-

Kjell Doksum

Mathematical Statistics: Basic Ideas and Selected Topics

http://www.stat.wisc.edu/~doksum/

Lecture Notes for Statistics 709, Fall 2008 - made by Jun Shao and Henry Zhang

Lecture 1: Measurable space, measure and probability
Lecture 2: Product measure, measurable function and distribution
Lecture 3: Integration
Lecture 4: Convergence theorems, change of variable, and Fubini's theorem
Lecture 5: Radon-Nikodym derivative
Lecture 6: p.d.f. and transformation
Lecture 7: Moments, inequalities, m.g.f. and ch.f.
Lecture 8: Conditional expectation
Lecture 9: Independence, conditional independence, conditional distribution
Lecture 10: Markov chains
Lecture 11: Convergence modes and stochastic orders
Lecture 12: Relationships and Uniform Integrabilty
Lecture 13: Wesk Convergence
Lecture 14: Transformations, Slutsky
Lecture 15: LLN; Lecture 16: CLT; Lecture 17: Samples
Lecture 18; Exponential Familes
Lecture 19: SuffStats, Factorization
Lecture 20: Minimal Sufficiency
Lecture 21: Complete statistics
Lecture 22: Decision rules, loss, and risk
Lecture 23: Sufficiency and Rao-Blackwell theorem, unbiasedness and invariance
Lecture 24: Bayes rules, minimax rules, point estimators, and hypothesis tests
Lecture 25: p-value, randomized tests, and confidence sets
Lecture 26: Asymptotic approach and consistency
Lecture 27: Asymptotic bias, variance, and mse
Lecture 28: Asymptotic inference
Lecture 29: UMVUE and the method of using the distribution of a sufficient and complete statistic
Lecture 30: UMVUE: the method of conditioning
Lecture 31: UMVUE: a necessary and sufficient condition
Lecture 32: Information inequality
Lecture 33: U-statistics and their variances
Lecture 34: The projection method
Lecture 35: The LSE and estimability
Lecture 36: The UMVUE and BLUE
Lecture 37: Robustness of LSE's
Lecture 38: Asymptotic properties of LSE's
Lecture 39: The method of moments
Lecture 40: V-statistics and the weighted LSE

Lecture 1: Measurable space, measure and probability

Random experiment: uncertainty in outcomes

Ω: sample space or outcome space; a set containing all possible outcomes

Definition 1.1. Let $\mathcal{F}$ be a collection of subsets of a sample space Ω. $\mathcal{F}$ is called a σ-field (or σ-algebra) if and only if it has the following properties.
(i) The empty set $\emptyset \in \mathcal{F}$.
(ii) If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$, then their union $\cup A_i \in \mathcal{F}$.

$\mathcal{F}$ is a set of sets
Two trivial examples: $\mathcal{F}$ contains $\emptyset$ and Ω only and $\mathcal{F}$ contains all subsets of Ω
Why do we need to consider other σ-field?
$\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$, where $A \subset \Omega$

$\mathcal{C}$ = a collection (set) of subsets of Ω
$\sigma(\mathcal{C})$: the smallest σ-field containing $\mathcal{C}$ (called the σ-field generated by $\mathcal{C}$)
$\sigma(\mathcal{C}) = \mathcal{C}$ if $\mathcal{C}$ itself is a σ-field
$\Gamma = \{\mathcal{F} : \mathcal{F}$ is a σ-field on Ω and $\mathcal{C} \subset \mathcal{F}\}$
$\sigma(\mathcal{C}) = \cap_{\mathcal{F} \in \Gamma} \mathcal{F}$

$\sigma(\{A\}) = \sigma(\{A, A^c\}) = \sigma(\{A, \Omega\}) = \sigma(\{A, \emptyset\}) = \{\emptyset, A, A^c, \Omega\}$

$\mathcal{R}^k$: the k-dimensional Euclidean space ($\mathcal{R}^1 = \mathcal{R}$ is the real line)
$\mathcal{B}^k$: the Borel σ-field on $\mathcal{R}^k$; $\mathcal{B}^k = \sigma(\mathcal{O})$, $\mathcal{O}$ is the collection of all open sets
$C \in \mathcal{B}^k$, $\mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$ is the Borel σ-field on C

Measure: length, area, volume...

Definition 1.2. Let $(\Omega, \mathcal{F})$ be a measurable space. A set function ν defined on $\mathcal{F}$ is called a *measure* if and only if it has the following properties.
(i) $0 \leq \nu(A) \leq \infty$ for any $A \in \mathcal{F}$.
(ii) $\nu(\emptyset) = 0$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$, and A_i's are disjoint, i.e., $A_i \cap A_j = \emptyset$ for any $i \neq j$, then

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \nu(A_i).$$

$(\Omega, \mathcal{F})$ a measurable space; $(\Omega, \mathcal{F}, \nu)$ a measure space

If $\nu(\Omega) = 1$, then ν is a probability measure (we usually use notation P instead of ν)

A measure ν may take ∞ as its value
(1) For any $x \in \mathcal{R}$, $\infty + x = \infty$, $x\,\infty = \infty$ if $x > 0$, $x\,\infty = -\infty$ if $x < 0$, and $0\,\infty = 0$;
(2) $\infty + \infty = \infty$;
(3) $\infty^a = \infty$ for any $a > 0$;
(4) $\infty - \infty$ or ∞/∞ is not defined

Examples:

$$\nu(A) = \begin{cases} \infty & A \in \mathcal{F}, A \neq \emptyset \\ 0 & A = \emptyset. \end{cases}$$

Counting measure. Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets, and $\nu(A)$ the number of elements in $A \in \mathcal{F}$ ($\nu(A) = \infty$ if A contains infinitely many elements). Then ν is a measure on $\mathcal{F}$ and is called the *counting measure*.
Lebesgue measure. There is a unique measure m on $(\mathcal{R}, \mathcal{B})$ that satisfies $m([a,b]) = b - a$ for every finite interval $[a,b]$, $-\infty < a \leq b < \infty$. This is called the *Lebesgue measure*. If we restrict m to the measurable space $([0,1], \mathcal{B}_{[0,1]})$, then m is a probability measure.

Proposition 1.1. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space.
(i) (Monotonicity). If $A \subset B$, then $\nu(A) \leq \nu(B)$.
(ii) (Subadditivity). For any sequence $A_1, A_2, ...,$

$$\nu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \nu(A_i).$$

(iii) (Continuity). If $A_1 \subset A_2 \subset A_3 \subset \cdots$ (or $A_1 \supset A_2 \supset A_3 \supset \cdots$ and $\nu(A_1) < \infty$), then

$$\nu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \nu\left(A_n\right),$$

where

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i \quad \left(\text{ or } = \bigcap_{i=1}^{\infty} A_i\right).$$

Let P be a probability measure. The *cumulative distribution function* (c.d.f.) of P is defined to be

$$F(x) = P\left((-\infty, x]\right), \quad x \in \mathcal{R}$$

Proposition 1.2. (i) Let F be a c.d.f. on $\mathcal{R}$. Then
(a) $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$;
(b) $F(\infty) = \lim_{x\to\infty} F(x) = 1$;
(c) F is nondecreasing, i.e., $F(x) \leq F(y)$ if $x \leq y$;
(d) F is right continuous, i.e., $\lim_{y\to x, y>x} F(y) = F(x)$.
(ii) Suppose that a real-valued function F on $\mathcal{R}$ satisfies (a)-(d) in part (i). Then F is the c.d.f. of a unique probability measure on $(\mathcal{R}, \mathcal{B})$.

Lecture 2: Product measure, measurable function and distribution

Product space
$\mathcal{I} = \{1, ..., k\}$, k is finite or ∞
Γ_i, $i \in \mathcal{I}$, are sets
$\prod_{i\in\mathcal{I}} \Gamma_i = \Gamma_1 \times \cdots \times \Gamma_k = \{(a_1, ..., a_k) : a_i \in \Gamma_i, i \in \mathcal{I}\}$
$\mathcal{R} \times \mathcal{R} = \mathcal{R}^2$, $\mathcal{R} \times \mathcal{R} \times \mathcal{R} = \mathcal{R}^3$

Let $(\Omega_i, \mathcal{F}_i)$, $i \in \mathcal{I}$, be measurable spaces
$\prod_{i\in\mathcal{I}} \mathcal{F}_i$ is not necessarily a σ-field
$\sigma\left(\prod_{i\in\mathcal{I}} \mathcal{F}_i\right)$ is called the *product σ-field* on the *product space* $\prod_{i\in\mathcal{I}} \Omega_i$
$\left(\prod_{i\in\mathcal{I}} \Omega_i, \sigma\left(\prod_{i\in\mathcal{I}} \mathcal{F}_i\right)\right)$ is denoted by $\prod_{i\in\mathcal{I}}(\Omega_i, \mathcal{F}_i)$

Example: $\prod_{i=1,...,k}(\mathcal{R}, \mathcal{B}) = (\mathcal{R}^k, \mathcal{B}^k)$

Product measure

Consider a rectangle $[a_1, b_1] \times [a_2, b_2] \subset \mathcal{R}^2$. The usual area of $[a_1, b_1] \times [a_2, b_2]$ is

$$(b_1 - a_1)(b_2 - a_2) = m([a_1, b_1])m([a_2, b_2])$$

Is $m([a_1, b_1])m([a_2, b_2])$ the same as the value of a measure defined on the product σ-field?

A measure ν on $(\Omega, \mathcal{F})$ is said to be *σ-finite* if and only if there exists a sequence $\{A_1, A_2, ...\}$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i
Any finite measure (such as a probability measure) is clearly σ-finite
The Lebesgue measure on $\mathcal{R}$ is σ-finite, since $\mathcal{R} = \cup A_n$ with $A_n = (-n, n)$, $n = 1, 2, ...$
The counting measure in is σ-finite if and only if Ω is countable

Proposition 1.3 (Product measure theorem). Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, be measure spaces with σ-finite measures, where $k \geq 2$ is an integer. Then there exists a unique σ-finite measure on the product σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$, called the *product measure* and denoted by $\nu_1 \times \cdots \times \nu_k$, such that

$$\nu_1 \times \cdots \times \nu_k(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu_k(A_k)$$

for all $A_i \in \mathcal{F}_i$, $i = 1, ..., k$.

Let P be a probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$. The c.d.f. (or *joint* c.d.f.) of P is defined by

$$F(x_1, ..., x_k) = P\left((-\infty, x_1] \times \cdots \times (-\infty, x_k]\right), \quad x_i \in \mathcal{R}$$

There is a one-to-one correspondence between probability measures and joint c.d.f.'s on $\mathcal{R}^k$

If $F(x_1, ..., x_k)$ is a joint c.d.f., then

$$F_i(x) = \lim_{x_j\to\infty, j=1,...,i-1,i+1,...,k} F(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_k)$$

is a c.d.f. and is called the ith *marginal* c.d.f.

Marginal c.d.f.'s are determined by their joint c.d.f.
But a joint c.d.f. cannot be determined by k marginal c.d.f.'s.
If

$$F(x_1, ..., x_k) = F_1(x_1) \cdots F_k(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k,$$

then the probability measure corresponding to F is the product measure $P_1 \times \cdots \times P_k$ with P_i being the probability measure corresponding to F_i

Measurable function

f: a function from Ω to Λ (often $\Lambda = \mathcal{R}^k$)
Inverse image of $B \subset \Lambda$ under f:

$$f^{-1}(B) = \{f \in B\} = \{\omega \in \Omega : f(\omega) \in B\}.$$

The inverse function f^{-1} need not exist for $f^{-1}(B)$ to be defined.
$f^{-1}(B^c) = (f^{-1}(B))^c$ for any $B \subset \Lambda$;
$f^{-1}(\cup B_i) = \cup f^{-1}(B_i)$ for any $B_i \subset \Lambda$, $i = 1, 2, ...$
Let $\mathcal{C}$ be a collection of subsets of Λ. Define $f^{-1}(\mathcal{C}) = \{f^{-1}(C) : C \in \mathcal{C}\}$

Definition 1.3. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and f a function from Ω to Λ. The function f is called a *measurable function* from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if $f^{-1}(\mathcal{G}) \subset \mathcal{F}$.

If f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then $f^{-1}(\mathcal{G})$ is a sub-σ-field of $\mathcal{F}$ (verify). It is called the σ-field generated by f and is denoted by $\sigma(f)$.

If f is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}, \mathcal{B})$, it is called a Borel function or a random variable
A random vector $(X_1, ..., X_n)$ is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}^n, \mathcal{B}^n)$ (each X_i is a random variable)

Examples
If $\mathcal{F}$ is the collection of all subsets of Ω, then any function f is measurable
Indicator function for $A \subset \Omega$:

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

For any $B \subset \mathcal{R}$,

$$I_A^{-1}(B) = \begin{cases} \emptyset & 0 \notin B, 1 \notin B \\ A & 0 \notin B, 1 \in B \\ A^c & 0 \in B, 1 \notin B \\ \Omega & 0 \in B, 1 \in B. \end{cases}$$

Then, $\sigma(I_A) = \{\emptyset, A, A^c, \Omega\}$ and I_A is Borel iff $A \in \mathcal{F}$
$\sigma(f)$ is much simpler than $\mathcal{F}$

Simple function

$$\varphi(\omega) = \sum_{i=1}^{k} a_i I_{A_i}(\omega),$$

where $A_1, ..., A_k$ are measurable sets on Ω and $a_1, ..., a_k$ are real numbers. Let $A_1, ..., A_k$ be a partition of Ω, i.e., A_i's are disjoint and $A_1 \cup \cdots \cup A_k = \Omega$. Then the simple function φ with distinct a_i's exactly characterizes this partition and $\sigma(\varphi) = \sigma(\{A_1, ..., A_k\})$.

Proposition 1.4. Let $(\Omega, \mathcal{F})$ be a measurable space.
(i) f is Borel if and only if $f^{-1}(a, \infty) \in \mathcal{F}$ for all $a \in \mathcal{R}$.
(ii) If f and g are Borel, then so are fg and $af + bg$, where a and b are real numbers; also, f/g is Borel provided $g(\omega) \neq 0$ for any $\omega \in \Omega$.
(iii) If $f_1, f_2, ...$ are Borel, then so are $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, and $\liminf_n f_n$. Furthermore, the set

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} f_n(\omega) \text{ exists}\right\}$$

is an event and the function

$$h(\omega) = \begin{cases} \lim_{n\to\infty} f_n(\omega) & \omega \in A \\ f_1(\omega) & \omega \notin A \end{cases}$$

is Borel.
(iv) Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and g is measurable from $(\Lambda, \mathcal{G})$ to $(\Delta, \mathcal{H})$. Then the composite function $g \circ f$ is measurable from $(\Omega, \mathcal{F})$ to $(\Delta, \mathcal{H})$.
(v) Let Ω be a Borel set in $\mathcal{R}^p$. If f is a continuous function from Ω to $\mathcal{R}^q$, then f is measurable.

Distribution (law)

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. The *induced measure* by f, denoted by $\nu \circ f^{-1}$, is a measure on $\mathcal{G}$ defined as

$$\nu \circ f^{-1}(B) = \nu(f \in B) = \nu\left(f^{-1}(B)\right), \quad B \in \mathcal{G}$$

If $\nu = P$ is a probability measure and X is a random variable or a random vector, then $P \circ X^{-1}$ is called the *law* or the *distribution* of X and is denoted by P_X.
The c.d.f. of P_X is also called the c.d.f. or joint c.d.f. of X and is denoted by F_X.

Examples 1.3 and 1.4

Lecture 3: Integration

Integration is a type of "average".

Definition 1.4
(a) The integral of a nonnegative simple function φ w.r.t. ν is defined as

$$\int \varphi d\nu = \sum_{i=1}^{k} a_i \nu(A_i).$$

(b) Let f be a nonnegative Borel function and let $\mathcal{S}_f$ be the collection of all nonnegative simple functions satisfying $\varphi(\omega) \leq f(\omega)$ for any $\omega \in \Omega$. The integral of f w.r.t. ν is defined as

$$\int f d\nu = \sup \left\{ \int \varphi d\nu : \ \varphi \in \mathcal{S}_f \right\}.$$

(Hence, for any Borel function $f \geq 0$, there exists a sequence of simple functions $\varphi_1, \varphi_2, ...$ such that $0 \leq \varphi_i \leq f$ for all i and $\lim_{n\to\infty} \int \varphi_n d\nu = \int f d\nu$.)
(c) Let f be a Borel function,

$$f_+(\omega) = \max\{f(\omega), 0\}$$

be the positive part of f, and

$$f_-(\omega) = \max\{-f(\omega), 0\}$$

be the negative part of f. (Note that f_+ and f_- are nonnegative Borel functions, $f(\omega) = f_+(\omega) - f_-(\omega)$, and $|f(\omega)| = f_+(\omega) + f_-(\omega)$.) We say that $\int f d\nu$ exists if and only if at least one of $\int f_+ d\nu$ and $\int f_- d\nu$ is finite, in which case

$$\int f d\nu = \int f_+ d\nu - \int f_- d\nu.$$

When both $\int f_+ d\nu$ and $\int f_- d\nu$ are finite, we say that f is integrable. Let A be a measurable set and I_A be its indicator function. The integral of f over A is defined as

$$\int_A f d\nu = \int I_A f d\nu.$$

A Borel function f is integrable if and only if $|f|$ is integrable.

For convenience, we define the integral of a measurable function f from $(\Omega, \mathcal{F}, \nu)$ to $(\bar{\mathcal{R}}, \bar{\mathcal{B}})$, where $\bar{\mathcal{R}} = \mathcal{R} \cup \{-\infty, \infty\}$, $\bar{\mathcal{B}} = \sigma(\mathcal{B} \cup \{\{\infty\}, \{-\infty\}\})$. Let $A_+ = \{f = \infty\}$ and $A_- = \{f = -\infty\}$. If $\nu(A_+) = 0$, we define $\int f_+ d\nu$ to be $\int I_{A_+^c} f_+ d\nu$; otherwise $\int f_+ d\nu = \infty$. $\int f_- d\nu$ is similarly defined. If at least one of $\int f_+ d\nu$ and $\int f_- d\nu$ is finite, then $\int f d\nu = \int f_+ d\nu - \int f_- d\nu$ is well defined.

Notation for integrals
$\int f d\nu = \int_\Omega f d\nu = \int f(\omega) d\nu = \int f(\omega) d\nu(\omega) = \int f(\omega)\nu(d\omega)$.
In probability and statistics, $\int X dP = EX = E(X)$ and is called the *expectation* or *expected value* of X.
If F is the c.d.f. of P on $(\mathcal{R}^k, \mathcal{B}^k)$, $\int f(x) dP = \int f(x) dF(x) = \int f dF$.

Example 1.5. Let Ω be a countable set, $\mathcal{F}$ be all subsets of Ω, and ν be the counting measure For any Borel function f,

$$\int f d\nu = \sum_{\omega \in \Omega} f(\omega).$$

Example 1.6. If $\Omega = \mathcal{R}$ and ν is the Lebesgue measure, then the Lebesgue integral of f over an interval $[a, b]$ is written as $\int_{[a,b]} f(x) dx = \int_a^b f(x) dx$, which agrees with the Riemann integral in calculus when the latter is well defined. However, there are functions for which the Lebesgue integrals are defined but not the Riemann integrals.

Properties

Proposition 1.5 (Linearity of integrals). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel functions.
(i) If $\int f d\nu$ exists and $a \in \mathcal{R}$, then $\int (af) d\nu$ exists and is equal to $a \int f d\nu$.
(ii) If both $\int f d\nu$ and $\int g d\nu$ exist and $\int f d\nu + \int g d\nu$ is well defined, then $\int (f + g) d\nu$ exists and is equal to $\int f d\nu + \int g d\nu$.

A statement holds a.e. ν (or simply a.e.) if it holds for all ω in N^c with $\nu(N) = 0$. If ν is a probability, then a.e. may be replaced by a.s.

Proposition 1.6. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel.
(i) If $f \le g$ a.e., then $\int f d\nu \le \int g d\nu$, provided that the integrals exist.
(ii) If $f \ge 0$ a.e. and $\int f d\nu = 0$, then $f = 0$ a.e.
Proof. (i) Exercise.
(ii) Let $A = \{f > 0\}$ and $A_n = \{f \ge n^{-1}\}$, $n = 1, 2, ...$ Then $A_n \subset A$ for any n and $\lim_{n\to\infty} A_n = \cup A_n = A$ (why?). By Proposition 1.1(iii), $\lim_{n\to\infty} \nu(A_n) = \nu(A)$. Using part (i) and Proposition 1.5, we obtain that

$$n^{-1}\nu(A_n) = \int n^{-1} I_{A_n} d\nu \le \int f I_{A_n} d\nu \le \int f d\nu = 0$$

for any n. Hence $\nu(A) = 0$ and $f = 0$ a.e.

Consequences:
$|\int f d\nu| \le \int |f| d\nu$
If $f \ge 0$ a.e., then $\int f d\nu \ge 0$
If $f = g$ a.e., then $\int f d\nu = \int g d\nu$.

Lecture 4: Convergence theorems, change of variable, and Fubini's theorem

$\{f_n : n = 1, 2, ...\}$: a sequence of Borel functions. Can we exchange the limit and integration, i.e.,

$$\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu?$$

Example 1.7. Consider $(\mathcal{R}, \mathcal{B})$ and the Lebesgue measure. Define $f_n(x) = nI_{[0,n^{-1}]}(x)$, $n = 1, 2, ...$ Then $\lim_{n\to\infty} f_n(x) = 0$ for all x but $x = 0$. Since the Lebesgue measure of a single point set is 0, $\lim_{n\to\infty} f_n(x) = 0$ a.e. and $\int \lim_{n\to\infty} f_n(x)dx = 0$. On the other hand, $\int f_n(x)dx = 1$ for any n and, hence, $\lim_{n\to\infty} \int f_n(x)dx = 1$.

Sufficient conditions

Theorem 1.1. Let $f_1, f_2, ...$ be a sequence of Borel functions on $(\Omega, \mathcal{F}, \nu)$.
(i) (Fatou's lemma). If $f_n \geq 0$, then $\int \liminf_n f_n d\nu \leq \liminf_n \int f_n d\nu$.
(ii) (Dominated convergence theorem). If $\lim_{n\to\infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then $\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu$.
(iii) (Monotone convergence theorem). If $0 \leq f_1 \leq f_2 \leq \cdots$ and $\lim_{n\to\infty} f_n = f$ a.e., then $\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu$.
Proof. (See the textbook).

Note
(a) To apply each part of the theorem, you need to check the conditions.
(b) If the conditions are not satisfied, you cannot apply the theorem, but it does not imply that you cannot exchange the limit and integration.

Example: Let $f_n(x) = \frac{n}{x+n}$, $x \in \Omega = [0, 1]$, $n = 1, 2, ...$ Then $\lim_n f_n(x) = 1$. To apply the DCT, note that $0 \leq f_n(x) \leq 1$. To apply the MCT, note that $0 \leq f_n(x) \leq f_{n+1}(x)$. Hence, $\lim_n \int f_n(x)dx = \int \lim_n f_n(x)dx = \int dx = 1$.

Example 1.8 (Interchange of differentiation and integration). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and, for any fixed $\theta \in \mathcal{R}$, let $f(\omega, \theta)$ be a Borel function on Ω. Suppose that $\partial f(\omega, \theta)/\partial\theta$ exists a.e. for $\theta \in (a, b) \subset \mathcal{R}$ and that $|\partial f(\omega, \theta)/\partial\theta| \leq g(\omega)$ a.e., where g is an integrable function on Ω. Then, for each $\theta \in (a, b)$, $\partial f(\omega, \theta)/\partial\theta$ is integrable and, by Theorem 1.1(ii),

$$\frac{d}{d\theta} \int f(\omega, \theta)d\nu = \int \frac{\partial f(\omega, \theta)}{\partial\theta} d\nu.$$

Theorem 1.2 (Change of variables). Let f be measurable from $(\Omega, \mathcal{F}, \nu)$ to $(\Lambda, \mathcal{G})$ and g be Borel on $(\Lambda, \mathcal{G})$. Then

$$\int_\Omega g \circ f d\nu = \int_\Lambda g d(\nu \circ f^{-1}),$$

i.e., if either integral exists, then so does the other, and the two are the same.

For Riemann integrals, $\int g(y)dy = \int g(f(x))f'(x)dx$, $y = f(x)$.

For a random variable X on $(\Omega, \mathcal{F}, P)$, $EX = \int_\Omega X dP = \int_\mathcal{R} x dP_X$, $P_X = P \circ X^{-1}$
Let Y be a random vector from Ω to $\mathcal{R}^k$ and g be Borel from $\mathcal{R}^k$ to $\mathcal{R}$.

$Eg(Y) = \int_{\mathcal{R}} x dP_{g(Y)} = \int_{\mathcal{R}^k} g(y) dP_Y$
Example: $Y = (X_1, X_2)$ and $g(Y) = X_1 + X_2$.
$E(X_1 + X_2) = EX_1 + EX_2$ (why?) $= \int_{\mathcal{R}} x dP_{X_1} + \int_{\mathcal{R}} x dP_{X_2}$.
We need to handle two integrals involving P_{X_1} and P_{X_2}. On the other hand, $E(X_1 + X_2) = \int_{\mathcal{R}} x dP_{X_1+X_2}$, which involves one integral w.r.t. $P_{X_1+X_2}$. Unless we have some knowledge about the joint c.d.f. of (X_1, X_2), it is not easy to obtain $P_{X_1+X_2}$.

Iterated integration on a product space

Theorem 1.3 (Fubini's theorem). Let ν_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$, $i = 1, 2$, and let f be a Borel function on $\prod_{i=1}^2 (\Omega_i, \mathcal{F}_i)$ whose integral w.r.t. $\nu_1 \times \nu_2$ exists. Then

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1$$

exists a.e. ν_2 and defines a Borel function on Ω_2 whose integral w.r.t. ν_2 exists, and

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2) d\nu_1 \times \nu_2 = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2) d\nu_1 \right] d\nu_2.$$

Note: If $f \geq 0$, then $\int f d\nu_1 \times \nu_2$ always exists. Extensions to $\prod_{i=1}^k (\Omega_i, \mathcal{F}_i)$ is straightforward.

Fubini's theorem is *very useful* in
(1) evaluating multi-dimensional integrals (exchanging the order of integrals);
(2) proving a function is measurable;
(3) proving some results by relating a one dimensional integral to a multi-dimensional integral

Example: Exercise 47
Let X and Y be random variables such that the joint c.d.f. of (X, Y) is $F_X(x)F_Y(y)$, where F_X and F_Y are marginal c.d.f.'s. Let $Z = X + Y$. Show that

$$F_Z(z) = \int F_Y(z - x) dF_X(x).$$

Note that

$$\begin{aligned} F_Z(z) &= \int_{x+y \leq z} dF_X(x) dF_Y(y) \\ &= \int \left(\int_{y \leq z-x} dF_Y(y) \right) dF_X(x) \\ &= \int F_Y(z - x) dF_X(x), \end{aligned}$$

where the second equality follows from Fubini's theorem.

Example 1.9. Let $\Omega_1 = \Omega_2 = \{0, 1, 2, ...\}$, and $\nu_1 = \nu_2$ be the counting measure. A function f on $\Omega_1 \times \Omega_2$ defines a double sequence. If $\int f d\nu_1 \times \nu_2$ exists, then

$$\int f d\nu_1 \times \nu_2 = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i, j) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} f(i, j)$$

(by Theorem 1.3 and Example 1.5). Thus, a double series can be summed in either order, if it is well defined.

Proof of Fubini's theorem

Lecture 5: Radon-Nikodym derivative

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a nonnegative Borel function. Note that

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F}$$

is a measure satisfying

$$\nu(A) = 0 \quad \text{implies} \quad \lambda(A) = 0.$$

(we say λ is *absolutely continuous* w.r.t. ν and write $\lambda \ll \nu$).

Comupting $\lambda(A)$ can be done through integration w.r.t. a well-known measure

$\lambda \ll \nu$ is also almost sufficient.

Theorem 1.4 (Radon-Nikodym theorem). Let ν and λ be two measures on $(\Omega, \mathcal{F})$ and ν be σ-finite. If $\lambda \ll \nu$, then there exists a nonnegative Borel function f on Ω such that

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F}.$$

Furthermore, f is unique a.e. ν, i.e., if $\lambda(A) = \int_A g d\nu$ for any $A \in \mathcal{F}$, then $f = g$ a.e. ν.

The function f is called the Radon-Nikodym *derivative* or *density* of λ w.r.t. ν and is denoted by $d\lambda/d\nu$.

Consequence: If f is Borel on $(\Omega, \mathcal{F})$ and $\int_A f d\nu = 0$ for any $A \in \mathcal{F}$, then $f = 0$ a.e.

If $\int f d\nu = 1$ for an $f \geq 0$ a.e. ν, then λ is a probability measure and f is called its *probability density function* (p.d.f.) w.r.t. ν. For any probability measure P on $(\mathcal{R}^k, \mathcal{B}^k)$ corresponding to a c.d.f. F or a random vector X, if P has a p.d.f. f w.r.t. a measure ν, then f is also called the p.d.f. of F or X w.r.t. ν.

Example 1.10 (Discrete c.d.f. and p.d.f.). Let $a_1 < a_2 < \cdots$ be a sequence of real numbers and let p_n, $n = 1, 2, ...$, be a sequence of positive numbers such that $\sum_{n=1}^{\infty} p_n = 1$. Then

$$F(x) = \begin{cases} \sum_{i=1}^{n} p_i & a_n \leq x < a_{n+1}, \quad n = 1, 2, ... \\ 0 & -\infty < x < a_1. \end{cases}$$

is a *stepwise* c.d.f. It has a jump of size p_n at each a_n and is flat between a_n and a_{n+1}, $n = 1, 2, ...$ Such a c.d.f. is called a *discrete* c.d.f. The corresponding probability measure is

$$P(A) = \sum_{i: a_i \in A} p_i, \quad A \in \mathcal{F},$$

where $\mathcal{F}$ = the set of all subsets (power set).
Let ν be the counting measure on the power set. Then

$$P(A) = \int_A f d\nu = \sum_{a_i \in A} f(a_i), \quad A \subset \Omega,$$

where $f(a_i) = p_i$, $i = 1, 2, ...$ That is, f is the p.d.f. of P or F w.r.t. ν. Hence, any discrete c.d.f. has a p.d.f. w.r.t. counting measure. A p.d.f. w.r.t. counting measure is called a *discrete* p.d.f.

Example 1.11. Let F be a c.d.f. Assume that F is differentiable in the usual sense in calculus. Let f be the derivative of F. From calculus,

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad x \in \mathcal{R}.$$

Let P be the probability measure corresponding to F.
Then $P(A) = \int_A f dm$ for any $A \in \mathcal{B}$, where m is the Lebesgue measure on $\mathcal{R}$.
f is the p.d.f. of P or F w.r.t. Lebesgue measure.
Radon-Nikodym derivative is the same as the usual derivative in calculus.

A continuous c.d.f. may not have a p.d.f. w.r.t. Lebesgue measure.
A necessary and sufficient condition for a c.d.f. F having a p.d.f. w.r.t. Lebesgue measure is that F is *absolute continuous* in the sense that for any $\epsilon > 0$, there exists a $\delta > 0$ such that for each finite collection of disjoint bounded open intervals (a_i, b_i), $\sum(b_i - a_i) < \delta$ implies $\sum[F(b_i) - F(a_i)] < \epsilon$.
Absolute continuity is weaker than differentiability, but is stronger than continuity.
Note that every c.d.f. is differentiable a.e. Lebesgue measure (Chung, 1974, Chapter 1).

A p.d.f. w.r.t. Lebesgue measure is called a Lebesgue p.d.f.

Proposition 1.7 (Calculus with Radon-Nikodym derivatives). Let ν be a σ-finite measure on a measure space $(\Omega, \mathcal{F})$. All other measures discussed in (i)-(iii) are defined on $(\Omega, \mathcal{F})$.
(i) If λ is a measure, $\lambda \ll \nu$, and $f \geq 0$, then

$$\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu.$$

(Notice how the $d\nu$'s "cancel" on the right-hand side.)
(ii) If λ_i, $i = 1, 2$, are measures and $\lambda_i \ll \nu$, then $\lambda_1 + \lambda_2 \ll \nu$ and

$$\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu} \quad \text{a.e. } \nu.$$

(iii) (Chain rule). If τ is a measure, λ is a σ-finite measure, and $\tau \ll \lambda \ll \nu$, then

$$\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda}\frac{d\lambda}{d\nu} \quad \text{a.e. } \nu.$$

In particular, if $\lambda \ll \nu$ and $\nu \ll \lambda$ (in which case λ and ν are *equivalent*), then

$$\frac{d\lambda}{d\nu} = \left(\frac{d\nu}{d\lambda}\right)^{-1} \quad \text{a.e. } \nu \text{ or } \lambda.$$

(iv) Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be a measure space and ν_i be σ-finite, $i = 1, 2$. Let λ_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$ and $\lambda_i \ll \nu_i$, $i = 1, 2$. Then $\lambda_1 \times \lambda_2 \ll \nu_1 \times \nu_2$ and

$$\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1)\frac{d\lambda_2}{d\nu_2}(\omega_2) \quad \text{a.e. } \nu_1 \times \nu_2.$$

Lecture 6: p.d.f. and transformation

Example 1.12. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ whose c.d.f. F_X has a Lebesgue p.d.f. f_X and $F_X(c) < 1$, where c is a fixed constant. Let $Y = \min\{X, c\}$, i.e., Y is the smaller of X and c. Note that $Y^{-1}((-\infty, x]) = \Omega$ if $x \geq c$ and $Y^{-1}((-\infty, x]) = X^{-1}((\infty, x])$ if $x < c$. Hence Y is a random variable and the c.d.f. of Y is

$$F_Y(x) = \begin{cases} 1 & x \geq c \\ F_X(x) & x < c. \end{cases}$$

This c.d.f. is discontinuous at c, since $F_X(c) < 1$. Thus, it does not have a Lebesgue p.d.f. It is not discrete either. Does P_Y, the probability measure corresponding to F_Y, have a p.d.f. w.r.t. some measure? Define a probability measure on $(\mathcal{R}, \mathcal{B})$, called *point mass* at c, by

$$\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A, \end{cases} \qquad A \in \mathcal{B}$$

Then $P_Y \ll m + \delta_c$, where m is the Lebesgue measure, and the p.d.f. of P_Y is

$$\frac{dP_Y}{d(m+\delta_c)}(x) = \begin{cases} 0 & x > c \\ 1 - F_X(c) & x = c \\ f_X(x) & x < c. \end{cases}$$

Example 1.14. Let X be a random variable with c.d.f. F_X and Lebesgue p.d.f. f_X, and let $Y = X^2$. Since $Y^{-1}((-\infty, x])$ is empty if $x < 0$ and equals $Y^{-1}([0, x]) = X^{-1}([-\sqrt{x}, \sqrt{x}\,])$ if $x \geq 0$, the c.d.f. of Y is

$$\begin{aligned} F_Y(x) &= P \circ Y^{-1}((-\infty, x]) \\ &= P \circ X^{-1}([-\sqrt{x}, \sqrt{x}\,]) \\ &= F_X(\sqrt{x}) - F_X(-\sqrt{x}) \end{aligned}$$

if $x \geq 0$ and $F_Y(x) = 0$ if $x < 0$. Clearly, the Lebesgue p.d.f. of F_Y is

$$f_Y(x) = \frac{1}{2\sqrt{x}}[f_X(\sqrt{x}) + f_X(-\sqrt{x})]I_{(0,\infty)}(x).$$

In particular, if

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

which is the Lebesgue p.d.f. of the standard normal distribution $N(0, 1)$, then

$$f_Y(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2} I_{(0,\infty)}(x),$$

which is the Lebesgue p.d.f. for the chi-square distribution χ^2_1 (Table 1.2). This is actually an important result in statistics.

Proposition 1.8. Let X be a random k-vector with a Lebesgue p.d.f. f_X and let $Y = g(X)$, where g is a Borel function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^k, \mathcal{B}^k)$. Let $A_1, ..., A_m$ be disjoint sets in $\mathcal{B}^k$ such that $\mathcal{R}^k - (A_1 \cup \cdots \cup A_m)$ has Lebesgue measure 0 and g on A_j is one-to-one with a nonvanishing Jacobian, i.e., the determinant $\det(\partial g(x)/\partial x) \neq 0$ on A_j, $j = 1, ..., m$. Then Y has the following Lebesgue p.d.f.:

$$f_Y(x) = \sum_{j=1}^{m} |\det(\partial h_j(x)/\partial x)| f_X(h_j(x)),$$

where h_j is the inverse function of g on A_j, $j = 1, ..., m$.

In Example 1.14, $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$, $g(x) = x^2$, $h_1(x) = -\sqrt{x}$, $h_2(x) = \sqrt{x}$, and $|dh_j(x)/dx| = 1/(2\sqrt{x})$.

Example 1.15. Let $X = (X_1, X_2)$ be a random 2-vector having a joint Lebesgue p.d.f. f_X. Consider first the transformation $g(x) = (x_1, x_1 + x_2)$. Using Proposition 1.8, one can show that the joint p.d.f. of $g(X)$ is

$$f_{g(X)}(x_1, y) = f_X(x_1, y - x_1),$$

where $y = x_1 + x_2$ (note that the Jacobian equals 1). The marginal p.d.f. of $Y = X_1 + X_2$ is then

$$f_Y(y) = \int f_X(x_1, y - x_1) dx_1.$$

In particular, if X_1 and X_2 are independent, then

$$f_Y(y) = \int f_{X_1}(x_1) f_{X_2}(y - x_1) dx_1.$$

Next, consider the transformation $h(x_1, x_2) = (x_1/x_2, x_2)$, assuming that $X_2 \neq 0$ a.s. Using Proposition 1.8, one can show that the joint p.d.f. of $h(X)$ is

$$f_{h(X)}(z, x_2) = |x_2| f_X(zx_2, x_2),$$

where $z = x_1/x_2$. The marginal p.d.f. of $Z = X_1/X_2$ is

$$f_Z(z) = \int |x_2| f_X(zx_2, x_2) dx_2.$$

In particular, if X_1 and X_2 are independent, then

$$f_Z(z) = \int |x_2| f_{X_1}(zx_2) f_{X_2}(x_2) dx_2.$$

Example 1.16 (t-distribution and F-distribution). Let X_1 and X_2 be independent random variables having the chi-square distributions $\chi^2_{n_1}$ and $\chi^2_{n_2}$ (Table 1.2), respectively. The p.d.f.

of $Z = X_1/X_2$ is

$$\begin{aligned} f_Z(z) &= \frac{z^{n_1/2-1} I_{(0,\infty)}(z)}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)} \int_0^\infty x_2^{(n_1+n_2)/2-1} e^{-(1+z)x_2/2} dx_2 \\ &= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\Gamma(n_2/2)} \frac{z^{n_1/2-1}}{(1+z)^{(n_1+n_2)/2}} I_{(0,\infty)}(z) \end{aligned}$$

Using Proposition 1.8, one can show that the p.d.f. of $Y = (X_1/n_1)/(X_2/n_2) = (n_2/n_1)Z$ is the p.d.f. of the F-distribution F_{n_1,n_2} given in Table 1.2.

Let U_1 be a random variable having the standard normal distribution $N(0,1)$ and U_2 a random variable having the chi-square distribution χ^2_n. Using the same argument, one can show that if U_1 and U_2 are independent, then the distribution of $T = U_1/\sqrt{U_2/n}$ is the t-distribution t_n given in Table 1.2.

Noncentral chi-square distribution

Let $X_1, ..., X_n$ be independent random variables and $X_i = N(\mu_i, \sigma^2)$, $i = 1, ..., n$. The distribution of $Y = (X_1^2 + \cdots + X_n^2)/\sigma^2$ is called the *noncentral chi-square* distribution and denoted by $\chi^2_n(\delta)$, where $\delta = (\mu_1^2 + \cdots + \mu_n^2)/\sigma^2$ is the noncentrality parameter.
$\chi^2_k(\delta)$ with $\delta = 0$ is called a *central* chi-square distribution.
It can be shown (exercise) that Y has the following Lebesgue p.d.f.:

$$e^{-\delta/2} \sum_{j=0}^{\infty} \frac{(\delta/2)^j}{j!} f_{2j+n}(x)$$

where $f_k(x)$ is the Lebesgue p.d.f. of the chi-square distribution χ^2_k.
If $Y_1, ..., Y_k$ are independent random variables and Y_i has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$, then $Y = Y_1 + \cdots + Y_k$ has the noncentral chi-square distribution $\chi^2_{n_1+\cdots+n_k}(\delta_1 + \cdots + \delta_k)$.

Noncentral t-distribution and F-distribution (in discussion)

Theorem 1.5. (Cochran's theorem). Suppose that $X = N_n(\mu, I_n)$ and

$$X^\tau X = X^\tau A_1 X + \cdots + X^\tau A_k X,$$

where I_n is the $n \times n$ identity matrix and A_i is an $n \times n$ symmetric matrix with rank n_i, $i = 1, ..., k$. A necessary and sufficient condition that $X^\tau A_i X$ has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$, and $X^\tau A_i X$'s are independent is $n = n_1 + \cdots + n_k$, in which case $\delta_i = \mu^\tau A_i \mu$ and $\delta_1 + \cdots + \delta_k = \mu^\tau \mu$.

Lecture 7: Moments, inequalities, m.g.f. and ch.f.

If EX^k is finite, where k is a positive integer, EX^k is called the kth *moment* of X or P_X.
If $E|X|^a < \infty$ for some real number a, $E|X|^a$ is called the ath *absolute moment* of X or P_X.
If $\mu = EX$ and $E(X-\mu)^k$ are finite for a positive integer k, $E(X-\mu)^k$ is called the kth *central moment* of X or P_X.
Variance: $E(X-EX)^2$

$X = (X_1, ..., X_k)$, $EX = (EX_1, ..., EX_k)$
$M = (M_{ij})$, $EM = (EM_{ij})$
Covariance matrix: $\mathrm{Var}(X) = E(X-EX)(X-EX)^\tau$
The (i,j)th element of $\mathrm{Var}(X)$, $i \neq j$, is $E(X_i - EX_i)(X_j - EX_j)$, which is called the *covariance* of X_i and X_j and is denoted by $\mathrm{Cov}(X_i, X_j)$.
$\mathrm{Var}(X)$ is nonnegative definite
$[\mathrm{Cov}(X_i, X_j)]^2 \leq \mathrm{Var}(X_i)\mathrm{Var}(X_j), \quad i \neq j$
If $\mathrm{Cov}(X_i, X_j) = 0$, then X_i and X_j are uncorrelated
Independence implies uncorrelation, not converse
If $Y = c^\tau X$, $c \in \mathcal{R}^k$, and X is a random k-vector, $EY = c^\tau EX$ and $\mathrm{Var}(Y) = c^\tau \mathrm{Var}(X)c$.

Three useful inequalities
Cauchy-Schwartz inequality: $[E(XY)]^2 \leq EX^2 EY^2$ for random variables X and Y
Jensen's inequality: $f(EX) \leq Ef(X)$ for a random vector X and convex function f ($f'' \geq 0$)
Chebyshev's inequality: Let X be a random variable and φ a nonnegative and nondecreasing function on $[0, \infty)$ satisfying $\varphi(-t) = \varphi(t)$. Then, for each constant $t \geq 0$,

$$\varphi(t) P(|X| \geq t) \leq \int_{\{|X| \geq t\}} \varphi(X) dP \leq E\varphi(X)$$

Example 1.18. If X is a nonconstant positive random variable with finite mean, then

$$(EX)^{-1} < E(X^{-1}) \quad \text{and} \quad E(\log X) < \log(EX),$$

since t^{-1} and $-\log t$ are convex functions on $(0, \infty)$. Let f and g be positive integrable functions on a measure space with a σ-finite measure ν. If $\int f d\nu \geq \int g d\nu > 0$, we want to show that

$$\int f \log\left(\frac{f}{g}\right) d\nu \geq 0.$$

Let $h = f / \int f d\nu$. Then h is a p.d.f. w.r.t. ν. Let $Y = g/f$ be a random variable defined on the probability space with P being the probability with p.d.f. h. By Jensen's inequality, $E\log(g/f) \leq \log(E(g/f))$. Note that

$$\log(E(g/f)) = \log\left(\int \frac{g}{f} h d\nu\right) = \log\left(\frac{\int g d\nu}{\int f d\nu}\right) \leq 0$$

and

$$E\log(g/f) = \int \log\left(\frac{g}{f}\right) h d\nu = \int \log\left(\frac{g}{f}\right) f d\nu \Big/ \int f d\nu$$

Moment generating and characteristic functions

Definition 1.5. Let X be a random k-vector.

(i) The *moment generating function* (m.g.f.) of X or P_X is defined as

$$\psi_X(t) = Ee^{t^\tau X}, \quad t \in \mathcal{R}^k.$$

(ii) The *characteristic function* (ch.f.) of X or P_X is defined as

$$\phi_X(t) = Ee^{\sqrt{-1}t^\tau X} = E[\cos(t^\tau X)] + \sqrt{-1}\, E[\sin(t^\tau X)], \quad t \in \mathcal{R}^k$$

If the m.g.f. is finite in a neighborhood of $0 \in \mathcal{R}^k$, then $\phi_X(t)$ can be obtained by replacing t in $\psi_X(t)$ by $\sqrt{-1}t$

If $Y = A^\tau X + c$, where A is a $k \times m$ matrix and $c \in \mathcal{R}^m$, it follows from Definition 1.5 that

$$\psi_Y(u) = e^{c^\tau u}\psi_X(Au) \quad \text{and} \quad \phi_Y(u) = e^{\sqrt{-1}c^\tau u}\phi_X(Au), \quad u \in \mathcal{R}^m$$

$X = (X_1, ..., X_k)$ with m.g.f. ψ_X finite in a neighborhood of 0

$$\psi_X(t) = \sum_{(r_1,...,r_k)} \frac{\mu_{r_1,...,r_k} t_1^{r_1} \cdots t_k^{r_k}}{r_1! \cdots r_k!} \qquad \mu_{r_1,...,r_k} = E(X_1^{r_1} \cdots X_k^{r_k})$$

Special case of $k = 1$:

$$\psi_X(t) = \sum_{i=0}^{\infty} \frac{E(X^i)t^i}{i!}$$

Consequently,

$$E(X_1^{r_1} \cdots X_k^{r_k}) = \frac{\partial^{r_1+\cdots+r_k}\psi_X(t)}{\partial t_1^{r_1} \cdots \partial t_k^{r_k}}\bigg|_{t=0} \qquad E(X^i) = \psi^{(i)}(0) = \frac{d\psi_X^i(t)}{dt^i}\bigg|_{t=0}$$

$$\frac{\partial \psi_X(t)}{\partial t}\bigg|_{t=0} = EX, \qquad \frac{\partial^2 \psi_X(t)}{\partial t \partial t^\tau}\bigg|_{t=0} = E(XX^\tau)$$

If $0 < \psi_X(t) < \infty$, then $\kappa_X(t) = \log \psi_X(t)$ is called the *cumulant generating function* of X or P_X.

If ψ_X is not finite and $E|X_1^{r_1} \cdots X_k^{r_k}| < \infty$ for some nonnegative integers $r_1, ..., r_k$, then

$$\frac{\partial^{r_1+\cdots+r_k}\phi_X(t)}{\partial t_1^{r_1} \cdots \partial t_k^{r_k}}\bigg|_{t=0} = (-1)^{(r_1+\cdots+r_k)/2} E(X_1^{r_1} \cdots X_k^{r_k})$$

$$\frac{\partial \phi_X(t)}{\partial t}\bigg|_{t=0} = \sqrt{-1}EX, \qquad \frac{\partial^2 \phi_X(t)}{\partial t \partial t^\tau}\bigg|_{t=0} = -E(XX^\tau), \qquad \phi_X^{(i)}(0) = (-1)^{i/2}E(X^i)$$

Example: a random variable X has finite $E(X^k)$ for $k = 1, 2...$ but $\psi_X(t) = \infty$, $t \neq 0$
P_n: the probability measure for $N(0, n)$ with p.d.f. f_n, $n = 1, 2, ...$
$P = \sum_{n=1}^{\infty} 2^{-n} P_n$ is a probability measure with Lebesgue p.d.f. $\sum_{n=1}^{\infty} 2^{-n} f_n$ (Exercise 35)
Let X be a random variable having distribution P.
It follows from Fubini's theorem that X has finite moments of any order; for even k,

$$E(X^k) = \int x^k dP = \int \sum_{n=1}^{\infty} x^k 2^{-n} dP_n = \sum_{n=1}^{\infty} 2^{-n} \int x^k dP_n = \sum_{n=1}^{\infty} 2^{-n}(k-1)(k-3)\cdots 1 n^{k/2} < \infty$$

and $E(X^k) = 0$ for odd k.
By Fubini's theorem,

$$\psi_X(t) = \int e^{tx} dP = \sum_{n=1}^{\infty} 2^{-n} \int e^{tx} dP_n = \sum_{n=1}^{\infty} 2^{-n} e^{nt^2/2} = \infty \quad t \neq 0$$

Since the ch.f. of $N(0, n)$ is $e^{-nt^2/2}$,

$$\phi_X(t) = \int e^{\sqrt{-1}tx} dP = \sum_{n=1}^{\infty} 2^{-n} \int e^{\sqrt{-1}tx} dP_n = \sum_{n=1}^{\infty} 2^{-n} e^{-nt^2/2} = (2e^{t^2/2} - 1)^{-1}$$

(Fubini's theorem)
Hence, the moments of X can be obtained by differentiating ϕ_X
For example, $\phi'_X(0) = 0$ and $\phi''_X(0) = -2$, which shows that $EX = 0$ and $EX^2 = 2$.

Theorem 1.6. (Uniqueness). Let X and Y be random k-vectors.
(i) If $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathcal{R}^k$, then $P_X = P_Y$.
(ii) If $\psi_X(t) = \psi_Y(t) < \infty$ for all t in a neighborhood of 0, then $P_X = P_Y$.

Another useful result: For independent X and Y,

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t) \quad \text{and} \quad \phi_{X+Y}(t) = \phi_X(t)\psi_Y(t), \quad t \in \mathcal{R}^k$$

Example 1.20. Let X_i, $i = 1, ..., k$, be independent random variables and X_i have the gamma distribution $\Gamma(\alpha_i, \gamma)$ (Table 1.2), $i = 1, ..., k$. From Table 1.2, X_i has the m.g.f. $\psi_{X_i}(t) = (1 - \gamma t)^{-\alpha_i}$, $t < \gamma^{-1}$, $i = 1, ..., k$. Then, the m.g.f. of $Y = X_1 + \cdots + X_k$ is equal to $\psi_Y(t) = (1 - \gamma t)^{-(\alpha_1 + \cdots + \alpha_k)}$, $t < \gamma^{-1}$. From Table 1.2, the gamma distribution $\Gamma(\alpha_1 + \cdots + \alpha_k, \gamma)$ has the m.g.f. $\psi_Y(t)$ and, hence, is the distribution of Y (by Theorem 1.6).

A random vector X is symmetric about 0 iff X and $-X$ have the same distribution
Show that: X is symmetric about 0 if and only if its ch.f. ϕ_X is real-valued.
If X and $-X$ have the same distribution, then by Theorem 1.6, $\phi_X(t) = \phi_{-X}(t)$.
But $\phi_{-X}(t) = \phi_X(-t)$. Then $\phi_X(t) = \phi_X(-t)$.
Note that $\sin(-t^\tau X) = -\sin(t^\tau X)$ and $\cos(t^\tau X) = \cos(-t^\tau X)$
Hence $E[\sin(t^\tau X)] = 0$ and, thus, ϕ_X is real-valued.
Conversely, if ϕ_X is real-valued, then $\phi_X(t) = E[\cos(t^\tau X)]$ and $\phi_{-X}(t) = \phi_X(-t) = \phi_X(t)$.
By Theorem 1.6, X and $-X$ must have the same distribution.

Lecture 8: Conditional expectation

Conditional probability $P(B|A) = P(A \cap B)/P(A)$ for events A and B with $P(A) > 0$
$P(X \in B|Y \in A)$
$P(X \in B|Y = y)$?

Definition 1.6. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$.
(i) Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. The *conditional expectation* of X given $\mathcal{A}$, denoted by $E(X|\mathcal{A})$, is the a.s.-unique random variable satisfying the following two conditions:
(a) $E(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathcal{R}, \mathcal{B})$;
(b) $\int_A E(X|\mathcal{A})dP = \int_A XdP$ for any $A \in \mathcal{A}$.
(Note that the existence of $E(X|\mathcal{A})$ follows from Theorem 1.4.)
(ii) Let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$ is defined to be $P(B|\mathcal{A}) = E(I_B|\mathcal{A})$.
(iii) Let Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. The conditional expectation of X given Y is defined to be $E(X|Y) = E[X|\sigma(Y)]$.

$\sigma(Y)$ contains "the information in Y"
$E(X|Y)$ is the "expectation" of X given the information provided by Y

Lemma 1.2. Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. Then Z is measurable from $(\Omega, \sigma(Y))$ to $(\mathcal{R}^k, \mathcal{B}^k)$ if and only if there is a measurable function h from $(\Lambda, \mathcal{G})$ to $(\mathcal{R}^k, \mathcal{B}^k)$ such that $Z = h \circ Y$.

The function h in $E(X|Y) = h \circ Y$ is a Borel function on $(\Lambda, \mathcal{G})$.
Let $y \in \Lambda$. We define

$$E(X|Y = y) = h(y)$$

to be the conditional expectation of X given $Y = y$.
Note that $h(y)$ is a function on Λ, whereas $h \circ Y = E(X|Y)$ is a function on Ω.

For a random vector X, $E(X|\mathcal{A})$ is defined as the vector of conditional expectations of components of X.

Example 1.21. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$, $A_1, A_2, ...$ be disjoint events on $(\Omega, \mathcal{F}, P)$ such that $\cup A_i = \Omega$ and $P(A_i) > 0$ for all i, and let $a_1, a_2, ...$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. We now show that

$$E(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} XdP}{P(A_i)} I_{A_i}.$$

We need to verify (a) and (b) in Definition 1.6 with $\mathcal{A} = \sigma(Y)$.
Since $\sigma(Y) = \sigma(\{A_1, A_2, ...\})$, it is clear that the function on the right-hand side is measurable on $(\Omega, \sigma(Y))$.
For any $B \in \mathcal{B}$, $Y^{-1}(B) = \cup_{i:a_i \in B} A_i$. Using properties of integrals, we obtain that

$$\begin{aligned}\int_{Y^{-1}(B)} X dP &= \sum_{i:a_i \in B} \int_{A_i} X dP \\ &= \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} P\left(A_i \cap Y^{-1}(B)\right) \\ &= \int_{Y^{-1}(B)} \left[\sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i}\right] dP.\end{aligned}$$

This verifies (b) and thus the result.
Let h be a Borel function on $\mathcal{R}$ satisfying $h(a_i) = \int_{A_i} X dP / P(A_i)$.
Then $E(X|Y) = h \circ Y$ and $E(X|Y = y) = h(y)$.

Proposition 1.9. Let X be a random n-vector and Y a random m-vector. Suppose that (X, Y) has a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, where ν and λ are σ-finite measures on $(\mathcal{R}^n, \mathcal{B}^n)$ and $(\mathcal{R}^m, \mathcal{B}^m)$, respectively. Let $g(x, y)$ be a Borel function on $\mathcal{R}^{n+m}$ for which $E|g(X, Y)| < \infty$. Then

$$E[g(X, Y)|Y] = \frac{\int g(x, Y) f(x, Y) d\nu(x)}{\int f(x, Y) d\nu(x)} \quad \text{a.s.}$$

Proof. Denote the right-hand side by $h(Y)$. By Fubini's theorem, h is Borel. Then, by Lemma 1.2, $h(Y)$ is Borel on $(\Omega, \sigma(Y))$. Also, by Fubini's theorem, $f_Y(y) = \int f(x, y) d\nu(x)$ is the p.d.f. of Y w.r.t. λ. For $B \in \mathcal{B}^m$,

$$\begin{aligned}\int_{Y^{-1}(B)} h(Y) dP &= \int_B h(y) dP_Y \\ &= \int_B \frac{\int g(x, y) f(x, y) d\nu(x)}{\int f(x, y) d\nu(x)} f_Y(y) d\lambda(y) \\ &= \int_{\mathcal{R}^n \times B} g(x, y) f(x, y) d\nu \times \lambda \\ &= \int_{\mathcal{R}^n \times B} g(x, y) dP_{(X,Y)} \\ &= \int_{Y^{-1}(B)} g(X, Y) dP,\end{aligned}$$

where the first and the last equalities follow from Theorem 1.2, the second and the next to last equalities follow from the definition of h and p.d.f.'s, and the third equality follows from Theorem 1.3 (Fubini's theorem).

(X, Y): a random vector with a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$
The *conditional* p.d.f. of X given $Y = y$: $f_{X|Y}(x|y) = f(x, y)/f_Y(y)$
$f_Y(y) = \int f(x, y) d\nu(x)$ is the marginal p.d.f. of Y w.r.t. λ.
For each fixed y with $f_Y(y) > 0$, $f_{X|Y}(x|y)$ is a p.d.f. w.r.t. ν.
Then Proposition 1.9 states that

$$E[g(X, Y)|Y] = \int g(x, Y) f_{X|Y}(x|Y) d\nu(x)$$

i.e., the conditional expectation of $g(X, Y)$ given Y is equal to the expectation of $g(X, Y)$ w.r.t. the conditional p.d.f. of X given Y.

Properties

Proposition 1.10. Let X, Y, $X_1, X_2, ...$ be integrable random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$.
(i) If $X = c$ a.s., $c \in \mathcal{R}$, then $E(X|\mathcal{A}) = c$ a.s.
(ii) If $X \le Y$ a.s., then $E(X|\mathcal{A}) \le E(Y|\mathcal{A})$ a.s.
(iii) If $a \in \mathcal{R}$ and $b \in \mathcal{R}$, then $E(aX + bY|\mathcal{A}) = aE(X|\mathcal{A}) + bE(Y|\mathcal{A})$ a.s.
(iv) $E[E(X|\mathcal{A})] = EX$.
(v) $E[E(X|\mathcal{A})|\mathcal{A}_0] = E(X|\mathcal{A}_0) = E[E(X|\mathcal{A}_0)|\mathcal{A}]$ a.s., where $\mathcal{A}_0$ is a sub-σ-field of $\mathcal{A}$.
(vi) If $\sigma(Y) \subset \mathcal{A}$ and $E|XY| < \infty$, then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$ a.s.
(vii) If X and Y are independent and $E|g(X, Y)| < \infty$ for a Borel function g, then $E[g(X, Y)|Y = y] = E[g(X, y)]$ a.s. P_Y.
(viii) If $EX^2 < \infty$, then $[E(X|\mathcal{A})]^2 \le E(X^2|\mathcal{A})$ a.s.
(ix) (Fatou's lemma). If $X_n \ge 0$ for any n, then $E(\liminf_n X_n|\mathcal{A}) \le \liminf_n E(X_n|\mathcal{A})$ a.s.
(x) (Dominated convergence theorem). Suppose that $|X_n| \le Y$ for any n and $X_n \to_{a.s.} X$. Then $E(X_n|\mathcal{A}) \to_{a.s.} E(X|\mathcal{A})$.

Example 1.22. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with $EX^2 < \infty$ and let Y be a measurable function from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. One may wish to predict the value of X based on an observed value of Y. Let $g(Y)$ be a predictor, i.e., $g \in \aleph = \{$all Borel functions g with $E[g(Y)]^2 < \infty\}$. Each predictor is assessed by the "mean squared prediction error" $E[X - g(Y)]^2$. We now show that $E(X|Y)$ is the best predictor of X in the sense that

$$E[X - E(X|Y)]^2 = \min_{g \in \aleph} E[X - g(Y)]^2.$$

First, Proposition 1.10(viii) implies $E(X|Y) \in \aleph$. Next, for any $g \in \aleph$,

$$\begin{aligned}
E[X - g(Y)]^2 &= E[X - E(X|Y) + E(X|Y) - g(Y)]^2 \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[X - E(X|Y)][E(X|Y) - g(Y)]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{E\{[X - E(X|Y)][E(X|Y) - g(Y)]|Y\}\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[E(X|Y) - g(Y)]E[X - E(X|Y)|Y]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\ge E[X - E(X|Y)]^2,
\end{aligned}$$

where the third equality follows from Proposition 1.10(iv), the fourth equality follows from Proposition 1.10(vi), and the last equality follows from Proposition 1.10(i), (iii), and (vi).

Lecture 9: Independence, conditional independence, conditional distribution

Definition 1.7. Let $(\Omega, \mathcal{F}, P)$ be a probability space.
(i) Let $\mathcal{C}$ be a collection of subsets in $\mathcal{F}$. Events in $\mathcal{C}$ are said to be *independent* if and only if for any positive integer n and distinct events A_1,...,A_n in $\mathcal{C}$,

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2)\cdots P(A_n).$$

(ii) Collections $\mathcal{C}_i \subset \mathcal{F}$, $i \in \mathcal{I}$ (an index set that can be uncountable), are said to be independent if and only if events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.
(iii) Random elements X_i, $i \in \mathcal{I}$, are said to be independent if and only if $\sigma(X_i)$, $i \in \mathcal{I}$, are independent.

A useful result for checking the independence of σ-fields.

Lemma 1.3. Let $\mathcal{C}_i$, $i \in \mathcal{I}$, be independent collections of events. Suppose that each $\mathcal{C}_i$ has the property that if $A \in \mathcal{C}_i$ and $B \in \mathcal{C}_i$, then $A \cap B \in \mathcal{C}_i$. Then $\sigma(\mathcal{C}_i)$, $i \in \mathcal{I}$, are independent.

Random variables X_i, $i = 1, ..., k$, are independent according to Definition 1.7 if and only if

$$F_{(X_1,...,X_k)}(x_1, ..., x_k) = F_{X_1}(x_1)\cdots F_{X_k}(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k$$

Take $\mathcal{C}_i = \{(a, b] : a \in \mathcal{R}, b \in \mathcal{R}\}$, $i = 1, ..., k$

If X and Y are independent random vectors, then so are $g(X)$ and $h(Y)$ for Borel functions g and h.

Two events A and B are independent if and only if $P(B|A) = P(B)$, which means that A provides no information about the probability of the occurrence of B.

Proposition 1.11. Let X be a random variable with $E|X| < \infty$ and let Y_i be random k_i-vectors, $i = 1, 2$. Suppose that (X, Y_1) and Y_2 are independent. Then

$$E[X|(Y_1, Y_2)] = E(X|Y_1) \text{ a.s.}$$

Proof. First, $E(X|Y_1)$ is Borel on $(\Omega, \sigma(Y_1, Y_2))$, since $\sigma(Y_1) \subset \sigma(Y_1, Y_2)$. Next, we need to show that for any Borel set $B \in \mathcal{B}^{k_1+k_2}$,

$$\int_{(Y_1,Y_2)^{-1}(B)} X dP = \int_{(Y_1,Y_2)^{-1}(B)} E(X|Y_1)dP. \tag{1}$$

If $B = B_1 \times B_2$, where $B_i \in \mathcal{B}^{k_i}$, then $(Y_1, Y_2)^{-1}(B) = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)$ and

$$\begin{aligned}
\int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} E(X|Y_1)dP &= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E(X|Y_1)dP \\
&= \int I_{Y_1^{-1}(B_1)} E(X|Y_1)dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} X dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} X dP \\
&= \int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} X dP,
\end{aligned}$$

where the second and the next to last equalities follow the independence of (X, Y_1) and Y_2, and the third equality follows from the fact that $E(X|Y_1)$ is the conditional expectation of X given Y_1. This shows that (1) holds for $B = B_1 \times B_2$. We can show that the collection $\mathcal{H} = \{B \subset \mathcal{R}^{k_1+k_2} : B \text{ satisfies (1) }\}$ is a σ-field. Since we have already shown that $\mathcal{B}^{k_1} \times \mathcal{B}^{k_2} \subset \mathcal{H}$, $\mathcal{B}^{k_1+k_2} = \sigma(\mathcal{B}^{k_1} \times \mathcal{B}^{k_2}) \subset \mathcal{H}$ and thus the result follows.

The result in Proposition 1.11 still holds if X is replaced by $h(X)$ for any Borel h and, hence,

$$P(A|Y_1, Y_2) = P(A|Y_1) \text{ a.s. for any } A \in \sigma(X), \tag{2}$$

if (X, Y_1) and Y_2 are independent.

We say that given Y_1, X and Y_2 are *conditionally independent* if and only if (2) holds.

Proposition 1.11 can be stated as: if Y_2 and (X, Y_1) are independent, then given Y_1, X and Y_2 are conditionally independent.

Conditional distribution

For random vectors X and Y, is $P[X^{-1}(B)|Y = y]$ a probability measure for given y?
The the following theorem shows that there exists a version of conditional probability such that $P[X^{-1}(B)|Y = y]$ is a probability measure for any fixed y.

Theorem 1.7. (i) (Existence of conditional distributions). Let X be a random n-vector on a probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. Then there exists a function $P(B, \omega)$ on $\mathcal{B}^n \times \Omega$ such that (a) $P(B, \omega) = P[X^{-1}(B)|\mathcal{A}]$ a.s. for any fixed $B \in \mathcal{B}^n$, and (b) $P(\cdot, \omega)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any fixed $\omega \in \Omega$.
Let Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. Then there exists $P_{X|Y}(B|y)$ such that (a) $P_{X|Y}(B|y) = P[X^{-1}(B)|Y = y]$ a.s. P_Y for any fixed $B \in \mathcal{B}^n$, and (b) $P_{X|Y}(\cdot|y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any fixed $y \in \Lambda$.
Furthermore, if $E|g(X, Y)| < \infty$ with a Borel function g, then

$$E[g(X, Y)|Y = y] = E[g(X, y)|Y = y] = \int_{\mathcal{R}^n} g(x, y) dP_{X|Y}(x|y) \text{ a.s. } P_Y.$$

(ii) Let $(\Lambda, \mathcal{G}, P_1)$ be a probability space. Suppose that P_2 is a function from $\mathcal{B}^n \times \Lambda$ to $\mathcal{R}$ and satisfies (a) $P_2(\cdot, y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any $y \in \Lambda$, and (b) $P_2(B, \cdot)$ is Borel for any $B \in \mathcal{B}^n$. Then there is a unique probability measure P on $(\mathcal{R}^n \times \Lambda, \sigma(\mathcal{B}^n \times \mathcal{G}))$ such that, for $B \in \mathcal{B}^n$ and $C \in \mathcal{G}$,

$$P(B \times C) = \int_C P_2(B, y) dP_1(y). \tag{3}$$

Furthermore, if $(\Lambda, \mathcal{G}) = (\mathcal{R}^m, \mathcal{B}^m)$, and $X(x, y) = x$ and $Y(x, y) = y$ define the coordinate random vectors, then $P_Y = P_1$, $P_{X|Y}(\cdot|y) = P_2(\cdot, y)$, and the probability measure in (3) is the joint distribution of (X, Y), which has the following joint c.d.f.:

$$F(x, y) = \int_{(-\infty, y]} P_{X|Y}((-\infty, x]|z) dP_Y(z), \quad x \in \mathcal{R}^n, y \in \mathcal{R}^m, \tag{4}$$

where $(-\infty, a]$ denotes $(-\infty, a_1] \times \cdots \times (-\infty, a_k]$ for $a = (a_1, ..., a_k)$.

For a fixed y, $P_{X|Y=y} = P_{X|Y}(\cdot|y)$ is called the conditional distribution of X given $Y = y$.

Two-stage experiment theorem:
If $Y \in \mathcal{R}^m$ is selected in stage 1 of an experiment according to its marginal distribution $P_Y = P_1$, and X is chosen afterward according to a distribution $P_2(\cdot, y)$, then the combined two-stage experiment produces a jointly distributed pair (X, Y) with distribution $P_{(X,Y)}$ given by (3) and $P_{X|Y=y} = P_2(\cdot, y)$.
This provides a way of generating dependent random variables.

Example 1.23. A market survey is conducted to study whether a new product is preferred over the product currently available in the market (old product). The survey is conducted by mail. Questionnaires are sent along with the sample products (both new and old) to N customers randomly selected from a population, where N is a positive integer. Each customer is asked to fill out the questionnaire and return it. Responses from customers are either 1 (new is better than old) or 0 (otherwise). Some customers, however, do not return the questionnaires. Let X be the number of ones in the returned questionnaires. What is the distribution of X?

If every customer returns the questionnaire, then (from elementary probability) X has the binomial distribution $Bi(p, N)$ in Table 1.1 (assuming that the population is large enough so that customers respond independently), where $p \in (0, 1)$ is the overall rate of customers who prefer the new product. Now, let Y be the number of customers who respond. Then Y is random. Suppose that customers respond independently with the same probability $\pi \in (0, 1)$. Then P_Y is the binomial distribution $Bi(\pi, N)$. Given $Y = y$ (an integer between 0 and N), $P_{X|Y=y}$ is the binomial distribution $Bi(p, y)$ if $y \geq 1$ and the point mass at 0 if $y = 0$. Using (4) and the fact that binomial distributions have p.d.f.'s w.r.t. counting measure, we obtain that the joint c.d.f. of (X, Y) is

$$\begin{aligned} F(x,y) &= \sum_{k=0}^{y} P_{X|Y=k}((-\infty, x]) \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \sum_{k=0}^{y} \sum_{j=0}^{\min\{x,k\}} \binom{k}{j} p^j (1-p)^{k-j} \binom{N}{k} \pi^k (1-\pi)^{N-k} \end{aligned}$$

for $x = 0, 1, ..., y$, $y = 0, 1, ..., N$. The marginal c.d.f. $F_X(x) = F(x, \infty) = F(x, N)$. The p.d.f. of X w.r.t. counting measure is

$$\begin{aligned} f_X(x) &= \sum_{k=x}^{N} \binom{k}{x} p^x (1-p)^{k-x} \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \sum_{k=x}^{N} \binom{N-x}{k-x} \left(\frac{\pi - \pi p}{1-\pi p}\right)^{k-x} \left(\frac{1-\pi}{1-\pi p}\right)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \end{aligned}$$

for $x = 0, 1, ..., N$. It turns out that the marginal distribution of X is the binomial distribution $Bi(\pi p, N)$.

Lecture 10: Markov chains

An important example of dependent sequence of random variables in statistical application

A sequence of random vectors $\{X_n : n = 1, 2, ...\}$ is a *Markov chain* or *Markov process* if and only if

$$P(B|X_1, ..., X_n) = P(B|X_n) \text{ a.s.}, \quad B \in \sigma(X_{n+1}), \ n = 2, 3, ... \tag{1}$$

X_{n+1} (tomorrow) is conditionally independent of $(X_1, ..., X_{n-1})$ (the past), given X_n (today). $(X_1, ..., X_{n-1})$ is not necessarily independent of (X_n, X_{n+1}).
A sequence of independent random vectors forms a Markov chain

Example 1.24 (First-order autoregressive processes). Let $\varepsilon_1, \varepsilon_2, ...$ be independent random variables defined on a probability space, $X_1 = \varepsilon_1$, and $X_{n+1} = \rho X_n + \varepsilon_{n+1}$, $n = 1, 2, ...$, where ρ is a constant in $\mathcal{R}$. Then $\{X_n\}$ is called a first-order autoregressive process. We now show that for any $B \in \mathcal{B}$ and $n = 1, 2, ...$,

$$P(X_{n+1} \in B|X_1, ..., X_n) = P_{\varepsilon_{n+1}}(B - \rho X_n) = P(X_{n+1} \in B|X_n) \text{ a.s.},$$

where $B - y = \{x \in \mathcal{R} : x + y \in B\}$, which implies that $\{X_n\}$ is a Markov chain. For any $y \in \mathcal{R}$,

$$P_{\varepsilon_{n+1}}(B - y) = P(\varepsilon_{n+1} + y \in B) = \int I_B(x + y) dP_{\varepsilon_{n+1}}(x)$$

and, by Fubini's theorem, $P_{\varepsilon_{n+1}}(B - y)$ is Borel. Hence, $P_{\varepsilon_{n+1}}(B - \rho X_n)$ is Borel w.r.t. $\sigma(X_n)$ and, thus, is Borel w.r.t. $\sigma(X_1, ..., X_n)$. Let $B_j \in \mathcal{B}$, $j = 1, ..., n$, and $A = \cap_{j=1}^n X_j^{-1}(B_j)$. Since $\varepsilon_{n+1} + \rho X_n = X_{n+1}$ and ε_{n+1} is independent of $(X_1, ..., X_n)$, it follows from Theorem 1.2 and Fubini's theorem that

$$\begin{aligned}
\int_A P_{\varepsilon_{n+1}}(B - \rho X_n) dP &= \int_{x_j \in B_j, j=1,...,n} \int_{t \in B - \rho x_n} dP_{\varepsilon_{n+1}}(t) dP_X(x) \\
&= \int_{x_j \in B_j, j=1,...,n, x_{n+1} \in B} dP_{(X, \varepsilon_{n+1})}(x, t) \\
&= P\left(A \cap X_{n+1}^{-1}(B)\right),
\end{aligned}$$

where X and x denote $(X_1, ..., X_n)$ and $(x_1, ..., x_n)$, respectively, and x_{n+1} denotes $\rho x_n + t$. Using this and the argument in the end of the proof for Proposition 1.11, we obtain $P(X_{n+1} \in B|X_1, ..., X_n) = P_{\varepsilon_{n+1}}(B - \rho X_n)$ a.s. The proof for $P_{\varepsilon_{n+1}}(B - \rho X_n) = P(X_{n+1} \in B|X_n)$ a.s. is similar and simpler.

Characterizations of Markov chains

Proposition 1.12. A sequence of random vectors $\{X_n\}$ is a Markov chain if and only if one of the following three conditions holds.
(a) For any $n = 2, 3, ...$ and any integrable $h(X_{n+1})$ with a Borel function h,

$$E[h(X_{n+1})|X_1, ..., X_n] = E[h(X_{n+1})|X_n] \text{ a.s.}$$

(b) For any $n = 1, 2, ...$ and $B \in \sigma(X_{n+1}, X_{n+2}, ...)$,

$$P(B|X_1, ..., X_n) = P(B|X_n) \text{ a.s.}$$

("the past and the future are conditionally independent given the present")
(c) For any $n = 2, 3, ...$, $A \in \sigma(X_1, ..., X_n)$, and $B \in \sigma(X_{n+1}, X_{n+2}, ...)$,

$$P(A \cap B|X_n) = P(A|X_n)P(B|X_n) \text{ a.s.}$$

Proof. (i) It is clear that (a) implies (1). If h is a simple function, then (1) and Proposition 1.10(iii) imply (a). If h is nonnegative, then there are nonnegative simple functions $h_1 \leq h_2 \leq \cdots \leq h$ such that $h_j \to h$. Then (1) together with Proposition 1.10(iii) and (x) imply (a). Since $h = h_+ - h_-$, we conclude that (1) implies (a).
(ii) It is also clear that (b) implies (1). We now show that (1) implies (b). Note that $\sigma(X_{n+1}, X_{n+2}, ...) = \sigma\left(\cup_{j=1}^{\infty}\sigma(X_{n+1}, ..., X_{n+j})\right)$ (Exercise 19). Hence, it suffices to show that $P(B|X_1, ..., X_n) = P(B|X_n)$ a.s. for $B \in \sigma(X_{n+1}, ..., X_{n+j})$ for any $j = 1, 2, ...$ We use induction. The result for $j = 1$ follows from (1). Suppose that the result holds for any $B \in \sigma(X_{n+1}, ..., X_{n+j})$. To show the result for any $B \in \sigma(X_{n+1}, ..., X_{n+j+1})$, it is enough (why?) to show that for any $B_1 \in \sigma(X_{n+j+1})$ and any $B_2 \in \sigma(X_{n+1}, ..., X_{n+j})$, $P(B_1 \cap B_2|X_1, ..., X_n) = P(B_1 \cap B_2|X_n)$ a.s. From the proof in (i), the induction assumption implies

$$E[h(X_{n+1}, ..., X_{n+j})|X_1, ..., X_n] = E[h(X_{n+1}, ..., X_{n+j})|X_n] \tag{2}$$

for any Borel function h. The result follows from

$$\begin{aligned}
E(I_{B_1}I_{B_2}|X_1, ..., X_n) &= E[E(I_{B_1}I_{B_2}|X_1, ..., X_{n+j})|X_1, ..., X_n] \\
&= E[I_{B_2}E(I_{B_1}|X_1, ..., X_{n+j})|X_1, ..., X_n] \\
&= E[I_{B_2}E(I_{B_1}|X_{n+j})|X_1, ..., X_n] \\
&= E[I_{B_2}E(I_{B_1}|X_{n+j})|X_n] \\
&= E[I_{B_2}E(I_{B_1}|X_n, ..., X_{n+j})|X_n] \\
&= E[E(I_{B_1}I_{B_2}|X_n, ..., X_{n+j})|X_n] \\
&= E(I_{B_1}I_{B_2}|X_n) \text{ a.s.},
\end{aligned}$$

where the first and last equalities follow from Proposition 1.10(v), the second and sixth equalities follow from Proposition 1.10(vi), the third and fifth equalities follow from (1), and the fourth equality follows from (2).

(iii) Let $A \in \sigma(X_1, ..., X_n)$ and $B \in \sigma(X_{n+1}, X_{n+2}, ...)$. If (b) holds, then

$$\begin{aligned} E(I_A I_B|X_n) &= E[E(I_A I_B|X_1, ..., X_n)|X_n] \\ &= E[I_A E(I_B|X_1, ..., X_n)|X_n] \\ &= E[I_A E(I_B|X_n)|X_n] \\ &= E(I_A|X_n)E(I_B|X_n), \end{aligned}$$

which is (c).

Assume that (c) holds. Let $A_1 \in \sigma(X_n)$, $A_2 \in \sigma(X_1, ..., X_{n-1})$, and $B \in \sigma(X_{n+1}, X_{n+2}, ...)$. Then

$$\begin{aligned} \int_{A_1 \cap A_2} E(I_B|X_n)dP &= \int_{A_1} I_{A_2} E(I_B|X_n)dP \\ &= \int_{A_1} E[I_{A_2} E(I_B|X_n)|X_n]dP \\ &= \int_{A_1} E(I_{A_2}|X_n)E(I_B|X_n)dP \\ &= \int_{A_1} E(I_{A_2} I_B|X_n)dP \\ &= P(A_1 \cap A_2 \cap B). \end{aligned}$$

Since disjoint unions of events of the form $A_1 \cap A_2$ as specified above generate $\sigma(X_1, ..., X_n)$, this shows that $E(I_B|X_n) = E(I_B|X_1, ..., X_n)$ a.s., which is (b).

Lecture 11: Convergence modes and stochastic orders

$c = (c_1, ..., c_k) \in \mathcal{R}^k$, $\|c\|_r = (\sum_{j=1}^k |c_j|^r)^{1/r}$, $r > 0$.
If $r \geq 1$, then $\|c\|_r$ is the L_r-distance between 0 and c.
When $r = 2$, $\|c\| = \|c\|_2 = \sqrt{c^\tau c}$.

Definition 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors defined on a probability space.
(i) We say that the sequence $\{X_n\}$ converges to X almost surely (a.s.) and write $X_n \to_{a.s.} X$ if and only if $\lim_{n\to\infty} X_n = X$ a.s.
(ii) We say that $\{X_n\}$ converges to X in probability and write $X_n \to_p X$ if and only if, for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\|X_n - X\| > \epsilon\right) = 0.$$

(iii) We say that $\{X_n\}$ converges to X in L_r (or in rth moment) and write $X_n \to_{L_r} X$ if and only if

$$\lim_{n\to\infty} E\|X_n - X\|_r^r = 0,$$

where $r > 0$ is a fixed constant.
(iv) Let F, F_n, $n = 1, 2, ...$, be c.d.f.'s on $\mathcal{R}^k$ and P, P_n, $n = 1, ...$, be their corresponding probability measures. We say that $\{F_n\}$ converges to F weakly (or $\{P_n\}$ converges to P weakly) and write $F_n \to_w F$ (or $P_n \to_w P$) if and only if, for each continuity point x of F,

$$\lim_{n\to\infty} F_n(x) = F(x).$$

We say that $\{X_n\}$ converges to X in distribution (or in law) and write $X_n \to_d X$ if and only if $F_{X_n} \to_w F_X$.

$\to_{a.s.}$, $\to_p$, $\to_{L_r}$: How close is between X_n and X as $n \to \infty$?

$F_{X_n} \to_w F_X$: X_n and X may not be close (they may be on different spaces)

Example 1.26. Let $\theta_n = 1 + n^{-1}$ and X_n be a random variable having the exponential distribution $E(0, \theta_n)$ (Table 1.2), $n = 1, 2, ...$ Let X be a random variable having the exponential distribution $E(0, 1)$. For any $x > 0$, as $n \to \infty$,

$$F_{X_n}(x) = 1 - e^{-x/\theta_n} \to 1 - e^{-x} = F_X(x)$$

Since $F_{X_n}(x) \equiv 0 \equiv F_X(x)$ for $x \leq 0$, we have shown that $X_n \to_d X$.

$X_n \to_p X$?
Need further information about the random variables X and X_n.
We consider two cases in which different answers can be obtained.
First, suppose that $X_n \equiv \theta_n X$ (then X_n has the given c.d.f.).
$X_n - X = (\theta_n - 1)X = n^{-1}X$, which has the c.d.f. $(1 - e^{-nx})I_{[0,\infty)}(x)$.

$$P\left(|X_n - X| \geq \epsilon\right) = e^{-n\epsilon} \to 0$$

for any $\epsilon > 0$. (In fact, by Theorem 1.8(v), $X_n \to_{a.s.} X$)
Since $E|X_n - X|^p = n^{-p} EX^p < \infty$ for any $p > 0$, $X_n \to_{L_p} X$ for any $p > 0$.

Next, suppose that X_n and X are independent random variables.
Since p.d.f.'s for X_n and $-X$ are $\theta_n^{-1}e^{-x/\theta_n}I_{(0,\infty)}(x)$ and $e^x I_{(-\infty,0)}(x)$, respectively, we have

$$P\left(|X_n - X| \le \epsilon\right) = \int_{-\epsilon}^{\epsilon}\int \theta_n^{-1}e^{-x/\theta_n}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy,$$

which converges to (by the dominated convergence theorem)

$$\int_{-\epsilon}^{\epsilon}\int e^{-x}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy = 1 - e^{-\epsilon}.$$

Thus, $P\left(|X_n - X| \ge \epsilon\right) \to e^{-\epsilon} > 0$ for any $\epsilon > 0$ and, therefore, $X_n \to_p X$ does not hold.

Proposition 1.16 (Pólya's theorem). If $F_n \to_w F$ and F is continuous on $\mathcal{R}^k$, then

$$\lim_{n\to\infty}\sup_{x\in\mathcal{R}^k}|F_n(x) - F(x)| = 0.$$

Lemma 1.4. For random k-vectors $X, X_1, X_2, \ldots$ on a probability space, $X_n \to_{a.s.} X$ if and only if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty}\{\|X_m - X\| > \epsilon\}\right) = 0. \tag{1}$$

Proof. Let $A_j = \cup_{n=1}^{\infty}\cap_{m=n}^{\infty}\{\|X_m - X\| \le j^{-1}\}$, $j = 1, 2, ...$
Then

$$\bigcap_{j=1}^{\infty} A_j = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$$

By Proposition 1.1(iii),

$$P(A_j) = \lim_{n\to\infty} P\left(\bigcap_{m=n}^{\infty}\{\|X_m - X\| \le j^{-1}\}\right) = 1 - \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty}\{\|X_m - X\| > j^{-1}\}\right)$$

(1) holds for every $\epsilon > 0$ if and only if $P(A_j) = 1$ for every j, i.e., $P(\cap_{j=1}^{\infty}A_j) = 1$

$$P(A_j) \ge P(\bigcap_{j=1}^{\infty} A_j) = 1 - P(\bigcup_{j=1}^{\infty} A_j^c) \ge 1 - \sum_{j=1}^{\infty} P(A_j^c)$$

Lemma 1.5. (Borel-Cantelli lemma). Let A_n be a sequence of events in a probability space and $\limsup_n A_n = \cap_{n=1}^{\infty}\cup_{m=n}^{\infty} A_m$.
(i) If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$.
(ii) If $A_1, A_2, ...$ are pairwise independent and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.
Proof. (i) By Proposition 1.1,

$$P\left(\limsup_{n\to\infty} A_n\right) = P\left(\bigcap_{n=1}^{\infty}\bigcup_{m=n}^{\infty} A_m\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) \le \lim_{n\to\infty}\sum_{m=n}^{\infty} P(A_n) = 0$$

if $\sum_{n=1}^{\infty} P(A_n) < \infty$.

(ii) We prove the case of independent A_n's.

$$P\left(\limsup_{n\to\infty} A_n\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) = 1 - \lim_{n\to\infty} P\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 1 - \lim_{n\to\infty} \prod_{m=n}^{\infty} P(A_m^c)$$

$$\prod_{m=n}^{n+k} P(A_m^c) = \prod_{m=n}^{n+k} [1 - P(A_m)] \le \prod_{m=n}^{n+k} \exp\{-P(A_m)\} = \exp\left\{-\sum_{m=n}^{n+k} P(A_m)\right\}$$

$(1 - t \le e^{-t} = \exp\{t\})$. Letting $k \to \infty$,

$$\prod_{m=n}^{\infty} P(A_m^c) = \lim_{k\to\infty} \prod_{m=n}^{n+k} P(A_m^c) \le \exp\left\{-\sum_{m=n}^{\infty} P(A_m)\right\} = 0.$$

See Chung (1974, pp. 76-78) for the pairwise independence A_n's.

The notion of $O(\cdot)$, $o(\cdot)$, and stochastic $O(\cdot)$ and $o(\cdot)$

In calculus, two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$, satisfy $a_n = O(b_n)$ if and only if $|a_n| \le c|b_n|$ for all n and a constant c
$a_n = o(b_n)$ if and only if $a_n/b_n \to 0$ as $n \to \infty$

Definition 1.9. Let $X_1, X_2, ...$ be random vectors and $Y_1, Y_2, ...$ be random variables defined on a common probability space.
(i) $X_n = O(Y_n)$ a.s. if and only if $P(\|X_n\| = O(|Y_n|)) = 1$.
(ii) $X_n = o(Y_n)$ a.s. if and only if $X_n/Y_n \to_{a.s.} 0$.
(iii) $X_n = O_p(Y_n)$ if and only if, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that $\sup_n P(\|X_n\| \ge C_\epsilon |Y_n|) < \epsilon$.
(iv) $X_n = o_p(Y_n)$ if and only if $X_n/Y_n \to_p 0$.

Since $a_n = O(1)$ means that $\{a_n\}$ is bounded, $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$.

$X_n = o_p(Y_n)$ implies $X_n = O_p(Y_n)$
$X_n = O_p(Y_n)$ and $Y_n = O_p(Z_n)$ implies $X_n = O_p(Z_n)$
$X_n = O_p(Y_n)$ does not imply $Y_n = O_p(X_n)$
If $X_n = O_p(Z_n)$, then $X_n Y_n = O_p(Y_n Z_n)$.
If $X_n = O_p(Z_n)$ and $Y_n = O_p(Z_n)$, then $X_n + Y_n = O_p(Z_n)$.
The same conclusion can be obtained if $O_p(\cdot)$ and $o_p(\cdot)$ are replaced by $O(\cdot)$ a.s. and $o(\cdot)$ a.s., respectively.
If $X_n \to_d X$ for a random variable X, then $X_n = O_p(1)$
If $E|X_n| = O(a_n)$, then $X_n = O_p(a_n)$, where $a_n \in (0, \infty)$.
If $X_n \to_{a.s.} X$, then $\sup_n |X_n| = O_p(1)$.

Lecture 12: Relationship among convergence modes and uniform integrability

Theorem 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) If $X_n \to_{a.s.} X$, then $X_n \to_p X$. (The converse is not true.)
(ii) If $X_n \to_{L_r} X$ for an $r > 0$, then $X_n \to_p X$. (The converse is not true.)
(iii) If $X_n \to_p X$, then $X_n \to_d X$. (The converse is not true.)
(iv) (Skorohod's theorem). If $X_n \to_d X$, then there are random vectors $Y, Y_1, Y_2, ...$ defined on a common probability space such that $P_Y = P_X$, $P_{Y_n} = P_{X_n}$, $n = 1, 2, ...$, and $Y_n \to_{a.s.} Y$. (A useful result; a conditional converse of (i)-(iii).)
(v) If, for every $\epsilon > 0$, $\sum_{n=1}^{\infty} P(\|X_n - X\| \geq \epsilon) < \infty$, then $X_n \to_{a.s.} X$.
(A conditional converse of (i): $P(\|X_n - X\| \geq \epsilon)$ tends to 0 fast enough.)
(vi) If $X_n \to_p X$, then there is a subsequence $\{X_{n_j}, j = 1, 2, ...\}$ such that $X_{n_j} \to_{a.s.} X$ as $j \to \infty$. (A partial converse of (i).)
(vii) If $X_n \to_d X$ and $P(X = c) = 1$, where $c \in \mathcal{R}^k$ is a constant vector, then $X_n \to_p c$. (A conditional converse of (i).)
(viii) Suppose that $X_n \to_d X$. Then, for any $r > 0$,

$$\lim_{n\to\infty} E\|X_n\|_r^r = E\|X\|_r^r < \infty \tag{1}$$

if and only if $\{\|X_n\|_r^r\}$ is *uniformly integrable* in the sense that

$$\lim_{t\to\infty} \sup_n E\left(\|X_n\|_r^r I_{\{\|X_n\|_r > t\}}\right) = 0. \tag{2}$$

(A conditional converse of (ii).)

Discussion on uniform integrability
If there is only one random vector, then (2) is

$$\lim_{t\to\infty} E\left(\|X\|_r^r I_{\{\|X\|_r > t\}}\right) = 0,$$

which is equivalent to the integrability of $\|X\|_r^r$ (dominated convergence theorem).
Sufficient conditions for uniform integrability:

$$\sup_n E\|X_n\|_r^{r+\delta} < \infty \quad \text{for a } \delta > 0$$

This is because

$$\begin{aligned}
\lim_{t\to\infty} \sup_n E\left(\|X_n\|_r^r I_{\{\|X_n\|_r > t\}}\right) &\leq \lim_{t\to\infty} \sup_n E\left(\|X_n\|_r^r I_{\{\|X_n\|_r > t\}} \frac{\|X_n\|_r^\delta}{t^\delta}\right) \\
&\leq \lim_{t\to\infty} \frac{1}{t^\delta} \sup_n E\left(\|X_n\|_r^{r+\delta}\right) \\
&= 0
\end{aligned}$$

Exercises 117-120.

Proof of Theorem 1.8. (i) The result follows from Lemma 1.4.
(ii) The result follows from Chebyshev's inequality with $\varphi(t) = |t|^r$.
(iii) Assume $k = 1$. (The general case is proved in the textbook.)
Let x be a continuity point of F_X and $\epsilon > 0$ be given. Then

$$\begin{aligned} F_X(x-\epsilon) &= P(X \le x-\epsilon) \\ &\le P(X_n \le x) + P(X \le x-\epsilon, X_n > x) \\ &\le F_{X_n}(x) + P\left(|X_n - X| > \epsilon\right). \end{aligned}$$

Letting $n \to \infty$, we obtain that

$$F_X(x-\epsilon) \le \liminf_n F_{X_n}(x).$$

Switching X_n and X in the previous argument, we can show that

$$F_X(x+\epsilon) \ge \limsup_n F_{X_n}(x).$$

Since ϵ is arbitrary and F_X is continuous at x, $F_X(x) = \lim_{n\to\infty} F_{X_n}(x)$.
(iv) The proof of this part can be found in Billingsley (1986, pp. 399-402).
(v) Let $A_n = \{\|X_n - X\| \ge \epsilon\}$. The result follows from Lemma 1.4, Lemma 1.5(i), and Proposition 1.1(iii).
(vi) $X_n \to_p X$ means $\lim_{n\to\infty} P(\|X_n - X\| > \epsilon) = 0$ for every $\epsilon > 0$.
That is, for every $\epsilon > 0$, $P(\|X_n - X\| > \epsilon) < \epsilon$ for $n > n_\epsilon$ (n_ϵ is an integer depending on ϵ).
For every $j = 1, 2, ...$, there is a positive integer n_j such that

$$P(\|X_{n_j} - X\| > 2^{-j}) < 2^{-j}.$$

For any $\epsilon > 0$, there is a k_ϵ such that for $j \ge k_\epsilon$, $P(\|X_{n_j} - X\| > \epsilon) < P(\|X_{n_j} - X\| > 2^{-j})$.
Since $\sum_{j=1}^{\infty} 2^{-j} = 1$, it follows from the result in (v) that $X_{n_j} \to_{a.s.} X$ as $j \to \infty$.
(vii) The proof for this part is left as an exercise.
(viii) First, by part (iv), we may assume that $X_n \to_{a.s.} X$ (why?).
Proof of (2) implies (1)
Note that (2) (the uniform integrability of $\{\|X_n\|_r^r\}$) implies that $\sup_n E\|X_n\|_r^r < \infty$ (why?)
By Fatou's lemma (Theorem 1.1(i)), $E\|X\|_r^r \le \liminf_n E\|X_n\|_r^r < \infty$.
Hence, (1) follows if we can show that

$$\limsup_n E\|X_n\|_r^r \le E\|X\|_r^r. \tag{3}$$

For any $\epsilon > 0$ and $t > 0$, let $A_n = \{\|X_n - X\|_r \le \epsilon\}$ and $B_n = \{\|X_n\|_r > t\}$. Then

$$\begin{aligned} E\|X_n\|_r^r &= E(\|X_n\|_r^r I_{A_n^c \cap B_n}) + E(\|X_n\|_r^r I_{A_n^c \cap B_n^c}) + E(\|X_n\|_r^r I_{A_n}) \\ &\le E(\|X_n\|_r^r I_{B_n}) + t^r P(A_n^c) + E\|X_n I_{A_n}\|_r^r. \end{aligned}$$

For $r \le 1$, $\|X_n I_{A_n}\|_r^r \le (\|X_n - X\|_r^r + \|X\|_r^r) I_{A_n}$ and

$$E\|X_n I_{A_n}\|_r^r \le E[(\|X_n - X\|_r^r + \|X\|_r^r) I_{A_n}] \le \epsilon^r + E\|X\|_r^r.$$

For $r > 1$, an application of Minkowski's inequality leads to

$$\begin{aligned} E\|X_n I_{A_n}\|_r^r &= E\|(X_n - X)I_{A_n} + XI_{A_n}\|_r^r \\ &\le E\left[\|(X_n - X)I_{A_n}\|_r + \|XI_{A_n}\|_r\right]^r \\ &\le \left\{[E\|(X_n - X)I_{A_n}\|_r^r]^{1/r} + [E\|XI_{A_n}\|_r^r]^{1/r}\right\}^r \\ &\le \left\{\epsilon + [E\|X\|_r^r]^{1/r}\right\}^r. \end{aligned}$$

In any case, since ϵ is arbitrary, $\limsup_n E\|X_n I_{A_n}\|_r^r \le E\|X\|_r^r$. This result and the previously established inequality imply that

$$\begin{aligned} \limsup_n E\|X_n\|_r^r &\le \limsup_n E(\|X_n\|_r^r I_{B_n}) + t^r \lim_{n\to\infty} P(A_n^c) \\ &\quad + \limsup_n E\|X_n I_{A_n}\|_r^r \\ &\le \sup_n E(\|X_n\|_r^r I_{\{\|X_n\|_r > t\}}) + E\|X\|_r^r, \end{aligned}$$

since $P(A_n^c) \to 0$. Since $\{\|X_n\|_r^r\}$ is uniformly integrable, letting $t \to \infty$ we obtain (3).

Proof of (1) impies (2)
Let $\xi_n = \|X_n\|_r^r I_{B_n^c} - \|X\|_r^r I_{B_n^c}$. Then $\xi_n \to_{a.s.} 0$ and $|\xi_n| \le t^r + \|X\|_r^r$, which is integrable. By the dominated convergence theorem, $E\xi_n \to 0$; this and (1) imply that

$$E(\|X_n\|_r^r I_{B_n}) - E(\|X\|_r^r I_{B_n}) \to 0.$$

From the definition of B_n, $B_n \subset \{\|X_n - X\|_r > t/2\} \cup \{\|X\|_r > t/2\}$.
Since $E\|X\|_r^r < \infty$, it follows from the dominated convergence theorem that

$$\lim_{n\to\infty} E(\|X\|_r^r I_{\{\|X_n - X\|_r > t/2\}}) = 0$$

Hence

$$\limsup_n E(\|X_n\|_r^r I_{B_n}) \le \limsup_n E(\|X\|_r^r I_{B_n}) \le E(\|X\|_r^r I_{\{\|X\|_r > t/2\}}).$$

Letting $t \to \infty$, it follows from the dominated convergence theorem that

$$\lim_{t\to\infty} \limsup_n E(\|X_n\|_r^r I_{B_n}) \le \lim_{t\to\infty} E(\|X\|_r^r I_{\{\|X\|_r > t/2\}}) = 0.$$

This proves (2).

Lecture 13: Weak convergence

A sequence $\{P_n\}$ of probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$ is *tight* if for every $\epsilon > 0$, there is a compact set $C \subset \mathcal{R}^k$ such that $\inf_n P_n(C) > 1 - \epsilon$.
If $\{X_n\}$ is a sequence of random k-vectors, then the tightness of $\{P_{X_n}\}$ is the same as the boundedness of $\{\|X_n\|\}$ in probability ($\|X_n\| = O_p(1)$).

Proposition 1.17. Let $\{P_n\}$ be a sequence of probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$.
(i) Tightness of $\{P_n\}$ is a necessary and sufficient condition that for every subsequence $\{P_{n_i}\}$ there exists a further subsequence $\{P_{n_j}\} \subset \{P_{n_i}\}$ and a probability measure P on $(\mathcal{R}^k, \mathcal{B}^k)$ such that $P_{n_j} \to_w P$ as $j \to \infty$.
(ii) If $\{P_n\}$ is tight and if each subsequence that converges weakly at all converges to the same probability measure P, then $P_n \to_w P$.
The proof can be found in Billingsley (1986, pp. 392-395).

The following result gives some useful sufficient and necessary conditions for convergence in distribution.

Theorem 1.9. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) $X_n \to_d X$ is equivalent to any one of the following conditions:
  (a) $E[h(X_n)] \to E[h(X)]$ for every bounded continuous function h;
  (b) $\limsup_n P_{X_n}(C) \le P_X(C)$ for any closed set $C \subset \mathcal{R}^k$;
  (c) $\liminf_n P_{X_n}(O) \ge P_X(O)$ for any open set $O \subset \mathcal{R}^k$.
(ii) (Lévy-Cramér continuity theorem). Let $\phi_X, \phi_{X_1}, \phi_{X_2}, \ldots$ be the ch.f.'s of $X, X_1, X_2, \ldots$, respectively. $X_n \to_d X$ if and only if $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathcal{R}^k$.
(iii) (Cramér-Wold device). $X_n \to_d X$ if and only if $c^\tau X_n \to_d c^\tau X$ for every $c \in \mathcal{R}^k$.
Proof. (i) First, we show $X_n \to_d X$ implies (a). By Theorem 1.8(iv) (Skorohod's theorem), there exists a sequence of random vectors $\{Y_n\}$ and a random vector Y such that $P_{Y_n} = P_{X_n}$ for all n, $P_Y = P_X$ and $Y_n \to_{a.s.} Y$. For bounded continuous h, $h(Y_n) \to_{a.s.} h(Y)$ and, by the dominated convergence theorem, $E[h(Y_n)] \to E[h(Y)]$. Then (a) follows from $E[h(X_n)] = E[h(Y_n)]$ for all n and $E[h(X)] = E[h(Y)]$.

Next, we show (a) implies (b). Let C be a closed set and $f_C(x) = \inf\{\|x - y\| : y \in C\}$. Then f_C is continuous. For $j = 1, 2, \ldots$, define $\varphi_j(t) = I_{(-\infty,0]} + (1 - jt)I_{(0,j^{-1}]}$. Then $h_j(x) = \varphi_j(f_C(x))$ is continuous and bounded, $h_j \ge h_{j+1}$, $j = 1, 2, \ldots$, and $h_j(x) \to I_C(x)$ as $j \to \infty$. Hence $\limsup_n P_{X_n}(C) \le \lim_{n\to\infty} E[h_j(X_n)] = E[h_j(X)]$ for each j (by (a)). By the dominated convergence theorem, $E[h_j(X)] \to E[I_C(X)] = P_X(C)$. This proves (b).

For any open set O, O^c is closed. Hence, (b) is equivalent to (c). Now, we show (b) and (c) imply $X_n \to_d X$. For $x = (x_1, \ldots, x_k) \in \mathcal{R}^k$, let $(-\infty, x] = (-\infty, x_1] \times \cdots \times (-\infty, x_k]$ and $(-\infty, x) = (-\infty, x_1) \times \cdots \times (-\infty, x_k)$. From (b) and (c), $P_X((-\infty, x)) \le \liminf_n P_{X_n}((-\infty, x)) \le \liminf_n F_{X_n}(x) \le \limsup_n F_{X_n}(x) = \limsup_n P_{X_n}((-\infty, x]) \le P_X((-\infty, x]) = F_X(x)$. If x is a continuity point of F_X, then $P_X((-\infty, x)) = F_X(x)$. This proves $X_n \to_d X$ and completes the proof of (i).

(ii) From (a) of part (i), $X_n \to_d X$ implies $\phi_{X_n}(t) \to \phi_X(t)$, since $e^{\sqrt{-1}t^\tau x} = \cos(t^\tau x) + \sqrt{-1}\sin(t^\tau x)$ and $\cos(t^\tau x)$ and $\sin(t^\tau x)$ are bounded continuous functions for any fixed t.

Suppose now that $k = 1$ and that $\phi_{X_n}(t) \to \phi_X(t)$ for every $t \in \mathcal{R}$.
We want to show that $\{P_{X_n}\}$ is tight. By Fubini's theorem,

$$\begin{aligned}\frac{1}{u}\int_{-u}^{u}[1-\phi_{X_n}(t)]dt &= \int_{-\infty}^{\infty}\left[\frac{1}{u}\int_{-u}^{u}(1-e^{\sqrt{-1}tx})dt\right]dP_{X_n}(x)\\ &= 2\int_{-\infty}^{\infty}\left(1-\frac{\sin ux}{ux}\right)dP_{X_n}(x)\\ &\geq 2\int_{\{|x|>2u^{-1}\}}\left(1-\frac{1}{|ux|}\right)dP_{X_n}(x)\\ &\geq P_{X_n}\left((-\infty,-2u^{-1})\cup(2u^{-1},\infty)\right)\end{aligned}$$

for any $u > 0$. Since ϕ_X is continuous at 0 and $\phi_X(0) = 1$, for any $\epsilon > 0$ there is a $u > 0$ such that $u^{-1}\int_{-u}^{u}[1-\phi_X(t)]dt < \epsilon/2$. Since $\phi_{X_n} \to \phi_X$, by the dominated convergence theorem, $\sup_n\{u^{-1}\int_{-u}^{u}[1-\phi_{X_n}(t)]dt\} < \epsilon$. Hence,

$$\inf_n P_{X_n}\left([-2u^{-1},2u^{-1}]\right) \geq 1-\sup_n\left\{\frac{1}{u}\int_{-u}^{u}[1-\phi_{X_n}(t)]dt\right\} \geq 1-\epsilon,$$

i.e., $\{P_{X_n}\}$ is tight.
Let $\{P_{X_{n_j}}\}$ be any subsequence that converges to a probability measure P.
By the first part of the proof, $\phi_{X_{n_j}} \to \phi$, which is the ch.f. of P.
By the convergence of ϕ_{X_n}, $\phi = \phi_X$. By the uniqueness theorem, $P = P_X$.
By Proposition 1.17(ii), $X_n \to_d X$.

Consider now the case where $k \geq 2$ and $\phi_{X_n} \to \phi_X$.
Let Y_{nj} be the jth component of X_n and Y_j be the jth component of X.
Then $\phi_{Y_{nj}} \to \phi_{Y_j}$ for each j.
By the proof for the case of $k = 1$, $Y_{nj} \to_d Y_j$.
By Proposition 1.17(i), $\{P_{Y_{nj}}\}$ is tight, $j = 1, ..., k$. This implies that $\{P_{X_n}\}$ is tight (why?).
Then the proof for $X_n \to_d X$ is the same as that for the case of $k = 1$.

(iii) Note that $\phi_{c^\tau X_n}(u) = \phi_{X_n}(uc)$ and $\phi_{c^\tau X}(u) = \phi_X(uc)$ for any $u \in \mathcal{R}$ and any $c \in \mathcal{R}^k$. Hence, convergence of ϕ_{X_n} to ϕ_X is equivalent to convergence of $\phi_{c^\tau X_n}$ to $\phi_{c^\tau X}$ for every $c \in \mathcal{R}^k$. Then the result follows from part (ii).

Example 1.28. Let $X_1, ..., X_n$ be independent random variables having a common c.d.f. and $T_n = X_1 + \cdots + X_n$, $n = 1, 2, ...$ Suppose that $E|X_1| < \infty$. It follows from a result in calculus that the ch.f. of X_1 satisfies

$$\phi_{X_1}(t) = \phi_{X_1}(0) + \sqrt{-1}\mu t + o(|t|)$$

as $|t| \to 0$, where $\mu = EX_1$. Then, the ch.f. of T_n/n is

$$\phi_{T_n/n}(t) = \left[\phi_{X_1}\left(\frac{t}{n}\right)\right]^n = \left[1+\frac{\sqrt{-1}\mu t}{n}+o\left(\frac{t}{n}\right)\right]^n$$

for any $t \in \mathcal{R}$, as $n \to \infty$. Since $(1 + c_n/n)^n \to e^c$ for any complex sequence $\{c_n\}$ satisfying $c_n \to c$, we obtain that $\phi_{T_n/n}(t) \to e^{\sqrt{-1}\mu t}$, which is the ch.f. of the distribution degenerated

at μ (i.e., the point mass probability measure at μ). By Theorem 1.9(ii), $T_n/n \to_d \mu$. From Theorem 1.8(vii), this also shows that $T_n/n \to_p \mu$.

Similarly, $\mu = 0$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$ imply

$$\phi_{T_n/\sqrt{n}}(t) = \left[1 - \frac{\sigma^2 t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n$$

for any $t \in \mathcal{R}$, which implies that $\phi_{T_n/\sqrt{n}}(t) \to e^{-\sigma^2 t^2/2}$, the ch.f. of $N(0, \sigma^2)$. Hence $T_n/\sqrt{n} \to_d N(0, \sigma^2)$. If $\mu \neq 0$, a transformation of $Y_i = X_i - \mu$ leads to $(T_n - n\mu)/\sqrt{n} \to_d N(0, \sigma^2)$.

Suppose now that $X_1, ..., X_n$ are random k-vectors and $\mu = EX_1$ and $\Sigma = \mathrm{Var}(X_1)$ are finite. For any fixed $c \in \mathcal{R}^k$, it follows from the previous discussion that $(c^\tau T_n - nc^\tau\mu)/\sqrt{n} \to_d N(0, c^\tau \Sigma c)$. From Theorem 1.9(iii) and a property of the normal distribution (Exercise 81), we conclude that $(T_n - n\mu)/\sqrt{n} \to_d N_k(0, \Sigma)$.

Example 1.29. Let $X_1, ..., X_n$ be independent random variables having a common Lebesgue p.d.f. $f(x) = (1 - \cos x)/(\pi x^2)$. Then the ch.f. of X_1 is $\max\{1 - |t|, 0\}$ (Exercise 73) and the ch.f. of $T_n/n = (X_1 + \cdots + X_n)/n$ is

$$\left(\max\left\{1 - \frac{|t|}{n}, 0\right\}\right)^n \to e^{-|t|}, \qquad t \in \mathcal{R}.$$

Since $e^{-|t|}$ is the ch.f. of the Cauchy distribution $C(0, 1)$ (Table 1.2), we conclude that $T_n/n \to_d X$, where X has the Cauchy distribution $C(0, 1)$.

Does this result contradict the first result in Example 1.28?

Other examples are given in Exercises 135-140.

The following result can be used to check whether $X_n \to_d X$ when X has a p.d.f. f and X_n has a p.d.f. f_n.

Proposition 1.18 (Scheffé's theorem). Let $\{f_n\}$ be a sequence of p.d.f.'s on $\mathcal{R}^k$ w.r.t. a measure ν. Suppose that $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. ν and $f(x)$ is a p.d.f. w.r.t. ν. Then $\lim_{n\to\infty} \int |f_n(x) - f(x)| d\nu = 0$.
Proof. Let $g_n(x) = [f(x) - f_n(x)]I_{\{f \geq f_n\}}(x)$, $n = 1, 2, ...$ Then

$$\int |f_n(x) - f(x)| d\nu = 2 \int g_n(x) d\nu.$$

Since $0 \leq g_n(x) \leq f(x)$ for all x and $g_n \to 0$ a.e. ν, the result follows from the dominated convergence theorem.

As an example, consider the Lebesgue p.d.f. f_n of the t-distribution t_n (Table 1.2), $n = 1, 2, ...$ One can show (exercise) that $f_n \to f$, where f is the standard normal p.d.f. This is an important result in statistics.

Lecture 14: Convergence of transformations, Slutsky's theorem and δ-method

Transformation is an important tool in statistics.
If X_n converges to X in some sense, is $g(X_n)$ converges to $g(X)$ in the same sense?
The following result (continuous mapping theorem) provides an answer to this question in many problems.

Theorem 1.10. Let $X, X_1, X_2, ...$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$;
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$.
Proof. (i) can be established using a result in calculus.
(iii) follows from Theorem 1.9(i): for any bounded and continuous h, $E[h(g(X_n))] \to E[h(g(X))]$, since $h \circ g$ is bounded and continuous.
To show (ii), we consider the special case of $X = c$ (a constant).
From the continuity of g, for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that $\|g(x) - g(c)\| < \epsilon$ whenever $\|x - c\| < \delta_\epsilon$. Hence,

$$\{\omega : \|g(X_n(\omega)) - g(c)\| < \epsilon\} \subset \{\omega : \|X_n(\omega) - c\| < \delta_\epsilon\}$$

and

$$P(\|g(X_n) - g(c)\| \geq \epsilon) \leq P(\|X_n - c\| \geq \delta_\epsilon).$$

Hence $g(X_n) \to_p g(c)$ follows from $X_n \to_p c$.
Is the previous arguement still valid when c is replaced by the random vector X in the general case? If not, how do we fix the proof?

Example 1.30. (i) Let $X_1, X_2, ...$ be random variables. If $X_n \to_d X$, where X has the $N(0,1)$ distribution, then $X_n^2 \to_d Y$, where Y has the chi-square distribution χ_1^2.
(ii) Let (X_n, Y_n) be random 2-vectors satisfying $(X_n, Y_n) \to_d (X, Y)$, where X and Y are independent random variables having the $N(0,1)$ distribution, then $X_n/Y_n \to_d X/Y$, which has the Cauchy distribution $C(0,1)$.
(iii) Under the conditions in part (ii), $\max\{X_n, Y_n\} \to_d \max\{X, Y\}$, which has the c.d.f. $[\Phi(x)]^2$ ($\Phi(x)$ is the c.d.f. of $N(0,1)$).

In Example 1.30(ii) and (iii), the condition that $(X_n, Y_n) \to_d (X, Y)$ cannot be relaxed to $X_n \to_d X$ and $Y_n \to_d Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of (X_n, Y_n). This is different when $\to_d$ is replaced by $\to_p$ or $\to_{a.s.}$. The following result, which plays an important role in probability and statistics, establishes the convergence in distribution of $X_n + Y_n$ or $X_n Y_n$ when no information regarding the joint c.d.f. of (X_n, Y_n) is provided.

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, ...,\ Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a constant. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n/Y_n \to_d X/c$ if $c \neq 0$.
Proof. We prove (i) only. The proofs of (ii) and (iii) are left as exercises.
Let $t \in \mathcal{R}$ and $\epsilon > 0$ be fixed constants. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \leq t) \\ &\leq P(\{X_n + Y_n \leq t\} \cap \{|Y_n - c| < \epsilon\}) + P(|Y_n - c| \geq \epsilon) \\ &\leq P(X_n \leq t - c + \epsilon) + P(|Y_n - c| \geq \epsilon) \end{aligned}$$

and, similarly,

$$F_{X_n+Y_n}(t) \geq P(X_n \leq t - c - \epsilon) - P(|Y_n - c| \geq \epsilon).$$

If $t - c$, $t - c + \epsilon$, and $t - c - \epsilon$ are continuity points of F_X, then it follows from the previous two inequalities and the hypotheses of the theorem that

$$F_X(t - c - \epsilon) \leq \liminf_n F_{X_n+Y_n}(t) \leq \limsup_n F_{X_n+Y_n}(t) \leq F_X(t - c + \epsilon).$$

Since ϵ can be arbitrary (why?),

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t - c).$$

The result follows from $F_{X+c}(t) = F_X(t - c)$.

An application of Theorem 1.11 is given in the proof of the following important result.

Theorem 1.12. Let $X_1, X_2, ...$ and Y be random k-vectors satisfying

$$a_n(X_n - c) \to_d Y, \tag{1}$$

where $c \in \mathcal{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let g be a function from $\mathcal{R}^k$ to $\mathcal{R}$.
(i) If g is differentiable at c, then

$$a_n[g(X_n) - g(c)] \to_d [\nabla g(c)]^\tau Y, \tag{2}$$

where $\nabla g(x)$ denotes the k-vector of partial derivatives of g at x.
(ii) Suppose that g has continuous partial derivatives of order $m > 1$ in a neighborhood of c, with all the partial derivatives of order j, $1 \leq j \leq m - 1$, vanishing at c, but with the mth-order partial derivatives not all vanishing at c. Then

$$a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}\bigg|_{x=c} Y_{i_1} \cdots Y_{i_m}, \tag{3}$$

where Y_j is the jth component of Y.

Proof. We prove (i) only. The proof of (ii) is similar. Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n[\nabla g(c)]^\tau (X_n - c).$$

If we can show that $Z_n = o_p(1)$, then by (1), Theorem 1.9(iii), and Theorem 1.11(i), result (2) holds.

The differentiability of g at c implies that for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - [\nabla g(c)]^\tau (x - c)| \le \epsilon \|x - c\| \tag{4}$$

whenever $\|x - c\| < \delta_\epsilon$. Let $\eta > 0$ be fixed. By (4),

$$P(|Z_n| \ge \eta) \le P(\|X_n - c\| \ge \delta_\epsilon) + P(a_n \|X_n - c\| \ge \eta/\epsilon).$$

Since $a_n \to \infty$, (1) and Theorem 1.11(ii) imply $X_n \to_p c$. By Theorem 1.10(iii), (1) implies $a_n\|X_n - c\| \to_d \|Y\|$. Without loss of generality, we can assume that η/ϵ is a continuity point of $F_{\|Y\|}$. Then

$$\begin{aligned} \limsup_n P(|Z_n| \ge \eta) &\le \lim_{n\to\infty} P(\|X_n - c\| \ge \delta_\epsilon) \\ &\quad + \lim_{n\to\infty} P(a_n\|X_n - c\| \ge \eta/\epsilon) \\ &= P(\|Y\| \ge \eta/\epsilon). \end{aligned}$$

The proof is complete since ϵ can be arbitrary.

In statistics, we often need a nondegenerated limiting distribution of $a_n[g(X_n) - g(c)]$ so that probabilities involving $a_n[g(X_n) - g(c)]$ can be approximated by the c.d.f. of $[\nabla g(c)]^\tau Y$, if (2) holds. Hence, result (2) is not useful for this purpose if $\nabla g(c) = 0$, and in such cases result (3) may be applied.

A useful method in statistics, called the *delta-method*, is based on the following corollary of Theorem 1.12.

Corollary 1.1. Assume the conditions of Theorem 1.12. If Y has the $N_k(0, \Sigma)$ distribution, then

$$a_n[g(X_n) - g(c)] \to_d N\left(0, [\nabla g(c)]^\tau \Sigma \nabla g(c)\right).$$

Example 1.31. Let $\{X_n\}$ be a sequence of random variables satisfying $\sqrt{n}(X_n - c) \to_d N(0, 1)$. Consider the function $g(x) = x^2$. If $c \ne 0$, then an application of Corollary 1.1 gives that $\sqrt{n}(X_n^2 - c^2) \to_d N(0, 4c^2)$. If $c = 0$, the first-order derivative of g at 0 is 0 but the second-order derivative of $g \equiv 2$. Hence, an application of result (3) gives that $nX_n^2 \to_d [N(0, 1)]^2$, which has the chi-square distribution χ_1^2 (Example 1.14). The last result can also be obtained by applying Theorem 1.10(iii).

Lecture 15: The law of large numbers

The law of large numbers concerns the limiting behavior of a sum of random variables.
The weak law of large numbers (WLLN) refers to convergence in probability.
Te strong law of large numbers (SLLN) refers to a.s. convergence.

Lemma 1.6. (Kronecker's lemma). Let $x_n \in \mathcal{R}$, $a_n \in \mathcal{R}$, $0 < a_n \leq a_{n+1}$, $n = 1, 2, ...$, and $a_n \to \infty$. If the series $\sum_{n=1}^{\infty} x_n/a_n$ converges, then $a_n^{-1} \sum_{i=1}^{n} x_i \to 0$.

Our first result gives the WLLN and SLLN for a sequence of independent and identically distributed (i.i.d.) random variables.

Theorem 1.13. Let $X_1, X_2, ...$ be i.i.d. random variables.
(i) (The WLLN). A necessary and sufficient condition for the existence of a sequence of real numbers $\{a_n\}$ for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to_p 0 \tag{1}$$

is that $nP(|X_1| > n) \to 0$, in which case we may take $a_n = E(X_1 I_{\{|X_1| \leq n\}})$.
(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c \tag{2}$$

is that $E|X_1| < \infty$, in which case $c = EX_1$ and

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - EX_1) \to_{a.s.} 0 \tag{3}$$

for any bounded sequence of real numbers $\{c_i\}$.
Proof. (i) We prove the sufficiency. The proof of necessity can be found in Petrov (1975). Consider a sequence of random variables obtained by truncating X_j's at n: $Y_{nj} = X_j I_{\{|X_j| \leq n\}}$. Let $T_n = X_1 + \cdots + X_n$ and $Z_n = Y_{n1} + \cdots + Y_{nn}$. Then

$$P(T_n \neq Z_n) \leq \sum_{j=1}^{n} P(Y_{nj} \neq X_j) = nP(|X_1| > n) \to 0. \tag{4}$$

For any $\epsilon > 0$, it follows from Chebyshev's inequality that

$$P\left(\left|\frac{Z_n - EZ_n}{n}\right| > \epsilon\right) \leq \frac{\mathrm{Var}(Z_n)}{\epsilon^2 n^2} = \frac{\mathrm{Var}(Y_{n1})}{\epsilon^2 n} \leq \frac{EY_{n1}^2}{\epsilon^2 n},$$

where the last equality follows from the fact that Y_{nj}, $j = 1, ..., n$, are i.i.d.
From integration by parts, we obtain that

$$\frac{EY_{n1}^2}{n} = \frac{1}{n}\int_{[0,n]} x^2 dF_{|X_1|}(x) = \frac{2}{n}\int_0^n xP(|X_1| > x)dx - nP(|X_1| > n),$$

which converges to 0 since $nP(|X_1| > n) \to 0$ (why?). This proves that $(Z_n - EZ_n)/n \to_p 0$, which together with (4) and the fact that $EY_{nj} = E(X_1 I_{\{|X_1| \leq n\}})$ imply the result.

(ii) The proof for sufficiency is given in the textbook.
We prove the necessity. Suppose that (2) holds for some $c \in \mathcal{R}$. Then

$$\frac{X_n}{n} = \frac{T_n}{n} - c - \frac{n-1}{n}\left(\frac{T_{n-1}}{n-1} - c\right) + \frac{c}{n} \to_{a.s.} 0.$$

From Exercise 114, $X_n/n \to_{a.s.} 0$ and the i.i.d. assumption on X_n's imply

$$\sum_{n=1}^{\infty} P(|X_n| \geq n) = \sum_{n=1}^{\infty} P(|X_1| \geq n) < \infty,$$

which implies $E|X_1| < \infty$ (Exercise 54). From the proved sufficiency, $c = EX_1$.

If $E|X_1| < \infty$, then a_n in (1) converges to EX_1 and result (1) is actually established in Example 1.28 in a much simpler way.
On the other hand, if $E|X_1| < \infty$, then the stronger result (2) can be obtained.
Some results for the case of $E|X_1| = \infty$ can be found in Exercise 148 and Theorem 5.4.3 in Chung (1974).

The next result is for sequences of independent but not necessarily identically distributed random variables.

Theorem 1.14. Let $X_1, X_2, ...$ be independent random variables with finite expectations.
(i) (The SLLN). If there is a constant $p \in [1, 2]$ such that

$$\sum_{i=1}^{\infty} \frac{E|X_i|^p}{i^p} < \infty, \tag{5}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{a.s.} 0. \tag{6}$$

(ii) (The WLLN). If there is a constant $p \in [1, 2]$ such that

$$\lim_{n\to\infty} \frac{1}{n^p} \sum_{i=1}^{n} E|X_i|^p = 0, \tag{7}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{p} 0. \tag{8}$$

Proof. See the textbook.

Note that (5) implies (7) (Lemma 1.6).
The result in Theorem 1.14(i) is called Kolmogorov's SLLN when $p = 2$ and is due to Marcinkiewicz and Zygmund when $1 \leq p < 2$.
An obvious sufficient condition for (5) with $p \in (1, 2]$ is $\sup_n E|X_n|^p < \infty$.

The WLLN and SLLN have many applications in probability and statistics.

Example 1.32. Let f and g be continuous functions on $[0,1]$ satisfying $0 \le f(x) \le Cg(x)$ for all x, where $C > 0$ is a constant. We now show that

$$\lim_{n\to\infty} \int_0^1 \int_0^1 \cdots \int_0^1 \frac{\sum_{i=1}^n f(x_i)}{\sum_{i=1}^n g(x_i)} dx_1 dx_2 \cdots dx_n = \frac{\int_0^1 f(x)dx}{\int_0^1 g(x)dx} \tag{9}$$

(assuming that $\int_0^1 g(x)dx \neq 0$). Let $X_1, X_2, ...$ be i.i.d. random variables having the uniform distribution on $[0,1]$. By Theorem 1.2, $E[f(X_1)] = \int_0^1 f(x)dx < \infty$ and $E[g(X_1)] = \int_0^1 g(x)dx < \infty$. By the SLLN (Theorem 1.13(ii)),

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \to_{a.s.} E[f(X_1)],$$

and the same result holds when f is replaced by g. By Theorem 1.10(i),

$$\frac{\sum_{i=1}^n f(X_i)}{\sum_{i=1}^n g(X_i)} \to_{a.s.} \frac{E[f(X_1)]}{E[g(X_1)]}. \tag{10}$$

Since the random variable on the left-hand side of (10) is bounded by C, result (9) follows from the dominated convergence theorem and the fact that the left-hand side of (9) is the expectation of the random variable on the left-hand side of (10).

Example: Let $T_n = \sum_{i=1}^n X_i$, where X_n's are independent random variables satisfying $P(X_n = \pm n^\theta) = 0.5$ and $\theta > 0$ is a constant.
We want to show that $T_n/n \to_{a.s.} 0.$ when $\theta < 0.5$.
When $\theta < 0.5$,

$$\sum_{n=1}^\infty \frac{EX_n^2}{n^2} = \sum_{n=1}^\infty \frac{n^{2\theta}}{n^2} < \infty.$$

By the Kolmogorov strong law of large numbers, $T_n/n \to_{a.s.} 0$.

Example (Exercise 165): Let $X_1, X_2, ...$ be independent random variables. Suppose that $\sum_{j=1}^n (X_j - EX_j)/\sigma_n \to_d N(0,1)$, where $\sigma_n^2 = \mathrm{Var}(\sum_{j=1}^n X_j)$.
We want to show that $n^{-1}\sum_{j=1}^n (X_j - EX_j) \to_p 0$ if and only if $\sigma_n/n \to 0$.
If $\sigma_n/n \to 0$, then by Slutsky's theorem,

$$\frac{1}{n}\sum_{j=1}^n (X_j - EX_j) = \frac{\sigma_n}{n}\frac{1}{\sigma_n}\sum_{j=1}^n (X_j - EX_j) \to_d 0.$$

Assume now σ_n/n does not converge to 0 but $n^{-1}\sum_{j=1}^n (X_j - EX_j) \to_p 0$. Without loss of generality, assume that $\sigma_n/n \to c \in (0,\infty]$. By Slutsky's theorem,

$$\frac{1}{\sigma_n}\sum_{j=1}^n (X_j - EX_j) = \frac{n}{\sigma_n}\frac{1}{n}\sum_{j=1}^n (X_j - EX_j) \to_p 0.$$

This contradicts the fact that $\sum_{j=1}^n (X_j - EX_j)/\sigma_n \to_d N(0,1)$. Hence, $n^{-1}\sum_{j=1}^n (X_j - EX_j)$ does not converge to 0 in probability.

Lecture 16: The central limit theorem

The WLLN and SLLN may not be useful in approximating the distributions of (normalized) sums of independent random variables.
We need to use the *central limit theorem* (CLT), which plays a fundamental role in statistical asymptotic theory.

Theorem 1.15 (Lindeberg's CLT). Let $\{X_{nj}, j = 1, ..., k_n\}$ be independent random variables with $0 < \sigma_n^2 = \text{Var}(\sum_{j=1}^{k_n} X_{nj}) < \infty$, $n = 1, 2,...$, and $k_n \to \infty$ as $n \to \infty$. If

$$\frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} E\left[(X_{nj} - EX_{nj})^2 I_{\{|X_{nj}-EX_{nj}|>\epsilon\sigma_n\}}\right] \to 0 \quad \text{for any } \epsilon > 0, \tag{1}$$

then

$$\frac{1}{\sigma_n} \sum_{j=1}^{k_n} (X_{nj} - EX_{nj}) \to_d N(0, 1). \tag{2}$$

Proof. Considering $(X_{nj} - EX_{nj})/\sigma_n$, without loss of generality we may assume $EX_{nj} = 0$ and $\sigma_n^2 = 1$ in this proof.
Let $t \in \mathcal{R}$ be given. From the inequality $|e^{\sqrt{-1}tx} - (1 + \sqrt{-1}tx - t^2x^2/2)| \le \min\{|tx|^2, |tx|^3\}$, the ch.f. of X_{nj} satisfies

$$\left|\phi_{X_{nj}}(t) - \left(1 - t^2\sigma_{nj}^2/2\right)\right| \le E\left(\min\{|tX_{nj}|^2, |tX_{nj}|^3\}\right), \tag{3}$$

where $\sigma_{nj}^2 = \text{Var}(X_{nj})$. For any $\epsilon > 0$, the right-hand side of (3) is bounded by

$$E(|tX_{nj}|^3 I_{\{|X_{nj}|<\epsilon\}}) + E(|tX_{nj}|^2 I_{\{|X_{nj}|\ge\epsilon\}}),$$

which is bounded by

$$\epsilon|t|^3\sigma_{nj}^2 + t^2 E(X_{nj}^2 I_{\{|X_{nj}|\ge\epsilon\}}).$$

Summing over j and using condition (1), we obtain that

$$\sum_{j=1}^{k_n} \left|\phi_{X_{nj}}(t) - \left(1 - t^2\sigma_{nj}^2/2\right)\right| \to 0. \tag{4}$$

By condition (1), $\max_{j\le k_n} \sigma_{nj}^2 \le \epsilon^2 + \max_{j\le k_n} E(X_{nj}^2 I_{\{|X_{nj}|>\epsilon\}}) \to \epsilon^2$ for arbitrary $\epsilon > 0$. Hence

$$\lim_{n\to\infty} \max_{j\le k_n} \frac{\sigma_{nj}^2}{\sigma_n^2} = 0. \tag{5}$$

(Note that $\sigma_n^2 = 1$ is assumed for convenience.) This implies that $1 - t^2\sigma_{nj}^2$ are all between 0 and 1 for large enough n. Using the inequality

$$|a_1 \cdots a_m - b_1 \cdots b_m| \le \sum_{j=1}^{m} |a_j - b_j|$$

for any complex numbers a_j's and b_j's with $|a_j| \le 1$ and $|b_j| \le 1$, $j = 1, ..., m$, we obtain that

$$\left| \prod_{j=1}^{k_n} e^{-t^2\sigma_{nj}^2/2} - \prod_{j=1}^{k_n} \left(1 - t^2\sigma_{nj}^2/2\right) \right| \le \sum_{j=1}^{k_n} \left| e^{-t^2\sigma_{nj}^2/2} - \left(1 - t^2\sigma_{nj}^2/2\right) \right|,$$

which is bounded by $t^4 \sum_{j=1}^{k_n} \sigma_{nj}^4 \le t^4 \max_{j \le k_n} \sigma_{nj}^2 \to 0$, since $|e^x - 1 - x| \le x^2/2$ if $|x| \le \frac{1}{2}$ and $\sum_{j=1}^{k_n} \sigma_{nj}^2 = \sigma_n^2 = 1$. Also,

$$\left| \prod_{j=1}^{k_n} \phi_{X_{nj}}(t) - \prod_{j=1}^{k_n} \left(1 - t^2\sigma_{nj}^2/2\right) \right|$$

is bounded by the quantity on the left-hand side of (4) and, hence, converges to 0 by (4). Thus,

$$\prod_{j=1}^{k_n} \phi_{X_{nj}}(t) = \prod_{j=1}^{k_n} e^{-t^2\sigma_{nj}^2/2} + o(1) = e^{-t^2/2} + o(1).$$

This shows that the ch.f. of $\sum_{j=1}^{k_n} X_{nj}$ converges to the ch.f. of $N(0,1)$ for every t. By Theorem 1.9(ii), the result follows.

Condition (1) is called Lindeberg's condition.
From the proof, Lindeberg's condition implies (5), which is called Feller's condition.
Feller's condition (5) means that all terms in the sum $\sigma_n^2 = \sum_{j=1}^{k_n} \sigma_{nj}^2$ are uniformly negligible as $n \to \infty$.
If Feller's condition is assumed, then Lindeberg's condition is not only sufficient but also necessary for result (2), which is the well-known Lindeberg-Feller CLT.
A proof can be found in Billingsley (1986, pp. 373-375).
Note that neither Lindeberg's condition nor Feller's condition is necessary for result (2) (Exercise 158).

A sufficient condition for Lindeberg's condition is the following Liapounov's condition, which is somewhat easier to verify:

$$\frac{1}{\sigma_n^{2+\delta}} \sum_{j=1}^{k_n} E|X_{nj} - EX_{nj}|^{2+\delta} \to 0 \quad \text{for some } \delta > 0. \tag{6}$$

Example 1.33. Let $X_1, X_2, ...$ be independent random variables. Suppose that X_i has the binomial distribution $Bi(p_i, 1)$, $i = 1, 2, ...$, and that $\sigma_n^2 = \sum_{i=1}^n \mathrm{Var}(X_i) = \sum_{i=1}^n p_i(1-p_i) \to \infty$ as $n \to \infty$. For each i, $EX_i = p_i$ and $E|X_i - EX_i|^3 = (1-p_i)^3 p_i + p_i^3(1-p_i) \le 2p_i(1-p_i)$. Hence $\sum_{i=1}^n E|X_i - EX_i|^3 \le 2\sigma_n^2$, i.e., Liapounov's condition (6) holds with $\delta = 1$. Thus, by Theorem 1.15,

$$\frac{1}{\sigma_n} \sum_{i=1}^n (X_i - p_i) \to_d N(0,1). \tag{7}$$

It can be shown (exercise) that the condition $\sigma_n \to \infty$ is also necessary for result (7).

Useful corollaries of Theorem 1.15 (and Theorem 1.9(iii))

Corollary 1.2 (Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random k-vectors with a finite $\Sigma = \mathrm{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - EX_1) \to_d N_k(0, \Sigma).$$

Corollary 1.3. Let $X_{ni} \in \mathcal{R}^{m_i}$, $i = 1, ..., k_n$, be independent random vectors with $m_i \le m$ (a fixed integer), $n = 1, 2, ...$, $k_n \to \infty$ as $n \to \infty$, and $\inf_{i,n} \lambda_-[\mathrm{Var}(X_{ni})] > 0$, where $\lambda_-[A]$ is the smallest eigenvalue of A. Let $c_{ni} \in \mathcal{R}^{m_i}$ be vectors such that

$$\lim_{n\to\infty} \left(\max_{1\le i\le k_n} \|c_{ni}\|^2 \Big/ \sum_{i=1}^{k_n} \|c_{ni}\|^2 \right) = 0.$$

(i) Suppose that $\sup_{i,n} E\|X_{ni}\|^{2+\delta} < \infty$ for some $\delta > 0$. Then

$$\sum_{i=1}^{k_n} c_{ni}^\tau (X_{ni} - EX_{ni}) \Big/ \left[\sum_{i=1}^{k_n} \mathrm{Var}(c_{ni}^\tau X_{ni}) \right]^{1/2} \to_d N(0, 1). \tag{8}$$

(ii) Suppose that whenever $m_i = m_j$, $1 \le i < j \le k_n$, $n = 1, 2, ...$, X_{ni} and X_{nj} have the same distribution with $E\|X_{ni}\|^2 < \infty$. Then (8) holds.

Proving Corollary 1.3 is a good exercise.

Applications of these corollaries can be found in later chapters.

More results on the CLT can be found, for example, in Serfling (1980) and Shorack and Wellner (1986).

Let Y_n be a sequence of random variables, $\{\mu_n\}$ and $\{\sigma_n\}$ be sequences of real numbers such that $\sigma_n > 0$ for all n, and $(Y_n - \mu_n)/\sigma_n \to_d N(0, 1)$. Then, by Proposition 1.16,

$$\lim_{n\to\infty} \sup_x |F_{(Y_n-\mu_n)/\sigma_n}(x) - \Phi(x)| = 0, \tag{9}$$

where Φ is the c.d.f. of $N(0, 1)$.
This implies that for any sequence of real numbers $\{c_n\}$, $\lim_{n\to\infty} |P(Y_n \le c_n) - \Phi(\frac{c_n-\mu_n}{\sigma_n})| = 0$, i.e., $P(Y_n \le c_n)$ can be approximated by $\Phi(\frac{c_n-\mu_n}{\sigma_n})$, regardless of whether $\{c_n\}$ has a limit.
Since $\Phi(\frac{t-\mu_n}{\sigma_n})$ is the c.d.f. of $N(\mu_n, \sigma_n^2)$, Y_n is said to be *asymptotically distributed* as $N(\mu_n, \sigma_n^2)$ or simply *asymptotically normal*.
For example, $\sum_{i=1}^{k_n} c_{ni}^\tau X_{ni}$ in Corollary 1.3 is asymptotically normal.
This can be extended to random vectors.
For example, $\sum_{i=1}^{n} X_i$ in Corollary 1.2 is asymptotically distributed as $N_k(nEX_1, n\Sigma)$.

Lecture 17: Populations, samples, models, and statistics

One or a series of random experiments is performed.
Some data from the experiment(s) are collected.
Planning experiments and collecting data (not discussed in the textbook).
Data analysis: extract information from the data, interpret the results, and draw some conclusions.

A descriptive data analysis: summary measures of the data, such as the mean, median, range, standard deviation, etc., and some graphical displays, such as the histogram and box-and-whisker diagram, etc.

It is simple and requires almost no assumptions, but may not allow us to gain enough insight into the problem.
We focus on more sophisticated methods of analyzing data: *statistical inference* and *decision theory*.

The data set is a realization of a random element defined on a probability space $(\Omega, \mathcal{F}, P)$
P is called the *population*.
The data set or the random element that produces the data is called a *sample* from P.
The size of the data set is called the *sample size*.
A population P is *known* if and only if $P(A)$ is a known value for every event $A \in \mathcal{F}$.
In a statistical problem, the population P is at least partially unknown.
We would like to deduce some properties of P based on the available sample.

Examples 2.1-2.3

A *statistical model* (a set of assumptions) on the population P in a given problem is often postulated to make the analysis possible or easy.
Although testing the correctness of postulated models is part of statistical inference and decision theory, postulated models are often based on knowledge of the problem under consideration.

Definition 2.1. A set of probability measures P_θ on $(\Omega, \mathcal{F})$ indexed by a *parameter* $\theta \in \Theta$ is said to be a *parametric family* if and only if $\Theta \subset \mathcal{R}^d$ for some fixed positive integer d *and* each P_θ is a *known* probability measure when θ is known. The set Θ is called the *parameter space* and d is called its *dimension*.

Parametric model: the population P is in a parametric family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$
$\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is *identifiable* if and only if $\theta_1 \neq \theta_2$ and $\theta_i \in \Theta$ imply $P_{\theta_1} \neq P_{\theta_2}$.
In most cases an identifiable parametric family can be obtained through reparameterization.

A family of populations $\mathcal{P}$ is dominated by ν (a σ-finite measure) if $P \ll \nu$ for all $P \in \mathcal{P}$
$\mathcal{P}$ can be identified by the family of densities $\{\frac{dP}{d\nu} : P \in \mathcal{P}\}$ or $\{\frac{dP_\theta}{d\nu} : \theta \in \Theta\}$.

Parametric methods: methods designed for parametric models

Example (The k-dimensional normal family).

$$\mathcal{P} = \{N_k(\mu, \Sigma) : \ \mu \in \mathcal{R}^k, \ \Sigma \in \mathcal{M}_k\},$$

where $\mathcal{M}_k$ is a collection of $k \times k$ symmetric positive definite matrices.
This family is dominated by the Lebesgue measure on $\mathcal{R}^k$.
When $k = 1$, $\mathcal{P} = \{N(\mu, \sigma^2) : \ \mu \in \mathcal{R}, \ \sigma^2 > 0\}$.

Nonparametric family: $\mathcal{P}$ is not parametric according to Definition 2.1.
A nonparametric model: the population P is in a given nonparametric family.

Examples of nonparametric family on $(\mathcal{R}^k, \mathcal{B}^k)$:
(1) The joint c.d.f.'s are continuous.
(2) The joint c.d.f.'s have finite moments of order $\leq$ a fixed integer.
(3) The joint c.d.f.'s have p.d.f.'s (e.g., Lebesgue p.d.f.'s).
(4) $k = 1$ and the c.d.f.'s are symmetric.
(5) The family of all probability measures on $(\mathcal{R}^k, \mathcal{B}^k)$.

Nonparametric methods: methods designed for nonparametric models

Semi-parametric models and methods

Statistics and their distributions

Our data set is a realization of a sample (random vector) X from an unknown population P
Statistic $T(X)$: A measurable function T of X; $T(X)$ is a known value whenever X is known.
Statistical analyses are based on various statistics, for various purposes.
X itself is a statistic, but it is a trivial statistic.
The range of a nontrivial statistic $T(X)$ is usually simpler than that of X.
For example, X may be a random n-vector and $T(X)$ may be a random p-vector with a p much smaller than n.
$\sigma(T(X)) \subset \sigma(X)$ and the two σ-fields are the same if and only if T is one-to-one.
Usually $\sigma(T(X))$ simplifies $\sigma(X)$, i.e., a statistic provides a "reduction" of the σ-field.

The "information" within the statistic $T(X)$ concerning the unknown distribution of X is contained in the σ-field $\sigma(T(X))$.
S is any other statistic for which $\sigma(S(X)) = \sigma(T(X))$.
Then, by Lemma 1.2, S is a measurable function of T, and T is a measurable function of S.
Thus, once the value of S (or T) is known, so is the value of T (or S).
It is not the particular values of a statistic that contain the information, but the generated σ-field of the statistic.
Values of a statistic may be important for other reasons.

A statistic $T(X)$ is a random element.
If the distribution of X is unknown, then the distribution of T may also be unknown, although T is a known function.
Finding the form of the distribution of T is one of the major problems in statistical inference and decision theory.

Since T is a transformation of X, tools we learn in Chapter 1 for transformations may be useful in finding the distribution or an approximation to the distribution of $T(X)$.

Example 2.8. Let $X_1, ..., X_n$ be i.i.d. random variables having a common distribution P and $X = (X_1, ..., X_n)$.
The sample mean $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ and sample variance $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ are two commonly used statistics.
Can we find the joint or the marginal distributions of $\bar{X}$ and S^2?
It depends on how much we know about P.

Moments of $\bar{X}$ and S^2
If P has a finite mean μ, then $E\bar{X} = \mu$.
If $P \in \{P_\theta : \theta \in \Theta\}$, then $E\bar{X} = \int x dP_\theta = \mu(\theta)$ for some function $\mu(\cdot)$.
Even if the form of μ is known, $\mu(\theta)$ is still unknown when θ is unknown.
If P has a finite variance σ^2, then $\mathrm{Var}(\bar{\mathrm{X}}) = \sigma^2/\mathrm{n}$, which equals $\sigma^2(\theta)/n$ for some function $\sigma^2(\cdot)$ if P is in a parametric family.
With a finite $\sigma^2 = \mathrm{Var}(\mathrm{X}_1)$, we can also obtain that $ES^2 = \sigma^2$.
With a finite $E|X_1|^3$, we can obtain $E(\bar{X})^3$ and $\mathrm{Cov}(\bar{\mathrm{X}}, \mathrm{S}^2)$.
With a finite $E|X_1|^4$, we can obtain $\mathrm{Var}(\mathrm{S}^2)$ (exercise).

The distribution of $\bar{X}$
If P is in a parametric family, we can often find the distribution of $\bar{X}$.
See Example 1.20 and some exercises in §1.6.
For example, $\bar{X}$ is $N(\mu, \sigma^2/n)$ if P is $N(\mu, \sigma^2)$;
$n\bar{X}$ has the gamma distribution $\Gamma(n, \theta)$ if P is the exponential distribution $E(0, \theta)$.
If P is not in a parametric family, then it is usually hard to find the exact form of the distribution of $\bar{X}$.
One can use the CLT to obtain an approximation to the distribution of $\bar{X}$.
Applying Corollary 1.2 (for the case of $k = 1$), we obtain that $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$, where μ and σ^2 are the mean and variance of P, respectively, and are assumed to be finite.
The distribution of $\bar{X}$ can be approximated by $N(\mu, \sigma^2/n)$

The distribution of S^2
If P is $N(\mu, \sigma^2)$, then $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1} (see Example 2.18).
An approximate distribution for S^2 can be obtained from the approximate joint distribution of $\bar{X}$ and S^2 discussed next.

Joint distribution of $\bar{X}$ and S^2
If P is $N(\mu, \sigma^2)$, then $\bar{X}$ and S^2 are independent (Example 2.18).
Hence, the joint distribution of $(\bar{X}, S^2)$ is the product of the marginal distributions of $\bar{X}$ and S^2 given in the previous discussion.
Without the normality assumption, an approximate joint distribution can be obtained.
Assume that $\mu = EX_1$, $\sigma^2 = \mathrm{Var}(\mathrm{X}_1)$, and $E|X_1|^4$ are finite.
Let $Y_i = (X_i - \mu, (X_i - \mu)^2)$, $i = 1, ..., n$.

$Y_1, ..., Y_n$ are i.i.d. random 2-vectors with $EY_1 = (0, \sigma^2)$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma^2 & E(X_1 - \mu)^3 \\ E(X_1 - \mu)^3 & E(X_1 - \mu)^4 - \sigma^4 \end{pmatrix}.$$

Note that $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i = (\bar{X} - \mu, \tilde{S}^2)$, where $\tilde{S}^2 = n^{-1} \sum_{i=1}^n (X_i - \mu)^2$.
Applying the CLT (Corollary 1.2) to Y_i's, we obtain that

$$\sqrt{n}(\bar{X} - \mu, \tilde{S}^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Since

$$S^2 = \frac{n}{n-1} \left[\tilde{S}^2 - (\bar{X} - \mu)^2 \right]$$

and $\bar{X} \to_{a.s.} \mu$ (the SLLN), an application of Slutsky's theorem leads to

$$\sqrt{n}(\bar{X} - \mu, S^2 - \sigma^2) \to_d N_2(0, \Sigma).$$

Example 2.9 (Order statistics). Let $X = (X_1, ..., X_n)$ with i.i.d. random components.
Let $X_{(i)}$ be the ith smallest value of $X_1, ..., X_n$.
The statistics $X_{(1)}, ..., X_{(n)}$ are called the *order statistics*.
Order statistics is a set of very useful statistics in addition to the sample mean and variance.
Suppose that X_i has a c.d.f. F having a Lebesgue p.d.f. f.
Then the joint Lebesgue p.d.f. of $X_{(1)}, ..., X_{(n)}$ is

$$g(x_1, x_2, ..., x_n) = \begin{cases} n! f(x_1) f(x_2) \cdots f(x_n) & x_1 < x_2 < \cdots < x_n \\ 0 & \text{otherwise.} \end{cases}$$

The joint Lebesgue p.d.f. of $X_{(i)}$ and $X_{(j)}$, $1 \le i < j \le n$, is

$$g_{i,j}(x, y) = \begin{cases} \frac{n![F(x)]^{i-1}[F(y)-F(x)]^{j-i-1}[1-F(y)]^{n-j} f(x) f(y)}{(i-1)!(j-i-1)!(n-j)!} & x < y \\ 0 & \text{otherwise} \end{cases}$$

and the Lebesgue p.d.f. of $X_{(i)}$ is

$$g_i(x) = \frac{n!}{(i-1)!(n-i)!} [F(x)]^{i-1} [1 - F(x)]^{n-i} f(x).$$

Lecture 18: Exponential and location-scale families

Two important types of parametric families

Definition 2.2 (Exponential families). A parametric family $\{P_\theta : \theta \in \Theta\}$ dominated by a σ-finite measure ν on $(\Omega, \mathcal{F})$ is called an *exponential family* if and only if

$$\frac{dP_\theta}{d\nu}(\omega) = \exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}h(\omega), \quad \omega \in \Omega, \tag{1}$$

where $\exp\{x\} = e^x$, T is a random p-vector with a fixed positive integer p, η is a function from Θ to $\mathcal{R}^p$, h is a nonnegative Borel function on $(\Omega, \mathcal{F})$, and

$$\xi(\theta) = \log\left\{\int_\Omega \exp\{[\eta(\theta)]^\tau T(\omega)\}h(\omega)d\nu(\omega)\right\}.$$

In Definition 2.2, T and h are functions of ω only, whereas η and ξ are functions of θ only.
The representation (1) of an exponential family is not unique.
$\tilde{\eta}(\theta) = D\eta(\theta)$ with a $p \times p$ nonsingular matrix D gives another representation (with T replaced by $\tilde{T} = (D^\tau)^{-1}T$).
A change of the measure that dominates the family also changes the representation.
If we define $\lambda(A) = \int_A h d\nu$ for any $A \in \mathcal{F}$, then we obtain an exponential family with densities

$$\frac{dP_\theta}{d\lambda}(\omega) = \exp\{[\eta(\theta)]^\tau T(\omega) - \xi(\theta)\}. \tag{2}$$

In an exponential family, consider the reparameterization $\eta = \eta(\theta)$ and

$$f_\eta(\omega) = \exp\{\eta^\tau T(\omega) - \zeta(\eta)\}h(\omega), \quad \omega \in \Omega, \tag{3}$$

where $\zeta(\eta) = \log\{\int_\Omega \exp\{\eta^\tau T(\omega)\}h(\omega)d\nu(\omega)\}$.
This is the *canonical form* for the family (not unique).
The new parameter η is called the *natural parameter*.
The new parameter space $\Xi = \{\eta(\theta) : \theta \in \Theta\}$, a subset of $\mathcal{R}^p$, is called the *natural parameter space*.
An exponential family in canonical form is called a *natural exponential family*.
If there is an open set contained in the natural parameter space of an exponential family, then the family is said to be of *full rank*.

Example 2.6. The normal family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}, \sigma > 0\}$ is an exponential family, since the Lebesgue p.d.f. of $N(\mu, \sigma^2)$ can be written as

$$\frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log\sigma\right\}.$$

Hence, $T(x) = (x, -x^2)$, $\eta(\theta) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$, $\theta = (\mu, \sigma^2)$, $\xi(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma$, and $h(x) = 1/\sqrt{2\pi}$.
Let $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Then $\Xi = \mathcal{R} \times (0, \infty)$ and we can obtain a natural exponential family of full rank with $\zeta(\eta) = \eta_1^2/(4\eta_2) + \log(1/\sqrt{2\eta_2})$.

A subfamily of the previous normal family, $\{N(\mu, \mu^2) : \mu \in \mathcal{R}, \mu \neq 0\}$, is also an exponential family with the natural parameter $\eta = (\frac{1}{\mu}, \frac{1}{2\mu^2})$ and natural parameter space $\Xi = \{(x, y) : y = 2x^2,\ x \in \mathcal{R},\ y > 0\}$. This exponential family is not of full rank.

For an exponential family, (2) implies that there is a nonzero measure λ such that

$$\frac{dP_\theta}{d\lambda}(\omega) > 0 \qquad \text{for all } \omega \text{ and } \theta. \tag{4}$$

We can use this fact to show that a family of distributions is not an exponential family.
Consider the family of uniform distributions, i.e., P_θ is $U(0, \theta)$ with an unknown $\theta \in (0, \infty)$.
If $\{P_\theta : \theta \in (0, \infty)\}$ is an exponential family, then (4) holds with a nonzero measure λ.
For any $t > 0$, there is a $\theta < t$ such that $P_\theta([t, \infty)) = 0$, which with (4) implies that $\lambda([t, \infty)) = 0$.
Also, for any $t \leq 0$, $P_\theta((-\infty, t]) = 0$, which with (4) implies that $\lambda((-\infty, t]) = 0$.
Since t is arbitrary, $\lambda \equiv 0$.
This contradiction implies that $\{P_\theta : \theta \in (0, \infty)\}$ cannot be an exponential family.

Which of the parametric families from Tables 1.1 and 1.2 are exponential families?

An important exponential family containing multivariate discrete distributions.

Example 2.7 (The multinomial family). Consider an experiment having $k + 1$ possible outcomes with p_i as the probability for the ith outcome, $i = 0, 1, ..., k$, $\sum_{i=0}^k p_i = 1$. In n independent trials of this experiment, let X_i be the number of trials resulting in the ith outcome, $i = 0, 1, ..., k$. Then the joint p.d.f. (w.r.t. counting measure) of $(X_0, X_1, ..., X_k)$ is

$$f_\theta(x_0, x_1, ..., x_k) = \frac{n!}{x_0! x_1! \cdots x_k!} p_0^{x_0} p_1^{x_1} \cdots p_k^{x_k} I_B(x_0, x_1, ..., x_k),$$

where $B = \{(x_0, x_1, ..., x_k) :\ x_i$'s are integers ≥ 0, $\sum_{i=0}^k x_i = n\}$ and $\theta = (p_0, p_1, ..., p_k)$. The distribution of $(X_0, X_1, ..., X_k)$ is called the *multinomial* distribution, which is an extension of the binomial distribution. In fact, the marginal c.d.f. of each X_i is the binomial distribution $Bi(p_i, n)$.
$\{f_\theta :\ \theta \in \Theta\}$ is the multinomial family, where $\Theta = \{\theta \in \mathcal{R}^{k+1} :\ 0 < p_i < 1, \sum_{i=0}^k p_i = 1\}$.
Let $x = (x_0, x_1, ..., x_k)$, $\eta = (\log p_0, \log p_1, ..., \log p_k)$, and $h(x) = [n!/(x_0! x_1! \cdots x_k!)] I_B(x)$.
Then

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{\eta^\tau x\}\, h(x), \qquad x \in \mathcal{R}^{k+1}. \tag{5}$$

Hence, the multinomial family is a natural exponential family with natural parameter η.
However, representation (5) does not provide an exponential family of full rank, since there is no open set of $\mathcal{R}^{k+1}$ contained in the natural parameter space.
A reparameterization leads to an exponential family with full rank.
Using the fact that $\sum_{i=0}^k X_i = n$ and $\sum_{i=0}^k p_i = 1$, we obtain that

$$f_\theta(x_0, x_1, ..., x_k) = \exp\{\eta_*^\tau x_* - \zeta(\eta_*)\}\, h(x), \qquad x \in \mathcal{R}^{k+1}, \tag{6}$$

where $x_* = (x_1, ..., x_k)$, $\eta_* = (\log(p_1/p_0), ..., \log(p_k/p_0))$, and $\zeta(\eta_*) = -n \log p_0$.
The η_*-parameter space is $\mathcal{R}^k$.
Hence, the family of densities given by (6) is a natural exponential family of full rank.

If $X_1, ..., X_m$ are independent random vectors with p.d.f.'s in exponential families, then the p.d.f. of $(X_1, ..., X_m)$ is again in an exponential family.
The following result summarizes some other useful properties of exponential families.
Its proof can be found in Lehmann (1986).

Theorem 2.1. Let $\mathcal{P}$ be a natural exponential family given by (3).
(i) Let $T = (Y, U)$ and $\eta = (\vartheta, \varphi)$, where Y and ϑ have the same dimension.
Then, Y has the p.d.f.

$$f_\eta(y) = \exp\{\vartheta^\tau y - \zeta(\eta)\}$$

w.r.t. a σ-finite measure depending on φ.
In particular, T has a p.d.f. in a natural exponential family.
Furthermore, the conditional distribution of Y given $U = u$ has the p.d.f. (w.r.t. a σ-finite measure depending on u)

$$f_{\vartheta,u}(y) = \exp\{\vartheta^\tau y - \zeta_u(\vartheta)\},$$

which is in a natural exponential family indexed by ϑ.
(ii) If η_0 is an interior point of the natural parameter space, then the m.g.f. ψ_{η_0} of $P_{\eta_0} \circ T^{-1}$ is finite in a neighborhood of 0 and is given by

$$\psi_{\eta_0}(t) = \exp\{\zeta(\eta_0 + t) - \zeta(\eta_0)\}.$$

Furthermore, if f is a Borel function satisfying $\int |f| dP_{\eta_0} < \infty$, then the function

$$\int f(\omega) \exp\{\eta^\tau T(\omega)\} h(\omega) d\nu(\omega)$$

is infinitely often differentiable in a neighborhood of η_0, and the derivatives may be computed by differentiation under the integral sign.

Example 2.5. Let P_θ be the binomial distribution $Bi(\theta, n)$ with parameter θ, where n is a fixed positive integer. Then $\{P_\theta : \theta \in (0, 1)\}$ is an exponential family, since the p.d.f. of P_θ w.r.t. the counting measure is

$$f_\theta(x) = \exp\left\{x \log \tfrac{\theta}{1-\theta} + n \log(1-\theta)\right\} \binom{n}{x} I_{\{0,1,...,n\}}(x)$$

($T(x) = x$, $\eta(\theta) = \log \frac{\theta}{1-\theta}$, $\xi(\theta) = -n \log(1-\theta)$, and $h(x) = \binom{n}{x} I_{\{0,1,...,n\}}(x)$). If we let $\eta = \log \frac{\theta}{1-\theta}$, then $\Xi = \mathcal{R}$ and the family with p.d.f.'s

$$f_\eta(x) = \exp\{x\eta - n \log(1 + e^\eta)\} \binom{n}{x} I_{\{0,1,...,n\}}(x)$$

is a natural exponential family of full rank.
Using Theorem 2.1(ii) and the result in Example 2.5, we obtain that the m.g.f. of the binomial distribution $Bi(\theta, n)$ is

$$\begin{aligned}
\psi_\eta(t) &= \exp\{n \log(1 + e^{\eta+t}) - n \log(1 + e^\eta)\} \\
&= \left(\frac{1 + e^\eta e^t}{1 + e^\eta}\right)^n \\
&= (1 - \theta + \theta e^t)^n.
\end{aligned}$$

Definition 2.3 (Location-scale families). Let P be a known probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$, $\mathcal{V} \subset \mathcal{R}^k$, and $\mathcal{M}_k$ be a collection of $k \times k$ symmetric positive definite matrices. The family

$$\{P_{(\mu,\Sigma)} : \ \mu \in \mathcal{V}, \ \Sigma \in \mathcal{M}_k\} \tag{7}$$

is called a *location-scale family* (on $\mathcal{R}^k$), where

$$P_{(\mu,\Sigma)}(B) = P\left(\Sigma^{-1/2}(B-\mu)\right), \quad B \in \mathcal{B}^k,$$

$\Sigma^{-1/2}(B-\mu) = \{\Sigma^{-1/2}(x-\mu) : \ x \in B\} \subset \mathcal{R}^k$, and $\Sigma^{-1/2}$ is the inverse of the "square root" matrix $\Sigma^{1/2}$ satisfying $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$. The parameters μ and $\Sigma^{1/2}$ are called the location and scale parameters, respectively.

The following are some important examples of location-scale families.
The family $\{P_{(\mu,I_k)} : \ \mu \in \mathcal{R}^k\}$ is a *location family*, where I_k is the $k \times k$ identity matrix.
The family $\{P_{(0,\Sigma)} : \ \Sigma \in \mathcal{M}_k\}$ is a *scale family*.
In some cases, we consider a location-scale family of the form $\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathcal{R}^k, \sigma > 0\}$.
If $X_1, ..., X_k$ are i.i.d. with a common distribution in the location-scale family $\{P_{(\mu,\sigma^2)} : \mu \in \mathcal{R}, \sigma > 0\}$, then the joint distribution of the vector $(X_1, ..., X_k)$ is in the location-scale family $\{P_{(\mu,\sigma^2 I_k)} : \ \mu \in \mathcal{V}, \sigma > 0\}$ with $\mathcal{V} = \{(x, ..., x) \in \mathcal{R}^k : \ x \in \mathcal{R}\}$.

A location-scale family can be generated as follows.
Let X be a random k-vector having a distribution P.
Then the distribution of $\Sigma^{1/2}X + \mu$ is $P_{(\mu,\Sigma)}$.

On the other hand, if X is a random k-vector whose distribution is in the location-scale family (7), then the distribution $DX + c$ is also in the same family, provided that $D\mu + c \in \mathcal{V}$ and $D\Sigma D^\tau \in \mathcal{M}_k$.

Let F be the c.d.f. of P.
Then the c.d.f. of $P_{(\mu,\Sigma)}$ is $F\left(\Sigma^{-1/2}(x-\mu)\right)$, $x \in \mathcal{R}^k$.
If F has a Lebesgue p.d.f. f, then the Lebesgue p.d.f. of $P_{(\mu,\Sigma)}$ is $\mathrm{Det}(\Sigma^{-1/2}) f\left(\Sigma^{-1/2}(x-\mu)\right)$, $x \in \mathcal{R}^k$ (Proposition 1.8).

Many families of distributions in Table 1.2 (§1.3.1) are location, scale, or location-scale families.
For example, the family of exponential distributions $E(a, \theta)$ is a location-scale family on $\mathcal{R}$ with location parameter a and scale parameter θ;
the family of uniform distributions $U(0, \theta)$ is a scale family on $\mathcal{R}$ with a scale parameter θ.
The k-dimensional normal family is a location-scale family on $\mathcal{R}^k$.

Lecture 19: Sufficient statistics and factorization theorem

A statistic $T(X)$ provides a reduction of the σ-field $\sigma(X)$
Does such a reduction results in any loss of information concerning the unknown population?
If a statistic $T(X)$ is fully as informative as the original sample X, then statistical analyses can be done using $T(X)$ that is simpler than X.
The next concept describes what we mean by fully informative.

Definition 2.4 (Sufficiency). Let X be a sample from an unknown population $P \in \mathcal{P}$, where $\mathcal{P}$ is a family of populations. A statistic $T(X)$ is said to be *sufficient* for $P \in \mathcal{P}$ (or for $\theta \in \Theta$ when $\mathcal{P} = \{P_\theta : \ \theta \in \Theta\}$ is a parametric family) if and only if the conditional distribution of X given T is *known* (does not depend on P or θ).

Once we observe X and compute a sufficient statistic $T(X)$, the original data X do not contain any further information concerning the unknown population P (since its conditional distribution is unrelated to P) and can be discarded.
A sufficient statistic $T(X)$ contains all information about P contained in X and provides a reduction of the data if T is not one-to-one.

The concept of sufficiency depends on the given family $\mathcal{P}$.
If T is sufficient for $P \in \mathcal{P}$, then T is also sufficient for $P \in \mathcal{P}_0 \subset \mathcal{P}$ but not necessarily sufficient for $P \in \mathcal{P}_1 \supset \mathcal{P}$.

Example 2.10. Suppose that $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ are i.i.d. from the binomial distribution with the p.d.f. (w.r.t. the counting measure)

$$f_\theta(z) = \theta^z (1-\theta)^{1-z} I_{\{0,1\}}(z), \quad z \in \mathcal{R}, \quad \theta \in (0,1).$$

For any realization x of X, x is a sequence of n ones and zeros.
Consider the statistic $T(X) = \sum_{i=1}^n X_i$, which is the number of ones in X.
T contains all information about θ, since θ is the probability of an occurrence of a one in x.
Given $T = t$ (the number of ones in x), what is left in the data set x is the redundant information about the positions of t ones.
Compute the conditional distribution of X given $T = t$.
$P(T = t) = \binom{n}{t} \theta^t (1-\theta)^{n-t} I_{\{0,1,...,n\}}(t)$.
Let x_i be the ith component of x.
If $t \neq \sum_{i=1}^n x_i$, then $P(X = x, T = t) = 0$. If $t = \sum_{i=1}^n x_i$, then

$$P(X = x, T = t) = \prod_{i=1}^n P(X_i = x_i) = \theta^t (1-\theta)^{n-t} \prod_{i=1}^n I_{\{0,1\}}(x_i).$$

Let $B_t = \{(x_1, ..., x_n) : \ x_i = 0, 1, \ \sum_{i=1}^n x_i = t\}$. Then

$$P(X = x | T = t) = \frac{P(X = x, T = t)}{P(T = t)} = \frac{1}{\binom{n}{t}} I_{B_t}(x)$$

is a known p.d.f. This shows that $T(X)$ is sufficient for $\theta \in (0,1)$, according to Definition 2.4 with the family $\{f_\theta : \theta \in (0,1)\}$.

Finding a sufficient statistic by means of the definition is not convenient
It involves guessing a statistic T that might be sufficient and computing the conditional distribution of X given $T = t$.
For families of populations having p.d.f.'s, a simple way of finding sufficient statistics is to use the factorization theorem.

Lemma 2.1. If a family $\mathcal{P}$ is dominated by a σ-finite measure, then $\mathcal{P}$ is dominated by a probability measure $Q = \sum_{i=1}^{\infty} c_i P_i$, where c_i's are nonnegative constants with $\sum_{i=1}^{\infty} c_i = 1$ and $P_i \in \mathcal{P}$.
Proof. See the textbook.

Theorem 2.2 (The factorization theorem). Suppose that X is a sample from $P \in \mathcal{P}$ and $\mathcal{P}$ is a family of probability measures on $(\mathcal{R}^n, \mathcal{B}^n)$ dominated by a σ-finite measure ν. Then $T(X)$ is sufficient for $P \in \mathcal{P}$ if and only if there are nonnegative Borel functions h (which does not depend on P) on $(\mathcal{R}^n, \mathcal{B}^n)$ and g_P (which depends on P) on the range of T such that

$$\frac{dP}{d\nu}(x) = g_P(T(x))h(x). \tag{1}$$

Proof. (i) Suppose that T is sufficient for $P \in \mathcal{P}$.
For any $A \in \mathcal{B}^n$, $P(A|T)$ does not depend on P.
Let Q be the probability measure in Lemma 2.1.
By Fubini's theorem and the result in Exercise 35 of §1.6,

$$\begin{aligned} Q(A \cap B) &= \sum_{j=1}^{\infty} c_j P_j(A \cap B) \\ &= \sum_{j=1}^{\infty} c_j \int_B P(A|T) dP_j \\ &= \int_B \sum_{j=1}^{\infty} c_j P(A|T) dP_j \\ &= \int_B P(A|T) dQ \end{aligned}$$

for any $B \in \sigma(T)$. Hence, $P(A|T) = E_Q(I_A|T)$ a.s. Q, where $E_Q(I_A|T)$ denotes the conditional expectation of I_A given T w.r.t. Q.
Let $g_P(T)$ be the Radon-Nikodym derivative dP/dQ on the space $(\mathcal{R}^n, \sigma(T), Q)$. Then

$$\begin{aligned} P(A) &= \int P(A|T) dP \\ &= \int E_Q(I_A|T) g_P(T) dQ \\ &= \int E_Q[I_A g_P(T)|T] dQ \\ &= \int_A g_P(T) \frac{dQ}{d\nu} d\nu \end{aligned}$$

for any $A \in \mathcal{B}^n$. Hence, (1) holds with $h = dQ/d\nu$.

(ii) Suppose that (1) holds. Then

$$\frac{dP}{dQ} = \frac{dP}{d\nu} \Big/ \sum_{i=1}^{\infty} c_i \frac{dP_i}{d\nu} = g_{_P}(T) \Big/ \sum_{i=1}^{\infty} g_{_{P_i}}(T) \quad \text{a.s. } Q, \tag{2}$$

where the second equality follows from the result in Exercise 35 of §1.6.
Let $A \in \sigma(X)$ and $P \in \mathcal{P}$.
The sufficiency of T follows from

$$P(A|T) = E_Q(I_A|T) \quad \text{a.s. } P, \tag{3}$$

where $E_Q(I_A|T)$ is given in part (i) of the proof.
This is because $E_Q(I_A|T)$ does not vary with $P \in \mathcal{P}$, and result (3) and Theorem 1.7 imply that the conditional distribution of X given T is determined by $E_Q(I_A|T)$, $A \in \sigma(X)$.
By the definition of conditional probability, (3) follows from

$$\int_B I_A dP = \int_B E_Q(I_A|T) dP \tag{4}$$

for any $B \in \sigma(T)$.
By (2), dP/dQ is a Borel function of T.
Then the right-hand side of (4) is equal to

$$\int_B E_Q(I_A|T) \frac{dP}{dQ} dQ = \int_B E_Q\left(I_A \frac{dP}{dQ} \Big| T \right) dQ = \int_B I_A \frac{dP}{dQ} dQ,$$

which equals the left-hand side of (4).
This proves (4) for any $B \in \sigma(T)$ and completes the proof.

If $\mathcal{P}$ is an exponential family, then Theorem 2.2 can be applied with

$$g_\theta(t) = \exp\{[\eta(\theta)]^\tau t - \xi(\theta)\},$$

i.e., T is a sufficient statistic for $\theta \in \Theta$.
In Example 2.10 the joint distribution of X is in an exponential family with $T(X) = \sum_{i=1}^n X_i$.
Hence, we can conclude that T is sufficient for $\theta \in (0, 1)$ without computing the conditional distribution of X given T.

Example 2.11 (Truncation families). Let $\phi(x)$ be a positive Borel function on $(\mathcal{R}, \mathcal{B})$ such that $\int_a^b \phi(x)dx < \infty$ for any a and b, $-\infty < a < b < \infty$. Let $\theta = (a, b)$, $\Theta = \{(a, b) \in \mathcal{R}^2 : a < b\}$, and

$$f_\theta(x) = c(\theta)\phi(x)I_{(a,b)}(x),$$

where $c(\theta) = \left[\int_a^b \phi(x)dx\right]^{-1}$. Then $\{f_\theta : \theta \in \Theta\}$, called a truncation family, is a parametric family dominated by the Lebesgue measure on $\mathcal{R}$. Let $X_1, ..., X_n$ be i.i.d. random variables having the p.d.f. f_θ. Then the joint p.d.f. of $X = (X_1, ..., X_n)$ is

$$\prod_{i=1}^n f_\theta(x_i) = [c(\theta)]^n I_{(a,\infty)}(x_{(1)})I_{(-\infty,b)}(x_{(n)}) \prod_{i=1}^n \phi(x_i), \tag{5}$$

where $x_{(i)}$ is the ith smallest value of $x_1, ..., x_n$. Let $T(X) = (X_{(1)}, X_{(n)})$, $g_\theta(t_1, t_2) = [c(\theta)]^n I_{(a,\infty)}(t_1)I_{(-\infty,b)}(t_2)$, and $h(x) = \prod_{i=1}^n \phi(x_i)$. By (5) and Theorem 2.2, $T(X)$ is sufficient for $\theta \in \Theta$.

Example 2.12 (Order statistics). Let $X = (X_1, ..., X_n)$ and $X_1, ..., X_n$ be i.i.d. random variables having a distribution $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. Let $X_{(1)}, ..., X_{(n)}$ be the order statistics given in Example 2.9. Note that the joint p.d.f. of X is

$$f(x_1) \cdots f(x_n) = f(x_{(1)}) \cdots f(x_{(n)}).$$

Hence, $T(X) = (X_{(1)}, ..., X_{(n)})$ is sufficient for $P \in \mathcal{P}$. The order statistics can be shown to be sufficient even when $\mathcal{P}$ is not dominated by any σ-finite measure, but Theorem 2.2 is not applicable (see Exercise 31 in §2.6).

Lecture 20: Minimal sufficiency

There are many sufficient statistics for a given family $\mathcal{P}$.
In fact, X (the whole data set) is sufficient.
If T is a sufficient statistic and $T = \psi(S)$, where ψ is measurable and S is another statistic, then S is sufficient.
This is obvious from Theorem 2.2 if the population has a p.d.f., but it can be proved directly from Definition 2.4 (Exercise 25).
For instance, if $X_1, ..., X_n$ are iid with $P(X_i = 1) = \theta$ and $P(X_i = 0) = 1 - \theta$, then $(\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i)$ is sufficient for θ, where m is any fixed integer between 1 and n.
If T is sufficient and $T = \psi(S)$ with a measurable ψ that is not one-to-one, then $\sigma(T) \subset \sigma(S)$ and T is more useful than S, since T provides a further reduction of the data (or σ-field) without loss of information.
Is there a sufficient statistic that provides "maximal" reduction of the data?

If a statement holds except for outcomes in an event A satisfying $P(A) = 0$ for all $P \in \mathcal{P}$, then we say that the statement holds a.s. $\mathcal{P}$.

Definition 2.5 (Minimal sufficiency). Let T be a sufficient statistic for $P \in \mathcal{P}$. T is called a *minimal sufficient* statistic if and only if, for any other statistic S sufficient for $P \in \mathcal{P}$, there is a measurable function ψ such that $T = \psi(S)$ a.s. $\mathcal{P}$.

If both T and S are minimal sufficient statistics, then by definition there is a one-to-one measurable function ψ such that $T = \psi(S)$ a.s. $\mathcal{P}$.
Hence, the minimal sufficient statistic is unique in the sense that two statistics that are one-to-one measurable functions of each other can be treated as one statistic.

Example 2.13. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(\theta, \theta + 1)$, $\theta \in \mathcal{R}$. Suppose that $n > 1$. The joint Lebesgue p.d.f. of $(X_1, ..., X_n)$ is

$$f_\theta(x) = \prod_{i=1}^n I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1, x_{(1)})}(\theta), \quad x = (x_1, ..., x_n) \in \mathcal{R}^n,$$

where $x_{(i)}$ denotes the ith smallest value of $x_1, ..., x_n$. By Theorem 2.2, $T = (X_{(1)}, X_{(n)})$ is sufficient for θ. Note that

$$x_{(1)} = \sup\{\theta : f_\theta(x) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : f_\theta(x) > 0\}.$$

If $S(X)$ is a statistic sufficient for θ, then by Theorem 2.2, there are Borel functions h and g_θ such that $f_\theta(x) = g_\theta(S(x))h(x)$. For x with $h(x) > 0$,

$$x_{(1)} = \sup\{\theta : g_\theta(S(x)) > 0\} \quad \text{and} \quad x_{(n)} = 1 + \inf\{\theta : g_\theta(S(x)) > 0\}.$$

Hence, there is a measurable function ψ such that $T(x) = \psi(S(x))$ when $h(x) > 0$. Since $h > 0$ a.s. $\mathcal{P}$, we conclude that T is minimal sufficient.

Minimal sufficient statistics exist under weak assumptions, e.g., $\mathcal{P}$ contains distributions on $\mathcal{R}^k$ dominated by a σ-finite measure (Bahadur, 1957).

Useful tools for finding minimal sufficient statistics.

Theorem 2.3. Let $\mathcal{P}$ be a family of distributions on $\mathcal{R}^k$.
(i) Suppose that $\mathcal{P}_0 \subset \mathcal{P}$ and a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. If T is sufficient for $P \in \mathcal{P}$ and minimal sufficient for $P \in \mathcal{P}_0$, then T is minimal sufficient for $P \in \mathcal{P}$.
(ii) Suppose that $\mathcal{P}$ contains p.d.f.'s $f_0, f_1, f_2, ...$, w.r.t. a σ-finite measure. Let $f_\infty(x) = \sum_{i=0}^{\infty} c_i f_i(x)$, where $c_i > 0$ for all i and $\sum_{i=0}^{\infty} c_i = 1$, and let $T_i(X) = f_i(x)/f_\infty(x)$ when $f_\infty(x) > 0$, $i = 0, 1, 2, ...$ Then $T(X) = (T_0, T_1, T_2, ...)$ is minimal sufficient for $P \in \mathcal{P}$. Furthermore, if $\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\}$ for all i, then we may replace f_∞ by f_0, in which case $T(X) = (T_1, T_2, ...)$ is minimal sufficient for $P \in \mathcal{P}$.
(iii) Suppose that $\mathcal{P}$ contains p.d.f.'s f_P w.r.t. a σ-finite measure and that there exists a sufficient statistic $T(X)$ such that, for any possible values x and y of X, $f_P(x) = f_P(y)\phi(x, y)$ for all P implies $T(x) = T(y)$, where ϕ is a measurable function. Then $T(X)$ is minimal sufficient for $P \in \mathcal{P}$.
Proof. (i) If S is sufficient for $P \in \mathcal{P}$, then it is also sufficient for $P \in \mathcal{P}_0$ and, therefore, $T = \psi(S)$ a.s. $\mathcal{P}_0$ holds for a measurable function ψ. The result follows from the assumption that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$.
(ii) Note that $f_\infty > 0$ a.s. $\mathcal{P}$. Let $g_i(T) = T_i$, $i = 0, 1, 2, ...$ Then $f_i(x) = g_i(T(x))f_\infty(x)$ a.s. $\mathcal{P}$. By Theorem 2.2, T is sufficient for $P \in \mathcal{P}$. Suppose that $S(X)$ is another sufficient statistic. By Theorem 2.2, there are Borel functions h and $\tilde{g}_i$ such that $f_i(x) = \tilde{g}_i(S(x))h(x)$, $i = 0, 1, 2, ...$ Then $T_i(x) = \tilde{g}_i(S(x))/\sum_{j=0}^{\infty} c_j \tilde{g}_j(S(x))$ for x's satisfying $f_\infty(x) > 0$. By Definition 2.5, T is minimal sufficient for $P \in \mathcal{P}$. The proof for the case where f_∞ is replaced by f_0 is the same.
(iii) From Bahadur (1957), there exists a minimal sufficient statistic $S(X)$. The result follows if we can show that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}$ for a measurable function ψ. By Theorem 2.2, there are Borel functions g_P and h such that $f_P(x) = g_P(S(x))h(x)$ for all P. Let $A = \{x : h(x) = 0\}$. Then $P(A) = 0$ for all P. For x and y such that $S(x) = S(y)$, $x \notin A$ and $y \notin A$,

$$\begin{aligned} f_P(x) &= g_P(S(x))h(x) \\ &= g_P(S(y))h(x)h(y)/h(y) \\ &= f_P(y)h(x)/h(y) \end{aligned}$$

for all P. Hence $T(x) = T(y)$. This shows that there is a function ψ such that $T(x) = \psi(S(x))$ except for $x \in A$. It remains to show that ψ is measurable. Since S is minimal sufficient, $g(T(X)) = S(X)$ a.s. $\mathcal{P}$ for a measurable function g. Hence g is one-to-one and $\psi = g^{-1}$. The measurability of ψ follows from Theorem 3.9 in Parthasarathy (1967).

Example 2.14. Let $\mathcal{P} = \{f_\theta : \theta \in \Theta\}$ be an exponential family with p.d.f.'s

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\}h(x)$$

Suppose that there exists $\Theta_0 = \{\theta_0, \theta_1, ..., \theta_p\} \subset \Theta$ such that the vectors $\eta_i = \eta(\theta_i) - \eta(\theta_0)$, $i = 1, ..., p$, are linearly independent in $\mathcal{R}^p$. (This is true if the family is of full rank.) We have shown that $T(X)$ is sufficient for $\theta \in \Theta$. We now show that T is in fact minimal sufficient for $\theta \in \Theta$. Let $\mathcal{P}_0 = \{f_\theta : \theta \in \Theta_0\}$. Note that the set $\{x : f_\theta(x) > 0\}$ does not depend on θ. It follows from Theorem 2.3(ii) with $f_\infty = f_{\theta_0}$ that

$$S(X) = \Big(\exp\{\eta_1^\tau T(x) - \xi_1\}, ..., \exp\{\eta_p^\tau T(x) - \xi_p\}\Big)$$

is minimal sufficient for $\theta \in \Theta_0$, where $\xi_i = \xi(\theta_i) - \xi(\theta_0)$. Since η_i's are linearly independent, there is a one-to-one measurable function ψ such that $T(X) = \psi(S(X))$ a.s. $\mathcal{P}_0$. Hence, T is minimal sufficient for $\theta \in \Theta_0$. It is easy to see that a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$. Thus, by Theorem 2.3(i), T is minimal sufficient for $\theta \in \Theta$.

The results in Examples 2.13 and 2.14 can also be proved by using Theorem 2.3(iii).

The sufficiency (and minimal sufficiency) depends on the postulated family $\mathcal{P}$ of populations (statistical models).
It may not be a useful concept if the proposed statistical model is wrong or at least one has some doubts about the correctness of the proposed model.
From the examples in this section and some exercises in §2.6, one can find that for a wide variety of models, statistics such as the sample mean $\bar{X}$, the sample variance S^2, $(X_{(1)}, X_{(n)})$ in Example 2.11, and the order statistics in Example 2.9 are sufficient.
Thus, using these statistics for data reduction and summarization does not lose any information when the true model is one of those models but we do not know exactly which model is correct.

A minimal statistic is not always the "simplest sufficient statistic".
For example, if $\bar{X}$ is minimal sufficent, then so is $(\bar{X}, \exp\{\bar{X}\})$.

Lecture 21: Complete statistics

A statistic $V(X)$ is *ancillary* if its distribution does not depend on the population P
$V(X)$ is *first-order ancillary* if $E[V(X)]$ is independent of P.
A trivial ancillary statistic is the constant statistic $V(X) \equiv c \in \mathcal{R}$.
If $V(X)$ is a nontrivial ancillary statistic, then $\sigma(V(X)) \subset \sigma(X)$ is a nontrivial σ-field that does not contain any information about P.
Hence, if $S(X)$ is a statistic and $V(S(X))$ is a nontrivial ancillary statistic, it indicates that $\sigma(S(X))$ contains a nontrivial σ-field that does not contain any information about P and, hence, the "data" $S(X)$ may be further reduced.
A sufficient statistic T appears to be most successful in reducing the data if no nonconstant function of T is ancillary or even first-order ancillary.

Definition 2.6 (Completeness). A statistic $T(X)$ is said to be *complete* for $P \in \mathcal{P}$ if and only if, for any Borel f, $E[f(T)] = 0$ for all $P \in \mathcal{P}$ implies $f = 0$ a.s. $\mathcal{P}$. T is said to be *boundedly complete* if and only if the previous statement holds for any bounded Borel f.

A complete statistic is boundedly complete.
If T is complete (or boundedly complete) and $S = \psi(T)$ for a measurable ψ, then S is complete (or boundedly complete).
Intuitively, a complete and sufficient statistic should be minimal sufficient (Exercise 48).
A minimal sufficient statistic is not necessarily complete; for example, the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ in Example 2.13 is not complete (Exercise 47).

Finding a complete and sufficient statistic

Proposition 2.1. If P is in an exponential family of full rank with p.d.f.'s given by

$$f_\eta(x) = \exp\{\eta^\tau T(x) - \zeta(\eta)\}h(x),$$

then $T(X)$ is complete and sufficient for $\eta \in \Xi$.
Proof. We have shown that T is sufficient. Suppose that there is a function f such that $E[f(T)] = 0$ for all $\eta \in \Xi$. By Theorem 2.1(i),

$$\int f(t) \exp\{\eta^\tau t - \zeta(\eta)\} d\lambda = 0 \quad \text{for all } \eta \in \Xi,$$

where λ is a measure on $(\mathcal{R}^p, \mathcal{B}^p)$. Let η_0 be an interior point of Ξ. Then

$$\int f_+(t) e^{\eta^\tau t} d\lambda = \int f_-(t) e^{\eta^\tau t} d\lambda \quad \text{for all } \eta \in N(\eta_0), \tag{1}$$

where $N(\eta_0) = \{\eta \in \mathcal{R}^p : \|\eta - \eta_0\| < \epsilon\}$ for some $\epsilon > 0$. In particular,

$$\int f_+(t) e^{\eta_0^\tau t} d\lambda = \int f_-(t) e^{\eta_0^\tau t} d\lambda = c.$$

If $c = 0$, then $f = 0$ a.e. λ. If $c > 0$, then $c^{-1} f_+(t) e^{\eta_0^\tau t}$ and $c^{-1} f_-(t) e^{\eta_0^\tau t}$ are p.d.f.'s w.r.t. λ and (1) implies that their m.g.f.'s are the same in a neighborhood of 0. By Theorem 1.6(ii), $c^{-1} f_+(t) e^{\eta_0^\tau t} = c^{-1} f_-(t) e^{\eta_0^\tau t}$, i.e., $f = f_+ - f_- = 0$ a.e. λ. Hence T is complete.

Example 2.15. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, $\mu \in \mathcal{R}$, $\sigma > 0$. From Example 2.6, the joint p.d.f. of $X_1, ..., X_n$ is

$$(2\pi)^{-n/2} \exp\left\{\eta_1 T_1 + \eta_2 T_2 - n\zeta(\eta)\right\},$$

where $T_1 = \sum_{i=1}^n X_i$, $T_2 = -\sum_{i=1}^n X_i^2$, and $\eta = (\eta_1, \eta_2) = \left(\frac{\mu}{\sigma^2}, \frac{1}{2\sigma^2}\right)$. Hence, the family of distributions for $X = (X_1, ..., X_n)$ is a natural exponential family of full rank ($\Xi = \mathcal{R} \times (0, \infty)$). By Proposition 2.1, $T(X) = (T_1, T_2)$ is complete and sufficient for η. Since there is a one-to-one correspondence between η and $\theta = (\mu, \sigma^2)$, T is also complete and sufficient for θ. It can be shown that any one-to-one measurable function of a complete and sufficient statistic is also complete and sufficient (exercise). Thus, $(\bar{X}, S^2)$ is complete and sufficient for θ, where $\bar{X}$ and S^2 are the sample mean and sample variance, respectively.

Example 2.16. Let $X_1, ..., X_n$ be i.i.d. random variables from P_θ, the uniform distribution $U(0, \theta)$, $\theta > 0$. The largest order statistic, $X_{(n)}$, is complete and sufficient for $\theta \in (0, \infty)$. The sufficiency of $X_{(n)}$ follows from the fact that the joint Lebesgue p.d.f. of $X_1, ..., X_n$ is $\theta^{-n} I_{(0,\theta)}(x_{(n)})$. From Example 2.9, $X_{(n)}$ has the Lebesgue p.d.f. $(nx^{n-1}/\theta^n) I_{(0,\theta)}(x)$ on $\mathcal{R}$. Let f be a Borel function on $[0, \infty)$ such that $E[f(X_{(n)})] = 0$ for all $\theta > 0$. Then

$$\int_0^\theta f(x) x^{n-1} dx = 0 \quad \text{for all } \theta > 0.$$

Let $G(\theta)$ be the left-hand side of the previous equation. Applying the result of differentiation of an integral (see, e.g., Royden (1968, §5.3)), we obtain that $G'(\theta) = f(\theta)\theta^{n-1}$ a.e. m_+, where m_+ is the Lebesgue measure on $([0, \infty), \mathcal{B}_{[0,\infty)})$. Since $G(\theta) = 0$ for all $\theta > 0$, $f(\theta)\theta^{n-1} = 0$ a.e. m_+ and, hence, $f(x) = 0$ a.e. m_+. Therefore, $X_{(n)}$ is complete and sufficient for $\theta \in (0, \infty)$.

Example 2.17. In Example 2.12, we showed that the order statistics $T(X) = (X_{(1)}, ..., X_{(n)})$ of i.i.d. random variables $X_1, ..., X_n$ is sufficient for $P \in \mathcal{P}$, where $\mathcal{P}$ is the family of distributions on $\mathcal{R}$ having Lebesgue p.d.f.'s. We now show that $T(X)$ is also complete for $P \in \mathcal{P}$. Let $\mathcal{P}_0$ be the family of Lebesgue p.d.f.'s of the form

$$f(x) = C(\theta_1, ..., \theta_n) \exp\{-x^{2n} + \theta_1 x + \theta_2 x^2 + \cdots + \theta_n x^n\},$$

where $\theta_j \in \mathcal{R}$ and $C(\theta_1, ..., \theta_n)$ is a normalizing constant such that $\int f(x)dx = 1$. Then $\mathcal{P}_0 \subset \mathcal{P}$ and $\mathcal{P}_0$ is an exponential family of full rank. Note that the joint distribution of $X = (X_1, ..., X_n)$ is also in an exponential family of full rank. Thus, by Proposition 2.1, $U = (U_1, ..., U_n)$ is a complete statistic for $P \in \mathcal{P}_0$, where $U_j = \sum_{i=1}^n X_i^j$. Since a.s. $\mathcal{P}_0$ implies a.s. $\mathcal{P}$, $U(X)$ is also complete for $P \in \mathcal{P}$.

The result follows if we can show that there is a one-to-one correspondence between $T(X)$ and $U(X)$. Let $V_1 = \sum_{i=1}^n X_i$, $V_2 = \sum_{i<j} X_i X_j$, $V_3 = \sum_{i<j<k} X_i X_j X_k$,..., $V_n = X_1 \cdots X_n$. From the identities

$$U_k - V_1 U_{k-1} + V_2 U_{k-2} - \cdots + (-1)^{k-1} V_{k-1} U_1 + (-1)^k k V_k = 0,$$

$k = 1, ..., n$, there is a one-to-one correspondence between $U(X)$ and $V(X) = (V_1, ..., V_n)$. From the identity

$$(t - X_1) \cdots (t - X_n) = t^n - V_1 t^{n-1} + V_2 t^{n-2} - \cdots + (-1)^n V_n,$$

there is a one-to-one correspondence between $V(X)$ and $T(X)$. This completes the proof and, hence, $T(X)$ is sufficient and complete for $P \in \mathcal{P}$. In fact, both $U(X)$ and $V(X)$ are sufficient and complete for $P \in \mathcal{P}$.

The relationship between an ancillary statistic and a complete and sufficient statistic is characterized in the following result.

Theorem 2.4 (Basu's theorem). Let V and T be two statistics of X from a population $P \in \mathcal{P}$. If V is ancillary and T is boundedly complete and sufficient for $P \in \mathcal{P}$, then V and T are independent w.r.t. any $P \in \mathcal{P}$.

Proof. Let B be an event on the range of V. Since V is ancillary, $P(V^{-1}(B))$ is a constant. Since T is sufficient, $E[I_B(V)|T]$ is a function of T (independent of P). Since

$$E\{E[I_B(V)|T] - P(V^{-1}(B))\} = 0 \quad \text{for all } P \in \mathcal{P},$$

$P(V^{-1}(B)|T) = E[I_B(V)|T] = P(V^{-1}(B))$ a.s. $\mathcal{P}$, by the bounded completeness of T. Let A be an event on the range of T. Then,

$$P(T^{-1}(A) \cap V^{-1}(B)) = E\{E[I_A(T) I_B(V)|T]\} = E\{I_A(T) E[I_B(V)|T]\}$$

$$= E\{I_A(T) P(V^{-1}(B))\} = P(T^{-1}(A)) P(V^{-1}(B)).$$

Hence T and V are independent w.r.t. any $P \in \mathcal{P}$.

Basu's theorem is useful in proving the independence of two statistics.

Example 2.18. Suppose that $X_1, ..., X_n$ are i.i.d. random variables having the $N(\mu, \sigma^2)$ distribution, with $\mu \in \mathcal{R}$ and a known $\sigma > 0$. It can be easily shown that the family $\{N(\mu, \sigma^2) : \mu \in \mathcal{R}\}$ is an exponential family of full rank with natural parameter $\eta = \mu/\sigma^2$. By Proposition 2.1, the sample mean $\bar{X}$ is complete and sufficient for η (and μ). Let S^2 be the sample variance. Since $S^2 = (n-1)^{-1} \sum_{i=1}^n (Z_i - \bar{Z})^2$, where $Z_i = X_i - \mu$ is $N(0, \sigma^2)$ and $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$, S^2 is an ancillary statistic (σ^2 is known). By Basu's theorem, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ with $\mu \in \mathcal{R}$. Since σ^2 is arbitrary, $\bar{X}$ and S^2 are independent w.r.t. $N(\mu, \sigma^2)$ for any $\mu \in \mathcal{R}$ and $\sigma^2 > 0$.

Using the independence of $\bar{X}$ and S^2, we now show that $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}. Note that

$$n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 + \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2.$$

From the properties of the normal distributions, $n(\bar{X} - \mu)^2/\sigma^2$ has the chi-square distribution χ^2_1 with the m.g.f. $(1 - 2t)^{-1/2}$ and $\sum_{i=1}^n (X_i - \mu)^2/\sigma^2$ has the chi-square distribution χ^2_n with

the m.g.f. $(1-2t)^{-n/2}$, $t < 1/2$. By the independence of $\bar{X}$ and S^2, the m.g.f. of $(n-1)S^2/\sigma^2$ is

$$(1-2t)^{-n/2}/(1-2t)^{-1/2} = (1-2t)^{-(n-1)/2}$$

for $t < 1/2$. This is the m.g.f. of the chi-square distribution χ^2_{n-1} and, therefore, the result follows.

Lecture 22: Decision rules, loss, and risk

Statistical decision theory

X: a sample from a population $P \in \mathcal{P}$
Decision: an action we take after observing X
$\mathcal{A}$: the set of allowable actions
$(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$: the action space
$\mathcal{X}$: the range of X
Decision rule: a measurable function (a statistic) T from $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ to $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$
If X is observed, then we take the action $T(X) \in \mathcal{A}$

Performance criterion: loss function $L(P, a)$ from $\mathcal{P} \times \mathcal{A}$ to $[0, \infty)$ and is Borel for each P
If $X = x$ is observed and our decision rule is T, then our "loss" is $L(P, T(x))$
It is difficult to compare $L(P, T_1(X))$ and $L(P, T_2(X))$ for two decision rules, T_1 and T_2, since both of them are random.

Risk: Average (expected) loss defined as

$$R_T(P) = E[L(P, T(X))] = \int_{\mathcal{X}} L(P, T(x)) dP_X(x).$$

If $\mathcal{P}$ is a parametric family indexed by θ, the loss and risk are denoted by $L(\theta, a)$ and $R_T(\theta)$

For decision rules T_1 and T_2, T_1 is *as good as* T_2 if and only if

$$R_{T_1}(P) \leq R_{T_2}(P) \quad \text{for any } P \in \mathcal{P},$$

and is *better* than T_2 if, in addition, $R_{T_1}(P) < R_{T_2}(P)$ for at least one $P \in \mathcal{P}$.

Two decision rules T_1 and T_2 are *equivalent* if and only if $R_{T_1}(P) = R_{T_2}(P)$ for all $P \in \mathcal{P}$.

Optimal rule: If T_* is as good as any other rule in $\Im$, a class of allowable decision rules, then T_* is $\Im$-*optimal* (or optimal if $\Im$ contains all possible rules).

Sometimes it is useful to consider *randomized decision rules*.
Randomized decision rule: a function δ on $\mathcal{X} \times \mathcal{F}_{\mathcal{A}}$ such that, for every $A \in \mathcal{F}_{\mathcal{A}}$, $\delta(\cdot, A)$ is a Borel function and, for every $x \in \mathcal{X}$, $\delta(x, \cdot)$ is a probability measure on $(\mathcal{A}, \mathcal{F}_{\mathcal{A}})$.
If $X = x$ is observed, our have a distribution of actions: $\delta(x, \cdot)$.
A nonrandomized decision rule T previously discussed can be viewed as a special randomized decision rule with $\delta(x, \{a\}) = I_{\{a\}}(T(x))$, $a \in \mathcal{A}$, $x \in \mathcal{X}$.

To choose an action in $\mathcal{A}$ when a randomized rule δ is used, we need to simulate a pseudo-random element of $\mathcal{A}$ according to $\delta(x, \cdot)$.
Thus, an alternative way to describe a randomized rule is to specify the method of simulating the action from $\mathcal{A}$ for each $x \in \mathcal{X}$.
For example, a randomized rule can be a discrete distribution $\delta(x, \cdot)$ assigning probability $p_j(x)$ to a nonrandomized decision rule $T_j(x)$, $j = 1, 2, ...$, in which case the rule δ can be

equivalently defined as a rule taking value $T_j(x)$ with probability $p_j(x)$, i.e.,

$$T(X) = \begin{cases} T_1(X) & \text{with probability } p_1(X) \\ \cdots & \cdots \\ T_k(X) & \text{with probability } p_k(X) \end{cases}$$

The loss function for a randomized rule δ is defined as

$$L(P, \delta, x) = \int_{\mathcal{A}} L(P, a) d\delta(x, a),$$

which reduces to the same loss function we discussed when δ is a nonrandomized rule.
The risk of a randomized rule δ is then

$$R_\delta(P) = E[L(P, \delta, X)] = \int_{\mathcal{X}} \int_{\mathcal{A}} L(P, a) d\delta(x, a) dP_X(x).$$

For $T(X)$ defined above,

$$L(P, T, x) = \sum_{j=1}^{k} L(P, T_j(x)) p_j(x)$$

and

$$R_T(P) = \sum_{j=1}^{k} E[L(P, T_j(X)) p_j(X)]$$

Example 2.19. Let $X = (X_1, ..., X_n)$ be a vector of iid measurements for a parameter $\theta \in \mathcal{R}$.
Action space: $(\mathcal{A}, \mathcal{F}_\mathcal{A}) = (\mathcal{R}, \mathcal{B})$.
A common loss function in this problem is the *squared error loss* $L(P, a) = (\theta - a)^2$, $a \in \mathcal{A}$.
Let $T(X) = \bar{X}$, the sample mean.
The loss for $\bar{X}$ is $(\bar{X} - \theta)^2$.
If the population has mean μ and variance $\sigma^2 < \infty$, then

$$\begin{aligned} R_{\bar{X}}(P) &= E(\theta - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + E(E\bar{X} - \bar{X})^2 \\ &= (\theta - E\bar{X})^2 + \mathrm{Var}(\bar{\mathrm{X}}) \\ &= (\mu - \theta)^2 + \tfrac{\sigma^2}{n}. \end{aligned}$$

If θ is in fact the mean of the population, then

$$R_{\bar{X}}(P) = \tfrac{\sigma^2}{n},$$

is an increasing function of the population variance σ^2 and a decreasing function of the sample size n.
Consider another decision rule $T_1(X) = (X_{(1)} + X_{(n)})/2$.

$R_{T_1}(P)$ does not have a simple explicit form if there is no further assumption on the population P.
Suppose that $P \in \mathcal{P}$. Then, for some $\mathcal{P}$, $\bar{X}$ (or T_1) is better than T_1 (or $\bar{X}$) (exercise), whereas for some $\mathcal{P}$, neither $\bar{X}$ nor T_1 is better than the other.
Consider a randomized rule:

$$T_2(X) = \begin{cases} \bar{X} & \text{with probability } p(X) \\ T_1(X) & \text{with probability } 1 - p(X) \end{cases}$$

The loss for $T_2(X)$ is

$$(\bar{X} - \theta)^2 p(X) + [T_1(X) - \theta]^2 [1 - p(X)]$$

and the risk of T_2 is

$$R_{T_2}(P) = E\{(\bar{X} - \theta)^2 p(X) + [T_1(X) - \theta]^2 [1 - p(X)]\}$$

In particular, if $p(X) = 0.5$, then

$$R_{T_2}(P) = \frac{R_{\bar{X}}(P) + R_{T_1}(P)}{2}.$$

The problem in Example 2.19 is a special case of a general problem called *estimation*.
In an estimation problem, a decision rule T is called an *estimator*.
The following example describes another type of important problem called *hypothesis testing*.

Example 2.20. Let $\mathcal{P}$ be a family of distributions, $\mathcal{P}_0 \subset \mathcal{P}$, and $\mathcal{P}_1 = \{P \in \mathcal{P} : P \notin \mathcal{P}_0\}$. A hypothesis testing problem can be formulated as that of deciding which of the following two statements is true:

$$H_0 : \ P \in \mathcal{P}_0 \qquad \text{versus} \qquad H_1 : \ P \in \mathcal{P}_1. \tag{1}$$

Here, H_0 is called the *null hypothesis* and H_1 is called the *alternative hypothesis*.
The action space for this problem contains only two elements, i.e., $\mathcal{A} = \{0, 1\}$, where 0 is the action of accepting H_0 and 1 is the action of rejecting H_0.
A decision rule is called a *test*.
Since a test $T(X)$ is a function from $\mathcal{X}$ to $\{0, 1\}$, $T(X)$ must have the form $I_C(X)$, where $C \in \mathcal{F}_{\mathcal{X}}$ is called the *rejection region* or *critical region* for testing H_0 versus H_1.
0-1 loss: $L(P, a) = 0$ if a correct decision is made and 1 if an incorrect decision is made, i.e., $L(P, j) = 0$ for $P \in \mathcal{P}_j$ and $L(P, j) = 1$ otherwise, $j = 0, 1$.
Under this loss, the risk is

$$R_T(P) = \begin{cases} P(T(X) = 1) = P(X \in C) & P \in \mathcal{P}_0 \\ P(T(X) = 0) = P(X \notin C) & P \in \mathcal{P}_1. \end{cases}$$

See Figure 2.2 on page 127 for an example of a graph of $R_T(\theta)$ for some T and P in a parametric family.

The 0-1 loss implies that the loss for two types of incorrect decisions (accepting H_0 when $P \in \mathcal{P}_1$ and rejecting H_0 when $P \in \mathcal{P}_0$) are the same.
In some cases, one might assume unequal losses: $L(P, j) = 0$ for $P \in \mathcal{P}_j$, $L(P, 0) = c_0$ when $P \in \mathcal{P}_1$, and $L(P, 1) = c_1$ when $P \in \mathcal{P}_0$.

Admissibility

Definition 2.7. Let $\Im$ be a class of decision rules (randomized or nonrandomized). A decision rule $T \in \Im$ is called $\Im$-*admissible* (or admissible when $\Im$ contains all possible rules) if and only if there does not exist any $S \in \Im$ that is better than T (in terms of the risk).

If a decision rule T is inadmissible, then there exists a rule better than T.
Thus, T should not be used in principle.
However, an admissible decision rule is not necessarily good.
For example, in an estimation problem a silly estimator $T(X) \equiv$ a constant may be admissible.

If T_* is $\Im$-optimal, then it is $\Im$-admissible.
If T_* is $\Im$-optimal and T_0 is $\Im$-admissible, then T_0 is also $\Im$-optimal and is equivalent to T_*.
If there are two $\Im$-admissible rules that are not equivalent, then there does not exist any $\Im$-optimal rule.

Lecture 23: Sufficiency and Rao-Blackwell theorem, unbiasedness and invariance

Suppose that we have a sufficient statistic $T(X)$ for $P \in \mathcal{P}$.
Intuitively, our decision rule should be a function of T.
This is not true in general, but the following result indicates that this is true if randomized decision rules are allowed.

Proposition 2.2. Suppose that $\mathcal{A}$ is a subset of $\mathcal{R}^k$. Let $T(X)$ be a sufficient statistic for $P \in \mathcal{P}$ and let δ_0 be a decision rule. Then

$$\delta_1(t, A) = E[\delta_0(X, A)|T = t],$$

which is a randomized decision rule depending only on T, is equivalent to δ_0 if $R_{\delta_0}(P) < \infty$ for any $P \in \mathcal{P}$.
Proof. Note that δ_1 is a decision rule since δ_1 does not depend on the unknown P by the sufficiency of T. Then

$$\begin{aligned} R_{\delta_1}(P) &= E\left\{\int_{\mathcal{A}} L(P, a)d\delta_1(X, a)\right\} \\ &= E\left\{E\left[\int_{\mathcal{A}} L(P, a)d\delta_0(X, a)\middle|T\right]\right\} \\ &= E\left\{\int_{\mathcal{A}} L(P, a)d\delta_0(X, a)\right\} \\ &= R_{\delta_0}(P), \end{aligned}$$

where the proof of the second equality is left to the reader.

Note that Proposition 2.2 does not imply that δ_0 is inadmissible.
If δ_0 is a nonrandomized rule,

$$\delta_1(t, A) = E[I_A(\delta_0(X))|T = t] = P(\delta_0(X) \in A|T = t)$$

is still a randomized rule, unless $\delta_0(X) = h(T(X))$ a.s. P for some Borel function h (Exercise 75).
Hence, Proposition 2.2 does not apply to situations where randomized rules are not allowed.

The following result tells us when nonrandomized rules are all we need and when decision rules that are not functions of sufficient statistics are inadmissible.

Theorem 2.5. Suppose that $\mathcal{A}$ is a convex subset of $\mathcal{R}^k$ and that for any $P \in \mathcal{P}$, $L(P, a)$ is a convex function of a.
(i) Let δ be a randomized rule satisfying $\int_{\mathcal{A}} \|a\|d\delta(x, a) < \infty$ for any $x \in \mathcal{X}$ and let $T_1(x) = \int_{\mathcal{A}} ad\delta(x, a)$. Then $L(P, T_1(x)) \le L(P, \delta, x)$ (or $L(P, T_1(x)) < L(P, \delta, x)$ if L is strictly convex in a) for any $x \in \mathcal{X}$ and $P \in \mathcal{P}$.
(ii) (Rao-Blackwell theorem). Let T be a sufficient statistic for $P \in \mathcal{P}$, $T_0 \in \mathcal{R}^k$ be a nonrandomized rule satisfying $E\|T_0\| < \infty$, and $T_1 = E[T_0(X)|T]$. Then $R_{T_1}(P) \le R_{T_0}(P)$

for any $P \in \mathcal{P}$. If L is strictly convex in a and T_0 is not a function of T, then T_0 is inadmissible.

The proof of Theorem 2.5 is an application of Jensen's inequality and is left to the reader.

The concept of admissibility helps us to eliminate some decision rules.
However, usually there are still too many rules left after the elimination of some rules according to admissibility and sufficiency.
Although one is typically interested in a $\Im$-optimal rule, frequently it does not exist, if $\Im$ is either too large or too small.

Example 2.22. Let $X_1, ..., X_n$ be i.i.d. random variables from a population $P \in \mathcal{P}$ that is the family of populations having finite mean μ and variance σ^2.
Consider the estimation of μ ($\mathcal{A} = \mathcal{R}$) under the squared error loss.
It can be shown that if we let $\Im$ be the class of all possible estimators, then there is no $\Im$-optimal rule (exercise).
Next, let $\Im_1$ be the class of all linear functions in $X = (X_1, ..., X_n)$, i.e., $T(X) = \sum_{i=1}^n c_i X_i$ with known $c_i \in \mathcal{R}$, $i = 1, ..., n$.
Then

$$R_T(P) = \mu^2 \left(\sum_{i=1}^n c_i - 1 \right)^2 + \sigma^2 \sum_{i=1}^n c_i^2. \tag{1}$$

We now show that there does not exist $T_* = \sum_{i=1}^n c_i^* X_i$ such that $R_{T_*}(P) \leq R_T(P)$ for any $P \in \mathcal{P}$ and $T \in \Im_1$.
If there is such a T_*, then $(c_1^*, ..., c_n^*)$ is a minimum of the function of $(c_1, ..., c_n)$ on the right-hand side of (1).
Then $c_1^*, ..., c_n^*$ must be the same and equal to $\mu^2/(\sigma^2 + n\mu^2)$, which depends on P.
Hence T_* is not a statistic.
This shows that there is no $\Im_1$-optimal rule.
Consider now a subclass $\Im_2 \subset \Im_1$ with c_i's satisfying $\sum_{i=1}^n c_i = 1$.
From (1), $R_T(P) = \sigma^2 \sum_{i=1}^n c_i^2$ if $T \in \Im_2$.
Minimizing $\sigma^2 \sum_{i=1}^n c_i^2$ subject to $\sum_{i=1}^n c_i = 1$ leads to an optimal solution of $c_i = n^{-1}$.
Thus, the sample mean $\bar{X}$ is $\Im_2$-optimal.
There may not be any optimal rule if we consider a small class of decision rules.
For example, if $\Im_3$ contains all the rules in $\Im_2$ except $\bar{X}$, then one can show that there is no $\Im_3$-optimal rule.

Example 2.23. Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$.
Consider the hypothesis testing problem described in Example 2.20 with $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value.
Suppose that we are only interested in the following class of nonrandomized decision rules: $\Im = \{T_j : j = 0, 1, ..., n-1\}$, where $T_j(X) = I_{\{j+1,...,n\}}(X)$.
From Example 2.20, the risk function for T_j under the 0-1 loss is

$$R_{T_j}(\theta) = P(X > j) I_{(0,\theta_0]}(\theta) + P(X \leq j) I_{(\theta_0,1)}(\theta).$$

For any integers k and j, $0 \le k < j \le n-1$,

$$R_{T_j}(\theta) - R_{T_k}(\theta) = \begin{cases} -P(k < X \le j) < 0 & 0 < \theta \le \theta_0 \\ P(k < X \le j) > 0 & \theta_0 < \theta < 1. \end{cases}$$

Hence, neither T_j nor T_k is better than the other.
This shows that every T_j is $\Im$-admissible and, thus, there is no $\Im$-optimal rule.

In view of the fact that an optimal rule often does not exist, statisticians adopt the following two approaches to choose a decision rule.
The first approach is to define a class $\Im$ of decision rules that have some desirable properties (statistical and/or nonstatistical) and then try to find the best rule in $\Im$.
In Example 2.22, for instance, any estimator T in $\Im_2$ has the property that T is linear in X and $E[T(X)] = \mu$.
In a general estimation problem, we can use the following concept.

Definition 2.8 (Unbiasedness). In an estimation problem, the *bias* of an estimator $T(X)$ of a real-valued parameter ϑ of the unknown population is defined to be $b_T(P) = E[T(X)] - \vartheta$ (which is denoted by $b_T(\theta)$ when P is in a parametric family indexed by θ). An estimator $T(X)$ is said to be *unbiased* for ϑ if and only if $b_T(P) = 0$ for any $P \in \mathcal{P}$.

Thus, $\Im_2$ in Example 2.22 is the class of unbiased estimators linear in X.
In Chapter 3, we discuss how to find a $\Im$-optimal estimator when $\Im$ is the class of unbiased estimators or unbiased estimators linear in X.

Another class of decision rules can be defined after we introduce the concept of *invariance*.

Definition 2.9 Let X be a sample from $P \in \mathcal{P}$.
(i) A class $\mathcal{G}$ of one-to-one transformations of X is called a *group* if and only if $g_i \in \mathcal{G}$ implies $g_1 \circ g_2 \in \mathcal{G}$ and $g_i^{-1} \in \mathcal{G}$.
(ii) We say that $\mathcal{P}$ is *invariant* under $\mathcal{G}$ if and only if $\bar{g}(P_X) = P_{g(X)}$ is a one-to-one transformation from $\mathcal{P}$ onto $\mathcal{P}$ for each $g \in \mathcal{G}$.
(iii) A decision problem is said to be *invariant* if and only if $\mathcal{P}$ is invariant under $\mathcal{G}$ and the loss $L(P, a)$ is invariant in the sense that, for every $g \in \mathcal{G}$ and every $a \in \mathcal{A}$, there exists a unique $g(a) \in \mathcal{A}$ such that $L(P_X, a) = L\left(P_{g(X)}, g(a)\right)$. (Note that $g(X)$ and $g(a)$ are different functions in general.)
(iv) A decision rule $T(x)$ is said to be *invariant* if and only if, for every $g \in \mathcal{G}$ and every $x \in \mathcal{X}$, $T(g(x)) = g(T(x))$.

Invariance means that our decision is not affected by one-to-one transformations of data.
In a problem where the distribution of X is in a location-scale family $\mathcal{P}$ on $\mathcal{R}^k$, we often consider location-scale transformations of data X of the form $g(X) = AX + c$, where $c \in \mathcal{C} \subset \mathcal{R}^k$ and $A \in \mathcal{T}$, a class of invertible $k \times k$ matrices.
In §4.2 and §6.3, we discuss the problem of finding a $\Im$-optimal rule when $\Im$ is a class of invariant decision rules.

Lecture 24: Bayes rules, minimax rules, point estimators, and hypothesis tests

The second approach to finding a good decision rule is to consider some characteristic R_T of $R_T(P)$, for a given decision rule T, and then minimize R_T over $T \in \Im$.
The following are two popular ways to carry out this idea.
The first one is to consider an average of $R_T(P)$ over $P \in \mathcal{P}$:

$$r_{_T}(\Pi) = \int_{\mathcal{P}} R_T(P) d\Pi(P),$$

where Π is a known probability measure on $(\mathcal{P}, \mathcal{F}_{\mathcal{P}})$ with an appropriate σ-field $\mathcal{F}_{\mathcal{P}}$.
$r_{_T}(\Pi)$ is called the *Bayes risk* of T w.r.t. Π.
If $T_* \in \Im$ and $r_{_{T_*}}(\Pi) \le r_{_T}(\Pi)$ for any $T \in \Im$, then T_* is called a $\Im$-*Bayes rule* (or Bayes rule when $\Im$ contains all possible rules) w.r.t. Π.
The second method is to consider the worst situation, i.e., $\sup_{P\in\mathcal{P}} R_T(P)$.
If $T_* \in \Im$ and

$$\sup_{P\in\mathcal{P}} R_{T_*}(P) \le \sup_{P\in\mathcal{P}} R_T(P)$$

for any $T \in \Im$, then T_* is called a $\Im$-*minimax* rule (or minimax rule when $\Im$ contains all possible rules).
Bayes and minimax rules are discussed in Chapter 4.

Example 2.25. We usually try to find a Bayes rule or a minimax rule in a parametric problem where $P = P_\theta$ for a $\theta \in \mathcal{R}^k$.
Consider the special case of $k = 1$ and $L(\theta, a) = (\theta - a)^2$, the squared error loss.
Note that

$$r_{_T}(\Pi) = \int_{\mathcal{R}} E[\theta - T(X)]^2 d\Pi(\theta),$$

which is equivalent to $E[\boldsymbol{\theta} - T(X)]^2$, where $\boldsymbol{\theta}$ is a random variable having the distribution Π and, given $\boldsymbol{\theta} = \theta$, the conditional distribution of X is P_θ.
Then, the problem can be viewed as a prediction problem for $\boldsymbol{\theta}$ using functions of X.
Using the result in Example 1.22, the best predictor is $E(\boldsymbol{\theta}|X)$, which is the $\Im$-Bayes rule w.r.t. Π with $\Im$ being the class of rules $T(X)$ satisfying $E[T(X)]^2 < \infty$ for any θ.

As a more specific example, let $X = (X_1, ..., X_n)$ with i.i.d. components having the $N(\mu, \sigma^2)$ distribution with an unknown $\mu = \theta \in \mathcal{R}$ and a known σ^2, and let Π be the $N(\mu_0, \sigma_0^2)$ distribution with known μ_0 and σ_0^2.
Then the conditional distribution of $\boldsymbol{\theta}$ given $X = x$ is $N(\mu_*(x), c^2)$ with

$$\mu_*(x) = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x} \qquad \text{and} \qquad c^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} \tag{1}$$

The Bayes rule w.r.t. Π is $E(\boldsymbol{\theta}|X) = \mu_*(X)$.

In this special case we can show that the sample mean $\bar{X}$ is minimax.
For any decision rule T,

$$\begin{aligned}\sup_{\theta\in\mathcal{R}} R_T(\theta) &\geq \int_{\mathcal{R}} R_T(\theta)d\Pi(\theta)\\ &\geq \int_{\mathcal{R}} R_{\mu_*}(\theta)d\Pi(\theta)\\ &= E\left\{[\boldsymbol{\theta}-\mu_*(X)]^2\right\}\\ &= E\left\{E\{[\boldsymbol{\theta}-\mu_*(X)]^2|X\}\right\}\\ &= E(c^2)\\ &= c^2,\end{aligned}$$

where $\mu_*(X)$ is the Bayes rule given in (1) and c^2 is also given in (1).
Since this result is true for any $\sigma_0^2 > 0$ and $c^2 \to \sigma^2/n$ as $\sigma_0^2 \to \infty$,

$$\sup_{\theta\in\mathcal{R}} R_T(\theta) \geq \frac{\sigma^2}{n} = \sup_{\theta\in\mathcal{R}} R_{\bar{X}}(\theta),$$

where the equality holds because the risk of $\bar{X}$ under the squared error loss is σ^2/n and independent of $\theta = \mu$.
Thus, $\bar{X}$ is minimax.

A minimax rule in a general case may be difficult to obtain. It can be seen that if both μ and σ^2 are unknown in the previous discussion, then

$$\sup_{\theta\in\mathcal{R}\times(0,\infty)} R_{\bar{X}}(\theta) = \infty, \tag{2}$$

where $\theta = (\mu, \sigma^2)$.
Hence $\bar{X}$ cannot be minimax unless (2) holds with $\bar{X}$ replaced by any decision rule T, in which case minimaxity becomes meaningless.

Statistical inference: Point estimators, hypothesis tests, and confidence sets

Point estimators

Let $T(X)$ be an estimator of $\vartheta \in \mathcal{R}$
Bias: $b_T(P) = E[T(X)] - \vartheta$
Mean squared error (mse):

$$\text{mse}_T(P) = E[T(X) - \vartheta]^2 = [b_T(P)]^2 + \text{Var}(T(X)).$$

Bias and mse are two common criteria for the performance of point estimators.

Example 2.26. Let $X_1, ..., X_n$ be i.i.d. from an unknown c.d.f. F.
Suppose that the parameter of interest is $\vartheta = 1 - F(t)$ for a fixed $t > 0$.
If F is not in a parametric family, then a *nonparametric* estimator of $F(t)$ is the *empirical* c.d.f.

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(X_i), \qquad t \in \mathcal{R}.$$

Since $I_{(-\infty,t]}(X_1), ..., I_{(-\infty,t]}(X_n)$ are i.i.d. binary random variables with $P(I_{(-\infty,t]}(X_i) = 1) = F(t)$, the random variable $nF_n(t)$ has the binomial distribution $Bi(F(t), n)$.
Consequently, $F_n(t)$ is an unbiased estimator of $F(t)$ and $\mathrm{Var}(F_n(t)) = \mathrm{mse}_{F_n(t)}(P) = F(t)[1 - F(t)]/n$.
Since any linear combination of unbiased estimators is unbiased for the same linear combination of the parameters (by the linearity of expectations), an unbiased estimator of ϑ is $U(X) = 1 - F_n(t)$, which has the same variance and mse as $F_n(t)$.
The estimator $U(X) = 1 - F_n(t)$ can be improved in terms of the mse if there is further information about F.
Suppose that F is the c.d.f. of the exponential distribution $E(0, \theta)$ with an unknown $\theta > 0$. Then $\vartheta = e^{-t/\theta}$.
The sample mean $\bar{X}$ is sufficient for $\theta > 0$.
Since the squared error loss is strictly convex, an application of Theorem 2.5(ii) (Rao-Blackwell theorem) shows that the estimator $T(X) = E[1-F_n(t)|\bar{X}]$, which is also unbiased, is better than $U(X)$ in terms of the mse.
Figure 2.1 shows graphs of the mse's of $U(X)$ and $T(X)$, as functions of θ, in the special case of $n = 10$, $t = 2$, and $F(x) = (1 - e^{-x/\theta})I_{(0,\infty)}(x)$.

Hypothesis tests

To test the hypotheses

$$H_0 : P \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : P \in \mathcal{P}_1,$$

there are two types of statistical errors we may commit: rejecting H_0 when H_0 is true (called the *type I error*) and accepting H_0 when H_0 is wrong (called the *type II error*).
A test T: a statistic from $\mathcal{X}$ to $\{0, 1\}$. Pprobabilities of making two types of errors:

$$\alpha_T(P) = P(T(X) = 1) \qquad P \in \mathcal{P}_0 \tag{3}$$

and

$$1 - \alpha_T(P) = P(T(X) = 0) \qquad P \in \mathcal{P}_1, \tag{4}$$

which are denoted by $\alpha_T(\theta)$ and $1 - \alpha_T(\theta)$ if P is in a parametric family indexed by θ.
Note that these are risks of T under the 0-1 loss in statistical decision theory.
Error probabilities in (3) and (4) cannot be minimized simultaneously.
Furthermore, these two error probabilities cannot be bounded simultaneously by a fixed $\alpha \in (0, 1)$ when we have a sample of a fixed size.

A common approach to finding an "optimal" test is to assign a small bound α to one of the error probabilities, say $\alpha_T(P)$, $P \in \mathcal{P}_0$, and then to attempt to minimize the other error probability $1 - \alpha_T(P)$, $P \in \mathcal{P}_1$, subject to

$$\sup_{P \in \mathcal{P}_0} \alpha_T(P) \le \alpha. \tag{5}$$

The bound α is called the *level of significance*.
The left-hand side of (5) is called the *size* of the test T.

The level of significance should be positive, otherwise no test satisfies (5) except the silly test $T(X) \equiv 0$ a.s. $\mathcal{P}$.

Example 2.28. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.
Consider the hypotheses $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$, where μ_0 is a fixed constant.
Since the sample mean $\bar{X}$ is sufficient for $\mu \in \mathcal{R}$, it is reasonable to consider the following class of tests: $T_c(X) = I_{(c,\infty)}(\bar{X})$, i.e., H_0 is rejected (accepted) if $\bar{X} > c$ ($\bar{X} \leq c$), where $c \in \mathcal{R}$ is a fixed constant.
Let Φ be the c.d.f. of $N(0, 1)$. Then, by the property of the normal distributions,

$$\alpha_{T_c}(\mu) = P(T_c(X) = 1) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).$$

Figure 2.2 provides an example of a graph of two types of error probabilities, with $\mu_0 = 0$.
Since $\Phi(t)$ is an increasing function of t,

$$\sup_{P \in \mathcal{P}_0} \alpha_{T_c}(\mu) = 1 - \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

In fact, it is also true that

$$\sup_{P \in \mathcal{P}_1} [1 - \alpha_{T_c}(\mu)] = \Phi\left(\frac{\sqrt{n}(c - \mu_0)}{\sigma}\right).$$

If we would like to use an α as the level of significance, then the most effective way is to choose a c_α (a test $T_{c_\alpha}(X)$) such that

$$\alpha = \sup_{P \in \mathcal{P}_0} \alpha_{T_{c_\alpha}}(\mu),$$

in which case c_α must satisfy

$$1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) = \alpha,$$

i.e., $c_\alpha = \sigma z_{1-\alpha}/\sqrt{n} + \mu_0$, where $z_a = \Phi^{-1}(a)$.
In Chapter 6, it is shown that for any test $T(X)$ satisfying (5),

$$1 - \alpha_T(\mu) \geq 1 - \alpha_{T_{c_\alpha}}(\mu), \qquad \mu > \mu_0.$$

Lecture 25: p-value, randomized tests, and confidence sets

The choice of a level of significance α is usually somewhat subjective.
In most applications there is no precise limit to the size of T that can be tolerated.
Standard values, such as 0.10, 0.05, or 0.01, are often used for convenience.

For most tests satisfying

$$\sup_{P\in\mathcal{P}_0} \alpha_T(P) \le \alpha. \tag{1}$$

a small α leads to a "small" rejection region.
It is good practice to determine not only whether H_0 is rejected or accepted for a given α and a chosen test T_α, but also the smallest possible level of significance at which H_0 would be rejected for the computed $T_\alpha(x)$, i.e.,

$$\hat{\alpha} = \inf\{\alpha \in (0,1) : T_\alpha(x) = 1\}.$$

Such an $\hat{\alpha}$, which depends on x and the chosen test and is a statistic, is called the *p-value* for the test T_α.

Example 2.29. Consider the problem in Example 2.28. Let us calculate the p-value for T_{c_α}. Note that

$$\alpha = 1 - \Phi\left(\frac{\sqrt{n}(c_\alpha - \mu_0)}{\sigma}\right) > 1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right)$$

if and only if $\bar{x} > c_\alpha$ (or $T_{c_\alpha}(x) = 1$). Hence

$$1 - \Phi\left(\frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}\right) = \inf\{\alpha \in (0,1) : T_{c_\alpha}(x) = 1\} = \hat{\alpha}(x)$$

is the p-value for T_{c_α}. It turns out that $T_{c_\alpha}(x) = I_{(0,\alpha)}(\hat{\alpha}(x))$.

With the additional information provided by p-values, using p-values is typically more appropriate than using fixed-level tests in a scientific problem.
However, a fixed level of significance is unavoidable when acceptance or rejection of H_0 implies an imminent concrete decision.

In Example 2.28, the equality in (1) can always be achieved by a suitable choice of c.
This is, however, not true in general.
We need to consider *randomized tests*.
Recall that a randomized decision rule is a probability measure $\delta(x, \cdot)$ on the action space for any fixed x.
Since the action space contains only two points, 0 and 1, for a hypothesis testing problem, any randomized test $\delta(X, A)$ is equivalent to a statistic $T(X) \in [0,1]$ with $T(x) = \delta(x, \{1\})$ and $1 - T(x) = \delta(x, \{0\})$.
A nonrandomized test is obviously a special case where $T(x)$ does not take any value in $(0,1)$.

For any randomized test $T(X)$, we define the type I error probability to be $\alpha_T(P) = E[T(X)]$, $P \in \mathcal{P}_0$, and the type II error probability to be $1 - \alpha_T(P) = E[1 - T(X)]$, $P \in \mathcal{P}_1$.
For a class of randomized tests, we would like to minimize $1 - \alpha_T(P)$ subject to (1).

Example 2.30. Assume that the sample X has the binomial distribution $Bi(\theta, n)$ with an unknown $\theta \in (0, 1)$ and a fixed integer $n > 1$.
Consider the hypotheses $H_0 : \theta \in (0, \theta_0]$ versus $H_1 : \theta \in (\theta_0, 1)$, where $\theta_0 \in (0, 1)$ is a fixed value.
Consider the following class of randomized tests:

$$T_{j,q}(X) = \begin{cases} 1 & X > j \\ q & X = j \\ 0 & X < j, \end{cases}$$

where $j = 0, 1, ..., n - 1$ and $q \in [0, 1]$. Then

$$\alpha_{T_{j,q}}(\theta) = P(X > j) + qP(X = j) \qquad 0 < \theta \leq \theta_0$$

and

$$1 - \alpha_{T_{j,q}}(\theta) = P(X < j) + (1 - q)P(X = j) \qquad \theta_0 < \theta < 1.$$

It can be shown that for any $\alpha \in (0, 1)$, there exist an integer j and $q \in (0, 1)$ such that the size of $T_{j,q}$ is α.

Confidence sets

ϑ: a k-vector of unknown parameters related to the unknown population $P \in \mathcal{P}$
$C(X)$ a Borel set (in the range of ϑ) depending only on the sample X
If

$$\inf_{P \in \mathcal{P}} P(\vartheta \in C(X)) \geq 1 - \alpha, \tag{2}$$

where α is a fixed constant in $(0, 1)$, then $C(X)$ is called a *confidence set* for ϑ with *level of significance* $1 - \alpha$.
The left-hand side of (2) is called the *confidence coefficient* of $C(X)$, which is the highest possible level of significance for $C(X)$.
A confidence set is a random element that covers the unknown ϑ with certain probability.
If (2) holds, then the *coverage probability* of $C(X)$ is at least $1 - \alpha$, although $C(x)$ either covers or does not cover ϑ whence we observe $X = x$.
The concepts of level of significance and confidence coefficient are very similar to the level of significance and size in hypothesis testing.
In fact, it is shown in Chapter 7 that some confidence sets are closely related to hypothesis tests.

Consider a real-valued ϑ.
If $C(X) = [\underline{\vartheta}(X), \overline{\vartheta}(X)]$ for a pair of real-valued statistics $\underline{\vartheta}$ and $\overline{\vartheta}$, then $C(X)$ is called a *confidence interval* for ϑ.

If $C(X) = (-\infty, \overline{\vartheta}(X)]$ (or $[\underline{\vartheta}(X), \infty)$), then $\overline{\vartheta}$ (or $\underline{\vartheta}$) is called an upper (or a lower) *confidence bound* for ϑ.

A confidence set (or interval) is also called a set (or an interval) estimator of ϑ, although it is very different from a point estimator (discussed in §2.4.1).

Example 2.31. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.
Suppose that a confidence interval for $\vartheta = \mu$ is needed.
We only need to consider $\underline{\vartheta}(\bar{X})$ and $\overline{\vartheta}(\bar{X})$, since the sample mean $\bar{X}$ is sufficient.
Consider confidence intervals of the form $[\bar{X} - c, \bar{X} + c]$, where $c \in (0, \infty)$ is fixed.
Note that

$$P\left(\mu \in [\bar{X} - c, \bar{X} + c]\right) = P\left(|\bar{X} - \mu| \le c\right) = 1 - 2\Phi\left(-\sqrt{n}c/\sigma\right),$$

which is independent of μ.
Hence, the confidence coefficient of $[\bar{X} - c, \bar{X} + c]$ is $1 - 2\Phi\left(-\sqrt{n}c/\sigma\right)$, which is an increasing function of c and converges to 1 as $c \to \infty$ or 0 as $c \to 0$.
Thus, confidence coefficients are positive but less than 1 except for silly confidence intervals $[\bar{X}, \bar{X}]$ and $(-\infty, \infty)$.
We can choose a confidence interval with an arbitrarily large confidence coefficient, but the chosen confidence interval may be so wide that it is practically useless.

If σ^2 is also unknown, then $[\bar{X} - c, \bar{X} + c]$ has confidence coefficient 0 and, therefore, is not a good inference procedure.
In such a case a different confidence interval for μ with positive confidence coefficient can be derived (Exercise 97 in §2.6).

This example tells us that a reasonable approach is to choose a level of significance $1 - \alpha \in (0, 1)$ (just like the level of significance in hypothesis testing) and a confidence interval or set satisfying (2).
In Example 2.31, when σ^2 is known and c is chosen to be $\sigma z_{1-\alpha/2}/\sqrt{n}$, where $z_a = \Phi^{-1}(a)$, the confidence coefficient of the confidence interval $[\bar{X} - c, \bar{X} + c]$ is *exactly* $1 - \alpha$ for any fixed $\alpha \in (0, 1)$.
This is desirable since, for all confidence intervals satisfying (2), the one with the shortest interval length is preferred.

For a general confidence interval $[\underline{\vartheta}(X), \overline{\vartheta}(X)]$, its length is $\overline{\vartheta}(X) - \underline{\vartheta}(X)$, which may be random.
We may consider the expected (or average) length $E[\overline{\vartheta}(X) - \underline{\vartheta}(X)]$.
The confidence coefficient and expected length are a pair of good measures of performance of confidence intervals.
Like the two types of error probabilities of a test in hypothesis testing, however, we cannot maximize the confidence coefficient and minimize the length (or expected length) simultaneously.
A common approach is to minimize the length (or expected length) subject to (2).

For an unbounded confidence interval, its length is ∞.

Hence we have to define some other measures of performance.
For an upper (or a lower) confidence bound, we may consider the distance $\overline{\vartheta}(X) - \vartheta$ (or $\vartheta - \underline{\vartheta}(X)$) or its expectation.

Example 2.32. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with both $\mu \in \mathcal{R}$ and $\sigma^2 > 0$ unknown.
Let $\theta = (\mu, \sigma^2)$ and $\alpha \in (0, 1)$ be given.
Let $\bar{X}$ be the sample mean and S^2 be the sample variance.
Since $(\bar{X}, S^2)$ is sufficient (Example 2.15), we focus on $C(X)$ that is a function of $(\bar{X}, S^2)$.
From Example 2.18, $\bar{X}$ and S^2 are independent and $(n-1)S^2/\sigma^2$ has the chi-square distribution χ^2_{n-1}.
Since $\sqrt{n}(\bar{X} - \mu)/\sigma$ has the $N(0, 1)$ distribution,

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha\right) = \sqrt{1-\alpha},$$

where $\tilde{c}_\alpha = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ (verify).
Since the chi-square distribution χ^2_{n-1} is a known distribution, we can always find two constants $c_{1\alpha}$ and $c_{2\alpha}$ such that

$$P\left(c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = \sqrt{1-\alpha}.$$

Then

$$P\left(-\tilde{c}_\alpha \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \tilde{c}_\alpha, c_{1\alpha} \le \frac{(n-1)S^2}{\sigma^2} \le c_{2\alpha}\right) = 1 - \alpha,$$

or

$$P\left(\frac{n(\bar{X} - \mu)^2}{\tilde{c}^2_\alpha} \le \sigma^2, \frac{(n-1)S^2}{c_{2\alpha}} \le \sigma^2 \le \frac{(n-1)S^2}{c_{1\alpha}}\right) = 1 - \alpha. \quad (3)$$

The left-hand side of (3) defines a set in the range of $\theta = (\mu, \sigma^2)$ bounded by two straight lines, $\sigma^2 = (n-1)S^2/c_{i\alpha}$, $i = 1, 2$, and a curve $\sigma^2 = n(\bar{X} - \mu)^2/\tilde{c}^2_\alpha$ (see the shadowed part of Figure 2.3).
This set is a confidence set for θ with confidence coefficient $1 - \alpha$, since (3) holds for any θ.

Lecture 26: Asymptotic approach and consistency

Asymptotic approach

In decision theory and inference, a key to the success of finding a good decision rule or inference procedure is being able to find some moments and/or distributions of various statistics.
There are many cases in which we are not able to find exactly the moments or distributions of given statistics, especially when the problem is complext.
When the sample size n is large, we may approximate the moments and distributions of statistics that are impossible to derive, using the asymptotic tools discussed in §1.5.
In an asymptotic analysis, we consider a sample $X = (X_1, ..., X_n)$ not for fixed n, but as a member of a sequence corresponding to $n = n_0, n_0 + 1, ...$, and obtain the limit of the distribution of an appropriately normalized statistic or variable $T_n(X)$ as $n \to \infty$.
The limiting distribution and its moments are used as approximations to the distribution and moments of $T_n(X)$ in the situation with a large but actually finite n.
This leads to some asymptotic statistical procedures and asymptotic criteria for assessing their performances.

The asymptotic approach is not only applied to the situation where no exact method is available, but also used to provide an inference procedure simpler (e.g., in terms of computation) than that produced by the exact approach (the approach considering a fixed n).

In addition to providing more theoretical results and/or simpler inference procedures, the asymptotic approach requires less stringent mathematical assumptions than does the exact approach.
The mathematical precision of the optimality results obtained in statistical decision theory tends to obscure the fact that these results are approximations in view of the approximate nature of the assumed models and loss functions.
As the sample size increases, the statistical properties become less dependent on the loss functions and models.

A major weakness of the asymptotic approach is that typically no good estimates for the precision of the approximations are available and, therefore, we cannot determine whether a particular n in a problem is large enough to safely apply the asymptotic results.
To overcome this difficulty, asymptotic results are frequently used in combination with some numerical/empirical studies for selected values of n to examine the *finite sample* performance of asymptotic procedures.

Consistency

A reasonable point estimator is expected to perform better, at least on the average, if more information about the unknown population is available.
With a fixed model assumption and sampling plan, more data (larger sample size n) provide more information about the unknown population.
Thus, it is distasteful to use a point estimator T_n which, if sampling were to continue indef-

initely, could possibly have a nonzero estimation error, although the estimation error of T_n for a fixed n may never equal 0.

Definition 2.10 (Consistency of point estimators). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a point estimator of ϑ for every n.
(i) $T_n(X)$ is called *consistent* for ϑ if and only if $T_n(X) \to_p \vartheta$ w.r.t. any $P \in \mathcal{P}$.
(ii) Let $\{a_n\}$ be a sequence of positive constants diverging to ∞. $T_n(X)$ is called a_n-*consistent* for ϑ if and only if $a_n[T_n(X) - \vartheta] = O_p(1)$ w.r.t. any $P \in \mathcal{P}$.
(iii) $T_n(X)$ is called *strongly consistent* for ϑ if and only if $T_n(X) \to_{a.s.} \vartheta$ w.r.t. any $P \in \mathcal{P}$.
(iv) $T_n(X)$ is called L_r-*consistent* for ϑ if and only if $T_n(X) \to_{L_r} \vartheta$ w.r.t. any $P \in \mathcal{P}$ for some fixed $r > 0$.

Consistency is actually a concept relating to a sequence of estimators, $\{T_n, n = n_0, n_0+1, ...\}$, but we usually just say "consistency of T_n" for simplicity.
Each of the four types of consistency in Definition 2.10 describes the convergence of $T_n(X)$ to ϑ in some sense, as $n \to \infty$.
In statistics, consistency according to Definition 2.10(i), which is sometimes called *weak consistency* since it is implied by any of the other three types of consistency, is the most useful concept of convergence of T_n to ϑ.
L_2-consistency is also called *consistency in mse*, which is the most useful type of L_r-consistency.

Example 2.33. Let $X_1, ..., X_n$ be i.i.d. from $P \in \mathcal{P}$.
If $\vartheta = \mu$, which is the mean of P and is assumed to be finite, then by the SLLN (Theorem 1.13), the sample mean $\bar{X}$ is strongly consistent for μ and, therefore, is also consistent for μ.
If we further assume that the variance of P is finite, then $\bar{X}$ is consistent in mse and is $\sqrt{n}$-consistent.
With the finite variance assumption, the sample variance S^2 is strongly consistent for the variance of P, according to the SLLN.

Consider estimators of the form $T_n = \sum_{i=1}^n c_{ni} X_i$, where $\{c_{ni}\}$ is a double array of constants.
If P has a finite variance, then T_n is consistent in mse if and only if $\sum_{i=1}^n c_{ni} \to 1$ and $\sum_{i=1}^n c_{ni}^2 \to 0$.
If we only assume the existence of the mean of P, then T_n with $c_{ni} = c_i/n$ satisfying $n^{-1} \sum_{i=1}^n c_i \to 1$ and $\sup_i |c_i| < \infty$ is strongly consistent (Theorem 1.13(ii)).

One or a combination of the law of large numbers, the CLT, Slutsky's theorem (Theorem 1.11), and the continuous mapping theorem (Theorems 1.10 and 1.12) are typically applied to establish consistency of point estimators.
In particular, Theorem 1.10 implies that if T_n is (strongly) consistent for ϑ and g is a continuous function of ϑ, then $g(T_n)$ is (strongly) consistent for $g(\vartheta)$.
For example, in Example 2.33 the point estimator $\bar{X}^2$ is strongly consistent for μ^2.
To show that $\bar{X}^2$ is $\sqrt{n}$-consistent under the assumption that P has a finite variance σ^2, we can use the identity

$$\sqrt{n}(\bar{X}^2 - \mu^2) = \sqrt{n}(\bar{X} - \mu)(\bar{X} + \mu)$$

and the fact that $\bar{X}$ is $\sqrt{n}$-consistent for μ and $\bar{X} + \mu = O_p(1)$.
$\bar{X}^2$ may not be consistent in mse since we do not assume that P has a finite fourth moment.

Alternatively, we can use the fact that $\sqrt{n}(\bar{X}^2 - \mu^2) \to_d N(0, 4\mu^2\sigma^2)$ (by the CLT and Theorem 1.12) to show the $\sqrt{n}$-consistency of $\bar{X}^2$.

The following example shows another way to establish consistency of some point estimators.

Example 2.34. Let $X_1, ..., X_n$ be i.i.d. from an unknown P with a continuous c.d.f. F satisfying $F(\theta) = 1$ for some $\theta \in \mathcal{R}$ and $F(x) < 1$ for any $x < \theta$.
Consider the largest order statistic $X_{(n)}$.
For any $\epsilon > 0$, $F(\theta - \epsilon) < 1$ and

$$P(|X_{(n)} - \theta| \geq \epsilon) = P(X_{(n)} \leq \theta - \epsilon) = [F(\theta - \epsilon)]^n,$$

which imply (according to Theorem 1.8(v)) $X_{(n)} \to_{a.s.} \theta$, i.e., $X_{(n)}$ is strongly consistent for θ.
If we assume that $F^{(i)}(\theta-)$, the ith-order left-hand derivative of F at θ, exists and vanishes for any $i \leq m$ and that $F^{(m+1)}(\theta-)$ exists and is nonzero, where m is a nonnegative integer, then

$$1 - F(X_{(n)}) = \frac{(-1)^m F^{(m+1)}(\theta-)}{(m+1)!}(\theta - X_{(n)})^{m+1} + o\left(|\theta - X_{(n)}|^{m+1}\right) \quad \text{a.s.}$$

This result and the fact that $P\left(n[1 - F(X_{(n)})] \geq s\right) = (1 - s/n)^n$ imply that $(\theta - X_{(n)})^{m+1} = O_p(n^{-1})$, i.e., $X_{(n)}$ is $n^{(m+1)^{-1}}$-consistent.
If $m = 0$, then $X_{(n)}$ is n-consistent, which is the most common situation.
If $m = 1$, then $X_{(n)}$ is $\sqrt{n}$-consistent.
The limiting distribution of $n^{(m+1)^{-1}}(X_{(n)} - \theta)$ can be derived as follows.
Let

$$h_n(\theta) = \left[\frac{(-1)^m(m+1)!}{nF^{(m+1)}(\theta-)}\right]^{(m+1)^{-1}}.$$

For $t \leq 0$, by Slutsky's theorem,

$$\begin{aligned}
\lim_{n\to\infty} P\left(\frac{X_{(n)} - \theta}{h_n(\theta)} \leq t\right) &= \lim_{n\to\infty} P\left(\left[\frac{\theta - X_{(n)}}{h_n(\theta)}\right]^{m+1} \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} P\left(n[1 - F(X_{(n)})] \geq (-t)^{m+1}\right) \\
&= \lim_{n\to\infty} \left[1 - (-t)^{m+1}/n\right]^n \\
&= e^{-(-t)^{m+1}}.
\end{aligned}$$

It can be seen from the previous examples that there are many consistent estimators.
Like the admissibility in statistical decision theory, consistency is a very essential requirement in the sense that any inconsistent estimators should not be used, but a consistent estimator is not necessarily good.
Thus, consistency should be used together with one or a few more criteria.

We discuss a situation in which finding a consistent estimator is crucial.
Suppose that an estimator T_n of ϑ satisfies

$$c_n[T_n(X) - \vartheta] \to_d \sigma Y, \tag{1}$$

where Y is a random variable with a known distribution, $\sigma > 0$ is an unknown parameter, and $\{c_n\}$ is a sequence of constants. For example, in Example 2.33, $\sqrt{n}(\bar{X} - \mu) \to_d N(0, \sigma^2)$; in Example 2.34, (1) holds with $c_n = n^{(m+1)^{-1}}$ and $\sigma = [(-1)^m(m+1)!/F^{(m+1)}(\theta-)]^{(m+1)^{-1}}$. If a consistent estimator $\hat{\sigma}_n$ of σ can be found, then, by Slutsky's theorem,

$$c_n[T_n(X) - \vartheta]/\hat{\sigma}_n \to_d Y$$

and, thus, we may approximate the distribution of $c_n[T_n(X) - \vartheta]/\hat{\sigma}_n$ by the known distribution of Y.

Lecture 27: Asymptotic bias, variance, and mse

Asymptotic bias

Unbiasedness as a criterion for point estimators is discussed in §2.3.2.
In some cases, however, there is no unbiased estimator.
Furthermore, having a "slight" bias in some cases may not be a bad idea.
Let $T_n(X)$ be a point estimator of ϑ for every n.
If ET_n exists for every n and $\lim_{n\to\infty} E(T_n - \vartheta) = 0$ for any $P \in \mathcal{P}$, then T_n is said to be *approximately unbiased*.

There are many reasonable point estimators whose expectations are not well defined.
It is desirable to define a concept of *asymptotic bias* for point estimators whose expectations are not well defined.

Definition 2.11. (i) Let $\xi, \xi_1, \xi_2, ...$ be random variables and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. If $a_n\xi_n \to_d \xi$ and $E|\xi| < \infty$, then $E\xi/a_n$ is called an *asymptotic expectation* of ξ_n.
(ii) Let T_n be a point estimator of ϑ for every n. An asymptotic expectation of $T_n - \vartheta$, if it exists, is called an asymptotic bias of T_n and denoted by $\tilde{b}_{T_n}(P)$ (or $\tilde{b}_{T_n}(\theta)$ if P is in a parametric family). If $\lim_{n\to\infty} \tilde{b}_{T_n}(P) = 0$ for any $P \in \mathcal{P}$, then T_n is said to be *asymptotically unbiased*.

Like the consistency, the asymptotic expectation (or bias) is a concept relating to sequences $\{\xi_n\}$ and $\{E\xi/a_n\}$ (or $\{T_n\}$ and $\{\tilde{b}_{T_n}(P)\}$).
The exact bias $b_{T_n}(P)$ is not necessarily the same as $\tilde{b}_{T_n}(P)$ when both of them exist.
Proposition 2.3 shows that the asymptotic expectation defined in Definition 2.11 is essentially unique.

Proposition 2.3. Let $\{\xi_n\}$ be a sequence of random variables. Suppose that both $E\xi/a_n$ and $E\eta/b_n$ are asymptotic expectations of ξ_n defined according to Definition 2.11(i). Then, one of the following three must hold: (a) $E\xi = E\eta = 0$; (b) $E\xi \neq 0$, $E\eta = 0$, and $b_n/a_n \to 0$; or $E\xi = 0$, $E\eta \neq 0$, and $a_n/b_n \to 0$; (c) $E\xi \neq 0$, $E\eta \neq 0$, and $(E\xi/a_n)/(E\eta/b_n) \to 1$.

If T_n is a consistent estimator of ϑ, then $T_n = \vartheta + o_p(1)$ and, by Definition 2.11(ii), T_n is asymptotically unbiased, although T_n may not be approximately unbiased.
In Example 2.34, $X_{(n)}$ has the asymptotic bias $\tilde{b}_{X_{(n)}}(P) = h_n(\theta)EY$, which is of order $n^{-(m+1)^{-1}}$.

When $a_n(T_n - \vartheta) \to_d Y$ with $EY = 0$ (e.g., $T_n = \bar{X}^2$ and $\vartheta = \mu^2$ in Example 2.33), a more precise order of the asymptotic bias of T_n may be obtained (for comparing different estimators in terms of their asymptotic biases).
Suppose that there is a sequence of random variables $\{\eta_n\}$ such that

$$a_n\eta_n \to_d Y \qquad \text{and} \qquad a_n^2(T_n - \vartheta - \eta_n) \to_d W, \tag{1}$$

where Y and W are random variables with finite means, $EY = 0$ and $EW \neq 0$.
Then we may define a_n^{-2} to be the order of $\tilde{b}_{T_n}(P)$ or define EW/a_n^2 to be the a_n^{-2} order

asymptotic bias of T_n.
However, η_n in (1) may not be unique.
Some regularity conditions have to be imposed so that the order of asymptotic bias of T_n can be uniquely defined.
We consider the case where $X_1, ..., X_n$ are i.i.d. random k-vectors with finite $\Sigma = \mathrm{Var}(\mathrm{X}_1)$. Let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, and $T_n = g(\bar{X})$, where g is a function on $\mathcal{R}^k$ that is second-order differentiable at $\mu = EX_1 \in \mathcal{R}^k$.
Consider T_n as an estimator of $\vartheta = g(\mu)$.
By Taylor's expansion,

$$T_n - \vartheta = [\nabla g(\mu)]^\tau(\bar{X} - \mu) + \frac{1}{2}(\bar{X} - \mu)^\tau \nabla^2 g(\mu)(\bar{X} - \mu) + o\left(\frac{1}{n}\right),$$

where ∇g is the k-vector of partial derivatives of g and $\nabla^2 g$ is the $k \times k$ matrix of second-order partial derivatives of g.
By the CLT and Theorem 1.10(iii),

$$\frac{n}{2}(\bar{X} - \mu)^\tau \nabla^2 g(\mu)(\bar{X} - \mu) \to_d \frac{Z_\Sigma^\tau \nabla^2 g(\mu) Z_\Sigma}{2},$$

where $Z_\Sigma = N_k(0, \Sigma)$. Thus,

$$\frac{E[Z_\Sigma^\tau \nabla^2 g(\mu) Z_\Sigma]}{2n} = \frac{\mathrm{tr}\,(\nabla^2 g(\mu)\Sigma)}{2n} \tag{2}$$

is the n^{-1} order asymptotic bias of $T_n = g(\bar{X})$, where $\mathrm{tr}(A)$ denotes the trace of the matrix A.

Example 2.35. Let $X_1, ..., X_n$ be i.i.d. binary random variables with $P(X_i = 1) = p$, where $p \in (0, 1)$ is unknown.
Consider first the estimation of $\vartheta = p(1 - p)$.
Since $\mathrm{Var}(\bar{\mathrm{X}}) = \mathrm{p(1-p)/n}$, the n^{-1} order asymptotic bias of $T_n = \bar{X}(1 - \bar{X})$ according to (2) with $g(x) = x(1 - x)$ is $-p(1 - p)/n$.
On the other hand, a direct computation shows $E[\bar{X}(1 - \bar{X})] = E\bar{X} - E\bar{X}^2 = p - (E\bar{X})^2 - \mathrm{Var}(\bar{\mathrm{X}}) = \mathrm{p(1-p) - p(1-p)/n}$.
Hence, the exact bias of T_n is the same as the n^{-1} order asymptotic bias.

Consider next the estimation of $\vartheta = p^{-1}$.
In this case, there is no unbiased estimator of p^{-1} (Exercise 84 in §2.6).
Let $T_n = \bar{X}^{-1}$.
Then, an n^{-1} order asymptotic bias of T_n according to (2) with $g(x) = x^{-1}$ is $(1 - p)/(p^2 n)$.
On the other hand, $ET_n = \infty$ for every n.

Asymptotic variance and mse

Like the bias, the mse of an estimator T_n of ϑ, $\mathrm{mse}_{T_n}(P) = E(T_n - \vartheta)^2$, is not well defined if the second moment of T_n does not exist.
We now define a version of *asymptotic mean squared error* (amse) and a measure of assessing different point estimators of a common parameter.

Definition 2.12. Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Assume that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$.
(i) The asymptotic mean squared error of T_n, denoted by $\mathrm{amse}_{\mathrm{T_n}}(\mathrm{P})$ or $\mathrm{amse}_{\mathrm{T_n}}(\theta)$ if P is in a parametric family indexed by θ, is defined to be the asymptotic expectation of $(T_n - \vartheta)^2$, i.e., $\mathrm{amse}_{\mathrm{T_n}}(\mathrm{P}) = \mathrm{EY^2/a_n^2}$. The asymptotic variance of T_n is defined to be $\sigma^2_{T_n}(P) = \mathrm{Var(Y)/a_n^2}$.
(ii) Let T_n' be another estimator of ϑ. The *asymptotic relative efficiency* of T_n' w.t.r. T_n is defined to be $e_{T_n', T_n}(P) = \mathrm{amse}_{\mathrm{T_n}}(\mathrm{P})/\mathrm{amse}_{\mathrm{T_n'}}(\mathrm{P})$.
(iii) T_n is said to be *asymptotically more efficient* than T_n' if and only if $\limsup_n e_{T_n', T_n}(P) \le 1$ for any P and < 1 for some P.

The amse and asymptotic variance are the same if and only if $EY = 0$.
By Proposition 2.3, the amse or the asymptotic variance of T_n is essentially unique and, therefore, the concept of asymptotic relative efficiency in Definition 2.12(ii)-(iii) is well defined.

In Example 2.33, $\mathrm{amse}_{\bar{\mathrm{X}}^2}(\mathrm{P}) = \sigma^2_{\bar{\mathrm{X}}^2}(\mathrm{P}) = 4\mu^2\sigma^2/\mathrm{n}$.
In Example 2.34, $\sigma^2_{X_{(n)}}(P) = [h_n(\theta)]^2\mathrm{Var(Y)}$ and $\mathrm{amse}_{\mathrm{X_{(n)}}}(\mathrm{P}) = [\mathrm{h_n}(\theta)]^2\mathrm{EY^2}$.

When both $\mathrm{mse}_{T_n}(P)$ and $\mathrm{mse}_{T_n'}(P)$ exist, one may compare T_n and T_n' by evaluating the relative efficiency $\mathrm{mse}_{T_n}(P)/\mathrm{mse}_{T_n'}(P)$.
However, this comparison may be different from the one using the asymptotic relative efficiency in Definition 2.12(ii), since the mse and amse of an estimator may be different (Exercise 115 in §2.6).
The following result shows that when the exact mse of T_n exists, it is no smaller than the amse of T_n.
It also provides a condition under which the exact mse and the amse are the same.

Proposition 2.4. Let T_n be an estimator of ϑ for every n and $\{a_n\}$ be a sequence of positive numbers satisfying $a_n \to \infty$ or $a_n \to a > 0$. Suppose that $a_n(T_n - \vartheta) \to_d Y$ with $0 < EY^2 < \infty$. Then
(i) $EY^2 \le \liminf_n E[a_n^2(T_n - \vartheta)^2]$ and
(ii) $EY^2 = \lim_{n\to\infty} E[a_n^2(T_n - \vartheta)^2]$ if and only if $\{a_n^2(T_n - \vartheta)^2\}$ is uniformly integrable.
Proof. (i) By Theorem 1.10(iii),

$$\min\{a_n^2(T_n - \vartheta)^2, t\} \to_d \min\{Y^2, t\}$$

for any $t > 0$. Since $\min\{a_n^2(T_n - \vartheta)^2, t\}$ is bounded by t,

$$\lim_{n\to\infty} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) = E(\min\{Y^2, t\})$$

(Theorem 1.8(viii)). Then

$$\begin{aligned} EY^2 &= \lim_{t\to\infty} E(\min\{Y^2, t\}) \\ &= \lim_{t\to\infty}\lim_{n\to\infty} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) \\ &= \liminf_{t,n} E(\min\{a_n^2(T_n - \vartheta)^2, t\}) \\ &\le \liminf_n E[a_n^2(T_n - \vartheta)^2], \end{aligned}$$

where the third equality follows from the fact that $E(\min\{a_n^2(T_n-\vartheta)^2, t\})$ is nondecreasing in t for any fixed n.
(ii) The result follows from Theorem 1.8(viii).

Example 2.36. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$.
Consider the estimation of $\vartheta = P(X_i = 0) = e^{-\theta}$.
Let $T_{1n} = F_n(0)$, where F_n is the empirical c.d.f.
Then T_{1n} is unbiased and has $\mathrm{mse}_{T_{1n}}(\theta) = e^{-\theta}(1-e^{-\theta})/n$.
Also, $\sqrt{n}(T_{1n}-\vartheta) \to_d N(0, e^{-\theta}(1-e^{-\theta}))$ by the CLT.
Thus, in this case $\mathrm{amse}_{\mathrm{T_{1n}}}(\theta) = \mathrm{mse}_{\mathrm{T_{1n}}}(\theta)$.
Consider $T_{2n} = e^{-\bar{X}}$.
Note that $ET_{2n} = e^{n\theta(e^{-1/n}-1)}$.
Hence $nb_{T_{2n}}(\theta) \to \theta e^{-\theta}/2$.
Using Theorem 1.12 and the CLT, we can show that $\sqrt{n}(T_{2n}-\vartheta) \to_d N(0, e^{-2\theta}\theta)$.
By Definition 2.12(i), $\mathrm{amse}_{\mathrm{T_{2n}}}(\theta) = \mathrm{e}^{-2\theta}\theta/\mathrm{n}$.
Thus, the asymptotic relative efficiency of T_{1n} w.r.t. T_{2n} is

$$e_{T_{1n},T_{2n}}(\theta) = \theta/(e^{\theta}-1),$$

which is always less than 1.
This shows that T_{2n} is asymptotically more efficient than T_{1n}.

The result for T_{2n} in Example 2.36 is a special case (with $U_n = \bar{X}$) of the following general result.

Theorem 2.6. Let g be a function on $\mathcal{R}^k$ that is differentiable at $\theta \in \mathcal{R}^k$ and let U_n be a k-vector of statistics satisfying $a_n(U_n-\theta) \to_d Y$ for a random k-vector Y with $0 < E\|Y\|^2 < \infty$ and a sequence of positive numbers $\{a_n\}$ satisfying $a_n \to \infty$. Let $T_n = g(U_n)$ be an estimator of $\vartheta = g(\theta)$. Then, the amse and asymptotic variance of T_n are, respectively, $E\{[\nabla g(\theta)]^\tau Y\}^2/a_n^2$ and $[\nabla g(\theta)]^\tau \mathrm{Var(Y)} \nabla \mathrm{g}(\theta)/\mathrm{a_n^2}$.

Lecture 28: Asymptotic inference

Statistical inference based on asymptotic criteria and approximations is called *asymptotic statistical inference* or simply *asymptotic inference*.
We have previously considered asymptotic estimation.
We now focus on asymptotic hypothesis tests and confidence sets.

Hypothesis tests

Definition 2.13. Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$ and $T_n(X)$ be a test for $H_0 : P \in \mathcal{P}_0$ versus $H_1 : P \in \mathcal{P}_1$.
(i) If $\limsup_n \alpha_{T_n}(P) \leq \alpha$ for any $P \in \mathcal{P}_0$, then α is an *asymptotic significance level* of T_n.
(ii) If $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} \alpha_{T_n}(P)$ exists, then it is called the *limiting size* of T_n.
(iii) T_n is called *consistent* if and only if the type II error probability converges to 0, i.e., $\lim_{n\to\infty}[1 - \alpha_{T_n}(P)] = 0$, for any $P \in \mathcal{P}_1$.
(iv) T_n is called *Chernoff-consistent* if and only if T_n is consistent *and* the type I error probability converges to 0, i.e., $\lim_{n\to\infty} \alpha_{T_n}(P) = 0$, for any $P \in \mathcal{P}_0$. T_n is called *strongly Chernoff-consistent* if and only if T_n is consistent and the limiting size of T_n is 0.

Obviously if T_n has size (or significance level) α for all n, then its limiting size (or asymptotic significance level) is α.
If the limiting size of T_n is $\alpha \in (0, 1)$, then for any $\epsilon > 0$, T_n has size $\alpha + \epsilon$ for all $n \geq n_0$, where n_0 is independent of P.
Hence T_n has level of significance $\alpha + \epsilon$ for any $n \geq n_0$.
However, if $\mathcal{P}_0$ is not a parametric family, it is likely that the limiting size of T_n is 1 (see, e.g., Example 2.37).
This is the reason why we consider the weaker requirement in Definition 2.13(i).
If T_n has asymptotic significance level α, then for any $\epsilon > 0$, $\alpha_{T_n}(P) < \alpha+\epsilon$ for all $n \geq n_0(P)$ but $n_0(P)$ depends on $P \in \mathcal{P}_0$; and there is no guarantee that T_n has significance level $\alpha + \epsilon$ for any n.

The consistency in Definition 2.13(iii) only requires that the type II error probability converge to 0.
We may define uniform consistency to be $\lim_{n\to\infty} \sup_{P\in\mathcal{P}_1}[1 - \alpha_{T_n}(P)] = 0$, but it is not satisfied in most problems.
If $\alpha \in (0, 1)$ is a pre-assigned level of significance for the problem, then a consistent test T_n having asymptotic significance level α is called *asymptotically correct*, and a consistent test having limiting size α is called *strongly* asymptotically correct.

The Chernoff-consistency (or strong Chernoff-consistency) in Definition 2.13(iv) requires that both types of error probabilities converge to 0.
Mathematically, Chernoff-consistency (or strong Chernoff-consistency) is better than asymptotic correctness (or strongly asymptotic correctness).
After all, both types of error probabilities should decrease to 0 if sampling can be continued indefinitely.
However, if α is chosen to be small enough so that error probabilities smaller than α can

be practically treated as 0, then the asymptotic correctness (or strongly asymptotic correctness) is enough, and is probably preferred, since requiring an unnecessarily small type I error probability usually results in an unnecessary increase in the type II error probability.

Example 2.37. Consider the testing problem $H_0 : \mu \le \mu_0$ versus $H_1 : \mu > \mu_0$ based on i.i.d. $X_1, ..., X_n$ with $EX_1 = \mu \in \mathcal{R}$. If each X_i has the $N(\mu, \sigma^2)$ distribution with a known σ^2, then the test $T_{c_\alpha} I_{(c_\alpha,\infty)}(\bar{X})$ with $c_\alpha = \sigma z_{1-\alpha}/\sqrt{n} + \mu_0$ and $\alpha \in (0, 1)$ has size α (and, therefore, limiting size α).
For any $\mu > \mu_0$,

$$1 - \alpha_{T_{c_\alpha}}(\mu) = \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) \to 0 \tag{1}$$

as $n \to \infty$.
This shows that T_{c_α} is consistent and, hence, is strongly asymptotically correct.
The convergence in (1) is not uniform in $\mu > \mu_0$, but is uniform in $\mu > \mu_1$ for any fixed $\mu_1 > \mu_0$.

Since the size of T_{c_α} is α for all n, T_{c_α} is not Chernoff-consistent.
A strongly Chernoff-consistent test can be obtained as follows.
Let

$$\alpha_n = 1 - \Phi(\sqrt{n} a_n), \tag{2}$$

where a_n's are positive numbers satisfying $a_n \to 0$ and $\sqrt{n} a_n \to \infty$.
Let T_n be T_{c_α} with $\alpha = \alpha_n$ for each n.
Then, T_n has size α_n.
Since $\alpha_n \to 0$, The limiting size of T_n is 0.
On the other hand, (1) still holds with α replaced by α_n.
This follows from the fact that

$$z_{1-\alpha_n} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma} = \sqrt{n}\left(a_n + \frac{\mu_0 - \mu}{\sigma}\right) \to -\infty$$

for any $\mu > \mu_0$.
Hence T_n is strongly Chernoff-consistent.
However, if $\alpha_n < \alpha$, then, from the left-hand side of (1), $1 - \alpha_{T_{c_\alpha}}(\mu) < 1 - \alpha_{T_n}(\mu)$ for any $\mu > \mu_0$.

We now consider the case where the population P is not in a parametric family.
We still assume that $\sigma^2 = \mathrm{Var(X_i)}$ is known.
Using the CLT, we can show that for $\mu > \mu_0$,

$$\lim_{n\to\infty}[1 - \alpha_{T_{c_\alpha}}(\mu)] = \lim_{n\to\infty} \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right) = 0,$$

i.e., T_{c_α} is still consistent.
For $\mu \le \mu_0$,

$$\lim_{n\to\infty} \alpha_{T_{c_\alpha}}(\mu) = 1 - \lim_{n\to\infty} \Phi\left(z_{1-\alpha} + \frac{\sqrt{n}(\mu_0 - \mu)}{\sigma}\right),$$

which equals α if $\mu = \mu_0$ and 0 if $\mu < \mu_0$.
Thus, the asymptotic significance level of T_{c_α} is α.
Combining these two results, we know that T_{c_α} is asymptotically correct.
However, if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments, then one can show that the limiting size of T_{c_α} is 1 (exercise).
For α_n defined by (2), we can show that $T_n = T_{c_\alpha}$ with $\alpha = \alpha_n$ is Chernoff-consistent (exercise).
But T_n is not strongly Chernoff-consistent if $\mathcal{P}$ contains all possible populations on $\mathcal{R}$ with finite second moments.

Example. Let $(X_1, ..., X_n)$ be a random sample from the exponential distribution $E(0, \theta)$, where $\theta \in (0, \infty)$.
Consider the hypotheses $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$, where $\theta_0 > 0$ is a fixed constant.
Let $T_c = I_{(c,\infty)}(\bar{X})$, where $\bar{X}$ is the sample mean.
$\bar{X}/\theta$ has the gamma distribution with shape parameter n and scale parameter θ/n.
Let $G_{n,\theta}$ denote the cumulative distribution function of this distribution and $c_{n,\alpha}$ be the constant satisfying $G_{n,\theta_0}(c_{n,\alpha}) = 1 - \alpha$.
Then,

$$\sup_{\theta \le \theta_0} P(T_{c_{n,\alpha}} = 1) = \sup_{\theta \le \theta_0} [1 - G_{n,\theta}(c_{n,\alpha})] = 1 - G_{n,\theta_0}(c_{n,\alpha}) = \alpha,$$

i.e., the size of $T_{c_{n,\alpha}}$ is α.
Since the power of $T_{c_{n,\alpha}}$ is $P(T_{c_{n,\alpha}} = 1) = P(\bar{X} > c_{n,\alpha})$ for $\theta > \theta_0$ and, by the law of large numbers, $\bar{X} \to_p \theta$, the consistency of $T_{c_{n,\alpha}}$ follows if we can show that $\lim_{n\to\infty} c_{n,\alpha} = \theta_0$.
By the central limit theorem, $\sqrt{n}(\bar{X} - \theta) \to_d N(0, \theta^2)$.
Hence, $\sqrt{n}(\frac{\bar{X}}{\theta} - 1) \to_d N(0, 1)$.
By Pólya's theorem (Proposition 1.16),

$$\lim_{n\to\infty} \sup_t \left| P\left(\sqrt{n}\left(\tfrac{\bar{X}}{\theta} - 1\right) \le t\right) - \Phi(t) \right| = 0,$$

where Φ is the cumulative distribution function of the standard normal distribution.
When $\theta = \theta_0$,

$$\alpha = P(\bar{X} \ge c_{n,\alpha}) = P\left(\sqrt{n}\left(\tfrac{\bar{X}}{\theta_0} - 1\right) \ge \sqrt{n}\left(\tfrac{c_{n,\alpha}}{\theta_0} - 1\right)\right).$$

Hence

$$\lim_{n\to\infty} \Phi\left(\sqrt{n}\left(\tfrac{c_{n,\alpha}}{\theta_0} - 1\right)\right) = 1 - \alpha,$$

which implies $\lim_{n\to\infty} \sqrt{n}(\frac{c_{n,\alpha}}{\theta_0} - 1) = \Phi^{-1}(1 - \alpha)$ and, thus, $\lim_{n\to\infty} c_{n,\alpha} = \theta_0$.

Let $\{a_n\}$ be a sequence of positive numbers such that $\lim_{n\to\infty} a_n = 0$ and $\lim_{n\to\infty} \sqrt{n} a_n = \infty$.
Let $\alpha_n = 1 - \Phi(\sqrt{n} a_n)$ and $b_n = c_{n,\alpha_n}$.
From the previous derivation, the size of T_{b_n} is α_n, which converges to 0 as $n \to \infty$ since $\lim_{n\to\infty} \sqrt{n} a_n = \infty$.
Using the previous argument, we can show that

$$\lim_{n\to\infty} \left| 1 - \alpha_n - \Phi\left(\sqrt{n}\left(\tfrac{c_{n,\alpha_n}}{\theta_0} - 1\right)\right) \right| = 0,$$

which implies that

$$\lim_{n\to\infty} \frac{\sqrt{n}}{\Phi^{-1}(1-\alpha_n)} \left(\frac{c_{n,\alpha_n}}{\theta_0} - 1\right) = 1.$$

Since $1 - \alpha_n = \Phi(\sqrt{n}a_n)$, this implies that $\lim_{n\to\infty} c_{n,\alpha_n} = \theta_0$.
Since $b_n = c_{n,\alpha_n}$, the test T_{b_n} is Chernoff-consistent.

Confidence sets

Definition 2.14. Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P}$, ϑ be a k-vector of parameters related to P, and $C(X)$ be a confidence set for ϑ.
(i) If $\liminf_n P(\vartheta \in C(X)) \geq 1 - \alpha$ for any $P \in \mathcal{P}$, then $1 - \alpha$ is an *asymptotic significance level* of $C(X)$.
(ii) If $\lim_{n\to\infty} \inf_{P\in\mathcal{P}} P(\vartheta \in C(X))$ exists, then it is called the *limiting confidence coefficient* of $C(X)$.

Note that the asymptotic significance level and limiting confidence coefficient of a confidence set are very similar to the asymptotic significance level and limiting size of a test, respectively. Some conclusions are also similar.
For example, in a parametric problem one can often find a confidence set having limiting confidence coefficient $1-\alpha \in (0, 1)$, which implies that for any $\epsilon > 0$, the confidence coefficient of $C(X)$ is $1-\alpha-\epsilon$ for all $n \geq n_0$, where n_0 is independent of P. In a nonparametric problem the limiting confidence coefficient of $C(X)$ might be 0, whereas $C(X)$ may have asymptotic significance level $1 - \alpha \in (0, 1)$, but for any fixed n, the confidence coefficient of $C(X)$ might be 0.

Lecture 29: UMVUE and the method of using the distribution of a sufficient and complete statistic

Unbiased or asymptotically unbiased estimation plays an important role in point estimation theory.
Unbiased estimators can be used as "building blocks" for the construction of better estimators.
Asymptotic unbiasedness is necessary for consistency.
How to derive unbiased estimators
How to find the best unbiased estimators

UMVUE

X: a sample from an unknown population $P \in \mathcal{P}$
ϑ: a real-valued parameter related to P.
An estimator $T(X)$ of ϑ is unbiased if and only if $E[T(X)] = \vartheta$ for any $P \in \mathcal{P}$.
If there exists an unbiased estimator of ϑ, then ϑ is called an *estimable* parameter.

Definition 3.1. An unbiased estimator $T(X)$ of ϑ is called the *uniformly minimum variance unbiased estimator* (UMVUE) if and only if $\mathrm{Var(T(X))} \leq \mathrm{Var(U(X))}$ for any $P \in \mathcal{P}$ and any other unbiased estimator $U(X)$ of ϑ.

Since the mse of any unbiased estimator is its variance, a UMVUE is $\Im$-optimal in mse with $\Im$ being the class of all unbiased estimators.
One can similarly define the uniformly minimum risk unbiased estimator in statistical decision theory when we use an arbitrary loss instead of the squared error loss that corresponds to the mse.

Sufficient and complete statistics

The derivation of a UMVUE is relatively simple if there exists a sufficient and complete statistic for $P \in \mathcal{P}$.

Theorem 3.1 (Lehmann-Scheffé theorem). Suppose that there exists a sufficient and complete statistic $T(X)$ for $P \in \mathcal{P}$. If ϑ is estimable, then there is a unique unbiased estimator of ϑ that is of the form $h(T)$ with a Borel function h. (Two estimators that are equal a.s. $\mathcal{P}$ are treated as one estimator.) Furthermore, $h(T)$ is the unique UMVUE of ϑ.

This theorem is a consequence of Theorem 2.5(ii) (Rao-Blackwell theorem).
One can easily extend this theorem to the case of the uniformly minimum risk unbiased estimator under any loss function $L(P, a)$ that is strictly convex in a.
The uniqueness of the UMVUE follows from the completeness of $T(X)$.

Two typical ways to derive a UMVUE when a sufficient and complete statistic T is available.

The 1st method: Directly solving for h
Need the distribution of T
Try some function h to see if $E[h(T)]$ is related to ϑ
If $E[h(T)] = \vartheta$ for all P, what should h be?

Example 3.1. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on $(0, \theta)$, $\theta > 0$.
Suppose that $\vartheta = \theta$.
Since the sufficient and complete statistic $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$,

$$EX_{(n)} = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

Hence an unbiased estimator of θ is $(n+1)X_{(n)}/n$, which is the UMVUE.
Suppose that $\vartheta = g(\theta)$, where g is a differentiable function on $(0, \infty)$.
An unbiased estimator $h(X_{(n)})$ of ϑ must satisfy

$$\theta^n g(\theta) = n\int_0^\theta h(x)x^{n-1}dx \qquad \text{for all } \theta > 0.$$

Differentiating both sizes of the previous equation and applying the result of differentiation of an integral (Royden (1968, §5.3)) lead to

$$n\theta^{n-1}g(\theta) + \theta^n g'(\theta) = nh(\theta)\theta^{n-1}.$$

Hence, the UMVUE of ϑ is $h(X_{(n)}) = g(X_{(n)}) + n^{-1}X_{(n)}g'(X_{(n)})$.
In particular, if $\vartheta = \theta$, then the UMVUE of θ is $(1 + n^{-1})X_{(n)}$.

Example 3.2. Let $X_1, ..., X_n$ be i.i.d. from the Poisson distribution $P(\theta)$ with an unknown $\theta > 0$.
Then $T(X) = \sum_{i=1}^n X_i$ is sufficient and complete for $\theta > 0$ and has the Poisson distribution $P(n\theta)$.
Since $E(T) = n\theta$, the UMVUE of θ is T/n.
Suppose that $\vartheta = g(\theta)$, where g is a smooth function such that $g(x) = \sum_{j=0}^\infty a_j x^j$, $x > 0$.
An unbiased estimator $h(T)$ of ϑ must satisfy

$$\begin{aligned}\sum_{t=0}^\infty \frac{h(t)n^t}{t!}\theta^t &= e^{n\theta}g(\theta)\\ &= \sum_{k=0}^\infty \frac{n^k}{k!}\theta^k \sum_{j=0}^\infty a_j\theta^j\\ &= \sum_{t=0}^\infty \left(\sum_{j,k:j+k=t} \frac{n^k a_j}{k!}\right)\theta^t\end{aligned}$$

for any $\theta > 0$.
Thus, a comparison of coefficients in front of θ^t leads to

$$h(t) = \frac{t!}{n^t}\sum_{j,k:j+k=t} \frac{n^k a_j}{k!},$$

i.e., $h(T)$ is the UMVUE of ϑ.
In particular, if $\vartheta = \theta^r$ for some fixed integer $r \geq 1$, then $a_r = 1$ and $a_k = 0$ if $k \neq r$ and

$$h(t) = \begin{cases} 0 & t < r \\ \frac{t!}{n^r(t-r)!} & t \geq r. \end{cases}$$

Example 3.5. Let $X_1, ..., X_n$ be i.i.d. from a power series distribution (see Exercise 13 in §2.6), i.e.,

$$P(X_i = x) = \gamma(x)\theta^x/c(\theta), \qquad x = 0, 1, 2, ...,$$

with a known function $\gamma(x) \geq 0$ and an unknown parameter $\theta > 0$.
It turns out that the joint distribution of $X = (X_1, ..., X_n)$ is in an exponential family with a sufficient and complete statistic $T(X) = \sum_{i=1}^n X_i$.
Furthermore, the distribution of T is also in a power series family, i.e.,

$$P(T = t) = \gamma_n(t)\theta^t/[c(\theta)]^n, \qquad t = 0, 1, 2, ...,$$

where $\gamma_n(t)$ is the coefficient of θ^t in the power series expansion of $[c(\theta)]^n$ (Exercise 13 in §2.6).
This result can help us to find the UMVUE of $\vartheta = g(\theta)$.
For example, by comparing both sides of

$$\sum_{t=0}^{\infty} h(t)\gamma_n(t)\theta^t = [c(\theta)]^{n-p}\theta^r,$$

we conclude that the UMVUE of $\theta^r/[c(\theta)]^p$ is

$$h(T) = \begin{cases} 0 & T < r \\ \frac{\gamma_{n-p}(T-r)}{\gamma_n(T)} & T \geq r, \end{cases}$$

where r and p are nonnegative integers.
In particular, the case of $p = 1$ produces the UMVUE $\gamma(r)h(T)$ of the probability $P(X_1 = r) = \gamma(r)\theta^r/c(\theta)$ for any nonnegative integer r.

Example 3.6. Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$.
We have discussed in §2.2 that in many cases the vector of order statistics, $T = (X_{(1)}, ..., X_{(n)})$, is sufficient and complete for $P \in \mathcal{P}$.
(For example, $\mathcal{P}$ is the collection of all Lebesgue p.d.f.'s.) Note that an estimator $\varphi(X_1, ..., X_n)$ is a function of T if and only if the function φ is symmetric in its n arguments.
Hence, if T is sufficient and complete, then a symmetric unbiased estimator of any estimable ϑ is the UMVUE.
For example,
$\bar{X}$ is the UMVUE of $\vartheta = EX_1$;
S^2 is the UMVUE of $\mathrm{Var}(\mathrm{X}_1)$;
$n^{-1}\sum_{i=1}^n X_i^2 - S^2$ is the UMVUE of $(EX_1)^2$;
$F_n(t)$ is the UMVUE of $P(X_1 \leq t)$ for any fixed t.

These conclusions are not true if T is *not* sufficient and complete for $P \in \mathcal{P}$.
For example, if $n > 1$ and $\mathcal{P}$ contains all symmetric distributions having Lebesgue p.d.f.'s and finite means, then there is no UMVUE for $\vartheta = EX_1$.
Suppose that T is a UMVUE of μ.

Let $\mathcal{P}_1 = \{N(\mu, 1) : \mu \in \mathcal{R}\}$.
Since the sample mean $\bar{X}$ is UMVUE when $\mathcal{P}_1$ is considered, and the Lebesgue measure is dominated by any $P \in \mathcal{P}_1$, we conclude that $T = \bar{X}$ a.e. Lebesgue measure.
Let $\mathcal{P}_2$ be the family of uniform distributions on $(\theta_1 - \theta_2, \theta_1 + \theta_2)$, $\theta_1 \in \mathcal{R}$, $\theta_2 > 0$.
Then $(X_{(1)} + X_{(n)})/2$ is the UMVUE when $\mathcal{P}_2$ is considered, where $X_{(j)}$ is the jth order statistic.
Then $\bar{X} = (X_{(1)} + X_{(n)})/2$ a.s. P for any $P \in \mathcal{P}_2$, which is impossible if $n > 1$.
Hence, there is no UMVUE of μ.

What if $n = 1$?
Consider the sub-family $\mathcal{P}_1 = \{N(\mu, 1) : \mu \in \mathcal{R}\}$.
Then X_1 is complete for $P \in \mathcal{P}_1$.
Hence, $E[h(X_1)] = 0$ for any $P \in \mathcal{P}$ implies that $E[h(X_1)] = 0$ for any $P \in \mathcal{P}_1$ and, thus, $h = 0$ a.e. Lebesgue measure.
This shows that X_1 is complete when the family $\mathcal{P}$ is considered.
Since $EX_1 = \mu$, X_1 is the UMVUE of μ.

Lecture 30: UMVUE: the method of conditioning

The 2nd method of deriving a UMVUE is conditioning on a sufficient and complete statistic $T(X)$,
i.e., if $U(X)$ is any unbiased estimator of ϑ, then $E[U(X)|T]$ is the UMVUE of ϑ.
We do not need the distribution of T.
But we need to work out the conditional expectation $E[U(X)|T]$.
From the uniqueness of the UMVUE, it does not matter which $U(X)$ is used.
Thus, we should choose $U(X)$ so as to make the calculation of $E[U(X)|T]$ as easy as possible.

Example 3.3. Let $X_1, ..., X_n$ be i.i.d. from the exponential distribution $E(0, \theta)$.
$F_\theta(x) = (1 - e^{-x/\theta})I_{(0,\infty)}(x)$.
Consider the estimation of $\vartheta = 1 - F_\theta(t)$.
$\bar{X}$ is sufficient and complete for $\theta > 0$.
$I_{(t,\infty)}(X_1)$ is unbiased for ϑ,

$$E[I_{(t,\infty)}(X_1)] = P(X_1 > t) = \vartheta.$$

Hence

$$T(X) = E[I_{(t,\infty)}(X_1)|\bar{X}] = P(X_1 > t|\bar{X})$$

is the UMVUE of ϑ. If the conditional distribution of X_1 given $\bar{X}$ is available, then we can calculate $P(X_1 > t|\bar{X})$ directly.
By Basu's theorem (Theorem 2.4), $X_1/\bar{X}$ and $\bar{X}$ are independent.
By Proposition 1.10(vii),

$$P(X_1 > t|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{X}|\bar{X} = \bar{x}) = P(X_1/\bar{X} > t/\bar{x}).$$

To compute this unconditional probability, we need the distribution of

$$X_1 \Big/ \sum_{i=1}^n X_i = X_1 \Big/ \left(X_1 + \sum_{i=2}^n X_i \right).$$

Using the transformation technique discussed in §1.3.1 and the fact that $\sum_{i=2}^n X_i$ is independent of X_1 and has a gamma distribution, we obtain that $X_1/\sum_{i=1}^n X_i$ has the Lebesgue p.d.f. $(n-1)(1-x)^{n-2}I_{(0,1)}(x)$.
Hence

$$P(X_1 > t|\bar{X} = \bar{x}) = (n-1)\int_{t/(n\bar{x})}^1 (1-x)^{n-2}dx = \left(1 - \frac{t}{n\bar{x}}\right)^{n-1}$$

and the UMVUE of ϑ is

$$T(X) = \left(1 - \frac{t}{n\bar{X}}\right)^{n-1}.$$

Example 3.4. Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$.
From Example 2.18, $T = (\bar{X}, S^2)$ is sufficient and complete for $\theta = (\mu, \sigma^2)$;
$\bar{X}$ and $(n-1)S^2/\sigma^2$ are independent;
$\bar{X}$ has the $N(\mu, \sigma^2/n)$ distribution;
S^2 has the chi-square distribution χ^2_{n-1}.
Using the method of solving for h directly, we find that
the UMVUE for μ is $\bar{X}$;
the UMVUE of μ^2 is $\bar{X}^2 - S^2/n$;
the UMVUE for σ^r with $r > 1 - n$ is $k_{n-1,r}S^r$, where

$$k_{n,r} = \frac{n^{r/2}\Gamma(n/2)}{2^{r/2}\Gamma\left(\frac{n+r}{2}\right)}$$

and the UMVUE of μ/σ is $k_{n-1,-1}\bar{X}/S$, if $n > 2$.

Suppose that ϑ satisfies $P(X_1 \le \vartheta) = p$ with a fixed $p \in (0, 1)$.
Let Φ be the c.d.f. of the standard normal distribution.
Then $\vartheta = \mu + \sigma\Phi^{-1}(p)$ and its UMVUE is $\bar{X} + k_{n-1,1}S\Phi^{-1}(p)$.

Let c be a fixed constant and $\vartheta = P(X_1 \le c) = \Phi\left(\frac{c-\mu}{\sigma}\right)$.
We can find the UMVUE of ϑ using the method of conditioning.
Since $I_{(-\infty,c)}(X_1)$ is an unbiased estimator of ϑ, the UMVUE of ϑ is

$$E[I_{(-\infty,c)}(X_1)|T] = P(X_1 \le c|T).$$

By Basu's theorem, the ancillary statistic $Z(X) = (X_1 - \bar{X})/S$ is independent of $T = (\bar{X}, S^2)$.
Then, by Proposition 1.10(vii),

$$\begin{aligned} P\left(X_1 \le c|T = (\bar{x}, s^2)\right) &= P\left(Z \le \frac{c - \bar{X}}{S}\middle| T = (\bar{x}, s^2)\right) \\ &= P\left(Z \le \frac{c - \bar{x}}{s}\right). \end{aligned}$$

It can be shown that Z has the Lebesgue p.d.f.

$$f(z) = \frac{\sqrt{n}\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}(n-1)\Gamma\left(\frac{n-2}{2}\right)}\left[1 - \frac{nz^2}{(n-1)^2}\right]^{(n/2)-2} I_{(0,(n-1)/\sqrt{n})}(|z|)$$

Hence the UMVUE of ϑ is

$$P(X_1 \le c|T) = \int_{-(n-1)/\sqrt{n}}^{(c-\bar{X})/S} f(z)dz$$

Suppose that we would like to estimate $\vartheta = \frac{1}{\sigma}\Phi'\left(\frac{c-\mu}{\sigma}\right)$, the Lebesgue p.d.f. of X_1 evaluated at a fixed c, where Φ' is the first-order derivative of Φ.

By the previous result, the conditional p.d.f. of X_1 given $\bar{X} = \bar{x}$ and $S^2 = s^2$ is $s^{-1} f\left(\frac{x-\bar{x}}{s}\right)$.
Let f_T be the joint p.d.f. of $T = (\bar{X}, S^2)$.
Then

$$\vartheta = \int\int \frac{1}{s} f\left(\frac{c-\bar{x}}{s}\right) f_T(t)dt = E\left[\frac{1}{S} f\left(\frac{c-\bar{X}}{S}\right)\right].$$

Hence the UMVUE of ϑ is

$$\frac{1}{S} f\left(\frac{c-\bar{X}}{S}\right).$$

Example. Let $X_1, ..., X_n$ be i.i.d. with Lebesgue p.d.f. $f_\theta(x) = \theta x^{-2} I_{(\theta,\infty)}(x)$, where $\theta > 0$ is unknown.
Suppose that $\vartheta = P(X_1 > t)$ for a constant $t > 0$.
The smallest order statistic $X_{(1)}$ is sufficient and complete for θ.
Hence, the UMVUE of ϑ is

$$\begin{aligned}
P(X_1 > t|X_{(1)}) &= P(X_1 > t|X_{(1)} = x_{(1)}) \\
&= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{X_{(1)}}\Big| X_{(1)} = x_{(1)}\right) \\
&= P\left(\frac{X_1}{X_{(1)}} > \frac{t}{x_{(1)}}\Big| X_{(1)} = x_{(1)}\right) \\
&= P\left(\frac{X_1}{X_{(1)}} > s\right)
\end{aligned}$$

(Basu's theorem), where $s = t/x_{(1)}$.
If $s \le 1$, this probability is 1.
Consider $s > 1$ and assume $\theta = 1$ in the calculation:

$$\begin{aligned}
P\left(\frac{X_1}{X_{(1)}} > s\right) &= \sum_{i=1}^n P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= \sum_{i=2}^n P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_i\right) \\
&= (n-1)P\left(\frac{X_1}{X_{(1)}} > s, X_{(1)} = X_n\right) \\
&= (n-1)P\left(X_1 > sX_n, X_2 > X_n, ..., X_{n-1} > X_n\right) \\
&= (n-1)\int_{x_1 > sx_n, x_2 > x_n, ..., x_{n-1} > x_n} \prod_{i=1}^n \frac{1}{x_i^2} dx_1 \cdots dx_n \\
&= (n-1)\int_1^\infty \left[\int_{sx_n}^\infty \prod_{i=2}^{n-1}\left(\int_{x_n}^\infty \frac{1}{x_i^2} dx_i\right)\frac{1}{x_1^2} dx_1\right] dx_n \\
&= (n-1)\int_1^\infty \frac{1}{s x_n^{n-1}} dx_n \\
&= \frac{(n-1)x_{(1)}}{nt}
\end{aligned}$$

This shows that the UMVUE of $P(X_1 > t)$ is

$$h(X_{(1)}) = \begin{cases} \frac{(n-1)X_{(1)}}{nt} & X_{(1)} < t \\ 1 & X_{(1)} \geq t \end{cases}$$

Another way of showing $h(X_{(1)})$ is the UMVUE.
Note that the Lebesgue p.d.f. of $X_{(1)}$ is

$$\frac{n\theta^n}{x^{n+1}} I_{(\theta,\infty)}(x).$$

If $\theta < t$,

$$\begin{aligned} E[h(X_{(1)})] &= \int_\theta^\infty h(x) \frac{n\theta^n}{x^{n+1}} dx \\ &= \int_\theta^t \frac{(n-1)x}{nt} \frac{n\theta^n}{x^{n+1}} dx + \int_t^\infty \frac{n\theta^n}{x^{n+1}} dx \\ &= \frac{\theta^n}{t\theta^{n-1}} - \frac{\theta^n}{t^n} + \frac{\theta^n}{t^n} \\ &= \frac{\theta}{t} \\ &= P(X_1 > t). \end{aligned}$$

If $\theta \geq t$, then $P(X_1 > t) = 1$ and $h(X_{(1)}) = 1$ a.s. P_θ since $P(t > X_{(1)}) = 0$.
Hence, for any $\theta > 0$,

$$E[h(X_{(1)}) = P(X_1 > t).$$

Lecture 31: UMVUE: a necessary and sufficient condition

When a complete and sufficient statistic is not available, it is usually very difficult to derive a UMVUE.
In some cases, the following result can be applied, if we have enough knowledge about unbiased estimators of 0.

Theorem 3.2. Let $\mathcal{U}$ be the set of all unbiased estimators of 0 with finite variances and T be an unbiased estimator of ϑ with $E(T^2) < \infty$.
(i) A necessary and sufficient condition for $T(X)$ to be a UMVUE of ϑ is that $E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}$ and any $P \in \mathcal{P}$.
(ii) Suppose that $T = h(\tilde{T})$, where $\tilde{T}$ is a sufficient statistic for $P \in \mathcal{P}$ and h is a Borel function.
Let $\mathcal{U}_{\tilde{T}}$ be the subset of $\mathcal{U}$ consisting of Borel functions of $\tilde{T}$.
Then a necessary and sufficient condition for T to be a UMVUE of ϑ is that $E[T(X)U(X)] = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and any $P \in \mathcal{P}$.
Proof. (i) Suppose that T is a UMVUE of ϑ.
Then $T_c = T + cU$, where $U \in \mathcal{U}$ and c is a fixed constant, is also unbiased for ϑ and, thus,

$$\mathrm{Var}(T_c) \geq \mathrm{Var}(T) \qquad c \in \mathcal{R},\ P \in \mathcal{P},$$

which is the same as

$$c^2\mathrm{Var}(U) + 2c\mathrm{Cov}(T, U) \geq 0 \qquad c \in \mathcal{R},\ P \in \mathcal{P}.$$

This is impossible unless $\mathrm{Cov}(T, U) = E(TU) = 0$ for any $P \in \mathcal{P}$.

Suppose now $E(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$.
Let T_0 be another unbiased estimator of ϑ with $\mathrm{Var}(T_0) < \infty$.
Then $T - T_0 \in \mathcal{U}$ and, hence,

$$E[T(T - T_0)] = 0 \qquad P \in \mathcal{P},$$

which with the fact that $ET = ET_0$ implies that

$$\mathrm{Var}(T) = \mathrm{Cov}(T, T_0) \qquad P \in \mathcal{P}.$$

Note that $[\mathrm{Cov}(T, T_0)]^2 \leq \mathrm{Var}(T)\mathrm{Var}(T_0)$.
Hence $\mathrm{Var}(T) \leq \mathrm{Var}(T_0)$ for any $P \in \mathcal{P}$.
(ii) It suffices to show that $E(TU) = 0$ for any $U \in \mathcal{U}_{\tilde{T}}$ and $P \in \mathcal{P}$ implies that $E(TU) = 0$ for any $U \in \mathcal{U}$ and $P \in \mathcal{P}$.
Let $U \in \mathcal{U}$. Then $E(U|\tilde{T}) \in \mathcal{U}_{\tilde{T}}$ and the result follows from the fact that $T = h(\tilde{T})$ and

$$E(TU) = E[E(TU|\tilde{T})] = E[E(h(\tilde{T})U|\tilde{T})] = E[h(\tilde{T})E(U|\tilde{T})].$$

Theorem 3.2 can be used
to find a UMVUE,
to check whether a particular estimator is a UMVUE, and
to show the nonexistence of any UMVUE.

If there is a sufficient statistic, then by Rao-Blackwell's theorem, we only need to focus on functions of the sufficient statistic and, hence, Theorem 3.2(ii) is more convenient to use.

As a consequence of Theorem 3.2, we have the following useful result.

Corollary 3.1. (i) Let T_j be a UMVUE of ϑ_j, $j = 1, ..., k$, where k is a fixed positive integer. Then $\sum_{j=1}^k c_j T_j$ is a UMVUE of $\vartheta = \sum_{j=1}^k c_j \vartheta_j$ for any constants $c_1, ..., c_k$.
(ii) Let T_1 and T_2 be two UMVUE's of ϑ. Then $T_1 = T_2$ a.s. P for any $P \in \mathcal{P}$.

Example 3.7. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on the interval $(0, \theta)$.
In Example 3.1, $(1 + n^{-1})X_{(n)}$ is shown to be the UMVUE for θ when the parameter space is $\Theta = (0, \infty)$.
Suppose now that $\Theta = [1, \infty)$.
Then $X_{(n)}$ is not complete, although it is still sufficient for θ.
Thus, Theorem 3.1 does not apply to $X_{(n)}$.
We now illustrate how to use Theorem 3.2(ii) to find a UMVUE of θ.
Let $U(X_{(n)})$ be an unbiased estimator of 0.
Since $X_{(n)}$ has the Lebesgue p.d.f. $n\theta^{-n}x^{n-1}I_{(0,\theta)}(x)$,

$$0 = \int_0^1 U(x)x^{n-1}dx + \int_1^\theta U(x)x^{n-1}dx$$

for all $\theta \geq 1$.
This implies that $U(x) = 0$ a.e. Lebesgue measure on $[1, \infty)$ and

$$\int_0^1 U(x)x^{n-1}dx = 0.$$

Consider $T = h(X_{(n)})$.
To have $E(TU) = 0$, we must have

$$\int_0^1 h(x)U(x)x^{n-1}dx = 0.$$

Thus, we may consider the following function:

$$h(x) = \begin{cases} c & 0 \leq x \leq 1 \\ bx & x > 1, \end{cases}$$

where c and b are some constants.
From the previous discussion,

$$E[h(X_{(n)})U(X_{(n)})] = 0, \qquad \theta \geq 1.$$

Since $E[h(X_{(n)})] = \theta$, we obtain that

$$\begin{aligned}\theta &= cP(X_{(n)} \le 1) + bE[X_{(n)} I_{(1,\infty)}(X_{(n)})] \\ &= c\theta^{-n} + [bn/(n+1)](\theta - \theta^{-n}).\end{aligned}$$

Thus, $c = 1$ and $b = (n+1)/n$. The UMVUE of θ is then

$$h(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ (1+n^{-1})X_{(n)} & X_{(n)} > 1. \end{cases}$$

This estimator is better than $(1+n^{-1})X_{(n)}$, which is the UMVUE when $\Theta = (0,\infty)$ and does not make use of the information about $\theta \ge 1$.

In fact, $h(X_{(n)})$ is complete and sufficient for θ.
It suffices to show that

$$g(X_{(n)}) = \begin{cases} 1 & 0 \le X_{(n)} \le 1 \\ X_{(n)} & X_{(n)} > 1. \end{cases}$$

is complete and sufficient for θ.
The sufficiency follows from the fact that the joint p.d.f. of $X_1, ..., X_n$ is

$$\frac{1}{\theta^n} I_{(0,\theta)}(X_{(n)}) = \frac{1}{\theta^n} I_{(0,\theta)}(g(X_{(n)})).$$

If $E[f(g(X_{(n)})] = 0$ for all $\theta > 1$, then

$$0 = \int_0^\theta f(g(x)) x^{n-1} dx = \int_0^1 f(1) x^{n-1} dx + \int_1^\theta f(x) x^{n-1} dx$$

for all $\theta > 1$.
Letting $\theta \to 1$ we obtain that $f(1) = 0$. Then

$$0 = \int_1^\theta f(x) x^{n-1} dx$$

for all $\theta > 1$, which implies $f(x) = 0$ a.e. for $x > 1$.
Hence, $g(X_{(n)})$ is complete.

Example 3.8. Let X be a sample (of size 1) from the uniform distribution $U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$, $\theta \in \mathcal{R}$.
We now apply Theorem 3.2 to show that there is no UMVUE of $\vartheta = g(\theta)$ for any nonconstant function g.
Note that an unbiased estimator $U(X)$ of 0 must satisfy

$$\int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} U(x) dx = 0 \qquad \text{for all } \theta \in \mathcal{R}.$$

Differentiating both sizes of the previous equation and applying the result of differentiation of an integral lead to $U(x) = U(x+1)$ a.e. m, where m is the Lebesgue measure on $\mathcal{R}$.
If T is a UMVUE of $g(\theta)$, then $T(X)U(X)$ is unbiased for 0 and, hence, $T(x)U(x) = T(x+1)U(x+1)$ a.e. m, where $U(X)$ is any unbiased estimator of 0.
Since this is true for all U, $T(x) = T(x+1)$ a.e. m. Since T is unbiased for $g(\theta)$,

$$g(\theta) = \int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} T(x)dx \qquad \text{for all } \theta \in \mathcal{R}.$$

Differentiating both sizes of the previous equation and applying the result of differentiation of an integral, we obtain that

$$g'(\theta) = T\left(\theta+\tfrac{1}{2}\right) - T\left(\theta-\tfrac{1}{2}\right) = 0 \quad \text{a.e. } m.$$

Lecture 32: Information inequality

Suppose that we have a lower bound for the variances of all unbiased estimators of ϑ.
There is an unbiased estimator T of ϑ whose variance is always the same as the lower bound.
Then T is a UMVUE of ϑ.
Although this is not an effective way to find UMVUE's, it provides a way of assessing the performance of UMVUE's.

Theorem 3.3 (Cramér-Rao lower bound). Let $X = (X_1, ..., X_n)$ be a sample from $P \in \mathcal{P} = \{P_\theta : \theta \in \Theta\}$, where Θ is an open set in $\mathcal{R}^k$. Suppose that $T(X)$ is an estimator with $E[T(X)] = g(\theta)$ being a differentiable function of θ; P_θ has a p.d.f. f_θ w.r.t. a measure ν for all $\theta \in \Theta$; and f_θ is differentiable as a function of θ and satisfies

$$\frac{\partial}{\partial\theta}\int h(x)f_\theta(x)d\nu = \int h(x)\frac{\partial}{\partial\theta}f_\theta(x)d\nu, \qquad \theta \in \Theta, \tag{1}$$

for $h(x) \equiv 1$ and $h(x) = T(x)$. Then

$$\mathrm{Var}(T(X)) \geq \left[\tfrac{\partial}{\partial\theta}g(\theta)\right]^\tau [I(\theta)]^{-1}\tfrac{\partial}{\partial\theta}g(\theta), \tag{2}$$

where

$$I(\theta) = E\left\{\frac{\partial}{\partial\theta}\log f_\theta(X)\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^\tau\right\} \tag{3}$$

is assumed to be positive definite for any $\theta \in \Theta$.
Proof. We prove the univariate case ($k = 1$) only.
When $k = 1$, (2) reduces to

$$\mathrm{Var}(T(X)) \geq \frac{[g'(\theta)]^2}{E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2}. \tag{4}$$

From the Cauchy-Schwartz inequality, we only need to show that

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right]^2 = \mathrm{Var}\left(\frac{\partial}{\partial\theta}\log f_\theta(X)\right)$$

and

$$g'(\theta) = \mathrm{Cov}\left(T(X), \frac{\partial}{\partial\theta}\log f_\theta(X)\right).$$

From condition (1) with $h(x) = 1$,

$$E\left[\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int\frac{\partial}{\partial\theta}f_\theta(X)d\nu = \frac{\partial}{\partial\theta}\int f_\theta(X)d\nu = 0.$$

From condition (1) with $h(x) = T(x)$,

$$E\left[T(X)\frac{\partial}{\partial\theta}\log f_\theta(X)\right] = \int T(x)\frac{\partial}{\partial\theta}f_\theta(X)d\nu = \frac{\partial}{\partial\theta}\int T(x)f_\theta(X)d\nu = g'(\theta).$$

The $k \times k$ matrix $I(\theta)$ in (3) is called the *Fisher information matrix*.
The greater $I(\theta)$ is, the easier it is to distinguish θ from neighboring values and, therefore, the more accurately θ can be estimated. Thus, $I(\theta)$ is a measure of the information that X contains about the unknown θ.
The inequalities in (2) and (4) are called *information inequalities*.
The following result is helpful in finding the Fisher information matrix.

Proposition 3.1. (i) Let X and Y be independent with the Fisher information matrices $I_X(\theta)$ and $I_Y(\theta)$, respectively. Then, the Fisher information about θ contained in (X, Y) is $I_X(\theta) + I_Y(\theta)$. In particular, if $X_1, ..., X_n$ are i.i.d. and $I_1(\theta)$ is the Fisher information about θ contained in a single X_i, then the Fisher information about θ contained in $X_1, ..., X_n$ is $nI_1(\theta)$.
(ii) Suppose that X has the p.d.f. f_θ that is twice differentiable in θ and that (1) holds with $h(x) \equiv 1$ and f_θ replaced by $\partial f_\theta / \partial\theta$. Then

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X)\right]. \tag{5}$$

Proof. Result (i) follows from the independence of X and Y and the definition of the Fisher information. Result (ii) follows from the equality

$$\frac{\partial^2}{\partial\theta\partial\theta^\tau} \log f_\theta(X) = \frac{\frac{\partial^2}{\partial\theta\partial\theta^\tau} f_\theta(X)}{f_\theta(X)} - \frac{\partial}{\partial\theta} \log f_\theta(X) \left[\frac{\partial}{\partial\theta} \log f_\theta(X)\right]^\tau.$$

Example 3.9. Let $X_1, ..., X_n$ be i.i.d. with the Lebesgue p.d.f. $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, where $f(x) > 0$ and $f'(x)$ exists for all $x \in \mathcal{R}$, $\mu \in \mathcal{R}$, and $\sigma > 0$ (a location-scale family). Let $\theta = (\mu, \sigma)$. Then, the Fisher information about θ contained in $X_1, ..., X_n$ is (exercise)

$$I(\theta) = \frac{n}{\sigma^2} \begin{pmatrix} \int \frac{[f'(x)]^2}{f(x)} dx & \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx \\ \int \frac{f'(x)[xf'(x)+f(x)]}{f(x)} dx & \int \frac{[xf'(x)+f(x)]^2}{f(x)} dx \end{pmatrix}.$$

Note that $I(\theta)$ depends on the particular parameterization.
If $\theta = \psi(\eta)$ and ψ is differentiable, then the Fisher information that X contains about η is

$$\frac{\partial}{\partial\eta}\psi(\eta) I(\psi(\eta)) \left[\frac{\partial}{\partial\eta}\psi(\eta)\right]^\tau.$$

However, the Cramér-Rao lower bound in (2) or (4) is not affected by any one-to-one reparameterization.

If we use inequality (2) or (4) to find a UMVUE $T(X)$, then we obtain a formula for $\mathrm{Var}(T(X))$ at the same time.
On the other hand, the Cramér-Rao lower bound in (2) or (4) is typically not sharp.

Under some regularity conditions, the Cramér-Rao lower bound is attained if and only if f_θ is in an exponential family; see Propositions 3.2 and 3.3 and the discussion in Lehmann (1983, p. 123).
Some improved information inequalities are available (see, e.g., Lehmann (1983, Sections 2.6 and 2.7)).

Proposition 3.2. Suppose that the distribution of X is from an exponential family $\{f_\theta : \theta \in \Theta\}$, i.e., the p.d.f. of X w.r.t. a σ-finite measure is

$$f_\theta(x) = \exp\{[\eta(\theta)]^\tau T(x) - \xi(\theta)\}c(x), \tag{6}$$

where Θ is an open subset of $\mathcal{R}^k$.
(i) The regularity condition (1) is satisfied for any h with $E|h(X)| < \infty$ and (5) holds.
(ii) If $\underline{I}(\eta)$ is the Fisher information matrix for the natural parameter η, then the variance-covariance matrix $\mathrm{Var}(T) = \underline{I}(\eta)$.
(iii) If $\overline{I}(\vartheta)$ is the Fisher information matrix for the parameter $\vartheta = E[T(X)]$, then $\mathrm{Var}(T) = [\overline{I}(\vartheta)]^{-1}$.
Proof. (i) This is a direct consequence of Theorem 2.1.
(ii) The p.d.f. under the natural parameter η is

$$f_\eta(x) = \exp\{\eta^\tau T(x) - \zeta(\eta)\}\, c(x).$$

From Theorem 2.1, $E[T(X)] = \frac{\partial}{\partial \eta}\zeta(\eta)$. The result follows from

$$\frac{\partial}{\partial \eta} \log f_\eta(x) = T(x) - \frac{\partial}{\partial \eta}\zeta(\eta).$$

(iii) Since $\vartheta = E[T(X)] = \frac{\partial}{\partial \eta}\zeta(\eta)$,

$$\underline{I}(\eta) = \frac{\partial \vartheta}{\partial \eta}\overline{I}(\vartheta)\left(\frac{\partial \vartheta}{\partial \eta}\right)^\tau = \frac{\partial^2}{\partial \eta \partial \eta^\tau}\zeta(\eta)\overline{I}(\vartheta)\left[\frac{\partial^2}{\partial \eta \partial \eta^\tau}\zeta(\eta)\right]^\tau.$$

By Theorem 2.1 and the result in (ii), $\frac{\partial^2}{\partial \eta \partial \eta^\tau}\zeta(\eta) = \mathrm{Var}(T) = \underline{I}(\eta)$. Hence

$$\overline{I}(\vartheta) = [\underline{I}(\eta)]^{-1}\underline{I}(\eta)[\underline{I}(\eta)]^{-1} = [\underline{I}(\eta)]^{-1} = [\mathrm{Var}(T)]^{-1}.$$

A direct consequence of Proposition 3.2(ii) is that the variance of any linear function of T in (6) attains the Cramér-Rao lower bound.
The following result gives a necessary condition for $\mathrm{Var}(U(X))$ of an estimator $U(X)$ to attain the Cramér-Rao lower bound.

Proposition 3.3. Assume that the conditions in Theorem 3.3 hold with $T(X)$ replaced by $U(X)$ and that $\Theta \subset \mathcal{R}$.
(i) If $\mathrm{Var}(U(X))$ attains the Cramér-Rao lower bound in (4), then

$$a(\theta)[U(X) - g(\theta)] = g'(\theta)\frac{\partial}{\partial \theta}\log f_\theta(X) \quad \text{a.s. } P_\theta$$

for some function $a(\theta)$, $\theta \in \Theta$.
(ii) Let f_θ and T be given by (6). If $\mathrm{Var}(U(X))$ attains the Cramér-Rao lower bound, then $U(X)$ is a linear function of $T(X)$ a.s. P_θ, $\theta \in \Theta$.

Example 3.10. Let $X_1, ..., X_n$ be i.i.d. from the $N(\mu, \sigma^2)$ distribution with an unknown $\mu \in \mathcal{R}$ and a known σ^2.
Let f_μ be the joint distribution of $X = (X_1, ..., X_n)$. Then

$$\frac{\partial}{\partial \mu} \log f_\mu(X) = \sum_{i=1}^{n} (X_i - \mu)/\sigma^2.$$

Thus, $I(\mu) = n/\sigma^2$.
It is obvious that $\mathrm{Var}(\bar{X})$ attains the Cramér-Rao lower bound in (4).
Consider now the estimation of $\vartheta = \mu^2$.
Since $E\bar{X}^2 = \mu^2 + \sigma^2/n$, the UMVUE of ϑ is $h(\bar{X}) = \bar{X}^2 - \sigma^2/n$.
A straightforward calculation shows that

$$\mathrm{Var}(h(\bar{X})) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

On the other hand, the Cramér-Rao lower bound in this case is $4\mu^2\sigma^2/n$.
Hence $\mathrm{Var}(h(\bar{X}))$ does not attain the Cramér-Rao lower bound.
The difference is $2\sigma^4/n^2$.

Condition (1) is a key regularity condition for the results in Theorem 3.3 and Proposition 3.3.
If f_θ is not in an exponential family, then (1) has to be checked.
Typically, it does not hold if the set $\{x : f_\theta(x) > 0\}$ depends on θ (Exercise 37).
More discussions can be found in Pitman (1979).

Lecture 33: U-statistics and their variances

Let $X_1, ..., X_n$ be i.i.d. from an unknown population P in a nonparametric family $\mathcal{P}$.
If the vector of order statistic is sufficient and complete for $P \in \mathcal{P}$, then a symmetric unbiased estimator of any estimable ϑ is the UMVUE of ϑ.
In a large class of problems, parameters to be estimated are of the form

$$\vartheta = E[h(X_1, ..., X_m)]$$

with a positive integer m and a Borel function h that is symmetric and satisfies

$$E|h(X_1, ..., X_m)| < \infty$$

for any $P \in \mathcal{P}$.
It is easy to see that a symmetric unbiased estimator of ϑ is

$$U_n = \binom{n}{m}^{-1} \sum_c h(X_{i_1}, ..., X_{i_m}), \tag{1}$$

where $\sum_c$ denotes the summation over the $\binom{n}{m}$ combinations of m distinct elements $\{i_1, ..., i_m\}$ from $\{1, ..., n\}$.

Definition 3.2. The statistic U_n in (1) is called a *U-statistic* with kernel h of order m.

The use of U-statistics is an effective way of obtaining unbiased estimators.
In nonparametric problems, U-statistics are often UMVUE's, whereas in parametric problems, U-statistics can be used as initial estimators to derive more efficient estimators.

If $m = 1$, U_n in (1) is simply a type of sample mean.
Examples include the empirical c.d.f. evaluated at a particular t and the *sample moments* $n^{-1} \sum_{i=1}^n X_i^k$ for a positive integer k.

Consider the estimation of $\vartheta = \mu^m$, where $\mu = EX_1$ and m is a positive integer. Using $h(x_1, ..., x_m) = x_1 \cdots x_m$, we obtain the following U-statistic unbiased for $\vartheta = \mu^m$:

$$U_n = \binom{n}{m}^{-1} \sum_c X_{i_1} \cdots X_{i_m}. \tag{2}$$

Consider the estimation of $\vartheta = \sigma^2 = \mathrm{Var}(X_1)$. Since

$$\sigma^2 = [\mathrm{Var}(X_1) + \mathrm{Var}(X_2)]/2 = E[(X_1 - X_2)^2/2],$$

we obtain the following U-statistic with kernel $h(x_1, x_2) = (x_1 - x_2)^2/2$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \frac{(X_i - X_j)^2}{2} = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right) = S^2,$$

which is the sample variance.

In some cases, we would like to estimate $\vartheta = E|X_1 - X_2|$, a measure of concentration. Using kernel $h(x_1, x_2) = |x_1 - x_2|$, we obtain the following U-statistic unbiased for $\vartheta = E|X_1 - X_2|$:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|,$$

which is known as *Gini's mean difference*.

Let $\vartheta = P(X_1 + X_2 \le 0)$.
Using kernel $h(x_1, x_2) = I_{(-\infty,0]}(x_1 + x_2)$, we obtain the following U-statistic unbiased for ϑ:

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} I_{(-\infty,0]}(X_i + X_j),$$

which is known as the *one-sample Wilcoxon statistic*.

If $E[h(X_1, ..., X_m)]^2 < \infty$, then the variance of U_n in (1) with kernel h has an explicit form. To derive $\mathrm{Var}(U_n)$, we need some notation.
For $k = 1, ..., m$, let

$$\begin{aligned} h_k(x_1, ..., x_k) &= E[h(X_1, ..., X_m)|X_1 = x_1, ..., X_k = x_k] \\ &= E[h(x_1, ..., x_k, X_{k+1}, ..., X_m)]. \end{aligned}$$

Note that $h_m = h$.
It can be shown that

$$h_k(x_1, ..., x_k) = E[h_{k+1}(x_1, ..., x_k, X_{k+1})]. \tag{3}$$

Define

$$\tilde{h}_k = h_k - E[h(X_1, ..., X_m)], \tag{4}$$

$k = 1, ..., m$, and $\tilde{h} = \tilde{h}_m$.
Then, for any U_n defined by (1),

$$U_n - E(U_n) = \binom{n}{m}^{-1} \sum_c \tilde{h}(X_{i_1}, ..., X_{i_m}). \tag{5}$$

Theorem 3.4 (Hoeffding's theorem). For a U-statistic U_n given by (1) with $E[h(X_1, ..., X_m)]^2 < \infty$,

$$\mathrm{Var}(U_n) = \binom{n}{m}^{-1} \sum_{k=1}^{m} \binom{m}{k} \binom{n-m}{m-k} \zeta_k,$$

where

$$\zeta_k = \mathrm{Var}(h_k(X_1, ..., X_k)).$$

Proof. Consider two sets $\{i_1, ..., i_m\}$ and $\{j_1, ..., j_m\}$ of m distinct integers from $\{1, ..., n\}$ with exactly k integers in common.

The number of distinct choices of two such sets is $\binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}$.
By the symmetry of $\tilde{h}_m$ and independence of $X_1, ..., X_n$,

$$E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] = \zeta_k \tag{6}$$

for $k = 1, ..., m$.
Then, by (5),

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{m}^{-2} \sum_c \sum_c E[\tilde{h}(X_{i_1}, ..., X_{i_m})\tilde{h}(X_{j_1}, ..., X_{j_m})] \\ &= \binom{n}{m}^{-2} \sum_{k=1}^{m} \binom{n}{m}\binom{m}{k}\binom{n-m}{m-k}\zeta_k.\end{aligned}$$

This proves the result.

Corollary 3.2. Under the condition of Theorem 3.4,
(i) $\frac{m^2}{n}\zeta_1 \le \mathrm{Var}(U_n) \le \frac{m}{n}\zeta_m$;
(ii) $(n+1)\mathrm{Var}(U_{n+1}) \le n\mathrm{Var}(U_n)$ for any $n > m$;
(iii) For any fixed m and $k = 1, ..., m$, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then

$$\mathrm{Var}(U_n) = \frac{k!\binom{m}{k}^2\zeta_k}{n^k} + O\left(\frac{1}{n^{k+1}}\right).$$

It follows from Corollary 3.2 that a U-statistic U_n as an estimator of its mean is consistent in mse (under the finite second moment assumption on h).
In fact, for any fixed m, if $\zeta_j = 0$ for $j < k$ and $\zeta_k > 0$, then the mse of U_n is of the order n^{-k} and, therefore, U_n is $n^{k/2}$-consistent.

Example 3.11. Consider first $h(x_1, x_2) = x_1x_2$, which leads to a U-statistic unbiased for μ^2, $\mu = EX_1$.
Note that $h_1(x_1) = \mu x_1$, $\tilde{h}_1(x_1) = \mu(x_1 - \mu)$, $\zeta_1 = E[\tilde{h}_1(X_1)]^2 = \mu^2\mathrm{Var}(X_1) = \mu^2\sigma^2$, $\tilde{h}(x_1, x_2) = x_1x_2 - \mu^2$, and $\zeta_2 = \mathrm{Var}(X_1X_2) = E(X_1X_2)^2 - \mu^4 = (\mu^2 + \sigma^2)^2 - \mu^4$.
By Theorem 3.4, for $U_n = \binom{n}{2}^{-1}\sum_{1\le i<j\le n} X_iX_j$,

$$\begin{aligned}\mathrm{Var}(U_n) &= \binom{n}{2}^{-1}\left[\binom{2}{1}\binom{n-2}{1}\zeta_1 + \binom{2}{2}\binom{n-2}{0}\zeta_2\right] \\ &= \frac{2}{n(n-1)}\left[2(n-2)\mu^2\sigma^2 + (\mu^2+\sigma^2)^2 - \mu^4\right] \\ &= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n(n-1)}.\end{aligned}$$

Comparing U_n with $\bar{X}^2 - \sigma^2/n$ in Example 3.10, which is the UMVUE under the normality and known σ^2 assumption, we find that

$$\mathrm{Var}(U_n) - \mathrm{Var}(\bar{X}^2 - \sigma^2/n) = \frac{2\sigma^4}{n^2(n-1)}.$$

Next, consider $h(x_1, x_2) = I_{(-\infty,0]}(x_1+x_2)$, which leads to the one-sample Wilcoxon statistic. Note that $h_1(x_1) = P(x_1 + X_2 \leq 0) = F(-x_1)$, where F is the c.d.f. of P. Then $\zeta_1 = \mathrm{Var}(F(-X_1))$.
Let $\vartheta = E[h(X_1, X_2)]$.
Then $\zeta_2 = \mathrm{Var}(h(X_1, X_2)) = \vartheta(1 - \vartheta)$.
Hence, for U_n being the one-sample Wilcoxon statistic,

$$\mathrm{Var}(U_n) = \frac{2}{n(n-1)} \left[2(n-2)\zeta_1 + \vartheta(1-\vartheta)\right].$$

If F is continuous and symmetric about 0, then ζ_1 can be simplified as

$$\zeta_1 = \mathrm{Var}(F(-X_1)) = \mathrm{Var}(1 - F(X_1)) = \mathrm{Var}(F(X_1)) = \tfrac{1}{12},$$

since $F(X_1)$ has the uniform distribution on $[0, 1]$.

Finally, consider $h(x_1, x_2) = |x_1 - x_2|$, which leads to Gini's mean difference.
Note that

$$h_1(x_1) = E|x_1 - X_2| = \int |x_1 - y| dP(y),$$

and

$$\zeta_1 = \mathrm{Var}(h_1(X_1)) = \int \left[\int |x - y| dP(y)\right]^2 dP(x) - \vartheta^2,$$

where $\vartheta = E|X_1 - X_2|$.

Lecture 34: The projection method

Since $\mathcal{P}$ is nonparametric, the exact distribution of any U-statistic is hard to derive.
We study asymptotic distributions of U-statistics by using the method of *projection*.

Definition 3.3. Let T_n be a given statistic based on $X_1, ..., X_n$. The projection of T_n on k_n random elements $Y_1, ..., Y_{k_n}$ is defined to be

$$\check{T}_n = E(T_n) + \sum_{i=1}^{k_n} [E(T_n|Y_i) - E(T_n)].$$

Let $\psi_n(X_i) = E(T_n|X_i)$.
If T_n is symmetric (as a function of $X_1, ..., X_n$), then $\psi_n(X_1), ..., \psi_n(X_n)$ are i.i.d. with mean $E[\psi_n(X_i)] = E[E(T_n|X_i)] = E(T_n)$.
If $E(T_n^2) < \infty$ and $\text{Var}(\psi_n(X_i)) > 0$, then

$$\frac{1}{\sqrt{n\text{Var}(\psi_n(X_1))}} \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)] \to_d N(0,1) \tag{1}$$

by the CLT.
Let $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$.
Then

$$T_n - \check{T}_n = T_n - E(T_n) - \sum_{i=1}^{n} [\psi_n(X_i) - E(T_n)]. \tag{2}$$

If we can show that $T_n - \check{T}_n$ has a negligible order of magnitude, then we can derive the asymptotic distribution of T_n by using (1)-(2) and Slutsky's theorem.
The order of magnitude of $T_n - \check{T}_n$ can be obtained with the help of the following lemma.

Lemma 3.1. Let T_n be a symmetric statistic with $\text{Var}(T_n) < \infty$ for every n and $\check{T}_n$ be the projection of T_n on $X_1, ..., X_n$. Then $E(T_n) = E(\check{T}_n)$ and

$$E(T_n - \check{T}_n)^2 = \text{Var}(T_n) - \text{Var}(\check{T}_n).$$

Proof. Since $E(T_n) = E(\check{T}_n)$,

$$E(T_n - \check{T}_n)^2 = \text{Var}(T_n) + \text{Var}(\check{T}_n) - 2\text{Cov}(T_n, \check{T}_n).$$

From Definition 3.3 with $Y_i = X_i$ and $k_n = n$,

$$\text{Var}(\check{T}_n) = n\text{Var}(E(T_n|X_i)).$$

The result follows from

$$\begin{aligned}
\text{Cov}(T_n, \check{T}_n) &= E(T_n\check{T}_n) - [E(T_n)]^2 \\
&= nE[T_nE(T_n|X_i)] - n[E(T_n)]^2 \\
&= nE\{E[T_nE(T_n|X_i)|X_i]\} - n[E(T_n)]^2 \\
&= nE\{[E(T_n|X_i)]^2\} - n[E(T_n)]^2 \\
&= n\text{Var}(E(T_n|X_i)) \\
&= \text{Var}(\check{T}_n).
\end{aligned}$$

This method of deriving the asymptotic distribution of T_n is known as the method of projection and is particularly effective for U-statistics.
For a U-statistic U_n, one can show (exercise) that

$$\check{U}_n = E(U_n) + \frac{m}{n}\sum_{i=1}^{n}\tilde{h}_1(X_i), \tag{3}$$

where $\check{U}_n$ is the projection of U_n on $X_1, ..., X_n$ and $\tilde{h}_1(x) = h_1(x) - E[h(X_1, ..., X_m)]$, $h_1(x) = E[h(x, X_2, ..., X_m)]$.
Hence

$$\mathrm{Var}(\check{U}_n) = m^2\zeta_1/n$$

and, by Corollary 3.2 and Lemma 3.1,

$$E(U_n - \check{U}_n)^2 = O(n^{-2}).$$

If $\zeta_1 > 0$, then (1) holds with $\psi_n(X_i) = mh_1(X_i)$, which leads to the result in Theorem 3.5(i) stated later.

If $\zeta_1 = 0$, then $\tilde{h}_1 \equiv 0$ and we have to use another projection of U_n.
Suppose that $\zeta_1 = \cdots = \zeta_{k-1} = 0$ and $\zeta_k > 0$ for an integer $k > 1$.
Consider the projection $\check{U}_{kn}$ of U_n on $\binom{n}{k}$ random vectors $\{X_{i_1}, ..., X_{i_k}\}$, $1 \le i_1 < \cdots < i_k \le n$.
We can establish a result similar to that in Lemma 3.1 and show that

$$E(U_n - \check{U}_n)^2 = O(n^{-(k+1)}).$$

Also, see Serfling (1980, §5.3.4).

With these results, we obtain the following theorem.

Theorem 3.5. Let U_n be a U-statistic with $E[h(X_1, ..., X_m)]^2 < \infty$.
(i) If $\zeta_1 > 0$, then

$$\sqrt{n}[U_n - E(U_n)] \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 > 0$, then

$$n[U_n - E(U_n)] \to_d \frac{m(m-1)}{2}\sum_{j=1}^{\infty}\lambda_j(\chi^2_{1j} - 1), \tag{4}$$

where χ^2_{1j}'s are i.i.d. random variables having the chi-square distribution χ^2_1 and λ_j's are some constants (which may depend on P) satisfying $\sum_{j=1}^{\infty}\lambda_j^2 = \zeta_2$.

We have actually proved Theorem 3.5(i).
A proof for Theorem 3.5(ii) is given in Serfling (1980, §5.5.2).
One may derive results for the cases where $\zeta_2 = 0$, but the case of either $\zeta_1 > 0$ or $\zeta_2 > 0$ is the most interesting case in applications.

If $\zeta_1 > 0$, it follows from Theorem 3.5(i) and Corollary 3.2(iii) that

$$\text{amse}_{U_n}(P) = m^2\zeta_1/n = \text{Var}(U_n) + O(n^{-2}).$$

By Proposition 2.4(ii), $\{n[U_n - E(U_n)]^2\}$ is uniformly integrable.

If $\zeta_1 = 0$ but $\zeta_2 > 0$, it follows from Theorem 3.5(ii) that $\text{amse}_{U_n}(P) = EY^2/n^2$, where Y denotes the random variable on the right-hand side of (4).
The following result provides the value of EY^2.

Lemma 3.2. Let Y be the random variable on the right-hand side of (4). Then $EY^2 = \frac{m^2(m-1)^2}{2}\zeta_2$.
Proof. Define

$$Y_k = \frac{m(m-1)}{2}\sum_{j=1}^{k}\lambda_j(\chi^2_{1j} - 1), \quad k = 1, 2, ...$$

It can be shown (exercise) that $\{Y_k^2\}$ is uniformly integrable.
Since $Y_k \to_d Y$ as $k \to \infty$, $\lim_{k\to\infty} EY_k^2 = EY^2$ (Theorem 1.8(viii)).
Since χ^2_{1j}'s are independent chi-square random variables with $E\chi^2_{1j} = 1$ and $\text{Var}(\chi^2_{1j}) = 2$, $EY_k = 0$ for any k and

$$\begin{aligned} EY_k^2 &= \frac{m^2(m-1)^2}{4}\sum_{j=1}^{k}\lambda_j^2\text{Var}(\chi^2_{1j}) \\ &= \frac{m^2(m-1)^2}{4}\left(2\sum_{j=1}^{k}\lambda_j^2\right) \\ &\to \frac{m^2(m-1)^2}{2}\zeta_2. \end{aligned}$$

It follows from Corollary 3.2(iii) and Lemma 3.2 that

$$\text{amse}_{U_n}(P) = \frac{m^2(m-1)^2}{2}\zeta_2/n^2 = \text{Var}(U_n) + O(n^{-3})$$

if $\zeta_1 = 0$.
Again, by Proposition 2.4(ii), the sequence $\{n^2[U_n - E(U_n)]^2\}$ is uniformly integrable.

We now apply Theorem 3.5 to the U-statistics in Example 3.11.
For $U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} X_iX_j$, $\zeta_1 = \mu^2\sigma^2$.
Thus, if $\mu \neq 0$, the result in Theorem 3.5(i) holds with $\zeta_1 = \mu^2\sigma^2$.
If $\mu = 0$, then $\zeta_1 = 0$, $\zeta_2 = \sigma^4 > 0$, and Theorem 3.5(ii) applies.
However, it is not convenient to use Theorem 3.5(ii) to find the limiting distribution of U_n.
We may derive this limiting distribution using the following technique, which is further discussed in §3.5.
By the CLT and Theorem 1.10,

$$n\bar{X}^2/\sigma^2 \to_d \chi^2_1$$

when $\mu = 0$, where χ_1^2 is a random variable having the chi-square distribution χ_1^2.
Note that

$$\frac{n\bar{X}^2}{\sigma^2} = \frac{1}{\sigma^2 n}\sum_{i=1}^{n} X_i^2 + \frac{(n-1)U_n}{\sigma^2}.$$

By the SLLN, $\frac{1}{\sigma^2 n}\sum_{i=1}^{n} X_i^2 \to_{a.s.} 1$.
An application of Slutsky's theorem leads to

$$nU_n/\sigma^2 \to_d \chi_1^2 - 1.$$

Since $\mu = 0$, this implies that the right-hand side of (4) is $\sigma^2(\chi_1^2 - 1)$, i.e., $\lambda_1 = \sigma^2$ and $\lambda_j = 0$ when $j > 1$.

For the one-sample Wilcoxon statistic, $\zeta_1 = \mathrm{Var}(F(-X_1)) > 0$ unless F is degenerate.
Similarly, for Gini's mean difference, $\zeta_1 > 0$ unless F is degenerate.
Hence Theorem 3.5(i) applies to these two cases.

Lecture 35: The LSE and estimability

One of the most useful statistical models

$$X_i = \beta^\tau Z_i + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where X_i is the ith observation and is often called the ith response;
β is a p-vector of unknown parameters (main parameters of interest), $p < n$;
Z_i is the ith value of a p-vector of explanatory variables (or covariates);
$\varepsilon_1, ..., \varepsilon_n$ are random errors (not observed).
Data: $(X_1, Z_1), ..., (X_n, Z_n)$.
Z_i's are nonrandom or given values of a random p-vector, in which case our analysis is conditioned on $Z_1, ..., Z_n$.
$X = (X_1, ..., X_n)$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$
$Z =$ the $n \times p$ matrix whose ith row is the vector Z_i, $i = 1, ..., n$
A matrix form of model (1) is

$$X = Z\beta + \varepsilon. \tag{2}$$

Definition 3.4. Suppose that the range of β in model (2) is $B \subset \mathcal{R}^p$. A *least squares estimator* (LSE) of β is defined to be any $\hat{\beta} \in B$ such that

$$\|X - Z\hat{\beta}\|^2 = \min_{b \in B} \|X - Zb\|^2. \tag{3}$$

For any $l \in \mathcal{R}^p$, $l^\tau \hat{\beta}$ is called an LSE of $l^\tau \beta$.

Throughout this book, we consider $B = \mathcal{R}^p$ unless otherwise stated.
Differentiating $\|X - Zb\|^2$ w.r.t. b, we obtain that any solution of

$$Z^\tau Z b = Z^\tau X \tag{4}$$

is an LSE of β.
If the rank of the matrix Z is p, in which case $(Z^\tau Z)^{-1}$ exists and Z is said to be of full rank, then there is a unique LSE, which is

$$\hat{\beta} = (Z^\tau Z)^{-1} Z^\tau X. \tag{5}$$

If Z is not of full rank, then there are infinitely many LSE's of β.
Any LSE of β is of the form

$$\hat{\beta} = (Z^\tau Z)^- Z^\tau X, \tag{6}$$

where $(Z^\tau Z)^-$ is called a *generalized inverse* of $Z^\tau Z$ and satisfies

$$Z^\tau Z (Z^\tau Z)^- Z^\tau Z = Z^\tau Z.$$

Generalized inverse matrices are not unique unless Z is of full rank, in which case $(Z^\tau Z)^- = (Z^\tau Z)^{-1}$ and (6) reduces to (5).

To study properties of LSE's of β, we need some assumptions on the distribution of X or ε (conditional on Z if Z is random).

Assumption A1: ε is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$.

Assumption A2: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ with an unknown $\sigma^2 > 0$.

Assumption A3: $E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon)$ is an unknown matrix.

Assumption A1 is the strongest and implies a parametric model.
We may assume a slightly more general assumption that ε has the $N_n(0, \sigma^2 D)$ distribution with unknown σ^2 but a known positive definite matrix D.
Let $D^{-1/2}$ be the inverse of the square root matrix of D.
Then model (2) with assumption A1 holds if we replace X, Z, and ε by the transformed variables $\tilde{X} = D^{-1/2}X$, $\tilde{Z} = D^{-1/2}Z$, and $\tilde{\varepsilon} = D^{-1/2}\varepsilon$, respectively.
A similar conclusion can be made for assumption A2.

Under assumption A1, the distribution of X is $N_n(Z\beta, \sigma^2 I_n)$, which is in an exponential family $\mathcal{P}$ with parameter $\theta = (\beta, \sigma^2) \in \mathcal{R}^p \times (0, \infty)$.
However, if the matrix Z is not of full rank, then $\mathcal{P}$ is not identifiable (see §2.1.2), since $Z\beta_1 = Z\beta_2$ does not imply $\beta_1 = \beta_2$.

Suppose that the rank of Z is $r \le p$.
Then there is an $n \times r$ submatrix Z_* of Z such that

$$Z = Z_* Q \tag{7}$$

and Z_* is of rank r, where Q is a fixed $r \times p$ matrix, and

$$Z\beta = Z_* Q\beta.$$

$\mathcal{P}$ is identifiable if we consider the reparameterization $\tilde{\beta} = Q\beta$.
The new parameter $\tilde{\beta}$ is in a subspace of $\mathcal{R}^p$ with dimension r.

In many applications, we are interested in estimating some linear functions of β, i.e., $\vartheta = l^\tau \beta$ for some $l \in \mathcal{R}^p$.
From the previous discussion, however, estimation of $l^\tau\beta$ is meaningless unless $l = Q^\tau c$ for some $c \in \mathcal{R}^r$ so that

$$l^\tau \beta = c^\tau Q\beta = c^\tau \tilde{\beta}.$$

The following result shows that $l^\tau\beta$ is estimable if $l = Q^\tau c$, which is also necessary for $l^\tau\beta$ to be estimable under assumption A1.

Theorem 3.6. Assume model (2) with assumption A3.
(i) A necessary and sufficient condition for $l \in \mathcal{R}^p$ being $Q^\tau c$ for some $c \in \mathcal{R}^r$ is $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$, where Q is given by (7) and $\mathcal{R}(A)$ is the smallest linear subspace containing all rows of A.
(ii) If $l \in \mathcal{R}(Z)$, then the LSE $l^\tau\hat{\beta}$ is unique and unbiased for $l^\tau\beta$.

(iii) If $l \notin \mathcal{R}(Z)$ and assumption A1 holds, then $l^\tau\beta$ is not estimable.

Proof. (i) Note that $a \in \mathcal{R}(A)$ if and only if $a = A^\tau b$ for some vector b. If $l = Q^\tau c$, then

$$l = Q^\tau c = Q^\tau Z_*^\tau Z_*(Z_*^\tau Z_*)^{-1}c = Z^\tau[Z_*(Z_*^\tau Z_*)^{-1}c].$$

Hence $l \in \mathcal{R}(Z)$. If $l \in \mathcal{R}(Z)$, then $l = Z^\tau\zeta$ for some ζ and

$$l = (Z_*Q)^\tau\zeta = Q^\tau c$$

with $c = Z_*^\tau\zeta$.

(ii) If $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$, then $l = Z^\tau Z\zeta$ for some ζ and by (6),

$$\begin{aligned} E(l^\tau\hat{\beta}) &= E[l^\tau(Z^\tau Z)^- Z^\tau X] \\ &= \zeta^\tau Z^\tau Z(Z^\tau Z)^- Z^\tau Z\beta \\ &= \zeta^\tau Z^\tau Z\beta \\ &= l^\tau\beta. \end{aligned}$$

If $\bar{\beta}$ is any other LSE of β, then, by (4),

$$l^\tau\hat{\beta} - l^\tau\bar{\beta} = \zeta^\tau(Z^\tau Z)(\hat{\beta} - \bar{\beta}) = \zeta^\tau(Z^\tau X - Z^\tau X) = 0.$$

(iii) Under assumption A1, if there is an estimator $h(X, Z)$ unbiased for $l^\tau\beta$, then

$$l^\tau\beta = \int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\tfrac{1}{2\sigma^2}\|x - Z\beta\|^2\right\}dx.$$

Differentiating w.r.t. β and applying Theorem 2.1 lead to

$$l^\tau = Z^\tau\int_{\mathcal{R}^n} h(x, Z)(2\pi)^{-n/2}\sigma^{-n-2}(x - Z\beta)\exp\left\{-\tfrac{1}{2\sigma^2}\|x - Z\beta\|^2\right\}dx,$$

which implies $l \in \mathcal{R}(Z)$.

Example 3.12 (Simple linear regression). Let $\beta = (\beta_0, \beta_1) \in \mathcal{R}^2$ and $Z_i = (1, t_i)$, $t_i \in \mathcal{R}$, $i = 1, ..., n$.

Then model (1) or (2) is called a *simple linear regression* model.

It turns out that

$$Z^\tau Z = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

This matrix is invertible if and only if some t_i's are different.

Thus, if some t_i's are different, then the unique unbiased LSE of $l^\tau\beta$ for any $l \in \mathcal{R}^2$ is $l^\tau(Z^\tau Z)^{-1}Z^\tau X$, which has the normal distribution if assumption A1 holds.

The result can be easily extended to the case of *polynomial regression* of order p in which $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$ and $Z_i = (1, t_i, ..., t_i^{p-1})$.

Example 3.13 (One-way ANOVA). Suppose that $n = \sum_{j=1}^m n_j$ with m positive integers $n_1, ..., n_m$ and that

$$X_i = \mu_j + \varepsilon_i, \qquad i = k_{j-1} + 1, ..., k_j, \ j = 1, ..., m,$$

where $k_0 = 0$, $k_j = \sum_{l=1}^j n_l$, $j = 1, ..., m$, and $(\mu_1, ..., \mu_m) = \beta$.
Let J_m be the m-vector of ones.
Then the matrix Z in this case is a block diagonal matrix with J_{n_j} as the jth diagonal column.
Consequently, $Z^\tau Z$ is an $m \times m$ diagonal matrix whose jth diagonal element is n_j.
Thus, $Z^\tau Z$ is invertible and the unique LSE of β is the m-vector whose jth component is $n_j^{-1} \sum_{i=k_{j-1}+1}^{k_j} X_i$, $j = 1, ..., m$.

Sometimes it is more convenient to use the following notation:

$$X_{ij} = X_{k_{i-1}+j}, \ \varepsilon_{ij} = \varepsilon_{k_{i-1}+j}, \qquad j = 1, ..., n_i, \ i = 1, ..., m,$$

and

$$\mu_i = \mu + \alpha_i, \qquad i = 1, ..., m.$$

Then our model becomes

$$X_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad j = 1, ..., n_i, i = 1, ..., m, \tag{8}$$

which is called a *one-way analysis of variance* (ANOVA) model.
Under model (8), $\beta = (\mu, \alpha_1, ..., \alpha_m) \in \mathcal{R}^{m+1}$.
The matrix Z under model (8) is not of full rank.
An LSE of β under model (8) is

$$\hat{\beta} = \left(\bar{X}, \bar{X}_{1\cdot} - \bar{X}, ..., \bar{X}_{m\cdot} - \bar{X}\right),$$

where $\bar{X}$ is still the sample mean of X_{ij}'s and $\bar{X}_{i\cdot}$ is the sample mean of the ith group $\{X_{ij}, j = 1, ..., n_i\}$.

The notation used in model (8) allows us to generalize the one-way ANOVA model to any s-way ANOVA model with a positive integer s under the so-called factorial experiments.

Example 3.14 (Two-way balanced ANOVA). Suppose that

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad i = 1, ..., a, j = 1, ..., b, k = 1, ..., c, \tag{9}$$

where a, b, and c are some positive integers.
Model (9) is called a two-way balanced ANOVA model.
If we view model (9) as a special case of model (2), then the parameter vector β is

$$\beta = (\mu, \alpha_1, ..., \alpha_a, \beta_1, ..., \beta_b, \gamma_{11}, ..., \gamma_{1b}, ..., \gamma_{a1}, ..., \gamma_{ab}). \tag{10}$$

One can obtain the matrix Z and show that it is $n \times p$, where $n = abc$ and $p = 1 + a + b + ab$, and is of rank $ab < p$.

It can also be shown that an LSE of β is given by the right-hand side of (10) with μ, α_i, β_j, and γ_{ij} replaced by $\hat{\mu}$, $\hat{\alpha}_i$, $\hat{\beta}_j$, and $\hat{\gamma}_{ij}$, respectively, where $\hat{\mu} = \bar{X}_{\cdots}$, $\hat{\alpha}_i = \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdots}$, $\hat{\beta}_j = \bar{X}_{\cdot j\cdot} - \bar{X}_{\cdots}$, $\hat{\gamma}_{ij} = \bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdots}$, and a dot is used to denote averaging over the indicated subscript, e.g.,

$$\bar{X}_{\cdot j\cdot} = \frac{1}{ac} \sum_{i=1}^{a} \sum_{k=1}^{c} X_{ijk}$$

with a fixed j.

Lecture 36: The UMVUE and BLUE

Theorem 3.7. Consider model

$$X = Z\beta + \varepsilon \tag{1}$$

with assumption A1 (ε is distributed as $N_n(0, \sigma^2 I_n)$ with an unknown $\sigma^2 > 0$).
(i) The LSE $l^\tau\hat{\beta}$ is the UMVUE of $l^\tau\beta$ for any estimable $l^\tau\beta$.
(ii) The UMVUE of σ^2 is $\hat{\sigma}^2 = (n-r)^{-1}\|X - Z\hat{\beta}\|^2$, where r is the rank of Z.
Proof. (i) Let $\hat{\beta}$ be an LSE of β. By $Z^\tau Zb = Z^\tau X$,

$$(X - Z\hat{\beta})^\tau Z(\hat{\beta} - \beta) = (X^\tau Z - X^\tau Z)(\hat{\beta} - \beta) = 0$$

and, hence,

$$\begin{aligned}\|X - Z\beta\|^2 &= \|X - Z\hat{\beta} + Z\hat{\beta} - Z\beta\|^2 \\ &= \|X - Z\hat{\beta}\|^2 + \|Z\hat{\beta} - Z\beta\|^2 \\ &= \|X - Z\hat{\beta}\|^2 - 2\beta^\tau Z^\tau X + \|Z\beta\|^2 + \|Z\hat{\beta}\|^2.\end{aligned}$$

Using this result and assumption A1, we obtain the following joint Lebesgue p.d.f. of X:

$$(2\pi\sigma^2)^{-n/2}\exp\left\{\frac{\beta^\tau Z^\tau x}{\sigma^2} - \frac{\|x - Z\hat{\beta}\|^2 + \|Z\hat{\beta}\|^2}{2\sigma^2} - \frac{\|Z\beta\|^2}{2\sigma^2}\right\}.$$

By Proposition 2.1 and the fact that $Z\hat{\beta} = Z(Z^\tau Z)^- Z^\tau X$ is a function of $Z^\tau X$, the statistic $(Z^\tau X, \|X - Z\hat{\beta}\|^2)$ is complete and sufficient for $\theta = (\beta, \sigma^2)$.
Note that $\hat{\beta}$ is a function of $Z^\tau X$ and, hence, a function of the complete sufficient statistic. If $l^\tau\beta$ is estimable, then $l^\tau\hat{\beta}$ is unbiased for $l^\tau\beta$ (Theorem 3.6) and, hence, $l^\tau\hat{\beta}$ is the UMVUE of $l^\tau\beta$.
(ii) From $\|X - Z\beta\|^2 = \|X - Z\hat{\beta}\|^2 + \|Z\hat{\beta} - Z\beta\|^2$ and $E(Z\hat{\beta}) = Z\beta$ (Theorem 3.6),

$$\begin{aligned}E\|X - Z\hat{\beta}\|^2 &= E(X - Z\beta)^\tau(X - Z\beta) - E(\beta - \hat{\beta})^\tau Z^\tau Z(\beta - \hat{\beta}) \\ &= \text{tr}\left(\text{Var}(X) - \text{Var}(Z\hat{\beta})\right) \\ &= \sigma^2[n - \text{tr}\left(Z(Z^\tau Z)^- Z^\tau Z(Z^\tau Z)^- Z^\tau\right)] \\ &= \sigma^2[n - \text{tr}\left((Z^\tau Z)^- Z^\tau Z\right)].\end{aligned}$$

Since each row of $Z \in \mathcal{R}(Z)$, $Z\hat{\beta}$ does not depend on the choice of $(Z^\tau Z)^-$ in $\hat{\beta} = (Z^\tau Z)^- Z^\tau X$ (Theorem 3.6).
Hence, we can evaluate $\text{tr}((Z^\tau Z)^- Z^\tau Z)$ using a particular $(Z^\tau Z)^-$.
From the theory of linear algebra, there exists a $p \times p$ matrix C such that $CC^\tau = I_p$ and

$$C^\tau(Z^\tau Z)C = \begin{pmatrix} \Lambda & 0 \\ 0 & 0 \end{pmatrix},$$

where Λ is an $r \times r$ diagonal matrix whose diagonal elements are positive.
Then, a particular choice of $(Z^\tau Z)^-$ is

$$(Z^\tau Z)^- = C \begin{pmatrix} \Lambda^{-1} & 0 \\ 0 & 0 \end{pmatrix} C^\tau \tag{2}$$

and

$$(Z^\tau Z)^- Z^\tau Z = C \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} C^\tau$$

whose trace is r.
Hence $\hat{\sigma}^2$ is the UMVUE of σ^2, since it is a function of the complete sufficient statistic and

$$E\hat{\sigma}^2 = (n-r)^{-1} E\|X - Z\hat{\beta}\|^2 = \sigma^2.$$

In general,

$$\mathrm{Var}(l^\tau \hat{\beta}) = l^\tau (Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) Z (Z^\tau Z)^- l. \tag{3}$$

If $l \in \mathcal{R}(Z)$ and $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ (assumption A2), then the use of the generalized inverse matrix in (2) leads to $\mathrm{Var}(l^\tau \hat{\beta}) = \sigma^2 l^\tau (Z^\tau Z)^- l$, which attains the Cramér-Rao lower bound under assumption A1 (Proposition 3.2).

The vector $X - Z\hat{\beta}$ is called the *residual vector* and $\|X - Z\hat{\beta}\|^2$ is called the *sum of squared residuals* and is denoted by SSR.
The estimator $\hat{\sigma}^2$ is then equal to $SSR/(n-r)$.

Since $X - Z\hat{\beta} = [I_n - Z(Z^\tau Z)^- Z^\tau]X$ and $l^\tau \hat{\beta} = l^\tau (Z^\tau Z)^- Z^\tau X$ are linear in X, they are normally distributed under assumption A1.
Also, using the generalized inverse matrix in (2), we obtain that

$$[I_n - Z(Z^\tau Z)^- Z^\tau] Z (Z^\tau Z)^- = Z(Z^\tau Z)^- - Z(Z^\tau Z)^- Z^\tau Z (Z^\tau Z)^- = 0,$$

which implies that $\hat{\sigma}^2$ and $l^\tau \hat{\beta}$ are independent (Exercise 58 in §1.6) for any estimable $l^\tau \beta$.
Furthermore,

$$[Z(Z^\tau Z)^- Z^\tau]^2 = Z(Z^\tau Z)^- Z^\tau$$

(i.e., $Z(Z^\tau Z)^- Z^\tau$ is a projection matrix) and

$$SSR = X^\tau [I_n - Z(Z^\tau Z)^- Z^\tau] X.$$

The rank of $Z(Z^\tau Z)^- Z^\tau$ is $\mathrm{tr}(Z(Z^\tau Z)^- Z^\tau) = r$.
Similarly, the rank of the projection matrix $I_n - Z(Z^\tau Z)^- Z^\tau$ is $n - r$.
From

$$X^\tau X = X^\tau [Z(Z^\tau Z)^- Z^\tau] X + X^\tau [I_n - Z(Z^\tau Z)^- Z^\tau] X$$

and Theorem 1.5 (Cochran's theorem), SSR/σ^2 has the chi-square distribution $\chi^2_{n-r}(\delta)$ with

$$\delta = \sigma^{-2} \beta^\tau Z^\tau [I_n - Z(Z^\tau Z)^- Z^\tau] Z\beta = 0.$$

Thus, we have proved the following result.

Theorem 3.8. Consider model (1) with assumption A1. For any estimable parameter $l^\tau\beta$, the UMVUE's $l^\tau\hat{\beta}$ and $\hat{\sigma}^2$ are independent; the distribution of $l^\tau\hat{\beta}$ is $N(l^\tau\beta, \sigma^2 l^\tau(Z^\tau Z)^- l)$; and $(n-r)\hat{\sigma}^2/\sigma^2$ has the chi-square distribution χ^2_{n-r}.

Example 3.15. In Examples 3.12-3.14, UMVUE's of estimable $l^\tau\beta$ are the LSE's $l^\tau\hat{\beta}$, under assumption A1. In Example 3.13,

$$SSR = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2;$$

in Example 3.14, if $c > 1$,

$$SSR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}(X_{ijk} - \bar{X}_{ij\cdot})^2.$$

We now study properties of $l^\tau\hat{\beta}$ and $\hat{\sigma}^2$ under assumption A2, i.e., without the normality assumption on ε.
From Theorem 3.6 and the proof of Theorem 3.7(ii), $l^\tau\hat{\beta}$ (with an $l \in \mathcal{R}(Z)$) and $\hat{\sigma}^2$ are still unbiased without the normality assumption.
In what sense are $l^\tau\hat{\beta}$ and $\hat{\sigma}^2$ optimal beyond being unbiased?
We have the following result for the LSE $l^\tau\hat{\beta}$.
Some discussion about $\hat{\sigma}^2$ can be found, for example, in Rao (1973, p. 228).

Theorem 3.9. Consider model (1) with assumption A2.
(i) A necessary and sufficient condition for the existence of a linear unbiased estimator of $l^\tau\beta$ (i.e., an unbiased estimator that is linear in X) is $l \in \mathcal{R}(Z)$.
(ii) (Gauss-Markov theorem). If $l \in \mathcal{R}(Z)$, then the LSE $l^\tau\hat{\beta}$ is the *best linear unbiased estimator* (BLUE) of $l^\tau\beta$ in the sense that it has the minimum variance in the class of linear unbiased estimators of $l^\tau\beta$.
Proof. (i) The sufficiency has been established in Theorem 3.6.
Suppose now a linear function of X, $c^\tau X$ with $c \in \mathcal{R}^n$, is unbiased for $l^\tau\beta$. Then

$$l^\tau\beta = E(c^\tau X) = c^\tau EX = c^\tau Z\beta.$$

Since this equality holds for all β, $l = Z^\tau c$, i.e., $l \in \mathcal{R}(Z)$.
(ii) Let $l \in \mathcal{R}(Z) = \mathcal{R}(Z^\tau Z)$.
Then $l = (Z^\tau Z)\zeta$ for some ζ and $l^\tau\hat{\beta} = \zeta^\tau(Z^\tau Z)\hat{\beta} = \zeta^\tau Z^\tau X$ by $Z^\tau Zb = Z^\tau X$.
Let $c^\tau X$ be any linear unbiased estimator of $l^\tau\beta$. From the proof of (i), $Z^\tau c = l$. Then

$$\begin{aligned}
\mathrm{Cov}(\zeta^\tau Z^\tau X, c^\tau X - \zeta^\tau Z^\tau X) &= E(X^\tau Z\zeta c^\tau X) - E(X^\tau Z\zeta\zeta^\tau Z^\tau X) \\
&= \sigma^2\mathrm{tr}(Z\zeta c^\tau) + \beta^\tau Z^\tau Z\zeta c^\tau Z\beta \\
&\quad - \sigma^2\mathrm{tr}(Z\zeta\zeta^\tau Z^\tau) - \beta^\tau Z^\tau Z\zeta\zeta^\tau Z^\tau Z\beta \\
&= \sigma^2\zeta^\tau l + (l^\tau\beta)^2 - \sigma^2\zeta^\tau l - (l^\tau\beta)^2 \\
&= 0.
\end{aligned}$$

Hence

$$\begin{aligned}
\mathrm{Var}(c^\tau X) &= \mathrm{Var}(c^\tau X - \zeta^\tau Z^\tau X + \zeta^\tau Z^\tau X) \\
&= \mathrm{Var}(c^\tau X - \zeta^\tau Z^\tau X) + \mathrm{Var}(\zeta^\tau Z^\tau X) \\
&\quad + 2\mathrm{Cov}(\zeta^\tau Z^\tau X, c^\tau X - \zeta^\tau Z^\tau X) \\
&= \mathrm{Var}(c^\tau X - \zeta^\tau Z^\tau X) + \mathrm{Var}(l^\tau \hat{\beta}) \\
&\geq \mathrm{Var}(l^\tau \hat{\beta}).
\end{aligned}$$

Lecture 37: Robustness of LSE's

Consider model

$$X = Z\beta + \varepsilon. \tag{1}$$

under assumption A3 ($E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon)$ is an unknown matrix).
An interesting question is under what conditions on $\mathrm{Var}(\varepsilon)$ is the LSE of $l^\tau\beta$ with $l \in \mathcal{R}(Z)$ still the BLUE.
If $l^\tau\hat{\beta}$ is still the BLUE, then we say that $l^\tau\hat{\beta}$, considered as a BLUE, is *robust* against violation of assumption A2.
A statistical procedure having certain properties under an assumption is said to be robust against violation of the assumption if and only if the statistical procedure still has the same properties when the assumption is (slightly) violated.
For example, the LSE of $l^\tau\beta$ with $l \in \mathcal{R}(Z)$, as an unbiased estimator, is robust against violation of assumption A1 or A2, since the LSE is unbiased as long as $E(\varepsilon) = 0$, which can be always assumed without loss of generality.
On the other hand, the LSE as a UMVUE may not be robust against violation of assumption A1.

Theorem 3.10. Consider model (1) with assumption A3. The following are equivalent.
(a) $l^\tau\hat{\beta}$ is the BLUE of $l^\tau\beta$ for any $l \in \mathcal{R}(Z)$.
(b) $E(l^\tau\hat{\beta}\eta^\tau X) = 0$ for any $l \in \mathcal{R}(Z)$ and any η such that $E(\eta^\tau X) = 0$.
(c) $Z^\tau\mathrm{Var}(\varepsilon)U = 0$, where U is a matrix such that $Z^\tau U = 0$ and $\mathcal{R}(U^\tau) + \mathcal{R}(Z^\tau) = \mathcal{R}^n$.
(d) $\mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau + U\Lambda_2 U^\tau$ for some Λ_1 and Λ_2.
(e) The matrix $Z(Z^\tau Z)^- Z^\tau\mathrm{Var}(\varepsilon)$ is symmetric.
Proof. We first show that (a) and (b) are equivalent, which is an analogue of Theorem 3.2(i).
Suppose that (b) holds.
Let $l \in \mathcal{R}(Z)$.
If $c^\tau X$ is unbiased for $l^\tau\beta$, then $E(\eta^\tau X) = 0$ with $\eta = c - Z(Z^\tau Z)^- l$.
Hence

$$\begin{aligned}
\mathrm{Var}(c^\tau X) &= \mathrm{Var}(c^\tau X - l^\tau\hat{\beta} + l^\tau\hat{\beta}) \\
&= \mathrm{Var}(c^\tau X - l^\tau(Z^\tau Z)^- Z^\tau X + l^\tau\hat{\beta}) \\
&= \mathrm{Var}(\eta^\tau X + l^\tau\hat{\beta}) \\
&= \mathrm{Var}(\eta^\tau X) + \mathrm{Var}(l^\tau\hat{\beta}) + 2\mathrm{Cov}(\eta^\tau X, l^\tau\hat{\beta}) \\
&= \mathrm{Var}(\eta^\tau X) + \mathrm{Var}(l^\tau\hat{\beta}) + 2E(l^\tau\hat{\beta}\eta^\tau X) \\
&= \mathrm{Var}(\eta^\tau X) + \mathrm{Var}(l^\tau\hat{\beta}) \\
&\geq \mathrm{Var}(l^\tau\hat{\beta}).
\end{aligned}$$

Suppose now that there are $l \in \mathcal{R}(Z)$ and η such that $E(\eta^\tau X) = 0$ but $\delta = E(l^\tau\hat{\beta}\eta^\tau X) \neq 0$.
Let $c_t = t\eta + Z(Z^\tau Z)^- l$.
From the previous proof,

$$\mathrm{Var}(c_t^\tau X) = t^2\mathrm{Var}(\eta^\tau X) + \mathrm{Var}(l^\tau\hat{\beta}) + 2\delta t.$$

As long as $\delta \neq 0$, there exists a t such that $\mathrm{Var}(c_t^\tau X) < \mathrm{Var}(l^\tau \hat{\beta})$.
This shows that $l^\tau \hat{\beta}$ cannot be a BLUE and, therefore, (a) implies (b).

We next show that (b) implies (c).
Suppose that (b) holds.
Since $l \in \mathcal{R}(Z)$, $l = Z^\tau \gamma$ for some γ.
Let $\eta \in \mathcal{R}(U^\tau)$.
Then $E(\eta^\tau X) = \eta^\tau Z\beta = 0$ and, hence,

$$0 = E(l^\tau \hat{\beta} \eta^\tau X) = E[\gamma^\tau Z(Z^\tau Z)^- Z^\tau X X^\tau \eta] = \gamma^\tau Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)\eta.$$

Since this equality holds for all $l \in \mathcal{R}(Z)$, it holds for all γ.
Thus,

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) U = 0,$$

which implies

$$Z^\tau Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) U = Z^\tau \mathrm{Var}(\varepsilon) U = 0,$$

since $Z^\tau Z(Z^\tau Z)^- Z^\tau = Z^\tau$.
Thus, (c) holds.

To show that (c) implies (d), we need to use the following facts from the theory of linear algebra: there exists a nonsingular matrix C such that $\mathrm{Var}(\varepsilon) = CC^\tau$ and $C = ZC_1 + UC_2$ for some matrices C_j (since $\mathcal{R}(U^\tau) + \mathcal{R}(Z^\tau) = \mathcal{R}^n$).
Let $\Lambda_1 = C_1 C_1^\tau$, $\Lambda_2 = C_2 C_2^\tau$, and $\Lambda_3 = C_1 C_2^\tau$.
Then

$$\mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau + U\Lambda_2 U^\tau + Z\Lambda_3 U^\tau + U\Lambda_3^\tau Z^\tau \tag{2}$$

and $Z^\tau \mathrm{Var}(\varepsilon) U = Z^\tau Z \Lambda_3 U^\tau U$, which is 0 if (c) holds.
Hence, (c) implies

$$0 = Z(Z^\tau Z)^- Z^\tau Z \Lambda_3 U^\tau U (U^\tau U)^- U^\tau = Z\Lambda_3 U^\tau,$$

which with (2) implies (d).

If (d) holds, then $Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) = Z\Lambda_1 Z^\tau$, which is symmetric.
Hence (d) implies (e).

To complete the proof, we need to show that (e) implies (b), which is left as an exercise.

As a corollary of this theorem, the following result shows when the UMVUE's in model (1) with assumption A1 are robust against the violation of $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$.

Corollary 3.3. Consider model (1) with a full rank Z, $\varepsilon = N_n(0, \Sigma)$, and an unknown positive definite matrix Σ. Then $l^\tau \hat{\beta}$ is a UMVUE of $l^\tau \beta$ for any $l \in \mathcal{R}^p$ if and only if one of (b)-(e) in Theorem 3.10 holds.

Example 3.16. Consider model (1) with β replaced by a random vector $\boldsymbol{\beta}$ that is independent of ε.

Such a model is called a linear model with random coefficients.
Suppose that $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$ and $E(\boldsymbol{\beta}) = \beta$. Then

$$X = Z\beta + Z(\boldsymbol{\beta} - \beta) + \varepsilon = Z\beta + e, \tag{3}$$

where $e = Z(\boldsymbol{\beta} - \beta) + \varepsilon$ satisfies $E(e) = 0$ and

$$\mathrm{Var}(e) = Z\mathrm{Var}(\boldsymbol{\beta})Z^\tau + \sigma^2 I_n.$$

Since

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(e) = Z\mathrm{Var}(\boldsymbol{\beta})Z^\tau + \sigma^2 Z(Z^\tau Z)^- Z^\tau$$

is symmetric, by Theorem 3.10, the LSE $l^\tau\hat{\beta}$ under model (3) is the BLUE for any $l^\tau\beta$, $l \in \mathcal{R}(Z)$.
If Z is of full rank and ε is normal, then, by Corollary 3.3, $l^\tau\hat{\beta}$ is the UMVUE of $l^\tau\beta$ for any $l \in \mathcal{R}^p$.

Example 3.17 (Random effects models). Suppose that

$$X_{ij} = \mu + A_i + e_{ij}, \quad j = 1, ..., n_i, i = 1, ..., m, \tag{4}$$

where $\mu \in \mathcal{R}$ is an unknown parameter, A_i's are i.i.d. random variables having mean 0 and variance σ_a^2, e_{ij}'s are i.i.d. random errors with mean 0 and variance σ^2, and A_i's and e_{ij}'s are independent.
Model (4) is called a one-way *random effects* model and A_i's are unobserved random effects.
Let $\varepsilon_{ij} = A_i + e_{ij}$.
Then (4) is a special case of the general model (1) with

$$\mathrm{Var}(\varepsilon) = \sigma_a^2\Sigma + \sigma^2 I_n,$$

where Σ is a block diagonal matrix whose ith block is $J_{n_i}J_{n_i}^\tau$ and J_k is the k-vector of ones.
Under this model, $Z = J_n$, $n = \sum_{i=1}^m n_i$, and $Z(Z^\tau Z)^- Z^\tau = n^{-1}J_n J_n^\tau$.
Note that

$$J_n J_n^\tau \Sigma = \begin{pmatrix} n_1 J_{n_1}J_{n_1}^\tau & n_2 J_{n_1}J_{n_2}^\tau & \cdots & n_m J_{n_1}J_{n_m}^\tau \\ n_1 J_{n_2}J_{n_1}^\tau & n_2 J_{n_2}J_{n_2}^\tau & \cdots & n_m J_{n_2}J_{n_m}^\tau \\ \cdots\cdots & \cdots\cdots & \cdots & \cdots\cdots \\ n_1 J_{n_m}J_{n_1}^\tau & n_2 J_{n_m}J_{n_2}^\tau & \cdots & n_m J_{n_m}J_{n_m}^\tau \end{pmatrix},$$

which is symmetric if and only if $n_1 = n_2 = \cdots = n_m$.
Since $J_n J_n^\tau \mathrm{Var}(\varepsilon)$ is symmetric if and only if $J_n J_n^\tau \Sigma$ is symmetric, a necessary and sufficient condition for the LSE of μ to be the BLUE is that all n_i's are the same.
This condition is also necessary and sufficient for the LSE of μ to be the UMVUE when ε_{ij}'s are normal.

In some cases, we are interested in some (not all) linear functions of β.
For example, consider $l^\tau\beta$ with $l \in \mathcal{R}(H)$, where H is an $n \times p$ matrix such that $\mathcal{R}(H) \subset \mathcal{R}(Z)$.

Proposition 3.4. Consider model (1) with assumption A3. Suppose that H is a matrix such that $\mathcal{R}(H) \subset \mathcal{R}(Z)$. A necessary and sufficient condition for the LSE $l^\tau\hat{\beta}$ to be the BLUE of $l^\tau\beta$ for any $l \in \mathcal{R}(H)$ is $H(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)U = 0$, where U is the same as that in (c) of Theorem 3.10.

Example 3.18. Consider model (1) with assumption A3 and $Z = (H_1\ H_2)$, where $H_1^\tau H_2 = 0$.
Suppose that under the reduced model

$$X = H_1\beta_1 + \varepsilon,$$

$l^\tau\hat{\beta}_1$ is the BLUE for any $l^\tau\beta_1$, $l \in \mathcal{R}(H_1)$, and that under the reduced model

$$X = H_2\beta_2 + \varepsilon,$$

$l^\tau\hat{\beta}_2$ is not a BLUE for some $l^\tau\beta_2$, $l \in \mathcal{R}(H_2)$, where $\beta = (\beta_1, \beta_2)$ and $\hat{\beta}_j$'s are LSE's under the reduced models.
Let $H = (H_1\ 0)$ be $n \times p$.
Note that

$$H(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)U = H_1(H_1^\tau H_1)^- H_1^\tau \mathrm{Var}(\varepsilon)U,$$

which is 0 by Theorem 3.10 for the U given in (c) of Theorem 3.10, and

$$Z(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon)U = H_2(H_2^\tau H_2)^- H_2^\tau \mathrm{Var}(\varepsilon)U,$$

which is not 0 by Theorem 3.10.
This implies that some LSE $l^\tau\hat{\beta}$ is not a BLUE of $l^\tau\beta$ but $l^\tau\hat{\beta}$ is the BLUE of $l^\tau\beta$ if $l \in \mathcal{R}(H)$.

Finally, we consider model (1) with $\mathrm{Var}(\varepsilon)$ being a diagonal matrix whose ith diagonal element is σ_i^2, i.e., ε_i's are uncorrelated but have unequal variances.
A straightforward calculation shows that condition (e) in Theorem 3.10 holds if and only if, for all $i \neq j$, $\sigma_i^2 \neq \sigma_j^2$ only when $h_{ij} = 0$, where h_{ij} is the (i,j)th element of the projection matrix $Z(Z^\tau Z)^- Z^\tau$.
Thus, an LSE is not a BLUE in general, although it is still unbiased for estimable $l^\tau\beta$.

Suppose that the unequal variances of ε_i's are caused by some small perturbations, i.e., $\varepsilon_i = e_i + u_i$, where $\mathrm{Var}(e_i) = \sigma^2$, $\mathrm{Var}(u_i) = \delta_i$, and e_i and u_i are independent so that $\sigma_i^2 = \sigma^2 + \delta_i$.

$$\mathrm{Var}(l^\tau\hat{\beta}) = l^\tau(Z^\tau Z)^- \sum_{i=1}^n \sigma_i^2 Z_i Z_i^\tau (Z^\tau Z)^- l.$$

If $\delta_i = 0$ for all i (no perturbations), then assumption A2 holds and $l^\tau\hat{\beta}$ is the BLUE of any estimable $l^\tau\beta$ with $\mathrm{Var}(l^\tau\hat{\beta}) = \sigma^2 l^\tau(Z^\tau Z)^- l$.
Suppose that $0 < \delta_i \leq \sigma^2\delta$. Then

$$\mathrm{Var}(l^\tau\hat{\beta}) \leq (1+\delta)\sigma^2 l^\tau(Z^\tau Z)^- l.$$

This indicates that the LSE is robust in the sense that its variance increases slightly when there is a slight violation of the equal variance assumption (small δ).

Lecture 38: Asymptotic properties of LSE's

We consider first the consistency of the LSE $l^\tau\hat{\beta}$ with $l \in \mathcal{R}(Z)$ for every n.

Theorem 3.11. Consider model

$$X = Z\beta + \varepsilon \tag{1}$$

under assumption A3 ($E(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon)$ is an unknown matrix).
Suppose that $\sup_n \lambda_+[\mathrm{Var}(\varepsilon)] < \infty$, where $\lambda_+[A]$ is the largest eigenvalue of the matrix A, and that $\lim_{n\to\infty} \lambda_+[(Z^\tau Z)^-] = 0$. Then $l^\tau\hat{\beta}$ is consistent in mse for any $l \in \mathcal{R}(Z)$.
Proof. The result follows from the fact that $l^\tau\hat{\beta}$ is unbiased and

$$\begin{aligned}\mathrm{Var}(l^\tau\hat{\beta}) &= l^\tau(Z^\tau Z)^- Z^\tau \mathrm{Var}(\varepsilon) Z(Z^\tau Z)^- l \\ &\le \lambda_+[\mathrm{Var}(\varepsilon)] l^\tau(Z^\tau Z)^- l.\end{aligned}$$

Without the normality assumption on ε, the exact distribution of $l^\tau\hat{\beta}$ is very hard to obtain. The asymptotic distribution of $l^\tau\hat{\beta}$ is derived in the following result.

Theorem 3.12. Consider model (1) with assumption A3. Suppose that $0 < \inf_n \lambda_-[\mathrm{Var}(\varepsilon)]$, where $\lambda_-[A]$ is the smallest eigenvalue of the matrix A, and that

$$\lim_{n\to\infty} \max_{1\le i\le n} Z_i^\tau (Z^\tau Z)^- Z_i = 0. \tag{2}$$

Suppose further that $n = \sum_{j=1}^k m_j$ for some integers k, m_j, $j = 1, ..., k$, with m_j's bounded by a fixed integer m, $\varepsilon = (\xi_1, ..., \xi_k)$, $\xi_j \in \mathcal{R}^{m_j}$, and ξ_j's are independent.
(i) If $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$, then for any $l \in \mathcal{R}(Z)$,

$$l^\tau(\hat{\beta} - \beta) \Big/ \sqrt{\mathrm{Var}(l^\tau\hat{\beta})} \to_d N(0, 1). \tag{3}$$

(ii) Suppose that when $m_i = m_j$, $1 \le i < j \le k$, ξ_i and ξ_j have the same distribution. Then result (3) holds for any $l \in \mathcal{R}(Z)$.
Proof. Let $l \in \mathcal{R}(Z)$. Then

$$l^\tau(Z^\tau Z)^- Z^\tau Z\beta - l^\tau\beta = 0$$

and

$$l^\tau(\hat{\beta} - \beta) = l^\tau(Z^\tau Z)^- Z^\tau\varepsilon = \sum_{j=1}^k c_{nj}^\tau \xi_j,$$

where c_{nj} is the m_j-vector whose components are $l^\tau(Z^\tau Z)^- Z_i$, $i = k_{j-1} + 1, ..., k_j$, $k_0 = 0$, and $k_j = \sum_{t=1}^j m_t$, $j = 1, ..., k$.
Note that

$$\sum_{j=1}^k \|c_{nj}\|^2 = l^\tau(Z^\tau Z)^- Z^\tau Z(Z^\tau Z)^- l = l^\tau(Z^\tau Z)^- l. \tag{4}$$

Also,

$$\begin{aligned}\max_{1\le j\le k} \|c_{nj}\|^2 &\le m \max_{1\le i\le n} [l^\tau(Z^\tau Z)^- Z_i]^2 \\ &\le m l^\tau(Z^\tau Z)^- l \max_{1\le i\le n} Z_i^\tau(Z^\tau Z)^- Z_i,\end{aligned}$$

which, together with (4) and condition (2), implies that

$$\lim_{n\to\infty}\left(\max_{1\le j\le k}\|c_{nj}\|^2\Big/\sum_{j=1}^{k}\|c_{nj}\|^2\right)=0.$$

The results then follow from Corollary 1.3.

Under the conditions of Theorem 3.12, $\mathrm{Var}(\varepsilon)$ is a diagonal block matrix with $\mathrm{Var}(\xi_j)$ as the jth diagonal block, which includes the case of independent ε_i's as a special case.

Exercise 80 shows that condition (2) is almost a necessary condition for the consistency of the LSE.

The following lemma tells us how to check condition (2).

Lemma 3.3. The following are sufficient conditions for (2).
(a) $\lambda_+[(Z^\tau Z)^-]\to 0$ and $Z_n^\tau(Z^\tau Z)^- Z_n\to 0$, as $n\to\infty$.
(b) There is an increasing sequence $\{a_n\}$ such that $a_n\to\infty$, $a_n/a_{n+1}\to 1$, and $Z^\tau Z/a_n$ converges to a positive definite matrix.
Proof. (a) Since $Z^\tau Z$ depends on n, we denote $(Z^\tau Z)^-$ by A_n.
Let i_n be the integer such that $h_{i_n}=\max_{1\le i\le n}h_i$.
If $\lim_{n\to\infty} i_n=\infty$, then

$$\lim_{n\to\infty}h_{i_n}=\lim_{n\to\infty}Z_{i_n}^\tau A_n Z_{i_n}\le\lim_{n\to\infty}Z_{i_n}^\tau A_{i_n}Z_{i_n}=0,$$

where the inequality follows from $i_n\le n$ and, thus, $A_{i_n}-A_n$ is nonnegative definite.
If $i_n\le c$ for all n, then

$$\lim_{n\to\infty}h_{i_n}=\lim_{n\to\infty}Z_{i_n}^\tau A_n Z_{i_n}\le\lim_{n\to\infty}\lambda_n\max_{1\le i\le c}\|Z_i\|^2=0.$$

Therefore, for any subsequence $\{j_n\}\subset\{i_n\}$ with $\lim_{n\to\infty}j_n=a\in(0,\infty]$, $\lim_{n\to\infty}h_{j_n}=0$.
This shows that $\lim_{n\to\infty}h_{i_n}=0$.
(b) Omitted.

If $n^{-1}\sum_{i=1}^n t_i^2\to c$ and $n^{-1}\sum_{i=1}^n t_i\to d$ in the simple linear regression model (Example 3.12), where c is positive and $c>d^2$, then condition (b) in Lemma 3.3 is satisfied with $a_n=n$ and, therefore, Theorem 3.12 applies.

In the one-way ANOVA model (Example 3.13),

$$\max_{1\le i\le n}Z_i^\tau(Z^\tau Z)^- Z_i=\lambda_+[(Z^\tau Z)^-]=\max_{1\le j\le m}n_j^{-1}.$$

Hence conditions related to Z in Theorem 3.12 are satisfied if and only if $\min_j n_j\to\infty$.
Some similar conclusions can be drawn in the two-way ANOVA model (Example 3.14).

Functions of unbiased estimators

If the parameter to be estimated is $\vartheta = g(\theta)$ with a vector-valued parameter θ and U_n is a vector of unbiased estimators of components of θ, then $T_n = g(U_n)$ is often asymptotically unbiased for ϑ.
Assume that g is differentiable and $c_n(U_n - \theta) \to_d Y$. Then

$$\text{amse}_{T_n}(P) = E\{[\nabla g(\theta)]^\tau Y\}^2/c_n^2$$

(Theorem 2.6). Hence, T_n has a good performance in terms of amse if U_n is optimal in terms of mse (such as the UMVUE or BLUE).

Example 3.22. Consider a polynomial regression of order p:

$$X_i = \beta^\tau Z_i + \varepsilon_i, \qquad i = 1, ..., n,$$

where $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})$, $Z_i = (1, t_i, ..., t_i^{p-1})$, and ε_i's are i.i.d. with mean 0 and variance $\sigma^2 > 0$.
Suppose that the parameter to be estimated is $t_\beta \in \mathcal{T} \subset \mathcal{R}$ such that

$$\sum_{j=0}^{p-1} \beta_j t_\beta^j = \max_{t \in \mathcal{T}} \sum_{j=0}^{p-1} \beta_j t^j.$$

Note that $t_\beta = g(\beta)$ for some function g.
Let $\hat{\beta}$ be the LSE of β.
Then the estimator $\hat{t}_\beta = g(\hat{\beta})$ is asymptotically unbiased and its amse can be derived under some conditions.

Example 3.23. In the study of the reliability of a system component, we assume that

$$X_{ij} = \boldsymbol{\theta}_i^\tau z(t_j) + \varepsilon_{ij}, \quad i = 1, ..., k, \ j = 1, ..., m.$$

Here X_{ij} is the measurement of the ith sample component at time t_j;
$z(t)$ is a q-vector whose components are known functions of the time t;
$\boldsymbol{\theta}_i$'s are unobservable random q-vectors that are i.i.d. from $N_q(\theta, \Sigma)$, where θ and Σ are unknown;
ε_{ij}'s are i.i.d. measurement errors with mean zero and variance σ^2;
$\boldsymbol{\theta}_i$'s and ε_{ij}'s are independent.
As a function of t, $\boldsymbol{\theta}^\tau z(t)$ is the degradation curve for a particular component and $\theta^\tau z(t)$ is the mean degradation curve.
Suppose that a component will fail to work if $\boldsymbol{\theta}^\tau z(t) < \eta$, a given critical value.
Assume that $\boldsymbol{\theta}^\tau z(t)$ is always a decreasing function of t.
Then the reliability function of a component is

$$R(t) = P(\boldsymbol{\theta}^\tau z(t) > \eta) = \Phi\left(\frac{\theta^\tau z(t) - \eta}{s(t)}\right),$$

where $s(t) = \sqrt{[z(t)]^\tau \Sigma z(t)}$ and Φ is the standard normal distribution function.
For a fixed t, estimators of $R(t)$ can be obtained by estimating θ and Σ, since Φ is a known function.
It can be shown (exercise) that the BLUE of θ is the LSE

$$\hat{\theta} = (Z^\tau Z)^{-1} Z^\tau \bar{X},$$

where Z is the $m \times q$ matrix whose jth row is the vector $z(t_j)$, $X_i = (X_{i1}, ..., X_{im})$, and $\bar{X}$ is the sample mean of X_i's.
The estimation of Σ is more difficult.
It can be shown (exercise) that a consistent (as $k \to \infty$) estimator of Σ is

$$\hat{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} (Z^\tau Z)^{-1} Z^\tau (X_i - \bar{X})(X_i - \bar{X})^\tau Z (Z^\tau Z)^{-1} - \hat{\sigma}^2 (Z^\tau Z)^{-1},$$

where

$$\hat{\sigma}^2 = \frac{1}{k(m-q)} \sum_{i=1}^{k} [X_i^\tau X_i - X_i^\tau Z (Z^\tau Z)^{-1} Z^\tau X_i].$$

Hence an estimator of $R(t)$ is

$$\hat{R}(t) = \Phi\left(\frac{\hat{\theta}^\tau z(t) - \eta}{\hat{s}(t)} \right),$$

where

$$\hat{s}(t) = \sqrt{[z(t)]^\tau \hat{\Sigma} z(t)}.$$

$Y_{i1} = X_i^\tau Z (Z^\tau Z)^{-1} z(t)$
$Y_{i2} = [X_i^\tau Z (Z^\tau Z)^{-1} z(t)]^2$
$Y_{i3} = [X_i^\tau X_i - X_i^\tau Z (Z^\tau Z)^{-1} Z^\tau X_i]/(m-q)$
$Y_i = (Y_{i1}, Y_{i2}, Y_{i3})$ It is apparent that $\hat{R}(t)$ can be written as $g(\bar{Y})$ for a function

$$g(y_1, y_2, y_3) = \Phi\left(\frac{y_1 - \eta}{\sqrt{y_2 - y_1^2 - y_3 [z(t)]^\tau (Z^\tau Z)^{-1} z(t)}} \right).$$

Suppose that ε_{ij} has a finite fourth moment, which implies the existence of $\mathrm{Var}(Y_i)$.
The amse of $\hat{R}(t)$ can be derived (exercise).

Lecture 39: The method of moments

The method of moments is the oldest method of deriving point estimators.
It almost always produces some asymptotically unbiased estimators, although they may not be the best estimators.

Consider a parametric problem where $X_1, ..., X_n$ are i.i.d. random variables from P_θ, $\theta \in \Theta \subset \mathcal{R}^k$, and $E|X_1|^k < \infty$.
Let $\mu_j = EX_1^j$ be the jth moment of P and let

$$\hat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

be the jth *sample moment*, which is an unbiased estimator of μ_j, $j = 1, ..., k$.
Typically,

$$\mu_j = h_j(\theta), \qquad j = 1, ..., k, \tag{1}$$

for some functions h_j on $\mathcal{R}^k$.
By substituting μ_j's on the left-hand side of (1) by the sample moments $\hat{\mu}_j$, we obtain a *moment estimator* $\hat{\theta}$, i.e., $\hat{\theta}$ satisfies

$$\hat{\mu}_j = h_j(\hat{\theta}), \qquad j = 1, ..., k,$$

which is a sample analogue of (1).
This method of deriving estimators is called the *method of moments*.
An important statistical principle, the *substitution principle*, is applied in this method.

Let $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_k)$ and $h = (h_1, ..., h_k)$.
Then $\hat{\mu} = h(\hat{\theta})$.
If the inverse function h^{-1} exists, then the unique moment estimator of θ is $\hat{\theta} = h^{-1}(\hat{\mu})$.
When h^{-1} does not exist (i.e., h is not one-to-one), any solution of $\hat{\mu} = h(\hat{\theta})$ is a moment estimator of θ;
if possible, we always choose a solution $\hat{\theta}$ in the parameter space Θ.
In some cases, however, a moment estimator does not exist (see Exercise 111).

Assume that $\hat{\theta} = g(\hat{\mu})$ for a function g.
If h^{-1} exists, then $g = h^{-1}$.
If g is continuous at $\mu = (\mu_1, ..., \mu_k)$, then $\hat{\theta}$ is strongly consistent for θ, since $\hat{\mu}_j \to_{a.s.} \mu_j$ by the SLLN.
If g is differentiable at μ and $E|X_1|^{2k} < \infty$, then $\hat{\theta}$ is asymptotically normal, by the CLT and Theorem 1.12, and

$$\text{amse}_{\hat{\theta}}(\theta) = n^{-1}[\nabla g(\mu)]^\tau V_\mu \nabla g(\mu),$$

where V_μ is a $k \times k$ matrix whose (i, j)th element is $\mu_{i+j} - \mu_i\mu_j$.
Furthermore, the n^{-1} order asymptotic bias of $\hat{\theta}$ is

$$(2n)^{-1}\text{tr}\left(\nabla^2 g(\mu) V_\mu\right).$$

Example 3.24. Let $X_1, ..., X_n$ be i.i.d. from a population P_θ indexed by the parameter $\theta = (\mu, \sigma^2)$, where $\mu = EX_1 \in \mathcal{R}$ and $\sigma^2 = \text{Var}(X_1) \in (0, \infty)$.
This includes cases such as the family of normal distributions, double exponential distributions, or logistic distributions (Table 1.2, page 20).
Since $EX_1 = \mu$ and $EX_1^2 = \text{Var}(X_1) + (EX_1)^2 = \sigma^2 + \mu^2$, setting $\hat{\mu}_1 = \mu$ and $\hat{\mu}_2 = \sigma^2 + \mu^2$ we obtain the moment estimator

$$\hat{\theta} = \left(\bar{X},\ \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right) = \left(\bar{X},\ \frac{n-1}{n}S^2\right).$$

Note that $\bar{X}$ is unbiased, but $\frac{n-1}{n}S^2$ is not.
If X_i is normal, then $\hat{\theta}$ is sufficient and is nearly the same as an optimal estimator such as the UMVUE.
On the other hand, if X_i is from a double exponential or logistic distribution, then $\hat{\theta}$ is not sufficient and can often be improved.

Consider now the estimation of σ^2 when we know that $\mu = 0$.
Obviously we cannot use the equation $\hat{\mu}_1 = \mu$ to solve the problem.
Using $\hat{\mu}_2 = \mu_2 = \sigma^2$, we obtain the moment estimator $\hat{\sigma}^2 = \hat{\mu}_2 = n^{-1}\sum_{i=1}^{n} X_i^2$.
This is still a good estimator when X_i is normal, but is not a function of sufficient statistic when X_i is from a double exponential distribution.
For the double exponential case one can argue that we should first make a transformation $Y_i = |X_i|$ and then obtain the moment estimator based on the transformed data.
The moment estimator of σ^2 based on the transformed data is $\bar{Y}^2 = (n^{-1}\sum_{i=1}^{n}|X_i|)^2$, which is sufficient for σ^2.
Note that this estimator can also be obtained based on absolute moment equations.

Example 3.25. Let $X_1, ..., X_n$ be i.i.d. from the uniform distribution on (θ_1, θ_2), $-\infty < \theta_1 < \theta_2 < \infty$.
Note that

$$EX_1 = (\theta_1 + \theta_2)/2$$

and

$$EX_1^2 = (\theta_1^2 + \theta_2^2 + \theta_1\theta_2)/3.$$

Setting $\hat{\mu}_1 = EX_1$ and $\hat{\mu}_2 = EX_1^2$ and substituting θ_1 in the second equation by $2\hat{\mu}_1 - \theta_2$ (the first equation), we obtain that

$$(2\hat{\mu}_1 - \theta_2)^2 + \theta_2^2 + (2\hat{\mu}_1 - \theta_2)\theta_2 = 3\hat{\mu}_2,$$

which is the same as

$$(\theta_2 - \hat{\mu}_1)^2 = 3(\hat{\mu}_2 - \hat{\mu}_1^2).$$

Since $\theta_2 > EX_1$, we obtain that

$$\hat{\theta}_2 = \hat{\mu}_1 + \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} + \sqrt{\frac{3(n-1)}{n}S^2}$$

and

$$\hat{\theta}_1 = \hat{\mu}_1 - \sqrt{3(\hat{\mu}_2 - \hat{\mu}_1^2)} = \bar{X} - \sqrt{\frac{3(n-1)}{n} S^2}.$$

These estimators are not functions of the sufficient and complete statistic $(X_{(1)}, X_{(n)})$.

Example 3.26. Let $X_1, ..., X_n$ be i.i.d. from the binomial distribution $Bi(p, k)$ with unknown parameters $k \in \{1, 2, ...\}$ and $p \in (0, 1)$.
Since

$$EX_1 = kp$$

and

$$EX_1^2 = kp(1-p) + k^2 p^2,$$

we obtain the moment estimators

$$\hat{p} = (\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2)/\hat{\mu}_1 = 1 - \frac{n-1}{n} S^2/\bar{X}$$

and

$$\hat{k} = \hat{\mu}_1^2/(\hat{\mu}_1 + \hat{\mu}_1^2 - \hat{\mu}_2) = \bar{X}/(1 - \frac{n-1}{n} S^2/\bar{X}).$$

The estimator $\hat{p}$ is in the range of $(0, 1)$.
But $\hat{k}$ may not be an integer.
It can be improved by an estimator that is $\hat{k}$ rounded to the nearest positive integer.

Example 3.27. Suppose that $X_1, ..., X_n$ are i.i.d. from the Pareto distribution $Pa(a, \theta)$ with unknown $a > 0$ and $\theta > 2$ (Table 1.2, page 20).
Note that

$$EX_1 = \theta a/(\theta - 1)$$

and

$$EX_1^2 = \theta a^2/(\theta - 2).$$

From the moment equation,

$$\frac{(\theta-1)^2}{\theta(\theta-2)} = \hat{\mu}_2/\hat{\mu}_1^2.$$

Note that $\frac{(\theta-1)^2}{\theta(\theta-2)} - 1 = \frac{1}{\theta(\theta-2)}$.
Hence

$$\theta(\theta - 2) = \hat{\mu}_1^2/(\hat{\mu}_2 - \hat{\mu}_1^2).$$

Since $\theta > 2$, there is a unique solution in the parameter space:

$$\hat{\theta} = 1 + \sqrt{\hat{\mu}_2/(\hat{\mu}_2 - \hat{\mu}_1^2)} = 1 + \sqrt{1 + \frac{n}{n-1}\bar{X}^2/S^2}$$

and

$$\begin{aligned}\hat{a} &= \frac{\hat{\mu}_1(\hat{\theta} - 1)}{\hat{\theta}} \\ &= \bar{X}\sqrt{1 + \frac{n}{n-1}\bar{X}^2/S^2} \Big/ \left(1 + \sqrt{1 + \frac{n}{n-1}\bar{X}^2/S^2}\right).\end{aligned}$$

Exercise 108. Let $X_1, ..., X_n$ be a random sample from the following discrete distribution:

$$P(X_1 = 1) = \frac{2(1-\theta)}{2-\theta}, \quad P(X_1 = 2) = \frac{\theta}{2-\theta},$$

where $\theta \in (0, 1)$ is unknown.
Note that

$$EX_1 = \frac{2(1-\theta)}{2-\theta} + \frac{2\theta}{2-\theta} = \frac{2}{2-\theta}.$$

Hence, a moment estimator of θ is $\hat{\theta} = 2(1 - \bar{X}^{-1})$, where $\bar{X}$ is the sample mean.
Note that

$$\mathrm{Var}(X_1) = \frac{2(1-\theta)}{2-\theta} + \frac{4\theta}{2-\theta} - \frac{4}{(2-\theta)^2} = \frac{4\theta - 2\theta^2 - 4}{(2-\theta)^2},$$

$$\theta = 2(1 - \mu^{-1}) = g(\mu),$$

$$g'(\mu) = 2/\mu^2 = 2/[2/(2-\theta)]^2 = (2-\theta)^2/2.$$

By the central limit theorem and δ-method,

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N\left(0, \frac{(2-\theta)^2(2\theta - \theta^2 - 2)}{2}\right).$$

The method of moments can also be applied to nonparametric problems.
Consider, for example, the estimation of the central moments

$$c_j = E(X_1 - \mu_1)^j, \qquad j = 2, ..., k.$$

Since

$$c_j = \sum_{t=0}^{j} \binom{j}{t} (-\mu_1)^t \mu_{j-t},$$

the moment estimator of c_j is

$$\hat{c}_j = \sum_{t=0}^{j} \binom{j}{t} (-\bar{X})^t \hat{\mu}_{j-t},$$

where $\hat{\mu}_0 = 1$.
It can be shown (exercise) that

$$\hat{c}_j = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^j, \qquad j = 2, ..., k, \tag{2}$$

which are sample central moments.
From the SLLN, $\hat{c}_j$'s are strongly consistent.
If $E|X_1|^{2k} < \infty$, then

$$\sqrt{n}\,(\hat{c}_2 - c_2, ..., \hat{c}_k - c_k) \to_d N_{k-1}(0, D) \tag{3}$$

where the (i, j)th element of the $(k-1) \times (k-1)$ matrix D is

$$c_{i+j+2} - c_{i+1}c_{j+1} - (i+1)c_i c_{j+2} - (j+1)c_{i+2}c_j + (i+1)(j+1)c_i c_j c_2.$$

Lecture 40: V-statistics and the weighted LSE

Let $X_1, ..., X_n$ be i.i.d. from P.
For every U-statistic U_n as an estimator of $\vartheta = E[h(X_1, ..., X_m)]$, there is a closely related *V-statistic* defined by

$$V_n = \frac{1}{n^m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} h(X_{i_1}, ..., X_{i_m}). \tag{1}$$

As an estimator of ϑ, V_n is biased; but the bias is small asymptotically as the following results show.
For a fixed sample size n, V_n may be better than U_n in terms of their mse's.

Proposition 3.5. Let V_n be defined by (1).
(i) Assume that $E|h(X_{i_1}, ..., X_{i_m})| < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.
Then the bias of V_n satisfies

$$b_{V_n}(P) = O(n^{-1}).$$

(ii) Assume that $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$. Then the variance of V_n satisfies

$$\mathrm{Var}(V_n) = \mathrm{Var}(U_n) + O(n^{-2}),$$

where U_n is the U-statistic corresponding to V_n.

To study the asymptotic behavior of a V-statistic, we consider the following representation of V_n in (1):

$$V_n = \sum_{j=1}^{m} \binom{m}{j} V_{nj},$$

where

$$V_{nj} = \vartheta + \frac{1}{n^j} \sum_{i_1=1}^{n} \cdots \sum_{i_j=1}^{n} g_j(X_{i_1}, ..., X_{i_j})$$

is a "V-statistic" with

$$\begin{aligned} g_j(x_1, ..., x_j) = & h_j(x_1, ..., x_j) - \sum_{i=1}^{j} \int h_j(x_1, ..., x_j) dP(x_i) \\ & + \sum_{1 \le i_1 < i_2 \le j} \int\int h_j(x_1, ..., x_j) dP(x_{i_1}) dP(x_{i_2}) - \cdots \\ & + (-1)^j \int \cdots \int h_j(x_1, ..., x_j) dP(x_1) \cdots dP(x_j) \end{aligned}$$

and $h_j(x_1, ..., x_j) = E[h(x_1, ..., x_j, X_{j+1}, ..., X_m)]$.
Using an argument similar to the proof of Theorem 3.4, we can show that

$$EV_{nj}^2 = O(n^{-j}), \qquad j = 1, ..., m, \tag{2}$$

provided that $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.
Thus,

$$V_n - \vartheta = mV_{n1} + \tfrac{m(m-1)}{2} V_{n2} + o_p(n^{-1}), \tag{3}$$

which leads to the following result similar to Theorem 3.5.

Theorem 3.16. Let V_n be given by (1) with $E[h(X_{i_1}, ..., X_{i_m})]^2 < \infty$ for all $1 \le i_1 \le \cdots \le i_m \le m$.
(i) If $\zeta_1 = \mathrm{Var}(h_1(X_1)) > 0$, then

$$\sqrt{n}(V_n - \vartheta) \to_d N(0, m^2\zeta_1).$$

(ii) If $\zeta_1 = 0$ but $\zeta_2 = \mathrm{Var}(h_2(X_1, X_2)) > 0$, then

$$n(V_n - \vartheta) \to_d \frac{m(m-1)}{2} \sum_{j=1}^{\infty} \lambda_j \chi_{1j}^2,$$

where χ_{1j}^2's and λ_j's are the same as those in Theorem 3.5.

Theorem 3.16 shows that if $\zeta_1 > 0$, then the amse's of U_n and V_n are the same. If $\zeta_1 = 0$ but $\zeta_2 > 0$, then an argument similar to that in the proof of Lemma 3.2 leads to

$$\begin{aligned} \mathrm{amse}_{V_n}(P) &= \frac{m^2(m-1)^2\zeta_2}{2n^2} + \frac{m^2(m-1)^2}{4n^2} \left(\sum_{j=1}^{\infty} \lambda_j \right)^2 \\ &= \mathrm{amse}_{U_n}(P) + \frac{m^2(m-1)^2}{4n^2} \left(\sum_{j=1}^{\infty} \lambda_j \right)^2 \end{aligned}$$

(see Lemma 3.2). Hence U_n is asymptotically more efficient than V_n, unless $\sum_{j=1}^{\infty} \lambda_j = 0$.

Example 3.28. Consider the estimation of μ^2, where $\mu = EX_1$.
From the results in §3.2, the U-statistic $U_n = \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} X_i X_j$ is unbiased for μ^2.
The corresponding V-statistic is simply $V_n = \bar{X}^2$.
If $\mu \neq 0$, then $\zeta_1 \neq 0$ and the asymptotic relative efficiency of V_n w.r.t. U_n is 1.
If $\mu = 0$, then

$$nV_n \to_d \sigma^2 \chi_1^2 \qquad \text{and} \qquad nU_n \to_d \sigma^2(\chi_1^2 - 1),$$

where χ_1^2 is a random variable having the chi-square distribution χ_1^2.
Hence the asymptotic relative efficiency of V_n w.r.t. U_n is

$$E(\chi_1^2 - 1)^2 / E(\chi_1^2)^2 = 2/3.$$

The weighted LSE

In the linear model

$$X = Z\beta + \varepsilon, \tag{4}$$

the unbiased LSE of $l^\tau\beta$ may be improved by a slightly biased estimator when $V = \mathrm{Var}(\varepsilon)$ is not $\sigma^2 I_n$ and the LSE is not BLUE.

Assume that Z is of full rank so that every $l^\tau\beta$ is estimable.
If V is known, then the BLUE of $l^\tau\beta$ is $l^\tau\breve{\beta}$, where

$$\breve{\beta} = (Z^\tau V^{-1} Z)^{-1} Z^\tau V^{-1} X \tag{5}$$

(see the discussion after the statement of assumption A3 in §3.3.1).
If V is unknown and $\hat{V}$ is an estimator of V, then an application of the substitution principle leads to a *weighted least squares estimator*

$$\hat{\beta}_w = (Z^\tau \hat{V}^{-1} Z)^{-1} Z^\tau \hat{V}^{-1} X. \tag{6}$$

The weighted LSE is not linear in X and not necessarily unbiased for β.
If the distribution of ε is symmetric about 0 and $\hat{V}$ remains unchanged when ε changes to $-\varepsilon$, then the distribution of $\hat{\beta}_w - \beta$ is symmetric about 0 and, if $E\hat{\beta}_w$ is well defined, $\hat{\beta}_w$ is unbiased for β.
In such a case the LSE $l^\tau\hat{\beta}$ may not be a UMVUE (when ε is normal), since $\mathrm{Var}(l^\tau\hat{\beta}_w)$ may be smaller than $\mathrm{Var}(l^\tau\hat{\beta})$.

Asymptotic properties of the weighted LSE depend on the asymptotic behavior of $\hat{V}$.
We say that $\hat{V}$ is consistent for V if and only if

$$\|\hat{V}^{-1}V - I_n\|_{\max} \to_p 0, \tag{7}$$

where $\|A\|_{\max} = \max_{i,j} |a_{ij}|$ for a matrix A whose (i,j)th element is a_{ij}.

Theorem 3.17. Consider model (4) with a full rank Z. Let $\breve{\beta}$ and $\hat{\beta}_w$ be defined by (5) and (6), respectively, with a $\hat{V}$ consistent in the sense of (7). Assume the conditions in Theorem 3.12. Then

$$l^\tau(\hat{\beta}_w - \beta)/a_n \to_d N(0,1),$$

where $l \in \mathcal{R}^p$, $l \neq 0$, and

$$a_n^2 = \mathrm{Var}(l^\tau\breve{\beta}) = l^\tau(Z^\tau V^{-1} Z)^{-1} l.$$

Proof. Using the same argument as in the proof of Theorem 3.12, we obtain that

$$l^\tau(\breve{\beta} - \beta)/a_n \to_d N(0,1).$$

By Slutsky's theorem, the result follows from

$$l^\tau\hat{\beta}_w - l^\tau\breve{\beta} = o_p(a_n).$$

Define

$$\xi_n = l^\tau (Z^\tau \hat{V}^{-1} Z)^{-1} Z^\tau (\hat{V}^{-1} - V^{-1})\varepsilon$$

and

$$\zeta_n = l^\tau [(Z^\tau \hat{V}^{-1} Z)^{-1} - (Z^\tau V^{-1} Z)^{-1}] Z^\tau V^{-1} \varepsilon.$$

Then

$$l^\tau \hat{\beta}_w - l^\tau \breve{\beta} = \xi_n + \zeta_n.$$

The result follows from $\xi_n = o_p(a_n)$ and $\zeta_n = o_p(a_n)$ (details are in the textbook).

Theorem 3.17 shows that as long as $\hat{V}$ is consistent in the sense of (7), the weighted LSE $\hat{\beta}_w$ is asymptotically as efficient as $\breve{\beta}$, which is the BLUE if V is known.
By Theorems 3.12 and 3.17, the asymptotic relative efficiency of the LSE $l^\tau \hat{\beta}$ w.r.t. the weighted LSE $l^\tau \hat{\beta}_w$ is

$$\frac{l^\tau (Z^\tau V^{-1} Z)^{-1} l}{l^\tau (Z^\tau Z)^{-1} Z^\tau V Z (Z^\tau Z)^{-1} l},$$

which is always less than 1 and equals 1 if $l^\tau \hat{\beta}$ is a BLUE (in which case $\hat{\beta} = \breve{\beta}$).

Finding a consistent $\hat{V}$ is possible when V has a certain type of structure.

Example 3.29. Consider model (4). Suppose that $V = \mathrm{Var}(\varepsilon)$ is a block diagonal matrix with the ith diagonal block

$$\sigma^2 I_{m_i} + U_i \Sigma U_i^\tau, \qquad i = 1, ..., k, \tag{8}$$

where m_i's are integers bounded by a fixed integer m, $\sigma^2 > 0$ is an unknown parameter, Σ is a $q \times q$ unknown nonnegative definite matrix, U_i is an $m_i \times q$ full rank matrix whose columns are in $\mathcal{R}(W_i)$, $q < \inf_i m_i$, and W_i is the $p \times m_i$ matrix such that $Z^\tau = (\ W_1\ W_2\ ...\ W_k\ )$.
Under (8), a consistent $\hat{V}$ can be obtained if we can obtain consistent estimators of σ^2 and Σ.
Let $X = (Y_1, ..., Y_k)$, where Y_i is an m_i-vector, and let R_i be the matrix whose columns are linearly independent rows of W_i. Then

$$\hat{\sigma}^2 = \frac{1}{n - kq} \sum_{i=1}^{k} Y_i^\tau [I_{m_i} - R_i (R_i^\tau R_i)^{-1} R_i^\tau] Y_i \tag{9}$$

is an unbiased estimator of σ^2.
Assume that Y_i's are independent and that $\sup_i E|\varepsilon_i|^{2+\delta} < \infty$ for some $\delta > 0$.
Then $\hat{\sigma}^2$ is consistent for σ^2 (exercise). Let $r_i = Y_i - W_i^\tau \hat{\beta}$ and

$$\hat{\Sigma} = \frac{1}{k} \sum_{i=1}^{k} \left[(U_i^\tau U_i)^{-1} U_i^\tau r_i r_i^\tau U_i (U_i^\tau U_i)^{-1} - \hat{\sigma}^2 (U_i^\tau U_i)^{-1} \right]. \tag{10}$$

It can be shown (exercise) that $\hat{\Sigma}$ is consistent for Σ in the sense that $\|\hat{\Sigma} - \Sigma\|_{\max} \to_p 0$ or, equivalently, $\|\hat{\Sigma} - \Sigma\| \to_p 0$ (see Exercise 116).