMVUE of

\[
E(T_n^r) = \frac{1}{\nu^r} \frac{\Gamma(r + n)}{\Gamma(n)}
\]

for \( r > -n \). In particular, \( \hat{\nu} = \frac{n}{T_n^r} \) is the MLE and \( \frac{n-1}{T_n^r} \) is the UMVUE of \( \nu \) for \( n > 1 \).
10.49 The Truncated Extreme Value Distribution

If $Y$ has a truncated extreme value distribution, $Y \sim \text{TEV}(\lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{1}{\lambda} \exp \left( y - \frac{e^y - 1}{\lambda} \right)$$

where $y > 0$ and $\lambda > 0$. The cdf of $Y$ is

$$F(y) = 1 - \exp \left[ -\frac{(e^y - 1)}{\lambda} \right]$$

for $y > 0$.

$\text{MED}(Y) = \log(1 + \lambda \log(2))$.

$W = e^Y - 1$ is EXP($\lambda$).

Notice that

$$f(y) = e^y I(y \geq 0) \frac{1}{\lambda} \exp \left[ -\frac{1}{\lambda} (e^y - 1) \right]$$

is a 1P–REF. Hence $\Theta = (0, \infty), \eta = -1/\lambda$ and $\Omega = (-\infty, 0)$.

If $Y_1, \ldots, Y_n$ are iid TEV($\lambda$), then

$$T_n = \sum_{i=1}^n (e^{Y_i} - 1) \sim G(n, \lambda).$$

The likelihood

$$L(\lambda) = c \frac{1}{\lambda^n} \exp \left[ -\frac{1}{\lambda} \sum_{i=1}^n \log(e^{y_i} - 1) \right],$$

and the log likelihood

$$\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^n \log(e^{y_i} - 1).$$

Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{\sum_{i=1}^n \log(e^{y_i} - 1)}{\lambda^2} = 0,$$
or \( \sum_{i=1}^{n} \log(e^{y_i} - 1) = n\lambda \), or

\[
\hat{\lambda} = -\frac{\sum_{i=1}^{n} \log(e^{y_i} - 1)}{n}.
\]

This solution is unique and

\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2\sum_{i=1}^{n} \log(e^{y_i} - 1)}{\lambda^3} \bigg|_{\lambda = \hat{\lambda}}
\]

\[
= \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = \frac{-n}{\lambda^2} < 0.
\]

Hence

\[
\hat{\lambda} = -\frac{\sum_{i=1}^{n} \log(e^{y_i} - 1)}{n}
\]

is the UMVUE and MLE of \( \lambda \).

If \( r > -n \), then \( T_n^r \) is the UMVUE of

\[
E(T_n^r) = \lambda^r \frac{\Gamma(r + n)}{\Gamma(n)}.
\]

A 100(1 - \( \alpha \))% CI for \( \lambda \) is

\[
\left( \frac{2T_n}{\lambda_{2n,1-\alpha/2}}, \frac{2T_n}{\lambda_{2n,\alpha/2}} \right). \tag{10.14}
\]

### 10.50 The Uniform Distribution

If \( Y \) has a uniform distribution, \( Y \sim U(\theta_1, \theta_2) \), then the pdf of \( Y \) is

\[
f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \leq y \leq \theta_2).
\]

The cdf of \( Y \) is \( F(y) = (y - \theta_1)/(\theta_2 - \theta_1) \) for \( \theta_1 \leq y \leq \theta_2 \).

This family is a location-scale family which is symmetric about \( (\theta_1 + \theta_2)/2 \).

By definition, \( m(0) = c(0) = 1 \). For \( t \neq 0 \), the mgf of \( Y \) is

\[
m(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{(\theta_2 - \theta_1)t}.
\]
and the characteristic function of $Y$ is

$$c(t) = \frac{e^{i\theta_2} - e^{i\theta_1}}{(\theta_2 - \theta_1)it}.$$ 

$E(Y) = (\theta_1 + \theta_2)/2$, and

$\text{MED}(Y) = (\theta_1 + \theta_2)/2$.

$\text{VAR}(Y) = (\theta_2 - \theta_1)^2/12$, and

$\text{MAD}(Y) = (\theta_2 - \theta_1)/4$.

Note that $\theta_1 = \text{MED}(Y) - 2\text{MAD}(Y)$ and $\theta_2 = \text{MED}(Y) + 2\text{MAD}(Y)$.

Some classical estimators are $\hat{\theta}_1 = Y^{(1)}$ and $\hat{\theta}_2 = Y^{(n)}$.

### 10.51 The Weibull Distribution

If $Y$ has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of $Y$ is

$$f(y) = \frac{\phi}{\lambda} y^{\phi-1} e^{-y^\phi/\lambda}$$

where $\lambda, y$, and $\phi$ are all positive. For fixed $\phi$, this is a scale family in $\sigma = \lambda^{1/\phi}$. The cdf of $Y$ is $F(y) = 1 - \exp(-y^\phi/\lambda)$ for $y > 0$.

$E(Y) = \lambda^{1/\phi} \Gamma(1 + 1/\phi)$.

$\text{VAR}(Y) = \lambda^{2/\phi} \Gamma(1 + 2/\phi) - (E(Y))^2$.

$E(Y^r) = \lambda^{r/\phi} \Gamma(1 + \frac{r}{\phi})$ for $r > -\phi$.

$\text{MED}(Y) = (\lambda \log(2))^{1/\phi}$. Note that

$$\lambda = \frac{(\text{MED}(Y))^{\phi}}{\log(2)}.$$

$W = Y^\phi$ is EXP$(\lambda)$.

$W = \log(Y)$ has a smallest extreme value SEV$(\theta = \log(\lambda^{1/\phi}), \sigma = 1/\phi)$ distribution.

Notice that

$$f(y) = y^\phi-1 I(y \geq 0) \frac{\phi}{\lambda} \exp\left[\frac{-1}{\lambda} y^\phi\right]$$

is a one parameter exponential family in $\lambda$ if $\phi$ is known.
If $Y_1, \ldots, Y_n$ are iid $W(\phi, \lambda)$, then

$$T_n = \sum_{i=1}^{n} Y_i^{\phi} \sim G(n, \lambda).$$

If $\phi$ is known, then the likelihood

$$L(\lambda) = c \frac{1}{\lambda^n} \exp \left[ -\frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi} \right],$$

and the log likelihood

$$\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} y_i^{\phi}.$$ 

Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} + \frac{\sum_{i=1}^{n} y_i^{\phi}}{\lambda^2} = 0,$$

or $\sum_{i=1}^{n} y_i^{\phi} = n \lambda$, or

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} y_i^{\phi}}{n}.$$

This solution was unique and

$$\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2 \sum_{i=1}^{n} y_i^{\phi}}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}}$$

$$= \frac{n}{\lambda^2} - \frac{2n \hat{\lambda}}{\lambda^3} = -\frac{n}{\lambda^2} < 0.$$ 

Hence

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i^{\phi}}{n}$$

is the UMVUE and MLE of $\lambda$.

If $r > -n$, then $T_n^{r}$ is the UMVUE of

$$E(T_n^{r}) = \lambda^r \frac{\Gamma(r + n)}{\Gamma(n)}.$$

MLEs and CIs for $\phi$ and $\lambda$ are discussed in Example 9.18.
10.52 The Zero Truncated Poisson Distribution

If $Y$ has a zero truncated Poisson distribution, $Y \sim ZTP(\theta)$, then the pmf of $Y$ is

$$f(y) = \frac{e^{-\theta} \theta^y}{(1 - e^{-\theta}) y!}$$

for $y = 1, 2, 3, \ldots$ where $\theta > 0$. This distribution is a 1P-REF. The mgf of $Y$ is

$$m(t) = \exp(\theta(e^t - 1)) \frac{1 - e^{-\theta e^t}}{1 - e^{-\theta}}$$

for all $t$.

$$E(Y) = \frac{\theta}{1 - e^{-\theta}}$$

and

$$V(Y) = \frac{\theta^2 + \theta}{1 - e^{-\theta}} - \left( \frac{\theta}{1 - e^{-\theta}} \right)^2.$$

The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. If $W \sim Poisson(\theta)$ with pmf $f_W(y)$, then $P(W = 0) = e^{-\theta}$, so $\sum_{y=1}^{\infty} f_W(y) = 1 - e^{-\theta} = \sum_{y=0}^{\infty} f_W(y) - \sum_{y=1}^{\infty} f_W(y)$. So the ZTP pmf $f(y) = f_W(y)/(1 - e^{-\theta})$ for $y \neq 0$.

Now $E(Y) = \sum_{y=1}^{\infty} y f(y) = \sum_{y=0}^{\infty} y f(y) = \sum_{y=0}^{\infty} y f_W(y)/(1 - e^{-\theta}) = E(W)/(1 - e^{-\theta}) = \theta/(1 - e^{-\theta})$.

Similarly, $E(Y^2) = \sum_{y=1}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f(y) = \sum_{y=0}^{\infty} y^2 f_W(y)/(1 - e^{-\theta}) = E(W^2)/(1 - e^{-\theta}) = [\theta^2 + \theta]/(1 - e^{-\theta})$. So

$$V(Y) = E(Y^2) - (E(Y))^2 = \frac{\theta^2 + \theta}{1 - e^{-\theta}} - \left( \frac{\theta}{1 - e^{-\theta}} \right)^2.$$

10.53 The Zeta Distribution

If $Y$ has a Zeta distribution, $Y \sim Zeta(\nu)$, then the pmf of $Y$ is

$$f(y) = P(Y = y) = \frac{1}{y^\nu \zeta(\nu)}$$
where $\nu > 1$ and $y = 1, 2, 3, \ldots$. Here the zeta function

$$\zeta(\nu) = \sum_{y=1}^{\infty} \frac{1}{y^\nu}$$

for $\nu > 1$. This distribution is a one parameter exponential family.

$$E(Y) = \frac{\zeta(\nu - 1)}{\zeta(\nu)}$$

for $\nu > 2$, and

$$\text{VAR}(Y) = \frac{\zeta(\nu - 2)}{\zeta(\nu)} - \left[ \frac{\zeta(\nu - 1)}{\zeta(\nu)} \right]^2$$

for $\nu > 3$.

$$E(Y^r) = \frac{\zeta(\nu - r)}{\zeta(\nu)}$$

for $\nu > r + 1$.

This distribution is sometimes used for count data, especially by linguistics for word frequency. See Lindsey (2004, p. 154).

### 10.54 The Zipf Distribution

$Y$ has a Zipf distribution, $Y \sim \text{Zipf}(\nu)$, if the pmf of $Y$ is

$$f(y) = \frac{1}{y^\nu z(\nu)}$$

where $y \in \{0, 1, \ldots, m\}$ and $m$ is known, $\nu$ is real and

$$z(\nu) = \sum_{y=1}^{m} \frac{1}{y^\nu}.$$

$$E(Y^r) = \frac{z(\nu - r)}{z(\nu)}$$

for real $r$. This family is a 1P-REF if $m$ is known.
CHAPTER 10. SOME USEFUL DISTRIBUTIONS

10.55 Complements


Abuassan and Olive (2008) discuss confidence intervals for the two parameter exponential, two parameter power, half normal and Pareto distributions.

A recent discussion of Burr TYPE III and XII distributions is given in Headrick, Pant and Sheng (2010).

Brownstein and Pensky (2008) show that if \( Y = t(X) \) where \( X \) has simple inference and \( t \) is a monotone transformation that does not depend on any unknown parameters, then \( Y \) also has simple inference.

Other distributions in this text (including the hburr, hlev, hpar, hpow, hray, hsev, htev, hweib and truncated exponential distributions) are the distributions in Problems 1.29, 1.30, 1.35, 2.52-2.59, 3.8, 3.9, 4.9a, and 6.25.
Chapter 11

Stuff for Students

To be blunt, many of us are lousy teachers, and our efforts to improve are feeble. So students frequently view statistics as the worst course taken in college.

Hogg (1991)

11.1 R/Splus Statistical Software

*R*/Splus are statistical software packages, and *R* is the free version of *Splus*. A very useful *R* link is (www.r-project.org/#doc).

As of May 2012, the author’s personal computer has Version 2.13.1 (July 8, 2011) of *R* and Splus–2000 (see Mathsoft 1999ab).

**Downloading the book’s R/Splus functions sipack.txt into R or Splus:**

In Chapter 9, several of the homework problems use *R*/Splus functions contained in the book’s website (www.math.siu.edu/olive/sipack.txt) under the file name sipack.txt. Suppose that you download sipack.txt onto a flash drive G. Enter *R* and wait for the cursor to appear. Then go to the File menu and drag down Source R Code. A window should appear. Navigate the Look in box until it says Removable Disk (G:). In the Files of type box choose All files(*.*) and then select sipack.txt. The following line should appear in the main *R* window.

```r
> source("G:/sipack.txt")
```

388
Type `ls()`. About 9 R/Splus functions from `sipack.txt` should appear.

Alternatively, from the website (www.math.siu.edu/olive/sipack.txt), go to the Edit menu and choose Select All, then go to the Edit menu and choose Copy. Next enter R, go to the Edit menu and choose Paste. These commands also enter the sipack functions into R.

When you finish your R/Splus session, enter the command `q()`. A window asking “Save workspace image?” will appear. Click on No if you do not want to save the programs in R. (If you do want to save the programs then click on Yes.)

If you use Splus, the command

```
> source("G:/sipack.txt")
```

will enter the functions into Splus. Creating a special workspace for the functions may be useful.

This section gives tips on using R/Splus, but is no replacement for books such as Becker, Chambers, and Wilks (1988), Chambers (1998), Crawley (2005, 2007), Dalgaard (2002) or Venables and Ripley (2010). Also see Mathsoft (1999ab) and use the website (www.google.com) to search for useful websites. For example enter the search words R documentation.

The command `q()` gets you out of R or Splus.

The commands `help(fn)` and `args(fn)` give information about the function fn, eg if fn = rnorm.

**Making functions in R and Splus is easy.**

For example, type the following commands.

```r
mysquare <- function(x){
  # this function squares x
  r <- x^2
  r }
```

The second line in the function shows how to put comments into functions.

**Modifying your function is easy.**

Use the fix command.

`fix(mysquare)`

This will open an editor such as Notepad and allow you to make changes.
In Splus, the command `Edit(mysquare)` may also be used to modify the function `mysquare`.

To save data or a function in R, when you exit, click on Yes when the “Save worksheet image?” window appears. When you reenter R, type `ls()`. This will show you what is saved. You should rarely need to save anything for this book. In Splus, data and functions are automatically saved. To remove unwanted items from the worksheet, eg `x`, type `rm(x)`, `pairs(x)` makes a scatterplot matrix of the columns of `x`, `hist(y)` makes a histogram of `y`, `boxplot(y)` makes a boxplot of `y`, `stem(y)` makes a stem and leaf plot of `y`, `scan()`, `source()`, and `sink()` are useful on a Unix workstation.

To type a simple list, use `y <- c(1,2,3.5)`. The commands `mean(y)`, `median(y)`, `var(y)` are self explanatory.

The following commands are useful for a scatterplot created by the command `plot(x,y)`.

- `lines(x,y)`, `lines(lowess(x,y,f=.2))`
- `identify(x,y)`
- `abline(out$coef), abline(0,1)`

The usual arithmetic operators are `2 + 4`, `3 - 7`, `8 * 4`, `8/4`, and `2^{10}`.

The ith element of vector `y` is `y[i]` while the ij element of matrix `x` is `x[i,j]`. The second row of `x` is `x[2,]` while the 4th column of `x` is `x[,4]`. The transpose of `x` is `t(x)`.

The command `apply(x,1,fn)` will compute the row means if `fn = mean`. The command `apply(x,2,fn)` will compute the column variances if `fn = var`. The commands `cbind` and `rbind` combine column vectors or row vectors with an existing matrix or vector of the appropriate dimension.

### 11.2 Hints and Solutions to Selected Problems

1.10. d) See Problem 1.19 with `Y = W` and `r = 1`. 
CHAPTER 11. STUFF FOR STUDENTS

f) Use the fact that \( E(Y^r) = E[(Y^r)^{r/\phi}] = E(W^{r/\phi}) \) where \( W \sim EXP(\lambda) \). Take \( r = 1 \).

1.11. d) Find \( E(Y^r) \) for \( r = 1, 2 \) using Problem 1.19 with \( Y = W \).

f) For \( r = 1, 2 \), find \( E(Y^r) \) using the fact that \( E(Y^r) = E[(Y^\phi)^{r/\phi}] = E(W^{r/\phi}) \) where \( W \sim EXP(\lambda) \).

1.12. a) 200
b) 0.9(10) + 0.1(200) = 29

1.13. a) 400(1) = 400
b) 0.9E(Z) + 0.1E(W) = 0.9(10) + 0.1(400) = 49

1.15. a) \( 1 \frac{A}{A+B} + 0 \frac{B}{A+B} = \frac{A}{A+B} \).
b) \( \frac{A}{A+B} \).

1.16. a) \( g(x_o)P(X = x_o) = g(x_o) \)
b) \( E(e^{tX}) = e^{tx_o} \) by a).
c) \( m'(t) = x_o e^{tx_o}, m''(t) = x_o^2 e^{tx_o}, m^{(n)}(t) = x_o^n e^{tx_o} \).

1.17. \( m(t) = E(e^{tX}) = e^t P(X = 1) + e^{-t} P(X = -1) = 0.5(e^t + e^{-t}) \).

1.18. a) \( \sum_{x=0}^n x e^{tx} f(x) \)
b) \( \sum_{x=0}^n x f(x) = E(X) \)
c) \( \sum_{x=0}^n x^2 e^{tx} f(x) \)
d) \( \sum_{x=0}^n x^2 f(x) = E(X^2) \)
e) \( \sum_{x=0}^n x^k e^{tx} f(x) \)

1.19. \( E(W^r) = E(e^{tX}) = m_X(r) = \exp(r\mu + r^2\sigma^2/2) \) where \( m_X(t) \) is the mgf of a \( N(\mu, \sigma^2) \) random variable.

1.20. a) \( E(X^2) = V(X) + (E(X))^2 = \sigma^2 + \mu^2 \).
b) \( E(X^3) = 2\sigma^2 E(X) + \mu E(X^2) = 2\sigma^2 \mu + \mu(\sigma^2 + \mu^2) = 3\sigma^2 \mu + \mu^3 \).

1.22. \( \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty \exp(-\frac{1}{2}y^2)dy = 1 \). So \( \int_{-\infty}^\infty \exp(-\frac{1}{2}y^2)dy = \sqrt{2\pi} \).

1.23. \( \int_\sigma^\infty f(x|\sigma, \theta)dx = 1 \), so
\[
\int_\sigma^\infty \frac{1}{x^{\theta+1}}dx = \frac{1}{\theta \sigma^\theta}.
\] (11.1)
CHAPTER 11. STUFF FOR STUDENTS

So
\[ EX^r = \int_0^\infty x^r \theta \sigma^\theta \frac{1}{x^\theta + 1} dx = \theta \sigma^\theta \int_\sigma^\infty \frac{1}{x^{\theta-r}+1} dx = \frac{\theta \sigma^\theta}{(\theta - r) \sigma^{\theta-r}} \]
by Equation (11.1). So
\[ EX^r = \frac{\theta \sigma^r}{\theta - r} \]
for \( \theta > r \).

1.24.
\[ EY^r = \int_0^1 y^r \frac{\Gamma(\delta + \nu)}{\Gamma(\delta) \Gamma(\nu)} y^{\delta-1}(1-y)^{\nu-1} dy = \]
\[ \frac{\Gamma(\delta + \nu) \Gamma(\delta + \rho) \Gamma(\nu)}{\Gamma(\delta) \Gamma(\nu) \Gamma(\delta + \rho + \nu)} \int_0^1 \frac{\Gamma(\delta + \rho + \nu)}{\Gamma(\delta + \rho + \nu)} y^{\delta + \rho - 1}(1-y)^{\nu-1} dy = \]
\[ \frac{\Gamma(\delta + \rho) \Gamma(\delta + \rho + \nu)}{\Gamma(\delta) \Gamma(\delta + \rho + \nu)} \]
for \( r > -\delta \) since \( 1 = \int_0^1 \text{beta}(\delta + r, \nu) \) pdf.

1.25. \[ E(e^{tY}) = \sum_{y=1}^\infty e^{ty} \frac{-1}{\log(1-\theta t)} y \exp[\log(\theta)y]. \]
But \( e^{ty} \exp[\log(\theta)y] = \exp[(\log(\theta) + t)y] = \exp[(\log(\theta) + \log(e^t))y] = \exp[\log(\theta e^t)y]. \)
So \( E(e^{tY}) = \frac{-1}{\log(1-\theta e^t)} \sum_{y=1}^\infty \frac{1}{\log(1-\theta e^t)} y \exp[\log(\theta e^t)y] = \sum_{y=1}^\infty \text{logarithmic}(\theta e^t) \) pmf if \( 0 < \theta e^t < 1 \) or \( 0 < e^t < 1/\theta \) or \( -\infty < t < -\log(\theta) \).

1.28. a) \( EX = 0.9EZ + 0.1EW = 0.9\nu \lambda + 0.1(10) = 0.9(3)(4) + 1 = 11.8. \)
b) \( EX^2 = 0.9[V(Z) + (E(Z))^2] + 0.1[V(W) + (E(W))^2] \)
\[ = 0.9[\nu \lambda^2 + (\nu \lambda)^2] + 0.1[10(10)] \]
\[ = 0.9[3(16) + 9(16)] + 0.1(110) = 0.9(192) + 11 = 183.8. \]

2.8. a) \( F_W(w) = P(W \leq w) = P(Y \leq w - \mu) = F_Y(w - \mu). \) So \( f_W(w) = \frac{d}{dw} F_Y(w - \mu) = f_Y(w - \mu). \)
b) \( F_W(w) = P(W \leq w) = P(Y \leq w/\sigma) = F_Y(w/\sigma). \) So \( f_W(w) = \frac{d}{dw} F_Y(w/\sigma) = f_Y(w/\sigma) \frac{1}{\sigma}. \)
c) \( F_W(w) = P(W \leq w) = P(\sigma Y \leq w - \mu) = F_Y(\frac{w - \mu}{\sigma}). \) So \( f_W(w) = \frac{d}{dw} F_Y(\frac{w - \mu}{\sigma}) = f_Y(\frac{w - \mu}{\sigma}) \frac{1}{\sigma}. \)

2.9. a) See Example 2.16.
2.11. \( W = Z^2 \sim \chi^2 \) where \( Z \sim N(0, 1) \). So the pdf of \( W \) is
\[
f(w) = \frac{w^{\frac{1}{2} - 1} e^{-\frac{w}{2}}}{2^{\frac{1}{2}} \Gamma(\frac{1}{2})} = \frac{1}{\sqrt{w} \sqrt{2\pi}} e^{-\frac{w}{2}}
\]
for \( w > 0 \).

2.12. \( (Y - \mu)/\sigma = |Z| \sim HN(0, 1) \) where \( Z \sim N(0, 1) \). So \((Y - \mu)^2 = \sigma^2 Z^2 \sim \sigma^2 \chi^2_1 \sim G(0.5, 2\sigma^2)\).

2.16. a) \( y = e^{-w} = t^{-1}(w) \), and
\[
\left| \frac{dt^{-1}(w)}{dw} \right| = |-e^{-w}| = e^{-w}.
\]
Now \( P(Y = 0) = 0 \) so \( 0 < Y \leq 1 \) implies that \( W = -\log(Y) > 0 \). Hence
\[
f_W(w) = f_Y(t^{-1}(w)) \left| \frac{dt^{-1}(w)}{dw} \right| = \frac{1}{\lambda} (e^{-w})^{\frac{1}{\lambda} - 1} e^{-w} = \frac{1}{\lambda} e^{-w/\lambda}
\]
for \( w > 0 \) which is the \( \text{EXP}(\lambda) \) pdf.

2.18. a) \[
f(y) = \frac{\phi y^{\phi - 1}}{\lambda (1 + y^\phi)^{\frac{1}{\lambda} + 1}}
\]
where \( y, \phi, \) and \( \lambda \) are all positive. Since \( Y > 0 \), \( W = \log(1 + Y^\phi) > \log(1) > 0 \) and the support \( W = (0, \infty) \). Now \( 1 + y^\phi = e^w \), so \( y = (e^w - 1)^{1/\phi} = t^{-1}(w) \). Hence
\[
\left| \frac{dt^{-1}(w)}{dw} \right| = \frac{1}{\phi} (e^w - 1)^{\frac{1}{\phi} - 1} e^w
\]
since \( w > 0 \). Thus
\[
f_W(w) = f_Y(t^{-1}(w)) \left| \frac{dt^{-1}(w)}{dw} \right| = \frac{1}{\lambda} \frac{\phi (e^w - 1)^{\frac{\phi - 1}{\phi}}}{(1 + (e^w - 1)^{\phi})^{\frac{1}{\lambda} + 1}} \frac{1}{\phi} (e^w - 1)^{\frac{1}{\phi} - 1} e^w
\]
\[
= \frac{1}{\lambda} \frac{(e^w - 1)^{1 - \frac{1}{\phi}}}{(e^w)^{\frac{1}{\phi} + 1}} \frac{1}{e^{w/\lambda}}
\]
for \( w > 0 \) which is the EXP(\( \lambda \)) pdf.

\[ f(y) = \frac{1}{\pi \sigma \left[ 1 + \left( \frac{y-\mu}{\sigma} \right)^2 \right]} \]

where \( y \) and \( \mu \) are real numbers and \( \sigma > 0 \). Now \( y = \log(w) = t^{-1}(w) \) and \( W = e^Y > 0 \) so the support \( \mathcal{W} = (0, \infty) \). Thus

\[
\left| \frac{dt^{-1}(w)}{dw} \right| = \frac{1}{w},
\]

and

\[
f_W(w) = f_Y(t^{-1}(w)) \left| \frac{dt^{-1}(w)}{dw} \right| = \frac{1}{\pi \sigma \left[ 1 + \left( \frac{\log(w)-\mu}{\sigma} \right)^2 \right]} \frac{1}{w} = \frac{1}{\pi \sigma w \left[ 1 + \left( \frac{\log(w)-\mu}{\sigma} \right)^2 \right]} \]

for \( w > 0 \) which is the \( \text{LC}(\mu, \sigma) \) pdf.

\[ 2.25. \text{ b)} \]

\[ \text{EX} = E[E[X|Y]] = E[\beta_0 + \beta_1 Y] = \beta_0 + 3\beta_1. \]

\[ E[V(X)] = E[V(X|Y)] + V[E(X|Y)] = E(Y^2) + V(\beta_0 + \beta_1 Y) = V(Y) + [E(Y)]^2 + \beta_1^2 V(Y) = 10 + 9 + \beta_1^2 10 = 19 + 10\beta_1^2. \]

\[ 2.63. \text{ a)} \] \( X \sim N(100, 6) \).

\[ b) \frac{1}{N_2 \left( \begin{pmatrix} 49 \\ 17 \end{pmatrix}, \begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \right)} \]

\[ c) X_1 \perp X_4 \text{ and } X_3 \perp X_4. \]

\[ d) \rho(X_1, X_2) = \frac{\text{Cov}(X_1, X_3)}{\sqrt{\text{VAR}(X_1) \text{VAR}(X_3)}} = \frac{-1}{\sqrt{3} \sqrt{4}} = -0.2887. \]

\[ 2.65. \text{ a)} Y|X \sim N(49, 16) \text{ since } Y \perp X. \](Or use \( E(Y|X) = \mu_Y + \sum_{12} \Sigma_{22}^{-1} (X - \mu_x) = 49 + 0(1/25)(X - 100) = 49 \) and \( \text{VAR}(Y|X) = \Sigma_{11} - \sum_{12} \Sigma_{22}^{-1} \Sigma_{21} = 16 - 0(1/25)0 = 16. \))

\[ b) E(Y|X) = \mu_Y + \sum_{12} \Sigma_{22}^{-1} (X - \mu_x) = 49 + 10(1/25)(X - 100) = 9 + 0.4X. \]
c) \( \text{VAR}(Y|X) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 16 - 10(1/25)10 = 16 - 4 = 12. \)

2.68. Note that the pdf for \( \lambda \) is the \text{EXP}(1) pdf, so \( \lambda \sim \text{EXP}(1) \).

a) \( E(Y) = E[E(Y|\lambda)] = E(\lambda) = 1. \)
b) \( V(Y) = E[V(Y|\lambda)] + V[E(Y|\lambda)] = E(\lambda) + V(\lambda) = 1 + 1^2 = 2. \)

2.71. \[
f_Y(y) = \frac{y}{2\pi} \exp\left[\frac{-1}{2}(y^2 + \frac{1}{\delta^2})\right]
\]

So \( m(t) = \sum_y e^{ty} f(y) = \sum_y e^{ty} P(Y = y) = e^{t0}0.76 + e^{t1}0.24 \)
\( = 0.76 + 0.24e^t. \)

2.72. No, \( f(x,y) \neq f_X(x)f_Y(y) = \frac{1}{2\pi} \exp\left[\frac{-1}{2}(x^2 + y^2)\right]. \)

2.73. a) \( E(Y) = E[E(Y|P)] = E(kP) = kE(P) = k\frac{\delta}{\delta + \nu} = k4/10 = 0.4k. \)
b) \( V(Y) = E[V(Y|P)] + V[E(Y|P)] = E[kP(1 - P)] + V(kP) \)
\( = kE(P) - kE(P^2) + k^2V(P) = \)
\[
k\frac{\delta}{\delta + \nu} - k \left[ \frac{\delta\nu}{(\delta + \nu)^2(\delta + \nu + 1)} + \left(\frac{\delta}{\delta + \nu}\right)^2 \right] + k^2\frac{\delta\nu}{(\delta + \nu)^2(\delta + \nu + 1)}
\]
\( = k0.4 - k[0.021818 + 0.16] + k^20.021818 = 0.021818k + 0.21818k. \)

2.74. a) \[
f_Y(y) = \frac{0.55}{0.55} \quad \frac{1}{0.16} \quad \frac{2}{0.29}
\]
b) \( f(y_1|2) = f(y_1, 2)/f_{Y2}(2) \) and \( f(0, 2)/f_{Y2}(2) = .24/.29 \) while \( f(1, 2)/f_{Y2}(2) = .05/.29 \)

<table>
<thead>
<tr>
<th>( y_1 )</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
</table>
| \( f_{Y1|Y2}(y_1|y_2 = 2) \) | 24/29 \( \approx 0.8276 \) | 5/29 \( \approx 0.1724 \)

3.1. a) See Section 10.3.
b) See Section 10.13.
c) See Section 10.42.
d) See Example 3.5.

d) See Section 10.35.

e) See Section 10.39.
3.3. b) See Section 10.19.
c) See Section 10.30.
d) See Section 10.38.
f) See Section 10.43.
g) See Section 10.49.
h) See Section 10.53.

3.4. a) See Section 10.39.
b) See Section 10.39.
c) See Section 10.16.

3.5. a) See Section 10.7.
b) See Section 10.12.
c) See Section 10.14.
d) See Section 10.29.
h) See Section 10.41.
i) See Section 10.45.
j) See Section 10.51.

3.10. Yes, the top version of the pdf multiplied on the left by $I(y > 0)$ is in the form $h(y)c(\nu)\exp[w(\nu)t(y)]$ where $t(Y)$ is given in the problem, $c(\nu) = 1/\nu$ and $w(\nu) = -1/(2\nu^2)$. Hence $\Omega = (-\infty, 0)$.

4.3. See the proof of Theorem 4.5b.


4.6. The appropriate section in Chapter 10 gives the 1P-REF parameterization. Then the complete minimal sufficient statistic is $T_n$ given by Theorem 3.6.

4.7. The appropriate section in Chapter 10 gives the 1P-REF parameterization. Then the complete minimal sufficient statistic is $T_n$ given by Theorem 3.7.

4.8. The appropriate section in Chapter 10 gives the 2P-REF parameterization. Then the complete minimal sufficient statistic is $T = (T_1(Y), T_2(Y))$ where $T_i(Y) = \sum_{i=1}^{n} t_i(Y_i)$ by Corollary 4.6.

4.10. b) The 2P-REF parameterization is given in Chapter 10, so $(\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} e^{Y_i})$ is the complete minimal sufficient statistic by Corollary 4.6.
4.26. See Example 4.11.

4.30. Method i): \( E_\lambda(\overline{X} - S^2) = \lambda - \lambda = 0 \) for all \( \lambda > 0 \), but \( P_\lambda(\overline{X} - S^2 = 0) < 1 \) so \( \mathbf{T} = (\overline{X}, S^2) \) is not complete.

Method ii): The Poisson distribution is a 1P–REF with complete sufficient statistic \( \sum X_i \), so \( \overline{X} \) is a minimal sufficient statistic. \( \mathbf{T} = (\overline{X}, S^2) \) is not a function of \( \overline{X} \), so \( \mathbf{T} \) is not minimal sufficient and hence not complete.

4.31.

\[
f(x) = \frac{\Gamma(2\theta)}{\Gamma(\theta)\Gamma(\theta)} x^{\theta-1}(1 - x)^{\theta-1} = \frac{\Gamma(2\theta)}{\Gamma(\theta)\Gamma(\theta)} \exp[(\theta - 1)(\log(x) + \log(1 - x))]
\]

for \( 0 < x < 1 \), a 1 parameter exponential family. Hence \( \sum_{i=1}^n \log(X_i) + \log(1 - X_i) \) is a complete minimal sufficient statistic.

4.32. a) and b)

\[
f(x) = I_{\{1, 2, \ldots\}}(x) \frac{1}{\zeta(\nu)} \exp[-\nu \log(x)]
\]
is a 1 parameter regular exponential family with \( \Omega = (-\infty, -1) \). Hence \( \sum_{i=1}^n \log(X_i) \) is a complete minimal sufficient statistic.

c) By the Factorization Theorem, \( \mathbf{W} = (X_1, \ldots, X_n) \) is sufficient, but \( \mathbf{W} \) is not minimal since \( \mathbf{W} \) is not a function of \( \sum_{i=1}^n \log(X_i) \).

4.33. \( f(x_1, \ldots, x_n) = \prod_{i=1}^n f(x_i) = \]

\[
\left[ \frac{2}{\sqrt{2\pi}} \right]^n \frac{1}{\sigma^n} \exp\left( \frac{-n\mu^2}{2\sigma^2} \right) I[x(1) > \mu] \exp\left( \frac{-1}{2\sigma^2} \sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n x_i \right)
\]

\[
= h(x)g(T(x)|\theta)
\]

where \( \theta = (\mu, \sigma) \). Hence \( \mathbf{T}(\mathbf{X}) = (X_1, \sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i) \) is a sufficient statistic for \( (\mu, \sigma) \).

4.34. Following the end of Example 4.4, \( X_{(1)} \sim \text{EXP}(\lambda/n) \) with \( \lambda = 1 \), so \( E[X_{(1)}] = 1/n \).

4.35. \( F_{X_{(n)}}(x) = [F(x)]^n = x^n \) for \( 0 < x < 1 \). Thus \( f_{X_{(n)}}(x) = nx^{n-1} \) for \( 0 < x < 1 \). This pdf is the beta(\( \delta = n, \nu = 1 \)) pdf. Hence

\[
E[X_{(n)}] = \int_0^1 xf_{X_{(n)}}(x)dx = \int_0^1 xn x^{n-1}dx = \int_0^1 n x^n dx = n \frac{x^{n+1}}{n+1} \bigg|_0^1 = \frac{n}{n+1}.
\]
4.36. Now
\[ f_X(x) = I(\theta < x < \theta + 1) \]
and
\[ \frac{f(x)}{f(y)} = \frac{I(\theta < x_{(1)} \leq x_{(n)} < \theta + 1)}{I(\theta < y_{(1)} \leq y_{(n)} < \theta + 1)} \]
which is constant for all real \( \theta \) iff \( (x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)}) \). Hence \( T = (X_{(1)}, X_{(n)}) \) is a minimal sufficient statistic by the LSM theorem. To show that \( T \) is not complete, first find \( E(T) \). Now
\[ F_X(t) = \int_t^\theta dx = t - \theta \]
for \( \theta < t < \theta + 1 \). Hence
\[ f_{X_{(n)}}(t) = n[F_X(t)]^{n-1}f_x(t) = n(t - \theta)^{n-1} \]
for \( \theta < t < \theta + 1 \) and
\[ E_\theta(X_{(n)}) = \int t f_{X_{(n)}}(t)dt = \int_\theta^{\theta+1} tn(t - \theta)^{n-1}dt. \]
Use u-substitution with \( u = t - \theta, t = u + \theta \) and \( dt = du \). Hence \( t = \theta \) implies \( u = 0 \), and \( t = \theta + 1 \) implies \( u = 1 \). Thus
\[ E_\theta(X_{(n)}) = \int_0^1 n(u + \theta)u^{n-1}du = \int_0^1 n^1 u^1 du + \int_0^1 n\theta u^{n-1}du =
\]
\[ \left. \frac{n}{n+1} u^{n+1} \right|_0^1 + \theta \left. \frac{n}{n} u^n \right|_0^1 = \frac{n}{n+1} + \frac{n\theta}{n} = \theta + \frac{n}{n+1}. \]
Now
\[ f_{X_{(1)}}(t) = n[1 - F_X(t)]^{n-1}f_x(t) = n(1 - t + \theta)^{n-1} \]
for \( \theta < t < \theta + 1 \) and thus
\[ E_\theta(X_{(1)}) = \int_\theta^{\theta+1} tn(1 - t + \theta)^{n-1}dt. \]
Use u-substitution with \( u = (1 - t + \theta) \) and \( t = 1 - u + \theta \) and \( du = -dt \). Hence \( t = \theta \) implies \( u = 1 \), and \( t = \theta + 1 \) implies \( u = 0 \). Thus
\[ E_\theta(X_{(1)}) = -\int_1^0 n(1 - u + \theta)u^{n-1}du = n(1 + \theta) \int_0^1 u^{n-1}du - n \int_0^1 u^n du =
\]
\[ n(1 + \theta) \frac{u^n}{n} \bigg|_0^1 - n \frac{u^{n+1}}{n+1} \bigg|_0^1 = (\theta + 1) \frac{n}{n} - \frac{n}{n+1} = \theta + \frac{1}{n+1}. \]

To show that \( T \) is not complete try showing \( E_\theta(aX_{(1)} + bX_{(n)} + c) = 0 \) for some constants \( a, b \) and \( c \). Note that \( a = -1, b = 1 \) and \( c = -\frac{n-1}{n+1} \) works. Hence
\[ E_\theta(-X_{(1)} + X_{(n)} - \frac{n-1}{n+1}) = 0 \]
for all real \( \theta \) but
\[ P_\theta(g(T) = 0) = P_\theta(-X_{(1)} + X_{(n)} - \frac{n-1}{n+1} = 0) = 0 < 1 \]
for all real \( \theta \). Hence \( T \) is not complete.

4.37. Note that
\[ f(y) = I(y > 0) \ 2y \ e^{-y^2} \tau \ \exp[(1 - \tau)(-\log(1 - e^{-y^2}))] \]
is a 1 parameter exponential family with minimal and complete sufficient statistic \(-\sum_{i=1}^{n} \log(1 - e^{-y_i^2})\).

5.2. The likelihood function \( L(\theta) = \)
\[ \frac{1}{(2\pi)^n} \exp\left(\frac{-1}{2} \left[ \sum (x_i - \rho \cos \theta)^2 + \sum (y_i - \rho \sin \theta)^2 \right]\right) = \]
\[ \frac{1}{(2\pi)^n} \exp\left(\frac{-1}{2} \left[ \sum x_i^2 - 2\rho \cos \theta \sum x_i + \rho^2 \cos^2 \theta + \sum y_i^2 - 2\rho \sin \theta \sum y_i + \rho^2 \sin^2 \theta \right]\right) \]
\[ = \frac{1}{(2\pi)^n} \exp\left(\frac{-1}{2} \left[ \sum x_i^2 + \sum y_i^2 + \rho^2 \right]\right) \exp(\rho \cos \theta \sum x_i + \rho \sin \theta \sum y_i). \]

Hence the log likelihood log \( L(\theta) \)
\[ = c + \rho \cos \theta \sum x_i + \rho \sin \theta \sum y_i. \]
The derivative with respect to \( \theta \) is
\[ -\rho \sin \theta \sum x_i + \rho \cos \theta \sum y_i. \]
Setting this derivative to zero gives
\[ \rho \sum y_i \cos \theta = \rho \sum x_i \sin \theta \]
or
\[ \frac{\sum y_i}{\sum x_i} = \tan \theta. \]
Thus
\[ \hat{\theta} = \tan^{-1}\left( \frac{\sum y_i}{\sum x_i} \right). \]

Now the boundary points are \( \theta = 0 \) and \( \theta = 2\pi \). Hence \( \hat{\theta}_{MLE} \) equals 0, 2\( \pi \), or \( \hat{\theta} \) depending on which value maximizes the likelihood.

5.6. See Section 10.7.
5.7. See Section 10.9.
5.8. See Section 10.12.
5.9. See Section 10.13.
5.10. See Section 10.16.
5.11. See Section 10.19.
5.15. See Section 10.38.
5.16. See Section 10.45.
5.17. See Section 10.51.
5.18. See Section 10.3.
5.20. See Section 10.49.

5.23. a) The log likelihood is
\[ \log L(\tau) = -\frac{n}{2} \log(2\pi\tau) - \frac{1}{2\tau} \sum_{i=1}^{n} (X_i - \mu)^2. \]

The derivative of the log likelihood is equal to
\[ -\frac{n}{2\tau^2} + \frac{1}{2\tau^2} \sum_{i=1}^{n} (X_i - \mu)^2. \]
Setting the derivative equal to 0 and solving for \( \tau \) gives the MLE \( \hat{\tau} = \frac{\sum_{i=1}^{n} (X_i - \mu)^2}{n} \).

Now the likelihood is only defined for \( \tau > 0 \). As \( \tau \) goes to 0 or \( \infty \), \( \log L(\tau) \) tends to \( -\infty \). Since there is only one critical point, \( \hat{\tau} \) is the MLE.
b) By the invariance principle, the MLE is $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n}}$.

5.28. This problem is nearly the same as finding the MLE of $\sigma^2$ when the data are iid $N(\mu, \sigma^2)$ when $\mu$ is known. See Problem 5.23 and Section 10.23. The MLE in a) is $\sum_{i=1}^{n}(X_i - \mu)^2/n$. For b) use the invariance principle and take the square root of the answer in a).

5.29. See Example 5.5.

5.30.

$$L(\theta) = \frac{1}{\theta\sqrt{2\pi}}e^{-(x-\theta)^2/2\theta^2}$$

$$\ln(L(\theta)) = -\ln(\theta) - \ln(\sqrt{2\pi}) - (x - \theta)^2/2\theta^2$$

$$\frac{d\ln(L(\theta))}{d\theta} = \frac{-1}{\theta} + \frac{x - \theta}{\theta^2} + \frac{(x - \theta)^2}{\theta^3}$$

$$= \frac{x^2}{\theta^3} - \frac{x}{\theta^2} - \frac{1}{\theta} \text{ set } = 0$$

by solving for $\theta$,

$$\theta = \frac{x}{2} \ast (-1 + \sqrt{5}),$$

and

$$\theta = \frac{x}{2} \ast (-1 - \sqrt{5}).$$

But, $\theta > 0$. Thus, $\hat{\theta} = \frac{x}{2} \ast (-1 + \sqrt{5})$, when $x > 0$, and $\hat{\theta} = \frac{x}{2} \ast (-1 - \sqrt{5})$, when $x < 0$.

To check with the second derivative

$$\frac{d^2\ln(L(\theta))}{d\theta^2} = \frac{-2\theta + x}{\theta^3} + \frac{3(\theta^2 + \theta x - x^2)}{\theta^4}$$

$$= \frac{\theta^2 + 2\theta x - 3x^2}{\theta^4}$$

but the sign of the $\theta^4$ is always positive, thus the sign of the second derivative depends on the sign of the numerator. Substitute $\hat{\theta}$ in the numerator and simplify, you get $\frac{x}{2}(-5 \pm \sqrt{5})$, which is always negative. Hence by the invariance principle, the MLE of $\theta^2$ is $\hat{\theta}^2$. 
5.31. a) For any $\lambda > 0$, the likelihood function

$$L(\sigma, \lambda) = \sigma^{n/\lambda} I[x(1) \geq \sigma] \frac{1}{\lambda^n} \exp \left[ -(1 + \frac{1}{\lambda}) \sum_{i=1}^{n} \log(x_i) \right]$$

is maximized by making $\sigma$ as large as possible. Hence $\hat{\sigma} = X_{(1)}$.

b) $L(\hat{\sigma}, \lambda) = \hat{\sigma}^{n/\lambda} I[x(1) \geq \hat{\sigma}] \frac{1}{\lambda^n} \exp \left[ -(1 + \frac{1}{\lambda}) \sum_{i=1}^{n} \log(x_i) \right]$.

Hence $\log L(\hat{\sigma}, \lambda) = \frac{n}{\lambda} \log(\hat{\sigma}) - n \log(\lambda) - (1 + \frac{1}{\lambda}) \sum_{i=1}^{n} \log(x_i)$.

Thus

$$\frac{d}{d\lambda} \log L(\hat{\sigma}, \lambda) = -\frac{n}{\lambda^2} \log(\hat{\sigma}) - \frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} \log(x_i) \equiv 0,$$

or $-n \log(\hat{\sigma}) + \sum_{i=1}^{n} \log(x_i) = n\lambda$. So

$$\hat{\lambda} = -\log(\hat{\sigma}) + \frac{\sum_{i=1}^{n} \log(x_i)}{n} = \frac{\sum_{i=1}^{n} \log(x_i/\hat{\sigma})}{n}.$$

Now

$$\frac{d^2}{d\lambda^2} \log L(\hat{\sigma}, \lambda) = \frac{2n}{\lambda^3} \log(\hat{\sigma}) + \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} \log(x_i) \bigg|_{\lambda = \hat{\lambda}}$$

$$= \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} \log(x_i/\hat{\sigma}) = \frac{-n}{\lambda^2} < 0.$$

Hence $(\hat{\sigma}, \hat{\lambda})$ is the MLE of $(\sigma, \lambda)$.

5.32. a) the likelihood

$$L(\lambda) = \frac{1}{\lambda^n} \exp \left[ -(1 + \frac{1}{\lambda}) \sum \log(x_i) \right].$$
and the log likelihood
\[
\log(L(\lambda)) = -n \log(\lambda) - (1 + \frac{1}{\lambda}) \sum \log(x_i).
\]
Hence
\[
\frac{d}{d\lambda} \log(L(\lambda)) = \frac{-n}{\lambda} + \frac{1}{\lambda^2} \sum \log(x_i) \stackrel{set}{=} 0,
\]
or \(\sum \log(x_i) = n\lambda\) or
\[
\hat{\lambda} = \frac{\sum \log(X_i)}{n}.
\]

Notice that
\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{2 \sum \log(x_i)}{\lambda^3} \bigg|_{\lambda=\hat{\lambda}} = \frac{n}{\lambda^2} - \frac{2 \hat{\lambda}}{\lambda^3} = \frac{-n}{\lambda^2} < 0.
\]
Hence \(\hat{\lambda}\) is the MLE of \(\lambda\).

b) By invariance, \(\hat{\lambda}_8\) is the MLE of \(\lambda_8\).

5.33. a) The likelihood
\[
L(\theta) = c e^{-n^2\theta} \exp[\log(2\theta) \sum x_i],
\]
and the log likelihood
\[
\log(L(\theta)) = d - n^2\theta + \log(2\theta) \sum x_i.
\]
Hence
\[
\frac{d}{d\theta} \log(L(\theta)) = -2n + 2 \frac{\sum x_i}{2\theta} \stackrel{set}{=} 0,
\]
or \(\sum x_i = 2n\theta\), or
\[
\hat{\theta} = \frac{X}{2}.
\]

Notice that
\[
\frac{d^2}{d\theta^2} \log(L(\theta)) = -\frac{\sum x_i}{\theta^2} < 0
\]
unless \(\sum x_i = 0\).

b) \((\hat{\theta})^4 = (\frac{X}{2})^4\) by invariance.
5.34. \( L(0|x) = 1 \) for \( 0 < x_i < 1 \), and \( L(1|x) = \prod_{i=1}^{n} \frac{1}{2\sqrt{x_i}} \) for \( 0 < x_i < 1 \). Thus the MLE is 0 if \( 1 \geq \prod_{i=1}^{n} \frac{1}{2\sqrt{x_i}} \) and the MLE is 1 if \( 1 < \prod_{i=1}^{n} \frac{1}{2\sqrt{x_i}} \).

5.35. a) Notice that \( \theta > 0 \) and 
\[
f(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\theta}} \exp \left( -\frac{(y - \theta)^2}{2\theta} \right).
\]
Hence the likelihood 
\[
L(\theta) = c \frac{1}{\theta^{n/2}} \exp \left[ -\frac{1}{2\theta} \sum (y_i - \theta)^2 \right]
\]
and the log likelihood 
\[
\log(L(\theta)) = d - \frac{n}{2} \log(\theta) - \frac{1}{2\theta} \sum (y_i - \theta)^2 =
\[
d - \frac{n}{2} \log(\theta) - \frac{1}{2} \sum_{i=1}^{n} \left( \frac{y_i^2}{\theta} - \frac{2y_i\theta}{\theta} + \frac{\theta^2}{\theta} \right)
\[
= d - \frac{n}{2} \log(\theta) - \frac{1}{2} \sum_{i=1}^{n} \frac{y_i^2}{\theta} + \sum_{i=1}^{n} y_i - \frac{1}{2} n\theta.
\]
Thus
\[
\frac{d}{d\theta} \log(L(\theta)) = -\frac{n}{2} \theta^{-1} + \frac{1}{2} \sum_{i=1}^{n} y_i^2 \theta^{-2} - \frac{n}{2} \frac{set}{2} = 0,
\]
or
\[
-\frac{n}{2} \theta^2 - \frac{n}{2} \theta + \frac{1}{2} \sum_{i=1}^{n} y_i^2 = 0,
\]
or
\[
\frac{n\theta^2 + n\theta - \sum_{i=1}^{n} y_i^2}{n} = 0. \tag{11.2}
\]

Now the quadratic formula states that for \( a \neq 0 \), the quadratic equation \( ay^2 + by + c = 0 \) has roots 
\[
\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.
\]
CHAPTER 11. STUFF FOR STUDENTS

Applying the quadratic formula to (11.3) gives

$$\theta = \frac{-n \pm \sqrt{n^2 + 4n \sum_{i=1}^{n} y_i^2}}{2n}.$$ 

Since $\theta > 0$, a candidate for the MLE is

$$\hat{\theta} = \frac{-n + \sqrt{n^2 + 4n \sum_{i=1}^{n} Y_i^2}}{2n} = \frac{-1 + \sqrt{1 + 4\left(\frac{1}{n} \sum_{i=1}^{n} Y_i^2\right)}}{2}.$$ 

Since $\hat{\theta}$ satisfies (11.3),

$$n\hat{\theta} - \sum_{i=1}^{n} y_i^2 = -n\hat{\theta}^2. \quad (11.3)$$

Note that

$$\frac{d^2}{d\theta^2} \log(L(\theta)) = \frac{n}{2\theta^2} - \frac{\sum_{i=1}^{n} y_i^2}{\theta^3} = \left. \frac{1}{2\theta^3}[n\theta - 2\sum_{i=1}^{n} y_i^2] \right|_{\theta=\hat{\theta}} =$$

$$\frac{1}{2\hat{\theta}^3}[n\hat{\theta} - \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} y_i^2] = \frac{1}{2\hat{\theta}^3}[-n\hat{\theta}^2 - \sum_{i=1}^{n} y_i^2] < 0$$

by (11.4). Since $L(\theta)$ is continuous with a unique root on $\theta > 0$, $\hat{\theta}$ is the MLE.

5.36. a) $L(\theta) = (\theta - x)^2/3$ for $x - 2 \leq \theta \leq x + 1$. Since $x = 7$, $L(5) = 4/3$, $L(7) = 0$, and $L(8) = 1/3$. So $L$ is maximized at an endpoint and the MLE $\hat{\theta} = 5$.

b) By invariance the MLE is $h(\hat{\theta}) = h(5) = 10 - e^{-25} \approx 10$.

5.37. a) $L(\lambda) = c \frac{1}{\lambda^n} \exp\left(\frac{1}{2\lambda^2} \sum_{i=1}^{n} (e^{x_i} - 1)^2\right)$.

Thus

$$\log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{2\lambda^2} \sum_{i=1}^{n} (e^{x_i} - 1)^2.$$ 

Hence

$$\frac{d\log(L(\lambda))}{d\lambda} = -\frac{n}{\lambda} + \frac{1}{\lambda^3} \sum_{i=1}^{n} (e^{x_i} - 1)^2 \overset{set}{=} 0,$$

or $n\lambda^2 = \sum (e^{x_i} - 1)^2$, or

$$\hat{\lambda} = \sqrt{\frac{\sum (e^{X_i} - 1)^2}{n}}.$$
Now
\[
\frac{d^2 \log(L(\lambda))}{d\lambda^2} = \frac{n}{\lambda^2} - \frac{3}{\lambda^4} \sum (e^{x_i} - 1)^2 \bigg|_{\lambda=\hat{\lambda}}
\]
\[
= \frac{n}{\lambda^2} - \frac{3n}{\lambda^4} \hat{\lambda}^2 = \frac{n}{\lambda^2} [1 - 3] < 0.
\]
So \(\hat{\lambda}\) is the MLE.

5.38. a) The likelihood
\[
L(\lambda) = \prod f(x_i) = c \left( \prod \frac{1}{x_i} \right) \frac{1}{\lambda^n} \exp \left[ \frac{\sum - (\log x_i)^2}{2\lambda^2} \right],
\]
and the log likelihood
\[
\log(L(\lambda)) = d - \sum \log(x_i) - n \log(\lambda) - \frac{\sum (\log x_i)^2}{2\lambda^2}.
\]
Hence
\[
\frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} + \frac{\sum (\log x_i)^2}{\lambda^3} \text{ set } = 0,
\]
or \(\sum (\log x_i)^2 = n\lambda^2\), or
\[
\hat{\lambda} = \sqrt{\frac{\sum (\log x_i)^2}{n}}.
\]
This solution is unique.

Notice that
\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \frac{3}{\lambda^4} \sum (\log x_i)^2 \bigg|_{\lambda=\hat{\lambda}}
\]
\[
= \frac{n}{\lambda^2} - \frac{3n}{\lambda^4} \hat{\lambda}^2 = \frac{-2n}{\lambda^2} < 0.
\]
Hence
\[
\hat{\lambda} = \sqrt{\frac{\sum (\log x_i)^2}{n}}
\]
is the MLE of \(\lambda\).

b)
\[
\hat{\lambda}^2 = \frac{\sum (\log x_i)^2}{n}
\]
is the MLE of $\lambda^2$ by invariance.

5.39. a) The plot of $L(\theta)$ should be 1 for $x_{(n)} - 1 \leq \theta \leq x_{(1)}$, and 0 otherwise.

b) $[c, d] = [x_{(n)} - 1, x_{(1)}].$

5.42. See Example 5.6.

5.43. a) $L(\theta) = \frac{[\cos(\theta)]^n \exp(\theta \sum x_i)}{\prod \cosh(\pi x_i/2)}.$ So $\log(L(\theta)) = c + n \log(\cos(\theta)) + \theta \sum x_i,$ and

$$\frac{d \log(L(\theta))}{d\theta} = n \frac{1}{\cos(\theta)} [-\sin(\theta)] + \sum x_i \stackrel{s} = 0,$$

or $\tan(\theta) = x$, or $\hat{\theta} = \tan^{-1}(X).$

Since

$$\frac{d^2 \log(L(\theta))}{d\theta^2} = -n \sec^2(\theta) < 0$$

for $|\theta| < 1/2$, $\hat{\theta}$ is the MLE.

b) The MLE is $\tan(\hat{\theta}) = \tan(\tan^{-1}(X)) = X$ by the invariance principle.

(By properties of the arctan function, $\hat{\theta} = \tan^{-1}(X)$ iff $\tan(\hat{\theta}) = X$ and $-\pi/2 < \hat{\theta} < \pi/2$.)

5.44. a) This is a two parameter exponential distribution. So see Section 10.14 where $\sigma = \lambda$ and $\mu = \theta$.

b) $1 - F(x) = \tau(\mu, \sigma) = \exp \left[-\left(\frac{x - \mu}{\sigma}\right)\right].$

By the invariance principle, the MLE of $\tau(\mu, \sigma) = \tau(\hat{\mu}, \hat{\sigma})$

$$= \exp \left[-\left(\frac{x - X_{(1)}}{\hat{\sigma}}\right)\right].$$

5.45. a) Let $w = t(y) = \frac{y}{\theta} + \frac{\theta}{y} - 2$.

Then the likelihood

$$L(\nu) = \frac{1}{\nu^n} \exp \left(\frac{-1}{2\nu^2} \sum_{i=1}^n w_i\right),$$
and the log likelihood

$$\log(L(\nu)) = c - n \log(\nu) - \frac{1}{2\nu^2} \sum_{i=1}^{n} w_i.$$ 

Hence

$$\frac{d}{d\nu} \log(L(\nu)) = -\frac{n}{\nu} + \frac{1}{\nu^3} \sum_{i=1}^{n} w_i \text{ set } = 0,$$

or

$$\hat{\nu} = \sqrt{\frac{\sum_{i=1}^{n} w_i}{n}}.$$

This solution is unique and

$$\frac{d^2}{d\nu^2} \log(L(\nu)) = \frac{n}{\nu^2} - \frac{3 \sum_{i=1}^{n} w_i}{\nu^4} \bigg|_{\nu=\hat{\nu}} = \frac{n}{\hat{\nu}^2} - \frac{3n\hat{\nu}^2}{\hat{\nu}^4} = -2n\hat{\nu}^2 < 0.$$ 

Thus

$$\hat{\nu} = \sqrt{\frac{\sum_{i=1}^{n} W_i}{n}}$$

is the MLE of \( \nu \) if \( \hat{\nu} > 0 \).

b) \( \hat{\nu}^2 = \frac{\sum_{i=1}^{n} W_i}{n} \) by invariance.

5.46. a) The likelihood

$$L(\lambda) = c \frac{1}{\lambda^n} \exp \left( -\frac{1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^{-\phi}) \right),$$

and the log likelihood \( \log(L(\lambda)) = d - n \log(\lambda) - \frac{1}{\lambda} \sum_{i=1}^{n} \log(1 + y_i^{-\phi}) \). Hence

$$\frac{d}{d\lambda} \log(L(\lambda)) = -\frac{n}{\lambda} + \frac{\sum_{i=1}^{n} \log(1 + y_i^{-\phi}) \text{ set}}{\lambda^2} = 0,$$

or \( \sum_{i=1}^{n} \log(1 + y_i^{-\phi}) = n\lambda \) or

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + y_i^{-\phi})}{n}.$$
This solution is unique and
\[
\frac{d^2}{d\lambda^2} \log(L(\lambda)) = \frac{n}{\lambda^2} - \left. \frac{2 \sum_{i=1}^{n} \log(1 + y_i^{-\phi})}{\lambda^3} \right|_{\lambda = \hat{\lambda}} = \frac{n}{\lambda^2} - \frac{2n\hat{\lambda}}{\lambda^3} = -\frac{n}{\lambda^2} < 0.
\]
Thus
\[
\hat{\lambda} = \frac{\sum_{i=1}^{n} \log(1 + Y_i^{-\phi})}{n}
\]
is the MLE of \(\lambda\) if \(\phi\) is known.

b) The MLE is \(\hat{\lambda}^2\) by invariance.

6.7. a) The joint density
\[
f(x) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum (x_i - \mu)^2\right]
= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \left( \sum x_i^2 - 2\mu \sum x_i + n\mu^2 \right) \right]
= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} \sum x_i^2 \right] \exp[n\mu x - n\mu^2/2].
\]
Hence by the factorization theorem \(\overline{X}\) is a sufficient statistic for \(\mu\).

b) \(\overline{X}\) is sufficient by a) and complete since the \(N(\mu, 1)\) family is a regular one parameter exponential family.

c) \(E(I_{-\infty,t}(X_1)|\overline{X} = \overline{x}) = P(X_1 \leq t|\overline{X} = \overline{x}) = \Phi(\frac{t-\overline{x}}{\sqrt{1-1/n}}).
\]
d) By the LSU theorem,
\[
\Phi\left(\frac{t - \overline{X}}{\sqrt{1 - 1/n}}\right)
\]
is the UMVUE.

6.14. Note that \(\sum X_i \sim G(n, \theta)\). Hence \(\text{MSE}(c) = \text{Var}(T_n(c)) + [E_0 T_n(c) - \theta]^2 = c^2 \text{Var}(\sum X_i) + [nc E_\theta X - \theta]^2 = c^2 n\theta^2 + [nc\theta - \theta]^2\).

So
\[
\frac{d}{dc} \text{MSE}(c) = 2cn\theta^2 + 2[nc\theta - \theta]n\theta.
\]
Set this equation to 0 to get \(2n\theta^2[c + nc - 1] = 0\) or \(c(n + 1) = 1\). So \(c = 1/(n + 1)\).
The second derivative is $2n\theta^2 + 2n^2\theta^2 > 0$ so the function is convex and the local min is in fact global.

6.17. a) Since this is an exponential family, $\log(f(x|\lambda)) = -\log(\lambda) - x/\lambda$ and
\[
\frac{\partial}{\partial \lambda} \log(f(x|\lambda)) = \frac{-1}{\lambda} + \frac{x}{\lambda^2}.
\]
Hence
\[
\frac{\partial^2}{\partial \lambda^2} \log(f(x|\lambda)) = \frac{1}{\lambda^2} - \frac{2x}{\lambda^3}
\]
and
\[
I_1(\lambda) = -E\left[\frac{\partial^2}{\partial \lambda^2} \log(f(x|\lambda))\right] = \frac{-1}{\lambda^2} + \frac{2\lambda}{\lambda^3} = \frac{1}{\lambda^2}.
\]

b) $FCRLB(\tau(\lambda)) = \frac{[\tau'(\lambda)]^2}{nI_1(\lambda)} = \frac{4\lambda^2}{n/\lambda^2} = 4\lambda^4/n.$

c) $(T = \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda))$ is a complete sufficient statistic. Now $E(T^2) = V(T) + [E(T)]^2 = n\lambda^2 + n^2\lambda^2$. Hence the UMVUE of $\lambda^2$ is $T^2/(n + n^2).$ No, $W$ is a nonlinear function of the complete sufficient statistic $T$.

6.19.
\[
W \equiv S^2(k)/\sigma^2 \sim \chi_n^2/k
\]
and
\[
MSE(S^2(k)) = MSE(W) = VAR(W) + (E(W) - \sigma^2)^2
\]
\[
= \frac{\sigma^4}{k^2}2n + \left(\frac{\sigma^2n}{k} - \sigma^2\right)^2
\]
\[
= \sigma^4\left[\frac{2n}{k^2} + \left(\frac{n}{k} - 1\right)^2\right] = \sigma^4\frac{2n + (n - k)^2}{k^2}.
\]

Now the derivative $\frac{d}{dk}MSE(S^2(k))/\sigma^4 = \frac{-2}{k^3}[2n + (n - k)k] + \frac{-2(n - k)}{k^2}.$

Set this derivative equal to zero. Then
\[
2k^2 - 2nk = 4n + 2(n - k)^2 = 4n + 2n^2 - 4nk + 2k^2.
\]

Hence
\[
2nk = 4n + 2n^2
\]
or $k = n + 2$.

Should also argue that $k = n + 2$ is the global minimizer. Certainly need $k > 0$ and the absolute bias will tend to $\infty$ as $k \to 0$ and the bias tends to $\sigma^2$ as $k \to \infty$, so $k = n + 2$ is the unique critical point and is the global minimizer.

6.20. a) Let $W = X^2$. Then $f(w) = f_X(\sqrt{w}) = 1/(2\sqrt{w}) = (1/\theta) \exp(-w/\theta)$ and $W \sim \exp(\theta)$. Hence $E_\theta(X^2) = E_\theta(W) = \theta$.

b) This is an exponential family and

$$\log(f(x|\theta)) = \log(2x) - \log(\theta) - \frac{1}{\theta}x^2$$

for $x > 0$. Hence

$$\frac{\partial}{\partial \theta} f(x|\theta) = -\frac{1}{\theta^2} + \frac{1}{\theta^2}x^2$$

and

$$\frac{\partial^2}{\partial \theta^2} f(x|\theta) = \frac{1}{\theta^2} - \frac{2}{\theta^3}x^2.$$ 

Hence

$$I_1(\theta) = -E_\theta[\frac{1}{\theta^2} + \frac{2}{\theta^3}x^2] = \frac{1}{\theta^2}$$

by a). Now

$$CRLB = \frac{[\tau'(\theta)]^2}{nI_1(\theta)} = \frac{\theta^2}{n}$$

where $\tau(\theta) = \theta$.

c) This is a regular exponential family so $\sum_{i=1}^{n} X_i^2$ is a complete sufficient statistic. Since

$$E_\theta[\frac{\sum_{i=1}^{n} X_i^2}{n}] = \theta,$$

the UMVUE is $\frac{\sum_{i=1}^{n} X_i^2}{n}$.

6.21. a) In normal samples, $\bar{X}$ and $S$ are independent, hence

$$Var_\theta[W(\alpha)] = \alpha^2 Var_\theta(T_1) + (1 - \alpha)^2 Var_\theta(T_2).$$
b) \( W(\alpha) \) is an unbiased estimator of \( \theta \). Hence \( \text{MSE}[W(\alpha)] \equiv \text{MSE}(\alpha) = \text{Var}_{\theta}[W(\alpha)] \) which is found in part a).

c) Now

\[
\frac{d}{d\alpha} \text{MSE}(\alpha) = 2\alpha \text{Var}_\theta(T_1) - 2(1 - \alpha) \text{Var}_\theta(T_2) = 0.
\]

Hence

\[
\hat{\alpha} = \frac{\text{Var}_\theta(T_2)}{\text{Var}_\theta(T_1) + \text{Var}_\theta(T_2)} \approx \frac{\frac{\theta^2}{2n}}{\frac{\theta^2}{2n} + \frac{\theta^2}{2n}} = 1/3
\]

using the approximation and the fact that \( \text{Var}(\bar{X}) = \theta^2/n \). Note that the second derivative

\[
\frac{d^2}{d\alpha^2} \text{MSE}(\alpha) = 2[\text{Var}_\theta(T_1) + \text{Var}_\theta(T_2)] > 0,
\]

so \( \alpha = 1/3 \) is a local min. The critical value was unique, hence \( 1/3 \) is the global min.

6.22. a) \( X_1 - X_2 \sim N(0, 2\sigma^2) \). Thus,

\[
E(T_1) = \int_0^\infty u \frac{1}{\sqrt{4\pi \sigma^2}} e^{-\frac{u^2}{4\sigma^2}} du = \frac{\sigma}{\sqrt{\pi}}.
\]

\[
E(T_1^2) = \frac{1}{2} \int_0^\infty u^2 \frac{1}{\sqrt{4\pi \sigma^2}} e^{-\frac{u^2}{4\sigma^2}} du = \frac{\sigma^2}{2}.
\]

\( V(T_1) = \sigma^2 \left( \frac{1}{2} - \frac{1}{\pi} \right) \) and

\[
\text{MSE}(T_1) = \sigma^2 \left( \frac{1}{\sqrt{\pi}} - 1 \right)^2 + \frac{1}{2} - \frac{1}{\pi} = \sigma^2 \left( \frac{3}{2} - \frac{2}{\sqrt{\pi}} \right).
\]

b) \( \frac{X_i}{\sigma} \) has a \( N(0,1) \) and \( \frac{\sum_{i=1}^n X_i^2}{\sigma^2} \) has a chi square distribution with \( n \) degrees of freedom. Thus

\[
E\left( \sqrt{\frac{\sum_{i=1}^n X_i^2}{\sigma^2}} \right) = \frac{\sqrt{2\Gamma\left( \frac{n+1}{2} \right)}}{\Gamma\left( \frac{n}{2} \right)},
\]
and

\[ E(T_2) = \frac{\sigma \sqrt{2 \Gamma\left(\frac{n+1}{2}\right)}}{\sqrt{n \Gamma\left(\frac{n}{2}\right)}}. \]

Therefore,

\[ E\left(\frac{\sqrt{n} \Gamma\left(\frac{n}{2}\right)}{\sqrt{2 \Gamma\left(\frac{n+1}{2}\right)}} T_2\right) = \sigma. \]

6.23. This is a regular one parameter exponential family with complete sufficient statistic \( T_n = \sum_{i=1}^{n} X_i \sim G(n, \lambda) \). Hence \( E(T_n) = n\lambda, E(T_n^2) = V(T_n) + (E(T_n))^2 = n\lambda^2 + n^2\lambda^2 \), and \( T_n^2/(n + n^2) \) is the UMVUE of \( \lambda^2 \).

6.24. \( \frac{1}{X_i} = \frac{W_i}{\sigma} \sim \frac{\chi_1^2}{\sigma} \).

Hence if

\[ T = \sum_{i=1}^{n} \frac{1}{X_i}, \text{ then } E\left(\frac{T}{n}\right) = \frac{n}{n\sigma}, \]

and \( T/n \) is the UMVUE since \( f(x) \) is an exponential family with complete sufficient statistic \( 1/X \).

6.25. The pdf of \( T \) is

\[ g(t) = \frac{2nt^{2n-1}}{\theta^{2n}} \]

for \( 0 < t < \theta \).

\( E(T) = \frac{2n}{2n+1} \theta \) and \( E(T^2) = \frac{2n}{2n+2} \theta^2 \).

\[ \text{MSE}(CT) = (C\frac{2n}{2n+1} \theta - \theta)^2 + C^2\left[\frac{2n}{2n+2} \theta^2 - \left(\frac{2n}{2n+1} \theta\right)^2\right] \]

\[ \frac{d\text{MSE}(CT)}{dC} = 2\left[\frac{2Cn\theta}{2n+1} - \theta\right][\frac{2n\theta}{2n+1}] + 2C\left[\frac{2n\theta^2}{2n+2} - \frac{4n^2\theta^2}{(2n+1)^2}\right]. \]

Solve \( \frac{d\text{MSE}(CT)}{dC} \) set \( \equiv 0 \) to get

\[ C = 2\frac{n+1}{2n+1}. \]

The MSE is a quadratic in \( C \) and the coefficient on \( C^2 \) is positive, hence the local min is a global min.
6.26. a) $E(X_i) = 2\theta/3$ and $V(X_i) = \theta^2/18$. So bias of $T = B(T) = E\bar{X} - \theta = c^2\theta/3 - \theta$ and $\text{Var}(T) = \frac{c^2}{n^2} \sum \text{Var}(X_i) = \frac{c^2 n \theta^2}{18}$.

So $\text{MSE} = \text{Var}(T) + [B(T)]^2 = \frac{c^2 \theta^2}{18n} + \left(\frac{2\theta}{3} c - \theta\right)^2$.

b) $\frac{d\text{MSE}(c)}{dc} = \frac{2c\theta^2}{18n} + 2\left(\frac{2\theta}{3} c - \theta\right) \frac{2\theta}{3}$.

Set this equation equal to 0 and solve, so

\[
\theta^2 \frac{2c}{18n} + \frac{4}{3} \left(\frac{2\theta}{3} c - \theta\right) = 0
\]

or

\[
c \left[\frac{2\theta^2}{18n} + \frac{8}{9} \theta^2\right] = \frac{4}{3} \theta^2
\]

or

\[
c \left(\frac{1}{9n} + \frac{8n}{9}\right) = \frac{4}{3}
\]

or

\[
c \left(\frac{1}{9n} + \frac{8n}{9n}\right) = \frac{4}{3}
\]

or

\[
c = \frac{9n}{1 + 8n} = \frac{12n}{1 + 8n}.
\]

This is a global min since the MSE is a quadratic in $c^2$ with a positive coefficient, or because

\[
\frac{d^2\text{MSE}(c)}{dc^2} = \frac{2\theta^2}{18n} + \frac{8\theta^2}{9} > 0.
\]

6.27. See Example 6.5.

6.30. See Example 6.3.

6.31. a) $E(T) = cE(Y) = c\alpha \beta = 10c\beta$.

$V(T) = c^2 V(\bar{Y}) = c^2 \alpha \beta^2 / n = 10c^2 \beta^2 / n$.

$MSE(T) = V(T) + [B(T)]^2 = 10c^2 \beta^2 / n + (10c\beta - \beta)^2$. 
CHAPTER 11. STUFF FOR STUDENTS

$\frac{d}{dc} \text{MSE}(c) = \frac{2c10\beta^2}{n} + 2(10c\beta - \beta)10^\text{set}$

or $[20\beta^2/n] \ c + 200\beta^2 \ c - 20\beta^2 = 0$

or $c/n + 10c - 1 = 0$ or $c(1/n + 10) = 1$

or $c = \frac{1}{n} + 10 = \frac{n}{10n + 1}$.

This value of $c$ is unique, and

$\frac{d^2 \text{MSE}(c)}{dc^2} = \frac{20\beta^2}{n} + 200\beta^2 > 0$,

so $c$ is the minimizer.

6.32. a) Since this distribution is a one parameter regular exponential family, $T_n = -\sum_{i=1}^{n} \log(2Y_i - Y_i^2)$ is complete.

b) Note that $\log(f(y|\nu)) = \log(\nu) + \log(2 - 2y) + (1 - \nu)[-\log(2y - y^2)]$.

Hence

$\frac{d}{d\nu} \log(f(y|\nu)) = \frac{1}{\nu} + \log(2y - y^2)$

and

$\frac{d^2 \log(f(y|\nu))}{d\nu^2} = -\frac{1}{\nu^2}$.

Since this family is a 1P-REF, $I_1(\nu) = -E \left( \frac{-1}{\nu^2} \right) = \frac{1}{\nu^2}$.

c) $\frac{[\tau'(\nu)]^2}{nI_1(\nu)} = \frac{\nu^2}{\nu^4/n} = \frac{1}{\nu^2}$.

d) $E[T_n^{-1}] = \frac{1}{\nu - 1} \frac{\Gamma(-1 + n)}{\Gamma(n)} = \frac{\nu}{n - 1}$. So $(n - 1)/T_n$ is the UMVUE of $\nu$ by LSU.

6.33. a) Since $f(y) = \frac{\theta}{2} \exp[-(\theta + 1)\log(1 + |y|)]$ is a 1P-REF,

$T = \sum_{i=1}^{n} \log(1 + |Y_i|)$ is a complete sufficient statistic.

b) Since this is an exponential family, $\log(f(y|\theta)) = \log(\theta/2) - (\theta + 1)\log(1 + |y|)$ and

$\frac{\partial}{\partial \theta} \log(f(y|\theta)) = \frac{1}{\theta} - \log(1 + |y|)$. 
\[ \frac{\partial^2}{\partial \theta^2} \log(f(y|\theta)) = -\frac{1}{\theta^2} \]

and

\[ I_1(\theta) = -E_\theta \left[ \frac{\partial^2}{\partial \theta^2} \log(f(Y|\theta)) \right] = \frac{1}{\theta^2}. \]

c) The complete sufficient statistic \( T \sim G(n, 1/\theta) \). Hence the UMVUE of \( \theta \) is \( (n - 1)/T \) since for \( r > -n \),

\[ E(T^r) = E(T^r) = \left( \frac{1}{\theta} \right)^r \frac{\Gamma(r + n)}{\Gamma(n)}. \]

So

\[ E(T^{-1}) = \theta \frac{\Gamma(n - 1)}{\Gamma(n)} = \theta/(n - 1). \]

6.34. a) Note that \( \overline{X} \) is a complete and sufficient statistic for \( \mu \) and \( \overline{X} \sim N(\mu, n^{-1}\sigma^2) \). We know that \( E(e^{2\overline{X}}) \), the mgf of \( \overline{X} \) when \( t = 2 \), is given by \( e^{2\mu+2n^{-1}\sigma^2} \). Thus the UMVUE of \( e^{2\mu} \) is \( e^{-2n^{-1}\sigma^2}e^{2\overline{X}} \).

b) The CRLB for the variance of unbiased estimator of \( g(\mu) \) is given by \( 4n^{-1}\sigma^2e^{4\mu} \) whereas

\[ V(e^{-2n^{-1}\sigma^2}e^{2\overline{X}}) = e^{-4n^{-1}\sigma^2}E(e^{4\overline{X}}) - e^{4\mu} \]
\[ = e^{-4n^{-1}\sigma^2}e^{4\mu+\frac{1}{2}16n^{-1}\sigma^2} - e^{4\mu} \]
\[ = e^{4\mu}[e^{4n^{-1}\sigma^2} - 1] \]
\[ > 4n^{-1}\sigma^2e^{4\mu} \]

since \( e^x > 1 + x \) for all \( x > 0 \). Hence the CRLB is not attained.

7.6. For both a) and b), the test is reject \( H_0 \) iff \( \prod_{i=1}^n x_i(1 - x_i) > c \) where
\[ P_{\theta=1}[\prod_{i=1}^n x_i(1 - x_i) > c] = \alpha. \]

7.10. \( H \) says \( f(x) = e^{-x} \) while \( K \) says

\[ f(x) = x^{\theta-1}e^{-x}/\Gamma(\theta). \]

The monotone likelihood ratio property holds for \( \prod x_i \) since then

\[ \frac{f_n(x, \theta_2)}{f_n(x, \theta_1)} = \frac{\prod_{i=1}^n x_i^{\theta_2-1}(\Gamma(\theta_1))^{n}}{\prod_{i=1}^n x_i^{\theta_1-1}(\Gamma(\theta_2))^{n}} = \left( \frac{\Gamma(\theta_1)}{\Gamma(\theta_2)} \right)^n \left( \prod_{i=1}^n x_i \right)^{\theta_2-\theta_1}. \]
which increases as $\prod_{i=1}^{n} x_i$ increases if $\theta_2 > \theta_1$. Hence the level $\alpha$ UMP test rejects $H$ if

$$\prod_{i=1}^{n} X_i > c$$

where

$$P_H(\prod_{i=1}^{n} X_i > c) = P_H(\sum \log(X_i) > \log(c)) = \alpha.$$

**7.11.** See Example 7.7.

**7.13.** Let $\theta_1 = 4$. By Neyman Pearson lemma, reject $H_0$ if

$$\frac{f(x|\theta_1)}{f(x|2)} = \left(\frac{\log(\theta_1)}{\theta_1 - 1}\right)^n \theta_1^{\sum x_i} \left(\frac{1}{\log(2)}\right)^n \frac{1}{2^{\sum x_i}} > k$$

iff

$$\left(\frac{\log(\theta_1)}{(\theta_1 - 1)\log(2)}\right)^n \left(\frac{\theta_1}{2}\right)^{\sum x_i} > k$$

iff

$$\left(\frac{\theta_1}{2}\right)^{\sum x_i} > k'$$

iff

$$\sum x_i \log(\theta_1/2) > c'.$$

So reject $H_0$ iff $\sum X_i > c$ where $P_{H=2}(\sum X_i > c) = \alpha$.

**7.14.** a) By NP lemma reject $H_0$ if

$$\frac{f(x|\sigma = 2)}{f(x|\sigma = 1)} > k'.$$

The LHS =

$$\frac{1}{2\pi} \exp[-\frac{1}{8} \sum x_i^2] \exp[\frac{1}{2\pi} \sum x_i^2]$$

So reject $H_0$ if

$$\frac{1}{2\pi} \exp[\sum x_i^2(\frac{1}{2} - \frac{1}{8})] > k'$$

or if $\sum x_i^2 > k$ where $P_{H_0}(\sum x_i^2 > k) = \alpha$. 


b) In the above argument, with any \( \sigma_1 > 1 \), get
\[
\sum x_i^2 \left( \frac{1}{2} - \frac{1}{2\sigma_1^2} \right)
\]
and
\[
\frac{1}{2} - \frac{1}{2\sigma_1^2} > 0
\]
for any \( \sigma_1^2 > 1 \). Hence the UMP test is the same as in a).

7.15. a) By NP lemma reject \( H_0 \) if
\[
\frac{f(x|\sigma = 2)}{f(x|\sigma = 1)} > k'.
\]
The LHS =
\[
\frac{1}{2^n} \exp \left[ -\frac{1}{8} \sum [\log(x_i)]^2 \right]
\]
\[\exp \left[ -\frac{1}{2} \sum [\log(x_i)]^2 \right]
\]
So reject \( H_0 \) if
\[
\frac{1}{2^n} \exp \left[ \sum [\log(x_i)]^2 \left( \frac{1}{2} - \frac{1}{8} \right) \right] > k'
\]
or if \( \sum [\log(X_i)]^2 > k \) where \( P_{H_0}(\sum [\log(X_i)]^2 > k) = \alpha \).

b) In the above argument, with any \( \sigma_1 > 1 \), get
\[
\sum [\log(x_i)]^2 \left( \frac{1}{2} - \frac{1}{2\sigma_1^2} \right)
\]
and
\[
\frac{1}{2} - \frac{1}{2\sigma_1^2} > 0
\]
for any \( \sigma_1^2 > 1 \). Hence the UMP test is the same as in a).

7.16. The most powerful test will have the following form.
Reject \( H_0 \) iff \( \frac{f_1(x)}{f_0(x)} > k \).

But \( \frac{f_1(x)}{f_0(x)} = 4x^{-\frac{3}{2}} \) and hence we reject \( H_0 \) iff \( X \) is small, i.e. reject \( H_0 \) is \( X < k \) for some constant \( k \). This test must also have the size \( \alpha \), that is we require:
\[
\alpha = P(X < k) \text{ when } f(x) = f_0(x) = \int_0^k \frac{3}{64} x^2 dx = \frac{1}{64} k^3,
\]
so that \( k = 4\alpha^{\frac{1}{3}} \).
For the power, when \( k = 4\alpha \)
\[
P[X < k \text{ when } f(x) = f_1(x)] = f_0^k \frac{3}{16} \sqrt{k} dx = \sqrt{\alpha}.
\]
When \( \alpha = 0.01 \), the power is \( 0.10 \).

**7.19.** See Example 7.5.

**7.20.** \( E[T(X)] = 1/\lambda_1 \) and the power \( = P(\text{test rejects } H_0) = P_{\lambda_1}(T(X) < \log(100/95)) = F_{\lambda_1}(\log(100/95)) \)
\[
= 1 - \exp(-\lambda_1 \log(100/95)) = 1 - (95/100)^{\lambda_1}.
\]
   a) Power \( = 1 - \exp(-\log(100/95)) = 1 - \exp(\log(95/100)) = 0.05. \)
   b) Power \( = 1 - (95/100)^{50} = 0.923055. \)
   c) Let \( T_0 \) be the observed value of \( T(X) \). Then pvalue \( = P(W \leq T_0) \) where \( W \sim \text{exponential}(1) \) since under \( H_0 \), \( T(X) \sim \text{exponential}(1) \). So pvalue \( = 1 - \exp(-T_0). \)

**7.21.** Note that
\[
f(x) = I(x > 0) \, 2x \, e^{-x^2} \, \tau \, \exp[(\tau - 1)(\log(1 - e^{-x^2}))]
\]
is a one parameter exponential family and \( w(\tau) = \tau - 1 \) is an increasing function of \( \tau \). Thus the UMP test rejects \( H_0 \) if \( T(x) = \sum_{i=1}^{n} \log(1 - e^{-x_i^2}) > k \) where \( \alpha = P_{\tau=2}(T(X) > k) \).

   Or use NP lemma.

   a) Reject Ho if
\[
f(x|\tau = 4) > f(x|\tau = 1).
\]
The LHS =
\[
\frac{4^n}{2^n} \prod_{i=1}^{n} (1 - e^{-x_i^2})^{4-1} = 2^n \prod_{i=1}^{n} (1 - e^{-x_i^2})^2.
\]
So reject Ho if
\[
\prod_{i=1}^{n} (1 - e^{-x_i^2})^2 > k'
\]
or
\[
\prod_{i=1}^{n} (1 - e^{-x_i^2}) > c
\]
or
\[
\sum_{i=1}^{n} \log(1 - e^{-x_i^2}) > d
\]
where
\[ \alpha = P_{\tau=2}(\prod_{i=1}^{n}(1 - e^{-x_i^2}) > c). \]

b) Replace 4 – 1 by \( \tau_1 - 1 \) where \( \tau_1 > 2 \). Then reject \( H_0 \) if
\[ \prod_{i=1}^{n}(1 - e^{-x_i^2})^{\tau_1-2} > k' \]
which gives the same test as in a).

7.22. By exponential family theory, the UMP test rejects \( H_0 \) if
T(\( x \)) = - \( \sum_{i=1}^{n} \frac{1}{x_i} \) > k where \( P_{\theta=1}(T(X) > k) = \alpha. \)
Alternatively, use the Neyman Pearson lemma:
a) reject Ho if
\[ \frac{f(\theta = 2)}{f(\theta = 1)} > k'. \]
The LHS =
\[ \frac{2^n \exp(-2 \sum \frac{1}{x_i})}{\exp(-\sum \frac{1}{x_i})}. \]
So reject Ho if
\[ 2^n \exp((-2 + 1) \sum \frac{1}{x_i}) > k' \]
or if - \( \sum \frac{1}{x_i} \) > k where \( P_1(- \sum \frac{1}{x_i} > k) = \alpha. \)
b) In the above argument, reject \( H_0 \) if
\[ 2^n \exp((-\theta_1 + 1) \sum \frac{1}{x_i}) > k' \]
or if - \( \sum \frac{1}{x_i} \) > k where \( P_1(- \sum \frac{1}{x_i} > k) = \alpha \) for any \( \theta_1 > 1 \). Hence the UMP test is the same as in a).

7.23 a) We reject \( H_0 \) iff \( \frac{f_1(x)}{f_0(x)} > k \). Thus we reject \( H_0 \) iff \( \frac{2x}{2(1-x)} > k \). That is \( \frac{1-x}{x} < k_1 \), that is \( \frac{1}{x} < k_2 \), that is \( x > k_3 \). Now 0.1 = \( P(X > k_3) \) when \( f(x) = f_0(x) \), so \( k_3 = 1 - \sqrt{0.1}. \)

7.24. a) Let \( k = [2\pi \sigma_1 \sigma_2 (1 - \rho^2)^{1/2}] \). Then the likelihood \( L(\theta) = \)
\[ \frac{1}{k^n} \exp \left( -1 \frac{1}{2(1-\rho^2)} \sum_{i=1}^{n} \left[ \left( \frac{x_i - \mu_1}{\sigma_1} \right)^2 - 2\rho \left( \frac{x_i - \mu_1}{\sigma_1} \right) \left( \frac{y_i - \mu_2}{\sigma_2} \right) + \left( \frac{y_i - \mu_2}{\sigma_2} \right)^2 \right] \right). \]
CHAPTER 11. STUFF FOR STUDENTS

Hence
\[
L(\hat{\theta}) = \frac{1}{[2\pi \hat{\sigma}_1 \hat{\sigma}_2 (1 - \hat{\rho}^2)^{1/2}]^n} \exp \left( -\frac{1}{2(1 - \hat{\rho}^2)} \left[ T_1 - 2\hat{\rho} T_2 + T_3 \right] \right)
\]
\[= \frac{1}{[2\pi \hat{\sigma}_1 \hat{\sigma}_2 (1 - \hat{\rho}^2)^{1/2}]^n} \exp(-n)\]
and
\[
L(\hat{\theta}_0) = \frac{1}{[2\pi \hat{\sigma}_1 \hat{\sigma}_2 ]^n} \exp \left( -\frac{1}{2} [T_1 + T_3] \right)
\]
\[= \frac{1}{[2\pi \hat{\sigma}_1 \hat{\sigma}_2 ]^n} \exp(-n).\]
Thus \(\lambda(x, y) = \frac{L(\hat{\theta}_0)}{L(\hat{\theta})} = (1 - \hat{\rho}^2)^{n/2} \).

So reject \(H_0\) if \(\lambda(x, y) \leq c\) where \(\alpha = \sup_{\theta \in \Theta_o} P(\lambda(X, Y) \leq c)\). Here \(\Theta_o\) is the set of \(\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\) such that the \(\mu_i\) are real, \(\sigma_i > 0\) and \(\rho = 0\), ie, such that \(X_i\) and \(Y_i\) are independent.

b) Since the unrestricted MLE has one more free parameter than the restricted MLE, \(-2\log(\lambda(X, Y)) \approx \chi^2_1\) and the approximate LRT rejects \(H_0\) if \(-2\log \lambda(x, y) > \chi^2_{1-\alpha}\) where \(P(\chi^2_1 > \chi^2_{1-\alpha}) = \alpha\).

8.1 c) The histograms should become more like a normal distribution as \(n\) increases from 1 to 200. In particular, when \(n = 1\) the histogram should be right skewed while for \(n = 200\) the histogram should be nearly symmetric. Also the scale on the horizontal axis should decrease as \(n\) increases.

d) Now \(\overline{Y} \sim N(0, 1/n)\). Hence the histograms should all be roughly symmetric, but the scale on the horizontal axis should be from about \(-3/\sqrt{n}\) to \(3/\sqrt{n}\).

8.3. a) \(E(X) = \frac{3\theta}{\theta + 1}\), thus
\[
\sqrt{n}(\overline{X} - E(X)) \overset{D}{\rightarrow} N(0, V(X)), \text{ where}
\]
\[V(X) = \frac{9\theta}{(\theta + 2)(\theta + 1)^2}.\]
Let \(g(y) = \frac{y}{3-y}\), thus \(g'(y) = \frac{3}{(3-y)^2}\). Using the delta method, \(\sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} N(0, \frac{g'(\theta)^2}{\theta + 2})\).

b) It is asymptotically efficient if \(\sqrt{n}(T_n - \theta) \overset{D}{\rightarrow} N(0, \nu(\theta))\), where
\[
\nu(\theta) = \frac{\frac{d}{d\theta}(\theta)}{-E(\frac{d^2}{d\theta^2} \ln f(x|\theta))}.
\]
CHAPTER 11. STUFF FOR STUDENTS

422

But, \( E(\frac{\partial}{\partial \theta} \ln f(x|\theta)) = \frac{1}{\theta} \). Thus \( \nu(\theta) = \theta^2 \neq \frac{\theta(\theta+1)^2}{\theta+2} \).

c) \( \bar{X} \to \frac{3\theta}{\theta+1} \) in probability. Thus \( T_n \to \theta \) in probability.

8.5. See Example 8.8.

8.7. a) See Example 8.7.

8.13. a) \( Y_n \overset{D}{=} \sum_{i=1}^{n} X_i \) where the \( X_i \) are iid \( \chi_1^2 \). Hence \( E(X_i) = 1 \) and \( \text{Var}(X_i) = 2 \). Thus by the CLT,

\[
\sqrt{n} \left( \frac{Y_n}{n} - 1 \right) \overset{D}{=} \sqrt{n} \left( \frac{\sum_{i=1}^{n} X_i}{n} - 1 \right) \overset{D}{=} N(0, 2).
\]

b) Let \( g(\theta) = \theta^3 \). Then \( g'(\theta) = 3\theta^2 \), \( g'(1) = 3 \), and by the delta method,

\[
\sqrt{n} \left[ \left( \frac{Y_n}{n} \right)^3 - 1 \right] \overset{D}{=} N(0, 2(g'(1))^2) = N(0, 18).
\]

8.23. See the proof of Theorem 6.3.

8.27. a) See Example 8.1b.

b) See Example 8.3.

c) See Example 8.14.

8.28. a) By the CLT, \( \sqrt{n}(\bar{X} - \lambda) / \sqrt{\lambda} \overset{D}{=} N(0, 1) \). Hence \( \sqrt{n}(\bar{X} - \lambda) \overset{D}{=} N(0, \lambda) \).

b) Let \( g(\lambda) = \lambda^3 \) so that \( g'(\lambda) = 3\lambda^2 \) then \( \sqrt{n}[\bar{X}^3 - (\lambda)^3] \overset{D}{=} N(0, \lambda[g'(\lambda)]^2) = N(0, 9\lambda^5) \).

8.29. a) \( \bar{X} \) is a complete sufficient statistic. Also, we have \( \frac{(n-1)S^2}{\sigma^2} \) has a chi square distribution with \( df = n-1 \), thus since \( \sigma^2 \) is known the distribution of \( S^2 \) does not depend on \( \mu \), so \( S^2 \) is ancillary. Thus, by Basu’s Theorem \( \bar{X} \) and \( S^2 \) are independent.

b) by CLT (\( n \) is large) \( \sqrt{n}(\bar{X} - \mu) \) has approximately normal distribution with mean 0 and variance \( \sigma^2 \). Let \( g(x) = x^3 \), thus \( g'(x) = 3x^2 \). Using delta method \( \sqrt{n}(g(\bar{X}) - g(\mu)) \) goes in distribution to \( N(0, \sigma^2(g'(\mu))^2) \) or \( \sqrt{n}(\bar{X}^3 - \mu^3) \) goes in distribution to \( N(0, \sigma^2(3\mu^2)^2) \).

8.30. a) According to the standard theorem, \( \sqrt{n}(\hat{\theta}_n - \theta) \to N(0, 3) \).

b) \( E(Y) = \theta, \text{Var}(Y) = \frac{\sigma^2}{3} \), according to CLT we have \( \sqrt{n}(\bar{Y}_n - \theta) \to N(0, \frac{\sigma^2}{3}) \).
CHAPTER 11. STUFF FOR STUDENTS

423

c) \( MED(Y) = \theta \), then \( \sqrt{n}(MED(n) - \theta) \to N(0, \frac{1}{4f^2(MED(Y))}) \) and
\( f(MED(Y)) = \frac{\exp\left(-\frac{(\theta-\theta)}{2}\right)}{[1+\exp\left(-\frac{(\theta-\theta)}{2}\right)]^2} = \frac{1}{4} \). Thus \( \sqrt{n}(MED(n) - \theta) \to N(0, \frac{1}{4f^2(MED(Y))}) \) and \( \sqrt{n}(MED(n) - \theta) \to N(0,4) \).

d) All three estimators are consistent, but \( 3 < \frac{\pi}{3} < 4 \), therefore the estimator \( \hat{\theta}_n \) is the best, and the estimator \( MED(n) \) is the worst.

8.32. a) \( F_n(y) = 0.5 + 0.5y/n \) for \( -n < y < n \), so \( F(y) \equiv 0.5 \).

b) No, since \( F(y) \) is not a cdf.

8.33. a) \( F_n(y) = y/n \) for \( 0 < y < n \), so \( F(y) \equiv 0 \).

b) No, since \( F(y) \) is not a cdf.

8.34. a) \( \sqrt{n} \left( \frac{1}{\rho} - \frac{\rho}{\rho} \right) \) \( \overset{D}{\to} N \left( 0, \frac{1}{\rho^2} \right) \)

by the CLT.

c) The method of moments estimator of \( \rho \) is \( \hat{\rho} = \frac{\sum}{1+y} \).

d) Let \( g(\theta) = 1 + \theta \) so \( g'(\theta) = 1 \). Then by the delta method,

\( \sqrt{n} \left( g(\bar{Y}) - g\left(\frac{1-\rho}{\rho}\right) \right) \) \( \overset{D}{\to} N \left( 0, \frac{1}{\rho^2} \right) \)

or

\( \sqrt{n} \left( (1+\bar{Y}) - \frac{1}{\rho} \right) \) \( \overset{D}{\to} N \left( 0, \frac{1}{\rho^2} \right) \).

This result could also be found with algebra since \( 1+\bar{Y} - \frac{1}{\rho} = \bar{Y} + 1 - \frac{1}{\rho} = \overline{\bar{Y}} + \frac{\rho-1}{\rho} = \bar{Y} - \frac{1-\rho}{\rho} \).

e) \( \bar{Y} \) is the method of moments estimator of \( E(Y) = (1-\rho)/\rho \), so \( 1+\bar{Y} \)

is the method of moments estimator of \( 1 + E(Y) = 1/\rho \).

8.35. a) \( \sqrt{n}(\bar{X} - \mu) \) is approximately \( N(0, \sigma^2) \) Define \( g(x) = \frac{1}{\bar{x}} \), \( g'(x) = \frac{-1}{\bar{x}^2} \). Using delta method \( \sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \) has approximately \( N(0, \frac{\sigma^2}{\mu^2}) \). Thus \( 1/\bar{X} \)

is approximately \( N\left(\frac{1}{\mu}, \frac{\sigma^2}{\mu^2}\right) \), provided \( \mu \neq 0 \).

b) Using part a)
\( \frac{1}{X} \) is asymptotically efficient for \( \frac{1}{\mu} \) if

\[
\frac{\sigma^2}{\mu^4} = \left[ \frac{(\tau'(\mu))^2}{E_\mu \left( \frac{\partial}{\partial \mu} \ln f(X/\mu) \right)^2} \right]
\]

\[
\tau(\mu) = \frac{1}{\mu}
\]

\[
\tau'(\mu) = -\frac{1}{\mu^2}
\]

\[
\ln f(x|\mu) = -\frac{1}{2} \ln 2\pi \sigma^2 - \frac{(x - \mu)^2}{2\sigma^2}
\]

\[
E \left[ \frac{\partial}{\partial \mu} \ln f(X/\mu) \right]^2 = \frac{E(X - \mu)^2}{\sigma^4} = \frac{1}{\sigma^2}
\]

Thus

\[
\frac{(\tau'(\mu))^2}{E_\mu \left[ \frac{\partial}{\partial \mu} \ln f(X/\mu) \right]^2} = \frac{\sigma^2}{\mu^4}.
\]

**8.36.** a) \( E(Y^k) = 2\theta^k/(k + 2) \) so \( E(Y) = 2\theta/3, \ E(Y^2) = \theta^2/2 \) and \( V(Y) = \theta^2/18. \) So \( \sqrt{n} \left( \overline{Y} - \frac{2\theta}{3} \right) \xrightarrow{D} N \left( 0, \frac{\theta^2}{18} \right) \) by the CLT.

b) Let \( g(\tau) = \log(\tau) \) so \( [g'(\tau)]^2 = 1/\tau^2 \) where \( \tau = 2\theta/3. \) Then by the delta method,

\[
\sqrt{n} \left( \log(\overline{Y}) - \log \left( \frac{2\theta}{3} \right) \right) \xrightarrow{D} N \left( 0, \frac{1}{8} \right).
\]

c) \( \hat{\theta}^k = \frac{k+2}{2n} \sum Y_i^k. \)

**9.1.** a) \( \sum_{i=1}^n X_i^b \) is minimal sufficient for \( a. \)

b) It can be shown that \( \frac{X_i^b}{a} \) has an exponential distribution with mean 1. Thus, \( \frac{2 \sum_{i=1}^n X_i^b}{a} \) is distributed \( \chi^2_{2n}. \) Let \( \chi^2_{2n,\alpha/2} \) be the upper 100(\(1/2\alpha\))% point of the chi-square distribution with \( 2n \) degrees of freedom. Thus, we can write

\[
1 - \alpha = P\left( \chi^2_{2n,1-\alpha/2} < \frac{2 \sum_{i=1}^n X_i^b}{a} < \chi^2_{2n,\alpha/2} \right)
\]
which translates into
\[
\left( \frac{2 \sum_{i=1}^{n} X_i^b}{\chi_{2n,\alpha/2}^2}, \frac{2 \sum_{i=1}^{n} X_i^b}{\chi_{2n,1-\alpha/2}^2} \right)
\]
as a two sided \((1 - \alpha)\) confidence interval for \(a\). For \(\alpha = 0.05\) and \(n = 20\), we have \(\chi_{2n,\alpha/2}^2 = 34.1696\) and \(\chi_{2n,1-\alpha/2}^2 = 9.59083\). Thus the confidence interval for \(a\) is
\[
\left( \frac{\sum_{i=1}^{n} X_i^b}{17.0848}, \frac{\sum_{i=1}^{n} X_i^b}{4.795415} \right)
\]

9.4c). Tables are from simulated data but should be similar to the table below.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>ccov</th>
<th>acov</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>.01</td>
<td>.4236</td>
<td>.9914</td>
<td>ACT CI better</td>
</tr>
<tr>
<td>100</td>
<td>.01</td>
<td>.6704</td>
<td>.9406</td>
<td>ACT CI better</td>
</tr>
<tr>
<td>150</td>
<td>.01</td>
<td>.8278</td>
<td>.9720</td>
<td>ACT CI better</td>
</tr>
<tr>
<td>200</td>
<td>.01</td>
<td>.9294</td>
<td>.9098</td>
<td>the CIs are about the same</td>
</tr>
<tr>
<td>250</td>
<td>.01</td>
<td>.9158</td>
<td>.8160</td>
<td>the CIs are about the same</td>
</tr>
<tr>
<td>300</td>
<td>.01</td>
<td>.9158</td>
<td>.9228</td>
<td>the CIs are about the same</td>
</tr>
<tr>
<td>350</td>
<td>.01</td>
<td>.9702</td>
<td>.8312</td>
<td>classical is better</td>
</tr>
<tr>
<td>400</td>
<td>.01</td>
<td>.9486</td>
<td>.6692</td>
<td>classical is better</td>
</tr>
<tr>
<td>450</td>
<td>.01</td>
<td>.9486</td>
<td>.6692</td>
<td>classical is better</td>
</tr>
</tbody>
</table>

9.11. The simulated coverages should be close to the values below. The pooled t CI has coverage that is too small.

<table>
<thead>
<tr>
<th>pcov</th>
<th>mpcov</th>
<th>wcov</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.847</td>
<td>0.942</td>
<td>0.945</td>
</tr>
</tbody>
</table>

9.12. a) Let \(W_i \sim U(0, 1)\) for \(i = 1, ..., n\) and let \(T_n = Y/\theta\). Then
\[
P\left(\frac{Y}{\theta} \leq t\right) = P\left(\max(W_1, ..., W_n) \leq t\right) =
\]
P(all \(W_i \leq t\)) = \([F_{W_i}(t)]^n = t^n\) for \(0 < t < 1\). So the pdf of \(T_n\) is
\[
f_{T_n}(t) = \frac{d}{dt} t^n = nt^{n-1}
\]
for \(0 < t < 1\).

b) Yes, the distribution of \(T_n = Y/\theta\) does not depend on \(\theta\) by a).
c) See Example 9.21.
11.3 Tables

Tabled values are $F(0.95,k,d)$ where $P(F < F(0.95,k,d)) = 0.95$.  

00 stands for $\infty$. Entries produced with the \texttt{qf(.95,k,d)} command in \textit{R}. The numerator degrees of freedom are $k$ while the denominator degrees of freedom are $d$.

<table>
<thead>
<tr>
<th>k</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>00</th>
</tr>
</thead>
<tbody>
<tr>
<td>d</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>161</td>
<td>200</td>
<td>216</td>
<td>225</td>
<td>230</td>
<td>234</td>
<td>237</td>
<td>239</td>
<td>241</td>
<td>254</td>
</tr>
<tr>
<td>2</td>
<td>18.5</td>
<td>19.0</td>
<td>19.2</td>
<td>19.3</td>
<td>19.3</td>
<td>19.4</td>
<td>19.4</td>
<td>19.4</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10.1</td>
<td>9.55</td>
<td>9.28</td>
<td>9.12</td>
<td>9.01</td>
<td>8.94</td>
<td>8.89</td>
<td>8.85</td>
<td>8.81</td>
<td>8.53</td>
</tr>
<tr>
<td>4</td>
<td>7.71</td>
<td>6.94</td>
<td>6.59</td>
<td>6.39</td>
<td>6.26</td>
<td>6.16</td>
<td>6.09</td>
<td>6.04</td>
<td>6.00</td>
<td>5.63</td>
</tr>
<tr>
<td>5</td>
<td>6.61</td>
<td>5.79</td>
<td>5.41</td>
<td>5.19</td>
<td>5.05</td>
<td>4.95</td>
<td>4.88</td>
<td>4.82</td>
<td>4.77</td>
<td>4.37</td>
</tr>
<tr>
<td>6</td>
<td>5.99</td>
<td>5.14</td>
<td>4.76</td>
<td>4.53</td>
<td>4.39</td>
<td>4.28</td>
<td>4.21</td>
<td>4.15</td>
<td>4.10</td>
<td>3.67</td>
</tr>
<tr>
<td>7</td>
<td>5.59</td>
<td>4.74</td>
<td>4.35</td>
<td>4.12</td>
<td>3.97</td>
<td>3.87</td>
<td>3.79</td>
<td>3.73</td>
<td>3.68</td>
<td>3.23</td>
</tr>
<tr>
<td>8</td>
<td>5.32</td>
<td>4.46</td>
<td>4.07</td>
<td>3.84</td>
<td>3.69</td>
<td>3.58</td>
<td>3.50</td>
<td>3.44</td>
<td>3.39</td>
<td>2.93</td>
</tr>
<tr>
<td>9</td>
<td>5.12</td>
<td>4.26</td>
<td>3.86</td>
<td>3.63</td>
<td>3.48</td>
<td>3.37</td>
<td>3.29</td>
<td>3.23</td>
<td>3.18</td>
<td>2.71</td>
</tr>
<tr>
<td>10</td>
<td>4.96</td>
<td>4.10</td>
<td>3.71</td>
<td>3.48</td>
<td>3.33</td>
<td>3.22</td>
<td>3.14</td>
<td>3.07</td>
<td>3.02</td>
<td>2.54</td>
</tr>
<tr>
<td>11</td>
<td>4.84</td>
<td>3.98</td>
<td>3.59</td>
<td>3.36</td>
<td>3.20</td>
<td>3.09</td>
<td>3.01</td>
<td>2.95</td>
<td>2.90</td>
<td>2.41</td>
</tr>
<tr>
<td>12</td>
<td>4.75</td>
<td>3.89</td>
<td>3.49</td>
<td>3.26</td>
<td>3.11</td>
<td>3.00</td>
<td>2.91</td>
<td>2.85</td>
<td>2.80</td>
<td>2.30</td>
</tr>
<tr>
<td>13</td>
<td>4.67</td>
<td>3.81</td>
<td>3.41</td>
<td>3.18</td>
<td>3.03</td>
<td>2.92</td>
<td>2.83</td>
<td>2.77</td>
<td>2.71</td>
<td>2.21</td>
</tr>
<tr>
<td>14</td>
<td>4.60</td>
<td>3.74</td>
<td>3.34</td>
<td>3.11</td>
<td>2.96</td>
<td>2.85</td>
<td>2.76</td>
<td>2.70</td>
<td>2.65</td>
<td>2.13</td>
</tr>
<tr>
<td>15</td>
<td>4.54</td>
<td>3.68</td>
<td>3.29</td>
<td>3.06</td>
<td>2.90</td>
<td>2.79</td>
<td>2.71</td>
<td>2.64</td>
<td>2.59</td>
<td>2.07</td>
</tr>
<tr>
<td>16</td>
<td>4.49</td>
<td>3.63</td>
<td>3.24</td>
<td>3.01</td>
<td>2.85</td>
<td>2.74</td>
<td>2.66</td>
<td>2.59</td>
<td>2.54</td>
<td>2.01</td>
</tr>
<tr>
<td>17</td>
<td>4.45</td>
<td>3.59</td>
<td>3.20</td>
<td>2.96</td>
<td>2.81</td>
<td>2.70</td>
<td>2.61</td>
<td>2.55</td>
<td>2.49</td>
<td>1.96</td>
</tr>
<tr>
<td>18</td>
<td>4.41</td>
<td>3.55</td>
<td>3.16</td>
<td>2.93</td>
<td>2.77</td>
<td>2.66</td>
<td>2.58</td>
<td>2.51</td>
<td>2.46</td>
<td>1.92</td>
</tr>
<tr>
<td>19</td>
<td>4.38</td>
<td>3.52</td>
<td>3.13</td>
<td>2.90</td>
<td>2.74</td>
<td>2.63</td>
<td>2.54</td>
<td>2.48</td>
<td>2.42</td>
<td>1.88</td>
</tr>
<tr>
<td>20</td>
<td>4.35</td>
<td>3.49</td>
<td>3.10</td>
<td>2.87</td>
<td>2.71</td>
<td>2.60</td>
<td>2.51</td>
<td>2.45</td>
<td>2.39</td>
<td>1.84</td>
</tr>
<tr>
<td>25</td>
<td>4.24</td>
<td>3.39</td>
<td>2.99</td>
<td>2.76</td>
<td>2.60</td>
<td>2.49</td>
<td>2.40</td>
<td>2.34</td>
<td>2.28</td>
<td>1.71</td>
</tr>
<tr>
<td>30</td>
<td>4.17</td>
<td>3.32</td>
<td>2.92</td>
<td>2.69</td>
<td>2.53</td>
<td>2.42</td>
<td>2.33</td>
<td>2.27</td>
<td>2.21</td>
<td>1.62</td>
</tr>
<tr>
<td>00</td>
<td>3.84</td>
<td>3.00</td>
<td>2.61</td>
<td>2.37</td>
<td>2.21</td>
<td>2.10</td>
<td>2.01</td>
<td>1.94</td>
<td>1.88</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Tabled values are $t_{\alpha,d}$ where $P(t < t_{\alpha,d}) = \alpha$ where $t$ has a $t$ distribution with $d$ degrees of freedom. If $d > 29$ use the $N(0,1)$ cutoffs $d = Z = \infty$.

<table>
<thead>
<tr>
<th>$d$</th>
<th>0.005</th>
<th>0.01</th>
<th>0.025</th>
<th>0.05</th>
<th>0.5</th>
<th>0.95</th>
<th>0.975</th>
<th>0.99</th>
<th>0.995 left tail</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-63.66</td>
<td>-31.82</td>
<td>-12.71</td>
<td>-6.314</td>
<td>0</td>
<td>6.314</td>
<td>12.71</td>
<td>31.82</td>
<td>63.66</td>
</tr>
<tr>
<td>3</td>
<td>-5.841</td>
<td>-4.541</td>
<td>-3.182</td>
<td>-2.353</td>
<td>0</td>
<td>2.353</td>
<td>3.182</td>
<td>4.541</td>
<td>5.841</td>
</tr>
<tr>
<td>5</td>
<td>-4.032</td>
<td>-3.365</td>
<td>-2.571</td>
<td>-2.015</td>
<td>0</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
</tr>
<tr>
<td>7</td>
<td>-3.499</td>
<td>-2.998</td>
<td>-2.365</td>
<td>-1.895</td>
<td>0</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
</tr>
<tr>
<td>9</td>
<td>-3.250</td>
<td>-2.821</td>
<td>-2.262</td>
<td>-1.833</td>
<td>0</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
</tr>
<tr>
<td>10</td>
<td>-3.169</td>
<td>-2.764</td>
<td>-2.228</td>
<td>-1.812</td>
<td>0</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
</tr>
<tr>
<td>11</td>
<td>-3.106</td>
<td>-2.718</td>
<td>-2.201</td>
<td>-1.796</td>
<td>0</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
</tr>
<tr>
<td>12</td>
<td>-3.055</td>
<td>-2.681</td>
<td>-2.179</td>
<td>-1.782</td>
<td>0</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
</tr>
<tr>
<td>13</td>
<td>-3.012</td>
<td>-2.650</td>
<td>-2.160</td>
<td>-1.771</td>
<td>0</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
</tr>
<tr>
<td>14</td>
<td>-2.977</td>
<td>-2.624</td>
<td>-2.145</td>
<td>-1.761</td>
<td>0</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
</tr>
<tr>
<td>15</td>
<td>-2.947</td>
<td>-2.602</td>
<td>-2.131</td>
<td>-1.753</td>
<td>0</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
</tr>
<tr>
<td>16</td>
<td>-2.921</td>
<td>-2.583</td>
<td>-2.120</td>
<td>-1.746</td>
<td>0</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
</tr>
<tr>
<td>17</td>
<td>-2.898</td>
<td>-2.567</td>
<td>-2.110</td>
<td>-1.740</td>
<td>0</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
</tr>
<tr>
<td>18</td>
<td>-2.878</td>
<td>-2.552</td>
<td>-2.101</td>
<td>-1.734</td>
<td>0</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
</tr>
<tr>
<td>19</td>
<td>-2.861</td>
<td>-2.539</td>
<td>-2.093</td>
<td>-1.729</td>
<td>0</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
</tr>
<tr>
<td>20</td>
<td>-2.845</td>
<td>-2.528</td>
<td>-2.086</td>
<td>-1.725</td>
<td>0</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
</tr>
<tr>
<td>21</td>
<td>-2.831</td>
<td>-2.518</td>
<td>-2.080</td>
<td>-1.721</td>
<td>0</td>
<td>1.721</td>
<td>2.080</td>
<td>2.518</td>
<td>2.831</td>
</tr>
<tr>
<td>22</td>
<td>-2.819</td>
<td>-2.508</td>
<td>-2.074</td>
<td>-1.717</td>
<td>0</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
</tr>
<tr>
<td>23</td>
<td>-2.807</td>
<td>-2.500</td>
<td>-2.069</td>
<td>-1.714</td>
<td>0</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
</tr>
<tr>
<td>24</td>
<td>-2.797</td>
<td>-2.492</td>
<td>-2.064</td>
<td>-1.711</td>
<td>0</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
</tr>
<tr>
<td>25</td>
<td>-2.787</td>
<td>-2.485</td>
<td>-2.060</td>
<td>-1.708</td>
<td>0</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
</tr>
<tr>
<td>26</td>
<td>-2.779</td>
<td>-2.479</td>
<td>-2.056</td>
<td>-1.706</td>
<td>0</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
</tr>
<tr>
<td>27</td>
<td>-2.771</td>
<td>-2.473</td>
<td>-2.052</td>
<td>-1.703</td>
<td>0</td>
<td>1.703</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
</tr>
<tr>
<td>28</td>
<td>-2.763</td>
<td>-2.467</td>
<td>-2.048</td>
<td>-1.701</td>
<td>0</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
</tr>
<tr>
<td>29</td>
<td>-2.756</td>
<td>-2.462</td>
<td>-2.045</td>
<td>-1.699</td>
<td>0</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
</tr>
<tr>
<td>Z</td>
<td>-2.576</td>
<td>-2.326</td>
<td>-1.960</td>
<td>-1.645</td>
<td>0</td>
<td>1.645</td>
<td>1.960</td>
<td>2.326</td>
<td>2.576</td>
</tr>
<tr>
<td>CI</td>
<td>90%</td>
<td>95%</td>
<td>99%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.995</td>
<td>0.99</td>
<td>0.975</td>
<td>0.95</td>
<td>0.5</td>
<td>0.05</td>
<td>0.025</td>
<td>0.01</td>
<td>0.005 right tail</td>
</tr>
<tr>
<td></td>
<td>0.01</td>
<td>0.02</td>
<td>0.05</td>
<td>0.10</td>
<td>1</td>
<td>0.10</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01 two tail</td>
</tr>
</tbody>
</table>


Index

Dalgaard, 389
actuaries, vi
Adell, 372
Agresti, 307
alternative hypothesis, 203
Amado, 307
Amer, 351
ancillary statistic, 134
Anderson, 78, 156
Apostol, 262
Arnold, vii
Ash, vii, 22, 262
asymptotic distribution, 228, 240
asymptotic pivot, 274
asymptotic relative efficiency, 237, 259
asymptotic theory, 228
asymptotic variance, 236, 277
asymptotically efficient, 239, 254, 259
Azzalini, vii
Basu, 138
Basu’s Theorem, 134
Bayes’ Theorem, 6
Becker, 389
Berger, vi, 97, 150, 166, 185, 231, 262, 365
Berk, ix, 166, 262
Bernoulli, 318
Berry, vii, 165
Bertsekas, 165
Besbeas, 368
beta distribution, 316
bias, 178, 192
Bibby, 78, 262
Bickel, vi, 57, 262, 296, 387
Bierens, vii
Billingsley, 22
binomial coefficient, 4
binomial distribution, 318
binomial theorem, 316
Birkes, 221
Birnbaum Saunders distribution, 112, 321
bivariate normal, 71
Boes, vii
Boltzmann distribution, 360
Bonferroni’s Inequality, 3
Boole’s Inequality, 3
Boos, 138
Bowman, 329, 338, 387
INDEX

Broffitt, 79
Brown, 110, 262, 307
Brownstein, 307, 387
Buckland, 302
Burr Type III distribution, 322
Burr Type X distribution, 112, 323
Burr Type XII distribution, 324
Byrne, 287, 307

Caffo, 307
Cai, 307
Cambanis, 79
Cartesian product, 36
Casella, vi, 97, 150, 166, 185, 262, 365
Castillo, 387
Cauchy distribution, 325
Cauchy–Schwarz Inequality, 185
cdf, 8, 326
central limit theorem, 237
cf, 11, 66
Chambers, 389
characteristic function, 11, 66
Chebyshev’s Inequality, 243
Chen, 338, 372
chi, 28
chi distribution, 326
chi–square distribution, 328
Chmielewski, 79
CI, 274, 277
Clarke, 379
Cohen, 327, 376, 387
combination, 4
complete sufficient statistic, 125
composite hypothesis, 207
conditional distribution, 71
conditional expected value, 46
conditional pdf, 34
conditional pmf, 34
conditional probability, 4
conditional variance, 46
certainty interval, 274
consistent, 242
consistent estimator, 236, 242, 260
Continuity Theorem, 247
Continuous Mapping Theorem, 246
converges almost everywhere, 245
converges in distribution, 240
converges in law, 240
converges in probability, 242
converges in quadratic mean, 242
Cook, 73, 74
Cooke, 379
Coull, 307
covariance, 43
covariance matrix, 65
Cox, vii, 130, 231, 262
Craig, vii
Cramér, vii, 261, 387
Cramér Rao lower bound, 185
Craven, 379
Crawley, 389
CRLB, 185
cross product, 36
Croux, 76, 335
cumulative distribution function, 8, 326
curved exponential family, 100
DasGupta, 236, 261, 289, 298, 301, 307
Datta, 138, 349
David, viii, 131
Davidson, 261
deCanis, 25
DeGroot, vii, 220, 387
INDEX

Dehon, 76
Dekking, vii
Delta Method, 230
DeMorgan’s Laws, 1
discrete random variable, 17
discrete uniform, 294
discrete uniform distribution, 329
disjoint, 2
Doksum, vi, 57, 262, 296, 387
Doleans-Dade, 22
double exponential distribution, 331
Dudley, 22
Durrett, 22

Eaton, 73, 79
Efron, 302
Ehsanes Saleh, vii
event, 2
expected value, 9, 16, 40, 65
exponential distribution, 333
exponential family, 92

F distribution, 337
Fang, 78
FCRLB, vi, 185
Feller, 22, 368
Ferguson, vii, 247, 261, 387
Fisher, viii, 138
Fisher Information, 182
fixed point algorithm, 291
Fréchet Cramér Rao lower bound, 185
full exponential family, 95
Gabel, 316
gamma distribution, 337
gamma function, 315
Gathwaite, vii
Gaughan, vii
Gaussian distribution, 364
generalized gamma distribution, 82,
340
generalized negative binomial distribution, 341
Gentle, 329, 365, 387
geometric distribution, 342
Gompertz distribution, 343
Gosset, viii
Grübel, 304
Graybill, vii
Greek letters, 315
Grosh, 287
Guenther, 193, 307
Gumbel distribution, 353
Gupta, 78, 110

Hahn, 307
half Cauchy distribution, 344
half logistic distribution, 344
half normal, 282
half normal distribution, 80, 99, 327,
345
Halmos, 138
Hamza, 319
Hanley, viii
Hastings, 387
Headrick, 387
hierarchical model, 48
Hinckley, vii, 130, 231
Hoel, vii
Hoffman, 262
Hogg, vii, 388
INDEX

Huang, 79
Hughes-Oliver, 138
hypergeometric distribution, 347
hypothesis test, 203, 215

iid, 59
incomplete gamma function, 340
increasing function, 207
independent, 5, 36
independent random vectors, 66
indicator function, 12
Information Inequality, 185
information matrix, 254
information number, 182
inverse exponential distribution, 81,
347, 352
inverse Gaussian distribution, 109, 348
inverse Weibull distribution, 81, 142,
351
inverted gamma distribution, 29, 82,
352

Jöckel, 302
Jacobian matrix, 253
Jiang, 261
Jodrá, 372
Johanson, 110, 127
Johnson, 70, 72, 78, 103, 110, 319,
338, 364, 387
joint cdf, 32
joint distribution, 70
joint pdf, 32
joint pmf, 31
Joliffe, vii
Jones, vii
Joshi, 193
Julien, viii
Kabaila, 287, 307

Kalbfleisch, 357
Kapadia, 324, 332, 334, 387
Karakostas, 193
Keener, vii
Kelker, 74
Kemp, 319
Kennedy, 329, 365, 387
Kent, 78, 262
kernel, 13, 95
kernel method, 13
Kiefer, 304
Knight, vii
Koehn, 138
Kolmogorov, 2
Kowalski, 79
Kraaij, vii
kurtosis, 296
Lévy distribution, 368
Ladella, 103
Lancaster, 79
Laplace distribution, 331
largest extreme value distribution, 82,
353
Larsen, vii
Law of Total Probability, 6
Lee, 79
Leemis, 387
Lehmann, vii, 102, 127, 138, 183,
220, 231, 237, 239, 249, 250,
254–256, 261, 305, 368, 387
level α test, 204
Levy, 148
likelihood ratio test, 212
limiting distribution, 229, 240
Lindgren, vii
Lindsey, vii
INDEX

linearity constraint, 94
Liu, 103
location family, 48
location–scale family, 48
log–Cauchy distribution, 82, 356
log–logistic distribution, 82, 356
log–gamma distribution, 29, 81, 142, 356
log–Weibull distribution, 378
logarithmic, 28
logarithmic distribution, 354
logistic distribution, 355
lognormal distribution, 286, 357
Lopuhaä, vii
LRT, 212, 218
Mahalanobis distance, 73
Mahmoud, 351
Mann, 119, 285, 336
Mardia, 78, 262
marginal pdf, 33
marginal pmf, 32
Markov’s Inequality, 243
Marsden, 262
Marshall, 316
Marx, vii
Massart, 303
Mathsoft, 389
maximum likelihood, 332
maximum likelihood estimator, 147
Maxwell distribution, 360
Maxwell–Boltzmann distribution, 82, 327, 360
McCulloch, 262
McKean, vii
McQueston, 387
mean squared error, 178
median, 115
median absolute deviation, 115
Meeker, 165, 307, 360, 387
Meester, vii
Melnick, 79
Mendenhall, vii
method of moments estimator, 161
mgf, 10, 66
minimal sufficient statistic, 125
mixture distribution, 16
MLE, vi, 147, 238
MLR, 219
MME, 320
modified DeMoivre’s law distribution, 361
moment generating function, 10, 66
monotone likelihood ratio, 219
Mood, vii
Moodie, viii
Moore, 148, 279
Morgan, 368
Morris, 29
MSE, 178
Muirhead, 79
Mukhopadhyay, vi, 23, 87, 88, 110, 199
multinomial theorem, 68
Multivariate Central Limit Theorem, 252
Multivariate Delta Method, 253
multivariate normal, 69, 73, 78
mutually exclusive, 2
Myers, vii
negative binomial distribution, 362
Newton’s method, 291, 307
Neyman, 138
Neyman Pearson Lemma, vi, 205
Ng, 78
normal distribution, 364
null hypothesis, 203

O’Reilly, 112
Olkin, 316
one sided stable distribution, 82, 368
one to one function, 126
order statistics, 115
Owen, 324, 332, 334, 387

p-dimensional Rayleigh distribution, 326
p-value, 279
Pal, 165
Panjer, 78
Pant, 387
parameter space, 12
Pareto, 28
Pareto distribution, 369
Parzen, vii
Pascal distribution, 362
Patel, 324, 332, 334, 387
Pawitan, 165
density function, 8, 316
Peacock, 387
Pearson, viii
Pedrosa De Lima, 261
Pensky, 307, 387
Peressini, 151, 164, 307
Perlman, 262
permutation, 3
Pewsey, 156, 283
Pfanzagl, 221, 262
Pires, 307
pivot, 274
probability density function, 8, 316
Poisson distribution, 372
Polansky, 261
Poor, vii
population correlation, 71
population mean, 65
Port, vii
Portnoy, 193, 262
Pourahmadi, 318, 372
power distribution, 286, 373
power function, 203
Pratt, 329
Prentice, 357
Press, 78
probability density function, 8, 316
probability mass function, 8, 317
profile likelihood function, 151
quadratic formula, 404
R, 151, 388
Rahman, 110
random sample, 59
random variable, 7
random vector, 64
Rao, vii, 69
Rayleigh distribution, 83, 327, 376
REF, 95
regular exponential family, 95
rejection region, 203
Resnick, 22
restricted MLE, 212
Rice, vii, 47, 116
Ripley, 389
Roberts, 316
Robinson, 303
Rohatgi, vii, 72, 221, 247, 262
Romano, vii, 220
Rosenlicht, vii
INDEX

Ross, vii
Roussas, vii
Rousseeuw, 76, 335
Rubin, 338, 372
Rudin, vii
Rueda, 112
rule of thumb, 130
RV, 7
Samanta, 349
sample mean, 228
sample space, 2, 12, 32
sampling distribution, 114
Samson, 138
Sankaran, 110
Sarker, 138
Savage, viii, 138
scale family, 48
scaled asymptotic length, 275
Schafer, 119, 285, 336
Scheaffer, vii
Scheffé, 138
Schervish, vii, 220, 231, 262, 387
Schwarz, 349
Scott, 220
SE, 228
Searle, 256
Seber, 79
Sehr, 110
Sen, 260–262, 305
Serfling, 258, 261, 303
Shao, vii, 305
Sheather, 237
Sheng, 387
Shenton, 329, 338, 387
Siegel, vii
Silvey, vii
Simons, 79
simple hypothesis, 207
Singer, 260–262, 305
Singpurwalla, 119, 285, 336
sipack, 388
size α test, 204
SLLN, 245
Slutsky’s Theorem, 245
smallest extreme value distribution, 83, 378
Solomen, 221
Spanos, vii
Spencer, 138
spherical, 73
Spiegel, vii, 112
standard deviation, 9, 115
standard error, 228, 236
statistic, 114
Staudte, 237
Stein, 365
Stigler, viii, 138
Stine, 25
Stone, vii
Strong Law of Large Numbers, 245
Student’s t distribution, 379
sufficient statistic, 121
Sullivan, 151, 164, 307
Sultan, 351
Sundaram, 165
support, 12, 32
symmetric distribution, 52
symmetrically trimmed mean, 279
Tanis, vii
Tenebien, 79
test function, 205, 208
Thom’s estimator, 338
Thomas, 138
Tibshirani, 302
Topp–Leone distribution, 55, 112, 379
truncated exponential, 112
truncated extreme value distribution, 381
Tucker, 22
two parameter power distribution, 81, 286, 375
type I error, 203
type II error, 203
Uhl, 151, 165, 307
UMP, 204
UMP test, 204
UMVUE, vi, 181, 192
unbiased estimator, 178, 192
uniform, 295
uniform distribution, 382
uniformly minimum variance unbiased estimator, 181
unrestricted MLE, 212
Van Aelst, 76
van der Vaart, 261
van Dorp, 112, 387
Vardeman, 307
Varga, 78
variance, 9, 115
Venables, 389
Wackerly, vii
Wade, 262
Wald, 262
Wald distribution, 348
Walpole, vii
Wasserman, vii
Weak Law of Large Numbers, 245
Weibull distribution, 253, 383
Welch, 278
Welch intervals, 278
Welsh, vii
White, 258, 261
Whitten, 327, 376, 387
Wichern, 70, 78
Wijsman, 193
Wilks, 389
Wilson–Hilferty approximation, 329, 339
WLLN, 245
Ye, vii
Yuen, 278
Zabell, viii
Zacks, vii
Zehna, ix, 138, 166
zero truncated Poisson distribution, 30, 385
Zeta distribution, 385
Zipf distribution, 30, 142, 386