# The Theory of Statistics and Its Applications

By Dennis D. Cox
Rice University

# Chapter 4

# Fundamental Concepts of Statistics.

In this chapter, we introduce some of the basic ideas of statistics making heavy use of the probability machinery that has been developed.

## 4.1 Basic Notions of Statistics.

In this section, we shall try to answer the question, "What is Statistics?" According to the *Encyclopedia Britannica*, "Statistics is the art and science of gathering, analyzing, and making inferences from data." The various activities subsumed under Statistics go by formal nomenclatures such as Design of Experiments, Descriptive Statistics, Data Analysis, Inference, and Decision Theory. In this overview, we shall describe each of these general areas and how mathematical methods come into play. We will also mention other tools, such as computational methods, knowledge outside of statistics, and intuition, which are indispensable to the practice of Statistics.

### 4.1.1 Gathering, Summarizing, and Describing Data.

The first activity in the *Encyclopedia Britannica* is "gathering data." In some sense, data is the raw material of statistics. What is it? How do we gather or collect it? The answer to the second question depends on the first, and also on what we wish to do with the data after it is collected.

In Table 4.1.1 we give an example of a data set. This is in fact a subset of a larger data set. We present only the subset for the purposes of discussion at this point. Many data sets, like this one, can be viewed as a matrix. The rows correspond to *cases* or *observations*, and the columns correspond to *variables*. We will sometimes refer to the cases as *observational units*. In this example, each case corresponds to a patient in a study of the effectiveness of an anti-siezure drug for epileptics. The variable $i$ (in the first column) is the number of the patient

| $i$ | $Y_1$ | $Y_2$ | $Y_3$ | $Y_4$ | Grp. | $X$ | Age |
|---|---|---|---|---|---|---|---|
| 1 | 5 | 3 | 3 | 3 | 0 | 11 | 31 |
| 2 | 3 | 5 | 3 | 3 | 0 | 11 | 30 |
| 3 | 2 | 4 | 0 | 5 | 0 | 6 | 25 |
| 4 | 4 | 4 | 1 | 4 | 0 | 8 | 36 |
| 5 | 7 | 18 | 9 | 21 | 0 | 66 | 22 |
| 29 | 11 | 14 | 9 | 8 | 1 | 76 | 18 |
| 30 | 8 | 7 | 9 | 4 | 1 | 38 | 32 |
| 31 | 0 | 4 | 3 | 0 | 1 | 19 | 20 |
| 32 | 3 | 6 | 1 | 3 | 1 | 10 | 30 |
| 33 | 2 | 6 | 7 | 4 | 1 | 19 | 18 |

Table 4.1: Example Data Set.

from the original data set (which contained 59 patients, of which we selected 10). It is not a particularly useful data set. Some of the patients (those in Group 0, as indicated by the column "Grp.") were given a placebo, and the others (Grp. = 1) were given the experimental treatment. The variable $X$ is the number of seizures the patients had in the 8 weeks prior to the study. They were observed for a total of 8 weeks after beginning the study, with the results broken down by 2 week periods. The variables $Y_1$ through $Y_4$ are the numbers of seizures in each of the 2 week periods after the study began.

This example data set has 8 variables. Many data sets have only 1 variable, but many one variable data sets are obtained by selecting out a single variable from a subset of the cases in a larger data set.

There are two general types of variables, and each of these is further broken down into sub-types:

**(i)** *categorical* or *qualitative* variables are those for which there are a set of discrete possible values. These may be either *nominal*, when there is no ordering, or *ordinal*, if there is an ordering among the categories. For instance, the racial or ethnic class of a patient is a nominal categorical variable. An ordinal variable would result if subjects in a survey were asked to reply to a question with "Poor," "Fair," "Good," or "Excellent." The possible values of a categorical variable may be numeric, but that in some sense would only result from coding.

**(i)** A *quantitative* variable is one which is inherently numeric. Such a variable may be either discrete (e.g. when a result of counting) or continuous (e.g. when a result of measurement). Further, a quantitative variable may have one of two scales: *interval* (when differences are meaningful, but there is no natural zero point) or *ratio* (when ratios are meaningful, and there is a

natural zero point). The sign of an interval measurement is more or less arbitrary, whereas a ratio measurement should always be positive.

In our example data set of Table 4.1.1, the Group variable is a nominal categorical variable (even though it is numeric). If we had recorded the sex of the patients, that would also be a nominal categorical variable. The other variables are quantitative variables. The $Y_i$'s and $X$ are clearly discrete, resulting from counting. Now Age appears discrete (as it only takes on whole number values), but we would generally prefer to think of it as a continuous variable which has just been rounded off (actually truncated – patient No. 1 is between 31.0000 and 31.9999.... years old; in truncation, the fractional part is thrown away).

In the ideal world, researchers needing statistical assistance would seek the advice of a trained statistician before collecting the data. The data in the example have the advantage of malice of forethought – the patients were assigned to a group (treatment or control, i.e. placebo) *randomly*.

### 4.1.2 Statistical Models.

One of the most important steps in addressing a statistical problem is the formulation of a model. We shall give a formal definition here, but we do not claim that this encompasses all "models" that arise in practice, only the models that are amenable to the mathematical methods described in this text. A *statistical model* consists of a measurable space $(\Omega, \mathcal{F})$, a collection of probability measures $\mathbf{P}$ on $(\Omega, \mathcal{F})$, and a collection of possible observable random vectors $\underline{\boldsymbol{X}}$. It is assumed that one of the probability measures in $\mathbf{P}$ is the "correct" or "true" one (or an approximation to the "true" one which is good enough for the purpose at hand), and that one can choose to observe a realization of one of the random vectors in $\underline{\boldsymbol{X}}$. Choice of which element of $\underline{\boldsymbol{X}}$ to observe is the *experimental design* problem. Many times the statistician does not have such a choice, as in a consulting problem where a client has already collected the data, in which case $\underline{\boldsymbol{X}}$ consists of just one random vector. The random vector may also have nonrandom components. Note that each $\underline{X}$ in $\underline{\boldsymbol{X}}$ will have an induced distribution under each probability measure in $\mathbf{P}$, and we usually just concentrate on these induced measures, especially if there is only one possible observable. As an example, suppose the underlying measurable space is a collection of laboratory mice which are treated with some chemical which will possibly cause cancer. The mice are to be given some amount of the chemical and then observed for a period of time. Then it is determined whether or not they have cancer. Let us assume the cancer researcher has already determined the amount of chemical to give and the length of time to wait after giving the chemical before testing for cancer, but she wishes to know how many mice she needs to test. In this case, the measurable space is a collection of mice, and the experiment consists of selecting mice at random, treating them with the chemical, and observing whether or not they develop cancer within the given period of time. The unknown probability distribution then

takes into account the probability of selecting mice, the probability of their subsequent development of cancer if they are selected, and the chance of observing that cancer (there is some chance it may be missed, or some other condition may be misdiagnosed as cancer). The exact specification of the probability measure and so forth is not required for the mathematical analysis (although we should be careful that the mice are truly selected at random from some well defined population of mice if we want to guarantee that the sample is *representative* of that population; more on this later). The space of observables is a collection $\underline{\boldsymbol{X}} = \{(n, X) : n \in \mathbb{N}, X \in \mathbb{N}\}$ where $X$ is a binomial random variable with $n$ trials and unknown "success probability" $p \in [0, 1]$, which is the probability of cancer in a single mouse. $X$ is the number of mice in the sample who develop cancer. (Alternatively, one could consider that observable $(n, Y_1, Y_2, \ldots, Y_n)$ where the $Y_i$'s are independent Bernoulli random variables defined by $Y_i = 1$ iff mouse $i$ develops cancer. Then $X$ above is given as the sum $X = \sum_{i=1}^{n} Y_i$. We shall see that it is good enough ("sufficient") to consider the simpler observable $(n, X)$.) Here, the sample size $n$ is a nonrandom component of the observation. The design problem consists of choosing $n$. Of course, a larger value of $n$ will result in more "accurate inferences", but there is a finite amount of money available for the entire experiment. Often, the client will ask for the minimum value of $n$ required to achieve a given "accuracy of inference", and then use that value if it is feasible or take the largest value she can afford. Determination of the value required to achieve a given "accuracy of inference" depends on the kind of inferences to be done, whether or not there is a reasonable estimate or guess at the value of $p$, or whether or not it is acceptable to use a "worst case" value of $p$. These issues are discussed at greater length in sections ???.

We will generally assume the experimental design is already decided upon so that there is only a single observable random $n$-vector $\underline{X}$ and a collection possible distributions $\mathbf{P}_{\underline{\mathbf{X}}}$. For instance, in the above example the cancer researcher may have already decided on the sample size $n$ and may already have collected the data so we need only concern ourselves with the collection of binomial distributions $\{B(n, p) : 0 \leq p \leq 1\}$. Then, at least as far as the mathematics of the inference goes, we can forget about the underlying measurable space and the probability measures thereon and simply concentrate on $\underline{X}$ and its distributions. It has been common practice in statistics to choose for $\mathbf{P}_{\underline{\mathbf{X}}}$ a *parametric family* meaning a collection $\{P_\theta : \theta \in \Theta\}$ where $\theta$ is a natural index or label for the probability measure. Here, the *parameter space* $\Theta$, is the collection of all possible values of the parameter $\theta$. In general, $\Theta$ is a "nice" subset of some finite dimensional space $\mathbb{R}^p$. For instance, in our mouse cancer example with $n$ fixed, we may index the possible distributions with the parameter $p$ (probability of cancer for a single mouse) and the parameter space is $[0, 1]$.

When our parameter is multidimensional, we may only be interested in some components of the parameter vector (called "*parameters of interest*") and the other parameters (referred to as "*nuisance parameters*") are not of interest. For

example, suppose we observe $X_1$, $X_2$, ..., $X_n$ which we model as i.i.d. $N(\mu, \sigma^2)$ where both $\mu$ and $\sigma^2$ are unknown. We may be interested only in $\mu$, but to completely specify the model, it is necessary to include $\sigma^2$. In this case, $\mu$ is the parameter of interest, and $\sigma^2$ is a nuisance parameter.

In the mouse cancer example, if we assume the observations of cancer or no cancer on each of the mice are independent and that the probability of cancer is the same from mouse to mouse (note that these are assumptions about the underlying measurable space and the possible probability measures on it), then there is really no other choice than this parametric family $\{B(n,p) : 0 \le p \le 1\}$. In many cases, however, such knowledge is not available. For instance, if instead of determining "cancer" or "no cancer" on each of our mice we made a measurement such as "mass of cancerous liver cells". (This may be done by sacrificing the mice, cutting out the livers, removing all tumors, and weighing their total mass.) Then our model is necessarily more complicated. We may model the measured masses as realizations of i.i.d. random variables with a distribution of the form

$$\text{Law}[X_i] \;=\; (1-p)\delta_0 \;+\; pP \tag{4.1}$$

where $p \in [0,1]$ is the probability of developing cancer (so that if the mouse has no cancer then the observed mass is 0) and $P$ is a probability distribution (for the tumor mass given that there is cancer) which is supported on the nonnegative $\mathbb{R}$. Since the tumor mass may in principle take any value in the continuum of positive real numbers, we would ordinarily assume $P$ has a Lebesgue density of the form

$$\frac{dP}{dm}(x) \;=\; g(x) \;,\; \text{where } g(x) = 0 \text{ for } x \le 0. \tag{4.2}$$

For various reasons, we may decide that a reasonable choice of possibilities for $g$ is the exponential family

$$g(x) \;=\; \lambda^{-1} e^{-x/\lambda} \quad , \quad x > 0, \tag{4.3}$$

where $\lambda > 0$ is unknown. Here, $\lambda$ has the interpretation as the mean tumor mass for mice with cancer. Now our parametric family has densities $\{f_\theta : \theta = (p,\lambda) \in \Theta\}$ where the parameter space $\Theta = [0,1] \times (0,\infty)$ and the dominating $\sigma$–finite measure is

$$\mu \;=\; \delta_0 \;+\; m \;. \tag{4.4}$$

For many reasons, we will usually want our family to have a $\sigma$–finite dominating measure $\mu$, in which case we say the family of probability measures is *dominated* and will write $\mathbf{P}_{\underline{\mathbf{X}}} \ll \mu$. Note that the statistical model is well defined by the family of densities and the dominating measure since we can determine the particular probability distributions therefrom. Note also that the parameterization is not unique. For instance, instead of using the mean $\lambda$ of the exponential, we could use the reciprocal mean $\alpha = 1/\lambda$, $0 < \alpha < \infty$, or the log mean $\xi = \log(\lambda)$, $-\infty < \xi < \infty$. Since the parameter is just a label for the distribution, choice

of a particular parameterization is often a matter of mathematical convenience. This will be discussed at greater length elsewhere.

In the above, we assumed the distribution for tumor mass given that the mouse had cancer was an exponential distribution with unknown mean $\lambda$. Of course, it is extremely doubtful that the exact density function has this form. There are two points to be made. First, we may wish to assume that the density $g$ in (4.2) is completely unknown, except possibly for very general assumptions, like that it is continuous on $(0, \infty)$ and decreasing (or maybe unimodal). Such a model is called *nonparametric* since there is no nice way of labelling or "parameterizing" such densities with a finite dimensional parameter. (However, we can "parameterize" any family of probability distributions – just use the probability distribution itself as "parameter". It has become traditional in statistics to speak of parametric families when referring to a model with a "nice" finite dimensional parameter even though everything is in principle parametric.) For whatever questions the researcher is interested, we may then want to use *nonparametric methods* which are appropriate and valid for such models. An alternative approach is to consider *robustness*. Based on past experience with similar chemicals, the researcher may believe that the exponential densities represent a reasonable approximation to the true density. *Robust methods* work well for the nominal (exponential) density and for other distributions which are "nearby". We shall give examples of robust and nonparametric procedures in sections ???.

### 4.1.3 Statistical Inference.

So far, we have talked about "inference" without saying what it is. Webster's New World Dictionary, Third College Edition, defines "infer" as "to conclude or decide from something known or assumed; derive by reasoning; draw as a conclusion." "Inference" then means "the act or process of inferring," or "a conclusion or opinion arrived at by inferring." In the case of statistical inference, what is known is the data and what is assumed is the model. We then try to reason to a conclusion, which would be the inference for that occasion (as opposed to the general subject of inference). In the mouse cancer example, the researcher may be interested in inferring the true value of the probability of cancer (referred to as an estimation problem), or in inferring whether or not the true value is above or below .001 (referred to as a hypothesis testing problem). The main point of statistical inference is to use the data to find out what the "truth" is in a scientific setting. Notice that this is in some sense dependent on the statistical model, as in the example where we refer to a probability which is a parameter in the model. Often a client will be interested in drawing some inferences but not have enough experience or knowledge of statistics to express the problem in a formal way. In these cases a good statistical consultant will try to understand the client's informal questions ("informal" means from the statistical point of view; they may be very "formal" from the client's point of view) and formulate a

model which facilitates a statistical statement of the question and is reasonably accurate for the data at hand. (Again, robustness becomes an issue here since no model can be reasonably expected to be exactly correct. George Box, a famous statistician, says "All models are false, but some are useful." We take "useful" to mean both that it fits the data and allows one to make the desired inferences.)

In the above, we talked about "accuracy of inference". There is no universally accepted way of defining this, but there are widely used methods for the two types of inference we talked about. For the mouse cancer example, if we want to estimate $p$, the probability of cancer for a single mouse, then we want to compute some function of the data, say $\hat{p}(X)$ which is called an *estimator*. When a particular realization $x$ is plugged in, we obtain the *estimate* $\hat{p}(x)$. (Note: an estimate $\hat{p}(x) \in \mathbb{R}$ in this case but an estimator $\hat{p} : \mathbb{N} \longrightarrow [0,1]$ is a function.) Of course, we require that $\hat{p}$ be measurable. The most common way (but certainly not the only way) of measuring accuracy is with *mean squared error*

$$\mathrm{MSE}(p, \hat{p}) \;=\; E_p[(\hat{p}(X) - p)^2] \;. \tag{4.5}$$

Here, $E_p$ means computing expectation using the (binomial) probability measure determined by the particular value $p$. Note that MSE is a function of both the value $p$ (which is assumed to be the "true" value for the purposes of computing MSE) and the estimator $\hat{p}$. MSE has a decomposition into "bias squared plus variance"

$$\mathrm{MSE}(p, \hat{p}) \;=\; \mathrm{Bias}^2(p, \hat{p}) \;+\; \mathrm{Var}_p(\hat{p}(X)) \;, \tag{4.6}$$

where the *bias* is given by

$$\mathrm{Bias}(p, \hat{p}) \;=\; E_p[\hat{p}(X)] - p \;. \tag{4.7}$$

Note in (4.6) we use $\mathrm{Var}_p$ to compute variance using the distribution with parameter $p$. It is common (but by no means unanimously accepted) to require that the estimator be *unbiased*, meaning that $\mathrm{Bias}(p, \hat{p}) = 0$ for all parameter values $p \in [0,1]$. In that case, MSE reduces to variance by (4.6).

For the problem of testing hypotheses there is given a value $p_0$ and one wishes to decide which of two statements is correct:

$$H_0 : p \leq p_0 \quad \text{vs.} \quad H_1 : p > p_0 \quad , \tag{4.8}$$

One approach would be to estimate $p$ with say $\hat{p}(X)$, and decide $H_0$ is true if $\hat{p}(X) \leq p_0$, and otherwise decide $H_1$ is true. However, the conceptual framework of hypothesis testing is sufficiently different from estimation to require seperate consideration. One distinguishes one of the two statements as being the *null hypothesis* $H_0$, and controls the probability of incorrectly deciding it is false. This probability of *rejecting* $H_0$ if in fact it is true is made to be small ("rejecting" $H_0$ means deciding it is false). The other statement $H_1$ is called the *alternative hypothesis*. Thus, the strategy is that if we reject the null hypothesis, we do so

only if there is "strong evidence" against it since there will be little chance of rejecting it if it is true. Thus, in some sense, we have given the "benefit of the doubt" to $H_0$. For this reason, one usually chooses for the null hypothesis the statement which one wishes to reject, since if that is done we can be reasonably sure of making the right decision, whereas accepting the null hypothesis may still leave some doubt. We will amplify on these issues in chapter ???.

Now we formalize the above. We are given a bound $\alpha$ on the probability of falsely rejecting $H_0$ called a *level of significance*, and we wish to have a *test function* $\phi : I\!N \longrightarrow \{0, 1\}$ where $\phi(X)$ is our decision (i.e. the subscript on $H$) such that

$$P_p[\phi(X) = 1] \ \leq \ \alpha \ , \text{ for all } p \leq p_0 \quad . \tag{4.9}$$

(Here, we have subscripted the probability with the parameter $p$. Since we have written $X$ for the r.v., this probability is a measure on the underlying space of mice, but since we are only concerned with the distribution of $X$, knowledge of $p$ is sufficient for all calculations so we need not concern ourself with other aspects of the unknown probability measure.) Thus, the probability of wrongly rejecting $H_0$ is $\leq \alpha$. Here, $\alpha$ is specified. For any given value $p$, we can compute the *power* of the test which is

$$\gamma(p) \ = \ P_p[\phi(X) = 1] \ = \ E_p[\phi(X)] \ . \tag{4.10}$$

Note that the latter equation follows since $\phi$ is just an indicator function, namely the indicator of the *critical region* $C = \{x \ : \ \phi(x) \ = \ 1\}$. Thus, we may be concerned with the power at a particular alternative value $p_1 > p_0$ and use this to compare various tests of the hypotheses. Thus, for hypothesis testing, we assess the "accuracy of the inference" by the power function.

In the context of hypothesis testing for the mouse cancer example, we introduce a notion intended for mathematical convenience that sometimes has practical uses, namely *randomized procedures*. Suppose the cancer researcher uses a sample of $n = 10$ mice and wishes to test the hypotheses

$$H_0 : \ p < .8 \qquad \text{vs.} \qquad H_1 : \ p \geq .8 \ .$$

Suppose also that the level of significance $\alpha = .05$ is required. (Incidentally, this is the most commonly used level of significance, and has become the "default" value, although there is no particularly good reason for this.) It is intuitively clear even if the student is unfamiliar with such testing problems that evidence against $H_0$ in favor of $H_1$ is a value of $X$ which is too large, so the critical region should take the form

$$C \ = \ \{x : \ x \ \geq \ x_0 \} \ .$$

Here, $x_0$ is to be determined so that (4.9) holds, i.e. that

$$P_p[X \ \geq \ x_0] \ \leq \ .05 \quad , \quad \text{for all } \ p \leq .8 \ . \tag{4.11}$$

Since $X$ takes nonnegative integer values, we may consider only these values for $x_0$. Notice that in general $P_p[X \geq x_0]$ is a increasing function of $p$. This is not easy to show, but is intuitively obvious. Hence, the l.h.s. of (4.11) is maximized at $p = .8$, so we need only ensure that (4.11) holds for this value of $p$. Now we have from a table of binomial probabilities and an obvious result (since $n = 10$) that

$$\begin{aligned} P_{.8}[X \geq 10] &= .134 \\ P_{.8}[X \geq 11] &= 0 \end{aligned}$$

Thus, in order to satisfy (4.11), we should take the critical region to be $[X \geq 11]$ and reject $H_0$ with probability 0. One sees how $H_0$ is given too much "benefit of the doubt". One technical way to rectify this is to allow us to do the following: if $X = 10$ is observed (which happens with probability .134) then reject $H_0$ with probability .05/.134 = .373. That is, perform some other random experiment with possible outcomes 0 or 1 where 1 has a probability of .373, and reject $H_0$ if 1 occurs on this auxiliary experiment. For instance, one may generate a uniform random number in $[0, 1]$ on a computer and reject $H_0$ if $X = 10$ and this uniform random number is $\leq .373$. Obviously, our client would probably not like this statistical procedure, but not much else can be done with such a small number of observations in respect to the question being asked.

A simple way to mathematically represent such a randomized test of hypotheses is to let the test function take values between 0 and 1. Henceforth, we define a test function as a measurable function $\phi$ on the range of values of the random observable $X$ into $[0, 1]$, and $\phi(x)$ represents the conditional probability of rejecting $H_0$ given that $X = x$ is observed. Then the total probability of rejecting $H_0$ is

$$E_p[\phi(X)] \; = \; \gamma(p) \; . \tag{4.12}$$

Notice that we have a two stage experiment here. The first stage is to observe $X$ according to its (unknown) binomial distribution, and the second stage is to generate a Bernoulli random variable $Z$ (a r.v. taking the values 0 or 1) with conditional probability $P[Z = 1|X = x] = \phi(x)$. We then decide in favor of the hypothesis $H_Z$. Of course, if $\phi(x)$ only takes values in $\{0, 1\}$ (i.e. is a nonrandomized test), then there is no point in performing the experiment to get $Z$ since it is just a degenerate r.v.

### 4.1.4 Statistical Decision Theory.

Both of the above kinds of inference problems (estimation and hypothesis testing) may be fit within a single framework called *(statistical) decision theory*. The statistical decision problem is often thought of as a game between two opponents, Nature and the Statistician. Nature chooses a probability measure from a parameterized family $\{P_\theta : \theta \in \Theta\}$ on the measurable observation space $(\Xi, \mathcal{G})$

with the choice being unknown to the Statistician. There is given a measurable *action space* $(A, \mathcal{D})$ of allowable *decisions* or *actions*, and the Statistician must choose a *decision rule* $\delta : (\Xi, \mathcal{G}) \longrightarrow (A, \mathcal{D})$ from a class $\Delta$ of *allowable decision rules*. There is also given a *loss function* $L : \Theta \times A \longrightarrow [0, \infty]$ such that $L(\theta, \cdot)$ is extended Borel measurable for each fixed $\theta \in \Theta$. If Nature chooses $\theta_0$, then there is generated an observation $x$ of $X$, a $\Xi$ valued random element with distribution $P_{\theta_0}$. If the Statistician uses the rule $\delta$, then he loses $L(\theta, \delta(x))$. The *risk* is the expected loss

$$R(\theta, \delta) \;=\; E_\theta[L(\theta, \delta(X))] \;. \tag{4.13}$$

This is a function of $\theta$ (the parameter of the probability distribution chosen by Nature) and $\delta$ (the decision rule chosen by the statistician). If the game is played repeatedly (with the same value of $\theta$ and the same decision rule), then the risk is the long run average loss. Note that the integrand in (4.13) is Borel measurable since for each $\theta \in \Theta$ it is a composition $L(\theta, \cdot) \circ \delta$ of two measurable functions (see Proposition 1.2.3). Also, the integral is defined since the integrand is nonnegative. Therefore, $R$ is a function on $\Theta \times \Delta$ taking values in $[0, \infty]$. One can allow more general loss functions, but we need not consider them here.

As with the hypothesis testing situtation, it will sometimes be useful to consider *randomized decision rules*. This means that the Statistician specifies not a mapping $\delta : (\Xi, \mathcal{G}) \longrightarrow (A, \mathcal{D})$ but rather a family of conditional distributions $\{P_{\delta|X}(B|x) : B \in \mathcal{D},\, x \in \Xi\}$. The interpretation is that we make a two stage experiment: after first observing $X = x$, we generate a decision $\delta \in A$ according to the conditional distribution $P_{\delta|X}(\cdot|x)$. Then the risk in (4.13) is computed by integrating over both the randomness in $X$ and the randomness in $\delta$, viz.

$$R(\theta, \delta) \;=\; \int_\Xi \int_A L(\theta, a)\, dP_{\delta|X}(a|x)\, dP_X(x) \;. \tag{4.14}$$

We have used $a$ as a dummy variable of integration to represent a particular value of the random element $\delta$. In general, we may still require that $\delta$ belong to an allowable class $\Delta$ of randomized decision rules. A nonrandomized decision rule is one for which all conditional distributions are degenerate, i.e. if $\delta_0(x)$ is a nonrandomized decision rule then the corresponding randomized decision rule is

$$P_{\delta|X}(B|x) \;=\; I_B(\delta_0(x)) \;, \tag{4.15}$$

where we have written the unit point mass measure $\delta_{\delta_0(x)}(B)$ as an indicator to avoid confusion with too many $\delta$'s.

Decision rules are compared by comparing their risk functions. Let $\Theta_1 \subset \Theta$. a decision rule $\delta_1$ is *as good as* another decision rule $\delta_2$ on $\Theta_1$ iff

$$R(\theta, \delta_1) \;\le\; R(\theta, \delta_2) \quad , \quad \text{for all } \; \theta \in \Theta_1 \;, \tag{4.16}$$

and we write $\delta_1 \preceq \delta_2$ on $\Theta_1$. It will usually be the case that $\Theta_1 = \Theta$, in which case we will drop the mention of $\Theta_1$. We say $\delta_1$ is *better than* $\delta_2$ on $\Theta_1$ if $\delta_1 \preceq \delta_2$

on $\Theta_1$ and for some $\theta \in \Theta_1$, $R(\theta, \delta_1) < R(\theta, \delta_2)$. We write $\delta_1 \prec \delta_2$ on $\Theta_1$ for this. Two decision rules are called *equivalent* on $\Theta_1$ if each is as good as the other on $\Theta_1$, and we write $\delta_1 \sim \delta_2$ on $\Theta_1$ in this case. We will be interested in optimal decision rules. A $\Theta_1$*–uniform minimum risk* $\Delta$*–rule* is a rule $\delta_*$ in $\Delta$ which is as good as any other rule in $\Delta$ on $\Theta_1$. This is clearly the best optimality one could hope for if $\Theta_1 = \Theta$ (given that the decision rules must come from $\Delta$), but it is frequently impossible to attain. This happens because some rules are better for some parameter values and other rules are better for other parameter values. One of the reasons for constraining the decision rule to lie in a particular $\Delta$ is so that there will exist a uniform minimum risk $\Delta$-rule, and if one does not exist for one $\Delta$, then it has been a common practice to make $\Delta$ even smaller to obtain such a best rule. Of course, this practice is highly questionable.

Now we show how the previous discussions of inference with the mouse cancer example are encompassed within decision theory. Nature "chooses" a probability $p$ that an individual mouse who has undergone the treatment develops cancer. Consider the point estimation problem. The action space is $[0, 1]$, the same as the parameter space. A decision rule is an estimator, which is simply a measurable map $\hat{p} : \{1, 2, ..., n\} \longrightarrow [0, 1]$. If we use *squared error loss*, given by

$$L(p, \hat{p}) \;=\; (\hat{p} - p)^2 \;, \tag{4.17}$$

then the risk is MSE. This choice of loss function is arbitrary but mathematically convenient. If it is required that the estimator be unbiased, then the class $\Delta$ of allowable decision rules is the class of measurable functions $\hat{p}$ satistifying the unbiasedness constraint

$$E_p[\delta(X)] \;=\; p \quad \text{for all } p \in [0, 1].$$

If an unbiased estimator has smallest variance (risk under squared error loss) among unbiased estimators for all parameter values, then it is called a *uniform minimum variance unbiased estimator*. In Chapter ?? we shall be concerned with finding such *UMVUE*'s.

To see why unbiasedness is used, recall that one estimator $\hat{p}_1$ is better than another $\hat{p}_2$ if $\text{MSE}(p, \hat{p}_1) \leq \text{MSE}(p, \hat{p}_2)$ for all $p \in [0, 1]$. Now let $\hat{p}_2$ be the "dumb" estimator

$$\hat{p}_2(x) \;=\; 1/2 \;. \tag{4.18}$$

Obviously, $R(1/2, \hat{p}_2) = 0$, and no other estimator can have risk this small at $p = 1/2$ unless it is almost surely equal to $\hat{p}_2$. On the other hand, $\hat{p}_2$ does not have good risk for values of $p$ very far from $1/2$. The point of requiring unbiasedness then is to rule out such foolish estimators which cannot be improved on for specific values of the parameter. This is an example of constraining the class of decision rules so as to find a uniformly best rule in the constrained class. However, we shall see that requiring the estimator to be unbiased also rules out good estimators sometimes.

The hypothesis testing problem is a little more complicated. The action space is very simple: $\{0, 1\}$ corresponding to choosing $H_0$ or $H_1$. A nonrandomized decision rule is a test $\phi : \{1, 2, ..., n\} \longrightarrow \{0, 1\}$. The collection of allowable tests $\Delta$ are those satisfying the level of significance constraint in (4.11). The loss function is so called $0 - 1$ loss

$$L(p, \phi) = \begin{cases} 1 & \text{if } \phi = i \text{ and } H_i \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.19}$$

That is, the loss is 1 if the right decision is made, and is 0 if the wrong decision is made. One can see then that the risk is

$$R(p, \phi) = \begin{cases} \gamma(p) & \text{if } p \leq p_0, \\ 1 - \gamma(p) & \text{if } p > p_0. \end{cases}$$

That is, the risk for a given $p$ is the probability of making the wrong decision for that $p$. Since the risk has already been constrained to be small for values of $p$ in $H_0$, we generally only concentrate on values of $p$ in $H_1$, so we want the power function to be large in $H_1$ so that the risk is small. Note that a randomized test is simply a randomized decision rule. A test which maximizes the power over $H_1$ subject to the level $\alpha$ constraint is called a *uniformly most powerful test of level* $\alpha$ or *UMP level* $\alpha$ test. Such a test is a $H_1$–uniform minimum risk significance level $\alpha$ rule, in the language of decision theory. Again, if one considers tests irrespective of the level $\alpha$ constraint, then there is a test which has no risk when $H_0$ is true (namely, always accept $H_0$) and another test with no risk when $H_1$ is true (always reject $H_0$). Thus, the level $\alpha$ constraint is a convenient way of constraining the class of decision rules so that one can sometimes find UMP tests.

Decision theory provides a nice mathematical framework within which to evaluate various procedures for statistical inference. It is also a subject of some interest in its own right. For instance, if making economic decisions for a business, one may be able to determine a loss (or profit) function for a decision rule in the face of uncertainty that can be formulated in terms of a parameterized family of probability measures. However, in applied statistics, decision theory is often times misleading. In the inference examples above, there is no good reason for choosing the particular loss functions, nor is it clear that even computing the risk is such a great idea, although it does provide some basis for comparing different inference methods. Other difficulties arise. For instance, it is assumed that the observable has a distribution in a prespecified parametric family. How does one choose this family? Is it an accurate representation of reality? Is the decision rule selected for doing the inference insensitive to departures of the true state of Nature from our idealization as a parametric model of probability distributions? What if we use the data in the choice of the parametric model? How does this affect the validity of the inferences? Some of these questions are very deep and it is fair to say that no one has given entirely satisfactory answers. In actually applying statistics to real world problems, it can be very misleading to confine oneself to thinking in terms of decision theory.

### 4.1.5 Bayesian Decision Theory.

The decision theory problem can be simplified by adding some more structure: suppose we believe that Nature chooses the parameter as well as the data at random, and we know that the distribution she uses for selecting $\theta$ is $\pi$, a probability measure on $\Theta$. The distribution $\pi$ is called the *prior distribution* for $\theta$. We can formulate a new problem of minimizing the *Bayes risk* which is given by

$$r(\delta) \;=\; \int_\Theta R(\theta,\delta)\,d\pi(\theta).$$

That is, we average the risk over the prior. In the problem of finding a uniform minimum risk estimator, we try to minimize $R(\theta,\delta)$ over $\delta$ simultaneously for all values of $\theta$, which is typically impossible, unless we restrict the set of allowable decision rules $\delta$. For the Bayesian problem we have a single function of $\delta$ we wish to minimize, which is a much more manageable problem, and in general can be accomplished without restrictions on $\delta$, as we now show.

Assume we have a dominated family as in

$$P_{X|\theta}(\cdot|\theta) \ll \mu \;\sigma\text{-finite},\; f(x|\theta) \;=\; \frac{dP_{X|\theta}(\cdot|\theta)}{d\mu}(x).$$

Note that we have changed our notation from $f_\theta(x)$ to $f(x|\theta)$ in order to reflect the fact that now $\theta$ is a random object, and we are conditioning on it. Also, we are using $\theta$ to denote both the random object and various values it might take (a very dangerous thing to do!), since we have already taken $\Theta$ for the parameter space. The student will have to be careful when interpreting the symbol $\theta$ – it should be clear from context what is meant. Inserting the definition of the risk function $R(\theta,\delta)$ (as the expected loss over the observations) and using Fubini's theorem, we obtain

$$\begin{aligned} r(\delta) &= \int_\Theta \int_\Xi L(\theta,\delta(x)) f(x|\theta)\,d\mu(x)\,d\pi(\theta) \\ &= \int_\Xi \int_\Theta L(\theta,\delta(x)) f(x|\theta)\,d\pi(\theta)\,d\mu(x). \end{aligned} \tag{4.20}$$

Now, define the inner integral using

$$\rho(x,a) \;=\; \int_\Theta L(\theta,a) f(x|\theta)\,d\pi(\theta). \tag{4.21}$$

Here, $a$ is any allowable action, but plugging in $\delta(x)$ for $a$ and integrating $d\mu(x)$ gives $r(\delta)$. Now suppose for each fixed $x \in \Xi$ we can find an action $a$ that minimizes $\rho(x,a)$. This minimizer will depend on $x$, of course, so denote it as

$$\delta^\star(x) \;:\; \rho(x,\delta^\star(x)) \;\le\; \rho(x,a),\; \forall a \in A. \tag{4.22}$$

Then, assuming measurability of $\delta^\star$, we see immediately that $r(\delta^\star) \le r(\delta)$ for any other decision rule $\delta$.

**Example 4.1.1** *Consider the problem of point estimation of $p$ in the $B(n,p)$ model ($n$ known) under squared error loss. The function $\rho$ in (4.21) is given by*

$$\rho(x,a) \;=\; \int_0^1 (p-a)^2 f(x|p)\, d\pi(p),$$

*where*

$$f(x|p) \;=\; \left( \begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}.$$

*Expanding the $(p-a)^2$ term in the integral expression for $\rho(x,a)$ and observing that it is a quadratic function of $a$, it is easy to see that the Bayes estimator under squared error loss is given by*

$$\delta^\star(x) \;=\; \frac{\int_0^1 p f(x|p) d\pi(p)}{\int_0^1 f(x|p) d\pi(p)} \;=\; \int_0^1 p f(x|p) d\pi(p|x),$$

*where $\pi(\cdot|x)$ is the conditional distribution of the parameter $p$ given $X = x$. The conditional distribution of the parameter given the data is known as the posterior distribution. In general, the optimal Bayes estimator under squared error loss is the posterior mean, as we see in this particular example.*

### 4.1.6 Data Analysis.

We briefly mention two other points of view which can be useful, although much of what we say in the remainder of the text is within the context of decision theory since the other topics do not lend themselves so nicely to mathematical analysis.

It could be said that statistics is the art and science of "data analysis" in its most general meaning. That is, other scientists generate data from their experiments or observatons and consult with the statistician to help them understand the data, determine whether it supports their hypotheses or not, and/or use the data to make predictions. Almost all of modern statistics is based on probability models for the data, even though presumably Nature does not roll dice to determine events in this world (recall probability models were developed as models for games of chance). However, Nature often seems to act as if she rolls dice. It is not clear whether this is an intrinsic property of nature (as quantum physicists think) or of probability models. The main point is that probability models seem to work for a wide variety of problems, if applied by a skilled person. Even in those statistical analyses where a probability model is not required, it is often useful to think of the data as generated by some random mechanism. One should always question "where does the randomness come from?" In the mouse cancer example, the randomness may come from the fact that the scientist actually selected the mice for the study by taking a random sample from a well defined population of mice. This is the best of all circumstances, although it is not clear

one can ever truly attain it since generation of the random numbers used to determine the sample is always questionable. (Computer generated random numbers are known to not truly be random, although they work well for most purposes. If the numbers were generated by rolling a die, how can we be sure that the roll of a die is a truly random experiment, and if so that that die is fair and the different rolls are independent?) More often than not, such sound experimental practices are not followed and one analyzes the data and draws the conclusions conditionally under the assumption that the probability model is correct, or at least close enough.

In the examples of inference given above we were mostly interested in "confirmatory" results rather than "exploratory" analysis. Often times, the statistician is asked to look for connections and patterns in the data. This is usually called *exploratory data analysis* or EDA. While this activity does not require one to assume a probability model for the data, there are at least two ways in which such an assumption is useful. For one, there always arises the question whether or not such structure as may be found is "truly" present or could have resulted from chance alone. For instance, if one sees an apparent difference between the cancer rates of mice treated with a chemical and a control group which is not treated, is it possible that the two groups' cancer rates are really the same but the apparent difference is caused by chance? A second use of probability models in EDA is for conceptualization of structures one might find. For instance, we may plot a histogram to get a convenient "picture" of the data, to see for instance if there are two or more "subpopulations" (as would be indicated by a multimodal histogram). A histogram may be also interpreted as an estimate of the underlying density, if we assume the observations are say i.i.d. with a Lebesgue density. This allows us for instance to compare various ways of selecting the width of the histogram class intervals (or bins) by looking at the mean squared error, whereas there is no obvious way to do this without assuming a probability model. It also allows us to assess the "accuracy" of the histogram as an estimate, and this can be helpful in assessing whether or not apparent multimodality is "really there".

Data analysis in the sense of looking of "looking at the data" is also useful after the inferences are drawn. In applied statistics, such techniques are often referred to as *diagnostics*. For instance, in the mouse cancer example we may look at the data as a function of observation number to see if there is an effect from time suggesting the data are not really i.i.d. Some of the same questions arise as in exploratory data analysis, such as if an observed trend is found could this be an illusion and such a trend was really generated by chance alone. Ideas from hypothesis testing (hence also decision theory) can be applied to such questions.

Finally, we mention *descriptive statistics*, which is usually taken to mean "summarizing" or "describing" the data. While in principle this is done under the presumption the data is just a "bunch" of numbers without any underlying probability distribution, probabilistic and decision theoretic ideas can be helpful. For instance, one can summarize the data by grouping it into classes and

determining class frequencies, which amounts to the same thing as forming a histogram, in which case the ideas above regarding the histogram as a nonparametric estimate of the probability density function are pertinent. One may also summarize the data by giving sample mean, sample variance, and/or sample quantiles, all of which have a probabilistic interpretation as estimates of functionals of an underlying distribution and hence can be studied from the point of view of decision theory.

**Exercises for Section 4.1.**

**4.1.1** Verify equation (4.6).

**4.1.2** Suppose $X$ is $B(n,p)$ for $n$ given and $p \in [0,1]$ unknown. For each of the following estimators of $p$, calculate the bias, variance, and MSE.

**(i)** $\hat{p}_1(x) = x/n$.

**(ii)** $\hat{p}_2(x) = (x + \sqrt{n}/2)/(n + \sqrt{n})$

Is either of these estimators better than the other? Is either of these estimators as good as the other?

**4.1.3** A test function $\phi : (\Xi, \mathcal{G}) \longrightarrow [0,1]$ is defined as the conditional probability of rejecting $H_0$ given $X$ (where $X : (\Omega, \mathcal{F}) \longrightarrow (\Xi, \mathcal{G})$ is the random observable). Let $\theta \in \Theta$ denote the parameter.

Show that the probability of rejecting $H_0$ if $\theta$ is true is the power

$$\gamma(\theta) \;=\; E_\theta[\phi(X)]$$

**4.1.4** Suppose $X : (\Omega, \mathcal{F}) \longrightarrow (\Xi, \mathcal{G})$ is a random observable with $\text{Law}[X] \in \{P_\theta : \theta \in \Theta\}$. The problem of *set estimation* (or *interval estimation* if $\Theta \subset I\!R$) involves finding a set $S(X)$ which contains the true $\theta$ with high probability. More specifically, a $1-\alpha$ *confidence set* satisfies

$$P_\theta[\theta \in S(X)] \;\geq\; 1-\alpha \quad , \qquad \text{for all } \theta \in \Theta \;. \tag{4.23}$$

Here, $\alpha \in (0,1)$ is given. The l.h.s. of (4.20) is called the *coverage probability*. Suppose $\Theta \subset I\!R^p$ for some $p$. $S^*$ is a *minimum volume* $1-\alpha$ *confidence set* if it satisfies (4.23) and for any other $1-\alpha$ confidence set $S(X)$,

$$E_\theta[m^p(S^*(X))] \;\leq\; E_\theta[m^p(S(X))] \quad , \qquad \text{for all } \;\theta \in \Theta \;.$$

Show how this notion of optimality of a confidence set can be expressed in terms of decision theory.

**Remark:** In practice, confidence sets are usually attached to point estimates as a measure of accuracy rather than as a separate kind of estimate in and of itself.

## 4.2 Sufficient Statistics.

In this section, we study statistics (more properly, $\sigma$–fields) which contain all pertinent information about a parameter. Here, a *statistic* is a mapping $T = h(X)$ of the observable **which doesn't depend on the unknown parameter**. For instance, if $X$ is real valued and $\theta$ is an unknown location parameter, then $X - \theta$ is not a statistic, but if $\theta_0$ is a fixed, known real number, then $X - \theta_0$ is a statistic. Since a statistic $T$ is a function of $X$, it follows that $\sigma(T) \subset \sigma(X)$. If $\sigma(T) \neq \sigma(X)$, i.e. if the inclusion is proper, then we will have lost "probabilistic" information if we only know $T$ but not $X$ (i.e. there will be events $E \in \sigma(X)$ for which we will not know if $E$ occurred or not). However, we may not have lost "statistical information" in that (intuitively speaking), we may still be able to do as well making inferences about an unknown parameter knowing only $T$. Further, since $T$ is in some sense "simpler" than $X$ (i.e. $\sigma(T)$ is smaller than $\sigma(X)$), if our only interest is in statistical inference, we have achieved some reduction of the data by going to $T$ and forgetting $X$. The following definition will make precise what we have been saying, in that a statistic $T$ is sufficient precisely when it contains as much "statistical information" about the unknown parameter as the original observable.

**Definition 4.2.1** *Let $X : (\Omega, \mathcal{F}) \longrightarrow (\Xi, \mathcal{G})$ be a random observable from a model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$. A statistic $T = T(X)$ is called* sufficient *for $\mathbf{P}$ (or for $\theta$) iff there is a version of $\{Law_\theta[X|T = t] : t \in T(\Xi)\}$ which is independent of $\theta$, i.e.*

$$Law_\theta[X|T = t] \;=\; Law_{\theta'}[X|T = t]$$

*for all $\theta$, $\theta' \in \Theta$, and we will write simply $Law[X|T = t]$.*

□

Intuitively, the reason $T$ contains as much "statistical information" about the parameter as $X$ is that we can recover a probabilistic replica of $X$ when we know $T$, using a two stage experiment. Given an observed value $t$ of $T$, we generate a random object (say $X^*$) from the conditional distribution $\mathrm{Law}[X|T = t]$ (note that we don't need to know $\theta$ to do this), and by the two stage experiment theorem (Theorem 1.5.10), $X^*$ has the same distribution as $X$, so we can use $X^*$ as if it were $X$. We will make these notions more precise using decision theory shortly, but first it is instructive to consider some examples.

**Example 4.2.1** Let $X_1, \ldots, X_n$ be i.i.d. with common distribution in $\{P_\theta : \theta \in \Theta\}$ and assume the common c.d.f. is continuous. We claim that the order statistics $\underline{Y} = \mathbf{Sort}(\underline{X})$ are sufficient. By Theorem 2.5.1,

$$\mathrm{Law}_\theta[\underline{X}|\underline{Y} = \underline{y}] \;=\; \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} \delta_{\tilde{\pi}\underline{y}} \ .$$

Notice that the r.h.s. doesn't depend on $\theta$. Thus, if we were only given the ordered sample, then we could reconstruct a probabilistic replica of the original sample by randomly permuting the ordered sample. Notice that we have not "reduced" the data much by going to the order statistics since they have the same dimensionality as the original data vector, but belong to a smaller subset of $I\!R^n$, namely $I\!P^n$.

□

**Example 4.2.2** Let $X_1, \ldots, X_n$ be i.i.d. with common distribution $Unif[a,b]$ where $a < b$ are unknown. We know from the previous example that the order statistics are sufficient, but here (with these specific distributional assumptions) we can make a further reduction. It is reasonable to think that the only observations which contain information about $a$ and $b$ are the minimum and maximum. Now by equation (2.121),

$$\begin{aligned} &\text{Law}_{(a,b)}[X_{(2)}, \ldots, X_{(n-1)} | (X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})] \qquad (4.24) \\ &= \text{Law}[X_{(2)}, \ldots, X_{(n-1)} | (X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})] \end{aligned}$$

since the l.h.s. has the same distribution as the order statistics from a random sample of size $n-2$ from $Unif[x_{(1)}, x_{(n)}]$. Thus, given $(X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})$, we can generate a probabilistic replica of the remaining order statistics $X_{(2)}, \ldots, X_{(n-1)}$, and as in Example 4.2.1, generate a probabilistic replica of the original sample.

Reasoning more formally, if $A \subset I\!R^n$ then by the law of successive conditioning (Theorem 1.5.7 (g)),

$$\begin{aligned} &P_{(a,b)}[\underline{X} \in A | (X_{(1)}, X_{(n)})] = \\ &E_{(a,b)}[\, P_{(a,b)}[\, \underline{X} \in A \,|\, \mathbf{Sort}(\underline{X})] \,|\, (X_{(1)}, X_{(n)}) \,] \ . \qquad (4.25) \end{aligned}$$

From Example 4.2.1, we know

$$\begin{aligned} &P_{(a,b)}[\, \underline{X} \in A \,|\, \mathbf{Sort}(\underline{X}) = (x_{(1)}, x_{(2)}, \ldots, x_{(n)}) \,] \\ &= P[\, \underline{X} \in A \,|\, \mathbf{Sort}(\underline{X}) = (x_{(1)}, x_{(2)}, \ldots, x_{(n)}) \,] \ . \end{aligned}$$

Now we use the last result in Theorem 1.5.6 to compute the conditional expectation on the r.h.s. of (4.25). This gives

$$\begin{aligned} &P_{(a,b)}[\, \underline{X} \in A \,|\, (X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)}) \,] = \\ &\int_{R^{n-2}} P[\, \underline{X} \in A \,|\, \mathbf{Sort}(\underline{X}) = (x_{(1)}, x_{(2)}, \ldots, x_{(n)}) \,] \\ &dP_{(X_{(2)}, \ldots, X_{(n-1)}) | (X_{(1)}, X_{(n)})}[(x_{(2)}, \ldots, x_{(n-1)}) | (x_{(1)}, x_{(n)})] \end{aligned}$$

In the last formula, we used (4.24) to eliminate the dependence on the unknown parameter $(a,b)$ in the conditional distribution w.r.t. which the integral is taken,

i.e. the conditional distribution of $(X_{(2)}, \ldots, X_{(n-1)})$ given $(X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})$. Since the r.h.s. of the last display is independent of the unknown parameter, it follows that $T = (X_{(1)}, X_{(n)})$ is sufficient for $(a, b)$. Note that in this case, unlike Example 4.2.1, we have achieved some reduction in dimensionality (if $n > 2$) in going from the original $n$-dimensional observable to the 2-dimensional sufficient statistic.

□

**Example 4.2.3** Let $X_1, \ldots, X_n$ be i.i.d. from $N(\mu, 1)$ where $\mu \in \mathbb{R}$ is unknown. We will show that the *sample mean*

$$\bar{X} \;=\; \frac{1}{n} \sum_{i=1}^{n} X_i$$

is a sufficient statistic for $\mu$. Note that the joint distribution $\mathrm{Law}_\mu[\, X_1, \ldots, X_n, \bar{X}]$ has a singular normal distribution on $\mathbb{R}^{n+1}$ since it satisfies the linear constraint defining $\bar{X}$. One can find the joint distribution $\mathrm{Law}_\mu[\, X_1, \ldots, X_{n-1}, \bar{X}]$ on $\mathbb{R}^n$ easily enough, obtain its Lebesgue density, and obtain the conditional density of $X_1, \ldots, X_{n-1}$, given $\bar{X}$, show it is independent of $\mu$, and observe that the joint conditional distribution of $X_1, \ldots, X_{n-1}, X_n$, can be obtained from that of $X_1, \ldots, X_{n-1}$, since $X_n$ is a function of $X_1, \ldots, X_{n-1}$, and $\bar{X}$. However, we use a trick which gives a simpler proof. We claim that the random $n$-vector

$$\underline{Y} \;=\; (X_1 - \bar{X}, \ldots, X_n - \bar{X}) \tag{4.26}$$

is independent of $\bar{X}$. To this end, notice that the random $(n+1)$-vector

$$\underline{W} \;=\; (\underline{Y}, \bar{X}) \;=\; A\underline{X}$$

where the $(n+1) \times n$ matrix $A$ is given by

$$A \;=\; \begin{bmatrix} 1-1/n & -1/n & -1/n & \ldots & -1/n \\ -1/n & 1-1/n & -1/n & \ldots & -1/n \\ -1/n & -1/n & 1-1/n & \ldots & -1/n \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ -1/n & -1/n & -1/n & \ldots & 1-1/n \\ 1/n & 1/n & 1/n & \ldots & 1/n \end{bmatrix}$$

$$= \begin{bmatrix} I - (1/n)J \\ (1/n)\underline{1}' \end{bmatrix}$$

where $\underline{1}$ denotes an $n$-vector of all 1's and $J$ denotes an $n \times n$ matrix of all 1's. Note that $\underline{1}\underline{1}' = J$. Also, $\underline{1}'\underline{1} = n$, which also implies that $J\underline{1} = (\underline{1}\underline{1}')\underline{1} = \underline{1}(\underline{1}'\underline{1})$

$= n\underline{1}$, by the associative law of matrix multiplication. Then since $\underline{X} \sim N(\mu\underline{1}, I)$, we have by $\underline{Y} \sim N(\mu A\underline{1}, AIA')$. This follows by a moment generating function calculation (left as an exercise). Let $\underline{0}$ be an $n$-vector of all 0's. Now

$$E_\mu[\underline{W}] \;=\; \mu A\underline{1} \;=\; \mu \begin{bmatrix} \underline{1} - (1/n)J\underline{1} \\ (1/n)\underline{1}'\underline{1} \end{bmatrix} \;=\; \begin{bmatrix} \underline{0} \\ \mu \end{bmatrix}$$

Also,

$$\mathrm{Cov}_\mu[\underline{W}] \;=\; AIA' \;=\; AA' \;=\; \begin{bmatrix} (I - (1/n)J)^2 & (I - (1/n)J)\underline{1} \\ \underline{1}'(I - (1/n)J) & (1/n)^2\underline{1}'\underline{1} \end{bmatrix}$$

$$= \begin{bmatrix} (I - (1/n)J)^2 & \underline{0} \\ \underline{0}' & 1/n \end{bmatrix}$$

It is not important to what follows, but one can check that

$$(I - (1/n)J)^2 \;=\; I - (1/n)J \tag{4.27}$$

Since $\mathrm{Cov}_\mu[\underline{Y}, \bar{X}] = \underline{0}$ and they have a joint normal distribution, they are independent (Exercise 4.2.6 (c)) as claimed.

From this and Exercise 4.2.2, it follows that

$$\mathrm{Law}_\mu[\underline{Y}|\bar{X} = \bar{x}] \;=\; \mathrm{Law}_\mu[\underline{Y}] \;=\; N(\underline{0}, I - (1/n)J)$$

Since $\underline{X} = \underline{Y} + \bar{X}\underline{1}$,

$$\mathrm{Law}_\mu[\underline{X}|\bar{X} = \bar{x}] \;=\; \mathrm{Law}_\mu[\underline{Y} + \bar{X}\underline{1}|\bar{X} = \bar{x}]$$

Now by Exercise 1.5.12,

$$\mathrm{Law}_\mu[(\underline{Y}, \bar{X})|\bar{X} = \bar{x}] \;=\; \mathrm{Law}_\mu[\underline{Y}|\bar{X} = \bar{x}] \times \delta_{\bar{x}}$$

$$= N\left( \begin{bmatrix} \underline{0} \\ \bar{x} \end{bmatrix}, \begin{bmatrix} I - (1/n)J & \underline{0} \\ \underline{0}' & 0 \end{bmatrix} \right)$$

The last equation is clear since $\delta_{\bar{x}} = N(\bar{x}, 0)$. Now since $\underline{X} = B(\underline{Y}, \bar{X})$ where $B$ is $n \times (n+1)$ and is given by $B = [\, I\, \underline{1}\,]$, we have by Exercise 4.2.6 (a),

$$\mathrm{Law}_\mu[\underline{X}|\bar{X} = \bar{x}] \;=\; N(\bar{x}\underline{1}, I - (1/n)J)$$

which is independent of $\mu$. Hence, $\bar{X}$ is sufficient for $\mu$, as claimed. Notice that for this problem, we have reduced the "statistical information" to a one dimensional sufficient statistic.

□

### 4.2.1 Rao–Blackwell Theorem.

Now we turn to the problem of a decision theoretic justification of the notion that a sufficient statistic contains all of the "statistical information". Given a class $\Delta$ of allowable decision rules and a loss function $L(\theta, a)$ where $a$ is an element of the action space $A$, we say $\Delta_0 \subset \Delta$ is a $\Delta$–*essentially complete class* if for every $\delta \in \Delta$ there is a $\delta_0 \in \Delta_0$ such that $\delta_0 \preceq \delta$. Thus, when looking for optimal decision rules, it is enough to look in the subclass of rules $\Delta_0$ (for given one outside of the $\Delta$–essentially complete class $\Delta_0$, we could find one in $\Delta_0$ as good). We say $\Delta_0 \subset \Delta$ is a $\Delta$–*complete class* iff it is $\Delta$-essentially complete and for every $\delta \in \Delta \setminus \Delta_0$ there is a $\delta_0 \in \Delta_0$ such that $\delta_0 \ll \delta$. (Note: we will not distinguish between decision rules that are equal except on a set of $X$ values which has $P_\theta$ measure 0 for all $\theta$. When we say $\delta \in \Delta \setminus \Delta_0$ we mean there is no $\delta_1 \in \Delta_0$ such that $\delta(X) = \delta_1(X)$ except on a set which has $P_\theta$ measure 0 for all $\theta$.) Thus, we know any optimal decision rule must be in a complete class. We say a decision rule $\delta(X)$ is *based on* a statistic $T$ iff it is a function of $T$, i.e. $\delta(X) = \delta_1(T)$ for some measurable $\delta_1$ (or what is the same, $\sigma(\delta(X)) \subset \sigma(T)$.).

Our previous discussion shows that the class of decision rules based on a sufficient statistic $T$ is a $\Delta$-essentially complete class, provided we are allowed to randomize. Thus, if $\delta(X)$ is an allowable decision rule and we only know $T = t$, then we can generate a probabilistic replica $X^*$ using the conditional distribution $\text{Law}[X|T = t]$ (which doesn't require knowledge of $\theta$), and compute $\delta(X^*)$. Then the risk of this new rule (call it $\delta^*$) is

$$R(\theta, \delta^*) \;=\; E_\theta[E[L(\theta, \delta^*)|T]]$$

by the law of successive conditioning and this

$$= E_\theta[E[L(\theta, \delta(X^*))|T]] \;=\; E_\theta[L(\theta, \delta(X^*))] \;=\; E_\theta[L(\theta, \delta(X))] \;=\; R(\theta, \delta)$$

since $\text{Law}_\theta[X^*] = \text{Law}_\theta[X]$ for all $\theta$.

In fact, we can use sufficient statistics to improve decision rules under certain assumptions. Assume the action space $A$ is a convex subset of $\mathbb{R}^k$ for some $k$. We say the *loss function* $L$ is convex *iff* $L(\theta, \cdot)$ *is a convex function on* $A$ *for every* $\theta \in \Theta$. *We say the* loss function $L$ *is strictly convex* iff $L(\theta, \cdot)$ is a strictly convex function on $A$ for every $\theta \in \Theta$. We say the space of allowable decision rules $\Delta$ is *closed under conditional expectation* if for any $\delta \in \Delta$ we have $E[\delta(X)|T] \in \Delta$ for any sufficient statistic $T$ (if $T$ is not a sufficient statistic, then $E[\delta(X)|T]$ may depend on $\theta$). Here, when we write $E[\delta(X)|T] \in \Delta$, we mean that if $h(t) = E[\delta(X)|T = t]$, then the decision rule $h \circ T \in \Delta$. Note that $h \circ T$ is a function defined on observation space taking values in the space of allowable actions.

**Theorem 4.2.1 (Rao-Blackwell Theorem.)** *Suppose a decision problem has convex action space, convex loss* $L$*, and allowable decision rules* $\Delta$ *closed under*

*conditional expectation. If $T$ is a sufficient statistic, the the class of nonrandomized rules $\Delta_0 = \{E[\delta|T] : \delta \in \Delta\}$ is a $\Delta$-essentially complete class. Furthermore, if $L$ is strictly convex, then $\Delta_0$ is a $\Delta$-complete class.*

**Proof.** (a) Given $\delta$, put $\delta^*(T) = E[\delta(X)|T]$. By the conditional version of Jensen's inequality (Theorem 2.4.3),

$$\begin{aligned} R(\theta, \delta^*) &= E_\theta[L(\theta, E[\delta(X)|T]] \\ &\leq E_\theta E[L(\theta, \delta(X))|T] \\ &= E_\theta L(\theta, \delta(X)) \;=\; R(\theta, \delta) \end{aligned}$$

so $\delta^* \preceq \delta$.

If $L$ is strictly convex, then strict inequality holds in the above unless $\text{Law}[\delta(X)|T]$ is degenerate with $\text{Law}_\theta[T]$-probability 1. Thus, $\delta^* \ll \delta$ unless $\text{Law}[\delta(X)|T]$ is degenerate with $\text{Law}_\theta[T]$-probability 1 for all $\theta \in \Theta$, i.e. $\text{Law}[\delta(X)|T = t]$ is a unit point mass at some $f(t)$ say, with $\text{Law}_\theta[T]$-probability 1 for all $\theta \in \Theta$. Thus, we see that $\delta(X)$ is already a function of $T$, namely $f(T)$, except on an event which has $P_\theta$ measure 0 for all $\theta$. But we may redefine $\delta(X)$ on this set so that it is a function of $T$, and thus in $\Delta_0$.

□

**Remarks 4.2.1** (a) The process of applying conditional expectation w.r.t. a sufficient statistic to a decision rule is sometimes known as *Rao-Blackwellization*. Thus, the theorem shows that in the context of a convex loss and convex action space, nothing is lost by Rao-Blackwellization. If the loss is strictly convex and the decision rule is not already a function of the sufficient statistic, then one uniformly improves the decision rule (in that the risk is made smaller at all values of $\theta$) through Rao-Blackwellization, so it is unwise to use a decision rule which is not a function of a sufficient statistic in this case.

(b) Referring back to the estimation problem discussed in Section 1, assume $\Theta$ is an interval in $\mathbb{R}$. The class of unbiased estimators is closed under conditional expectation. Indeed, if $\delta(X)$ is unbiased and $\delta^*(T)$ is obtained by Rao-Blackwellization, then by the law of successive conditioning,

$$E_\theta[\delta^*(T)] \;=\; E_\theta E[\delta(X)|T] \;=\; E_\theta[\delta(X)] \;=\; \theta\,.$$

(c) Under the same setup as (b), it is easy to see that that squared error loss is strictly convex, so one improves the variance of an unbiased estimator by Rao-Blackwellization. For instance, suppose that we wish to estimate $\mu$ from a random sample which is from the $N(\mu, 1)$ distribution. Then we already know from Example 4.2.1 that the order statistics are sufficient. It is easy to see that $X_1$ is an unbiased estimator of $\mu$. From Theorem 2.5.1, if $h(\underline{x}) = x_1$ is projection of a vector $\underline{x}$ onto its first coordinate, then

$$E[X_1|\mathbf{Sort}(\underline{X})] \;=\; \frac{1}{n!} \sum_{\pi \in \mathbf{Perm}} h(\tilde{\pi}\mathbf{Sort}(\underline{X})) \tag{4.28}$$

$$= \frac{1}{n!}\sum_{i=1}^{n}\sum_{\pi:h(\tilde{\pi}\underline{y})=y_i} h(\tilde{\pi}\mathbf{Sort}(\underline{X}))$$

$$= \frac{1}{n!}\sum_{i=1}^{n}(n-1)!X_{(i)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}X_{(i)} \;=\; \frac{1}{n}\sum_{i=1}^{n}X_i \;=\; \bar{X}\ ,$$

which is the sample mean. Note that we used that for each $i$ there are $(n-1)!$ permutations $\pi$ such that $h(\tilde{\pi}\mathbf{Sort}(\underline{X})) = X_{(i)}$, so each $X_i$ appears $(n-1)!$ times in the summation over $\pi$. Of course, the same result would be obtained if we Rao-Blackwellized using the statistic $\bar{X}$ which we know from Example 4.2.3 is sufficient.

In general, we see that if we believe the data are i.i.d. from a continuous distribution, then we need only know the order statistics to do inferences, and in fact any procedure which is not a function of the order statistics can be improved on, at least from the point of view of risk under a convex loss. Thus, if someone proposes a procedure for such data which depend on the order, we should be suspicious. In general one will find such procedures which depend on the order of the data proposed only for the purpose of computational efficiency in large data sets. When such a rule is proposed, it is usually necessary to show that for such large data sets the order dependent procedure is not much different from a procedure which is independent of order.

□

### 4.2.2 The Factorization Theorem.

In Examples 4.2.1 through 4.2.3, we found sufficient statistics somewhat laboriously by "guessing" that a particular statistic was sufficient and computing the conditional distribution of the data given the statistic. We would like to have a method for finding sufficient statistics which is more automatic, and does not require calculation of Law$[X|T=t]$. The next result provides this.

**Theorem 4.2.2 (Fisher-Neyman Factorization Theorem.)** *Assume the observation space $(\Xi, \mathcal{G})$ is Euclidean, say $\Xi \subset \mathbb{R}^n$, and suppose $\{P_\theta : \theta \in \Theta\}$ is dominated by a $\sigma$–finite Borel measure $\mu$. Then a $\mathbb{R}^k$ valued statistic $T = T(X)$ is sufficient for $\theta$ if and only if there are a nonnegative Borel functions $g(\cdot,\theta)$ and $h$ such that the densities factor*

$$f_\theta(x) \;=\; g(T(x),\theta)h(x)\ , \tag{4.29}$$

*i.e. all of the functional dependence of the density on the parameter can be concentrated in a factor that depends only on $T$.*

**Partial Proof.** We will prove here the theorem only in the case where the observable is discrete (so that ordinary conditional probability calculations can be made). This will hopefully provide some insight into the result. The general proof is given in the final subsection of this section. Suppose that $f_\theta(x)$ is the density w.r.t counting measure on $\{x_1, x_2, \dots\} = \Xi$ which may be finite or infinite. For any statistic $T$, the set of possible values is also discrete, say $\{t_1, t_2, \dots\}$. If the factorization (4.29) holds then

$$P_\theta[T = t_j] \;=\; \sum_{\{i: x_i \in T^{-1}(t_j)\}} P_\theta[X = x_i]$$

$$= \sum_{\{i: x_i \in T^{-1}(t_j)\}} f_\theta(x_i) \;=\; \sum_{\{i: x_i \in T^{-1}(t_j)\}} g(T(x_i), \theta) h(x_i)$$

$$= \sum_{\{i: x_i \in T^{-1}(t_j)\}} g(t_j, \theta) h(x_i) \;=\; g(t_j, \theta) \sum_{\{i: x_i \in T^{-1}(t_j)\}} h(x_i)$$

Hence, assuming $P_\theta[T = t_j] > 0$, which is the only case we care about, we have

$$P_\theta[X = x_k | T = t_j] \;=\; \frac{P_\theta[X = x_k \& T = t_j]}{P_\theta[T = t_j]} \;.$$

Now the event whose probability is in the numerator can be simplified

$$[X = x_k \;\&\; T = t_j] \;=\; \begin{cases} [X = x_k] & \text{if } T(x_k) = t_j, \\ \\ \emptyset & \text{otherwise.} \end{cases}$$

Note that the value of $\theta$ doesn't enter into this. Assuming $T(x_k) = t_j$ then for all $\theta$,

$$\begin{aligned} P_\theta[X = x_k | T = t_j] \quad &= \quad \frac{g(T(x_k), \theta) h(x_k)}{g(t_j, \theta) \sum_{\{i: x_i \in T^{-1}(t_j)\}} h(x_i)} \\ &= \quad \frac{h(x_k)}{\sum_{\{i: x_i \in T^{-1}(t_j)\}} h(x_i)} \end{aligned}$$

which doesn't depend on $\theta$. This shows existence of the factorization implies sufficiency of $T$, for the discrete case.

Conversely, assume $T$ is sufficient. Then if $t_j = T(x_i)$, we have

$$\begin{aligned} f_\theta(x_i) \;=\; P_\theta[X = x_i] \quad &= \quad \sum_k P_\theta[X = x_i | T = t_k] P_\theta[T = t_k] \\ &= \quad P_\theta[X = x_i | T = t_j] P_\theta[T = t_j] \;=\; P[X = x_i | T = t_j] P_\theta[T = t_j] \end{aligned}$$

where we have used sufficiency of $T$ at the last step. Now writing

$$g(t_j, \theta) \;=\; P_\theta[T = t_j]$$

$$h(x_i) \;=\; P[X = x_i | T = t_j]$$

(recall $t_j = T(x_i)$ so the latter depends only on $x_i$), we have

$$f_\theta(x) \;=\; g(T(x), \theta) h(x)$$

for all $x \in \Xi$. This completes the proof for the discrete case.

□

**Corollary 4.2.3** *Let $\{P_\theta : \theta \in \Theta\}$ be an exponential family on a Euclidean space $(\Xi, \mathcal{G})$ with dominating Borel measure $\mu$ and densities*

$$f_\theta(x) \;=\; \exp[\eta(\theta)' T(x) - B(\theta)] h(x) \; .$$

*Then $T$ is sufficient for $\theta$.*

**Proof.** Take $g(T(x), \theta) = \exp[\eta(\theta)' T(x) - B(\theta)]$ in the factorization theorem.

□

Now we revisit Examples 4.2.1 through 4.2.3 and show how the same results may be obtained via the factorization theorem. For Example 4.2.1, assume $X_1$, $X_2$, ..., $X_n$ are i.i.d. with Lebesgue density $f_\theta$. Then the joint density is

$$f_{\theta, \underline{X}}(\underline{x}) \;=\; \prod_{i=1}^{n} f_\theta(x_i) \;=\; \prod_{i=1}^{n} f_\theta(x_{(i)})$$

where $(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = \mathbf{Sort}(\underline{x}) = T(\underline{x})$. The last equation follows since it makes no difference which order we take the product in by commutativity and associativity of ordinary multiplication. Thus,

$$f_{\theta, \underline{X}}(\underline{x}) \;=\; f_{\theta, \underline{X}}(T(\underline{x}))$$

and we may take the factor $h \equiv 1$ in the factorization theorem.

For Example 4.2.2 where the $X_i$'s are i.i.d. $Unif[a, b]$ and the unknown parameter is $\theta = (a, b)$, we have

$$f_{\theta, \underline{X}}(\underline{x}) \;=\; \prod_{i=1}^{n} (b - a)^{-1} I_{(a,b)}(x_i) \; .$$

Now the product on the r.h.s. is nonzero just in case all $x_i$'s are between $a$ and $b$, and this is the same as $\min\{x_1, x_2, \ldots, x_n\} = x_{(1)}$ and $\max\{x_1, x_2, \ldots, x_n\} = x_{(n)}$ between $a$ and $b$, so

$$f_{\theta, \underline{X}}(\underline{x}) \;=\; (b - a)^{-n} I_{(a,b)}(x_{(1)}) I_{(a,b)}(x_{(n)}) \; ,$$

and so with $T(\underline{x}) = (x_{(1)}, x_{(n)})$ we may take $g(T(\underline{x}), \theta) = f_\theta(\underline{x})$ and $h \equiv 1$ in the factorization theorem.

Finally, for Example 4.2.3 where the $X_i$'s are i.i.d. $N(\mu, 1)$ where $\theta = \mu$ is the unknown parameter, we have that the joint Lebesgue density is

$$f_\theta \;=\; (2\pi)^{-n/2} \exp\left[ -\frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right]$$

$$= \exp\left[ n\mu \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) - \frac{n}{2}\mu^2 \right] (2\pi)^{-n/2} \exp\left[ -\frac{1}{2} \sum_{i=1}^{n} x_i^2 \right]$$

which is an exponential family with $\eta(\mu) = n\mu$ and $T(\underline{x}) = (1/n) \sum x_i$. Thus, $T(\underline{X}) = \bar{X}$ is sufficient.

The student will get the experience of applying the factorization theorem to find sufficient statistics in several of the exercises at the end of the section. In general, if one has a "nice" $p$-dimensional parameter and the parameterization is identifiable, then one will be able to get a $p$-dimensional sufficient statistic, although there are many exceptions to this rule. One should always seek to reduce the sufficient statistic to one as "small" as possible in the sense of having the smallest possible generated $\sigma$–field. Thus, for an exponential family one should seek to put it into minimal form before reading off the sufficient statistic. We will next consider formalizing these notions.

### 4.2.3 Minimal Sufficiency.

**Definition 4.2.2** *If $T$ is a sufficient statistic for a family $\mathbf{P}$, then $T$ is* minimal sufficient *if and only if for any other sufficient statistic $S$, $T = h(S)$ $\mathbf{P}$-a.s.*

*By $\mathbf{P}$-a.s., we mean $P$-a.s. for all $P \in \mathbf{P}$. By Theorem 1.5.1, $T = h(S)$ is essentially the same as $\sigma(T) \subset \sigma(S)$ (actually $\bar{\sigma}(T) \subset \bar{\sigma}(S)$ where $\bar{\sigma}(T)$ denotes the completion of $\sigma(T)$ w.r.t. $\mathbf{P}$).*

□

Our next result provides useful criteria for checking for minimal sufficiency.

**Proposition 4.2.4** *Let $\mathbf{P}$ be a family of statistical models on a Euclidean space.*

*(a) Suppose $\mathbf{P_0} \subset \mathbf{P}$ is a submodel, and that every $\mathbf{P_0}$ null set is also a $\mathbf{P}$ null set (i.e. $P(N) = 0$ for all $P \in \mathbf{P_0} \Rightarrow P(N) = 0$ for all $P \in \mathbf{P}$). If $T$ is minimal sufficient for $\mathbf{P_0}$ and sufficient for $\mathbf{P}$, then $T$ is minimal sufficient for $\mathbf{P}$.*

*(b) Suppose $\mathbf{P}$ is finite with densities $f_i$, $0 \le i \le k$, and suppose*

$$\{x : f_i(x) > 0\} \subset \{x : f_0(x) > 0\} \quad , \quad 1 \le i \le k \ . \tag{4.30}$$

*Then the vector of likelihood ratios*

$$T(x) \;=\; \left( \frac{f_1(x)}{f_0(x)}, \frac{f_2(x)}{f_0(x)}, \ldots, \frac{f_k(x)}{f_0(x)} \right)$$

*is minimal sufficient for* $\mathbf{P}$.

*(c) If* $h$ *is one to one and bimeasurable, and if* $T$ *is minimal sufficient, then* $h(T)$ *is minimal sufficient.*

**Proof.** (a) Suppose $T$ is minimal sufficient for $\mathbf{P_0}$ and sufficient for $\mathbf{P}$. Then $U$ sufficient for $\mathbf{P} \Rightarrow U$ sufficient for $\mathbf{P_0}$ (check the definition of sufficiency for $\mathbf{P_0}$) $\Rightarrow T = h(U)$ for some $h$. Since $T$ is sufficient for $\mathbf{P}$ and is a function of any other sufficient statistic, it follows that $T$ is minimal sufficient for $\mathbf{P}$.

(b) Let $T$ be the vector of likelihood ratios given in the statement of part (b) with $i^{\text{th}}$ component $T_i = f_i(x)/f_0(x)$. According to the factorization theorem, if $U$ is sufficient, then $f_i(x) = g_i(U(x))h(x)$, and so

$$T_i(x) \;=\; \frac{f_i(x)}{f_0(x)} \;=\; \frac{g_i(U(x))}{g_0(U(x))} \;,$$

which shows that $T$ is a function of $U$. To show that $T$ is sufficient, note that

$$f_i(x) \;=\; T_i(x) f_0(x) \;=\; g_i(T(x))h(x)$$

where $g_i(t) = t_i$ is a projection map ($t_i$ is the $i^{\text{th}}$ component of $t$) and $h = f_0$. Hence, $T$ is sufficient by the factorization theorem.

(c) Now, $\sigma(h(T)) = \sigma(T)$ since $h$ is one to one so $P_\theta[X \in B|h(T)] = P_\theta[X \in B|T] = P[X \in B|T]$, and we can find a version of the conditional distribution of $X$ given $h(T)$ which doesn't depend on $\theta$, i.e. $h(T)$ is sufficient. If $T$ is minimal sufficient, then for $U$ sufficient we have $T = h_1(U)$ for some $h_1$ and so $h(T) = h(h_1(U)) = (h \circ h_1)(U)$. This shows $h(T)$ is minimal sufficient.

□

**Proposition 4.2.5** *Let* $\mathbf{P}$ *be an exponential family on a Euclidean space with densities*

$$f_\theta(x) \;=\; \exp\left[ \underline{\eta}(\theta)' \underline{T}(x) \;-\; B(\theta) \right] h(x) \;, \tag{4.31}$$

*where* $\underline{\eta}$ *and* $\underline{T}$ *are* $p$*-dimensional. Suppose there exists* $\{\theta_0, \theta_1, \ldots, \theta_p\} = \Theta_0 \subset \Theta$ *such that the vectors*

$$\underline{\zeta}_i \;=\; \underline{\eta}(\theta_i) - \underline{\eta}(\theta_0) \quad , \quad 1 \le i \le p$$

*are linearly independent in* $\mathbb{R}^p$. *Then* $\underline{T} = \underline{T}(X)$ *is minimal sufficient for* $\theta$. *In particular, if (4.31) is full rank, then* $\underline{T}$ *is minimal sufficient.*

**Proof.** By Remark 2.3.1 (c), $\{x : f_\theta(x) > 0\}$ doesn't depend on $\theta$. Taking the finite subfamily $\Theta_0$, the vector of likelihood ratios as in Proposition 4.2.4 (b) has $i$'th component

$$\begin{aligned} R_i &= \frac{f_{\theta_i}(x)}{f_{\theta_0}(x)} \\ &= \exp\left[\underline{\zeta}_i'\underline{T}(x) - (B(\theta_i) - B(\theta_0))\right] . \end{aligned}$$

Now $\underline{\zeta}_i'\underline{T}(x) = \log R_i - (B(\theta_0) - B(\theta_i))$, so $Z\underline{T}$ is a one to one function of $\underline{R}$ where $Z$ is the $p \times p$ matrix with $i$'th row equal to $\underline{\zeta}_i$, and $Z\underline{T}$ is minimal sufficient by Proposition 4.2.4 (c). The assumptions of linear independence of the $\underline{\zeta}_i$ guarantee that $Z$ is nonsingular, so it follows that $\underline{T}$ is minimal sufficient.

If the family is full rank, then there is a $\theta_0$ such that there is a neighborhood $B(\underline{\eta}(\theta_0), \epsilon)$ contained in $\underline{\eta}(\Theta)$, and one can find $p$ values $\underline{\eta}(\theta_1)$, $\underline{\eta}(\theta_2)$, ..., $\underline{\eta}(\theta_p)$ in $B(\underline{\eta}(\theta_0), \epsilon)$ such that the corresponding $\zeta_i$ as defined above are linearly independent.

□

**Example 4.2.4** Suppose $P_\theta$ with $\theta = (b, s)$, $b \in I\!R$ and $s > 0$, is the location-scale family generated by the Lebesgue density

$$f(x) = Ce^{-x^4}, \quad x \in I\!R,$$

that is

$$f_\theta(x) = Cs^{-1} \exp\left[\left(\frac{x-b}{s}\right)\right] .$$

Here, $C$ is a positive constant which makes $f$ integrate to 1. Suppose $X_1$, $X_2$, ..., $X_n$ are i.i.d. from $f_\theta$. The joint density is

$$\begin{aligned} f_\theta(\underline{x}) = \exp\Bigg[ &-\frac{1}{s^4}\sum x_i^4 + \frac{4b}{s^4}\sum x_i^3 - \frac{6b^2}{s^4}\sum x_i^2 \\ &+ \frac{4b^3}{s^4}\sum x_i - n\frac{b^4}{s^4} - n\log s - n\log C\Bigg] \end{aligned}$$

Take $\theta_0 = (0, 1)$ and then

$$\underline{\eta}(\theta) - \underline{\eta}(\theta_0) = \frac{1}{s^4}\begin{bmatrix} s^4 - 1 \\ 4b \\ -6b^2 \\ 4b^3 \end{bmatrix}$$

Now let $s_i = 2^{1/4}$ for $i = 1, 2, 3, 4$ and $b_i = i - 1$, $i = 1, 2, 3, 4$. The the matrix $Z$ in Proposition 4.2.5 is

$$Z = \frac{1}{2}\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 4 & -6 & 4 \\ 1 & 8 & -24 & 32 \\ 1 & 12 & -54 & 108 \end{bmatrix}$$

Direct computation shows $\det(Z) = -576/16 \neq 0$. Hence, $T = (\sum X_i^4, \sum X_i^3, \sum X_i^2, \sum X_i)$ is minimal sufficient. Note that $T$ is 4 dimensional whereas $\theta$ is 2 dimensional. Also, the family is not full rank (Exercise 4.2.13).

□

**Example 4.2.5** We consider an instance where Proposition 4.2.4 is not applicable. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $Unif[\theta - 1/2, \theta + 1/2]$, $\theta \in I\!R$. Then the joint density w.r.t. $m^n$ is

$$f_\theta(\underline{x}) \;=\; \prod_{i=1}^{n} I_{(\theta-1/2,\theta+1/2)}(x_i)$$

$$=\; I_{(\theta-1/2,\theta+1/2)}(x_{(1)}) I_{(\theta-1/2,\theta+1/2)}(x_{(n)})$$

where $x_{(1)}$ and $x_{(n)}$ are the minimum and maximum of the $x_i$. Thus, $T = (X_{(1)}, X_{(n)})$ is sufficient. If $U$ is any sufficient statistic, then by the Fisher-Neyman factorization theorem,

$$f_\theta(\underline{x}) \;=\; g(U(\underline{x}), \theta) h(\underline{x}) \;.$$

Now for any $\underline{x} \in I\!R^n$ satisfying $\max\{x_i\} - \min\{x_i\} = x_{(n)} - x_{(1)} < 1$, we must have $h(\underline{x}) > 0$ since $f_\theta(\underline{x}) > 0$ for some $\theta$ (namely $\theta = (x_{(n)} + x_{(1)})/2$). For such an $\underline{x}$, $f_\theta(\underline{x}) > 0$ if and only if $\theta \in (x_{(n)} - 1/2,\, x_{(1)} + 1/2)$, so

$$x_{(n)} \;=\; \inf\{\theta + 1/2 : g(U, \theta) > 0\}$$

$$x_{(1)} \;=\; \sup\{\theta - 1/2 : g(U, \theta) > 0\}$$

which shows $(X_{(1)}, X_{(n)})$ is a function of $U(\underline{X})$.

□

### 4.2.4 Proof of Factorization Theorem.

Here we prove the general case of Theorem 4.2.2. First, assume $T$ satisfies the factorization in (4.29), and we will show $T$ is sufficient. We will need the following Lemma, which is proved below:

**Lemma 4.2.6** *Suppose* $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ *is a family of measures dominated by a* $\sigma$*–finite measure* $\mu$*. Then there is a* $\{\theta_1, \theta_2, \ldots\} \subset \Theta$ *and a sequence of positive real numbers* $a_1, a_2, \ldots$ *with* $\sum_n a_n = 1$ *such that*

$$Q \;=\; \sum_n a_n P_{\theta_n}$$

*is a probability measure satisfying* $Q(A) = 0$ *if and only if* $P_\theta(A) = 0$ *for all* $\theta \in \Theta$.

In most examples, one can show directly the existence of such a $Q$. (See Exercises 4.2.10 and 4.2.11.) Clearly $Q \ll \mu$ since $\mu(A) = 0 \Rightarrow P_\theta(A) = 0$ for all $\theta \in \Theta$. Also, by an obvious and easy extension of Proposition 1.4.5(b) (Exercise 4.2.8),

$$\frac{dQ}{d\mu}(x) \;=\; \sum_n a_n \frac{dP_{\theta_n}}{d\mu}(x)$$

$$= \left[ \sum_n a_n g(T(x), \theta_n) \right] h(x) \;=\; \gamma(T(x))h(x) \;,$$

where we have used (4.29) at the last step. By the chain rule for Radon-Nikodym derivatives,

$$\frac{dP_\theta}{dQ}(x) \;=\; \left[ \frac{dP_\theta}{d\mu} \div \frac{dQ}{d\mu} \right](x) \tag{4.32}$$

$$= \frac{g(T(x), \theta)h(x)}{\gamma(T(x))h(x)} \;=\; \frac{g(T(x), \theta)}{\gamma(T(x))} \;=\; \tilde{g}(T(x), \theta) \quad , \quad Q - a.s.$$

We will now show that a version of $\mathrm{Law}_Q[X|T = t]$ is a version of $\mathrm{Law}_\theta[X|T = t](A)$ for all $\theta$, where by $\mathrm{Law}_Q[X|T = t]$ we mean the conditional distribution of $X$ given $T$ under $\mathrm{Law}[X] = Q$, i.e. the conditional distribution of $X$ given $T = t$ computed under the assumption that the true distribution for $X$ is $Q$.

Fix a Borel set $A \subset I\!R^n$, and a $\theta \in \Theta$. Define a Borel measure $\nu$ on $I\!R^n$ by

$$d\nu \;=\; I_A dP_\theta \tag{4.33}$$

Let $\tilde{\nu}$ and $\tilde{P}_\theta$ be $\nu$ and $P_\theta$ restricted to $\sigma(T)$, as in the proof of Theorem 2.3.4. Then as in the proof of that theorem,

$$\frac{d\tilde{\nu}}{d\tilde{P}_\theta} \;=\; E_\theta[I_A(X)|T] \;=\; P_\theta[X \in A|T] \quad , \quad \tilde{P}_\theta \; a.s. \tag{4.34}$$

Let $\tilde{Q}$ be the restriction of $Q$ to $\sigma(T)$. Since $\tilde{g}(T(\cdot), \theta)$ in (4.32) is already $\sigma(T)$ measurable,

$$\frac{d\tilde{P}_\theta}{d\tilde{Q}} \;=\; \tilde{g}(T(\cdot), \theta) \;=\; \frac{dP_\theta}{dQ} \quad , \quad \tilde{Q} \; a.s. \tag{4.35}$$

To explain (4.35), we can obtain the measures of $\sigma(T)$ measurable sets by integrating over the set the $\sigma(T)$ measurable function $\tilde{g}(T(\cdot), \theta)$, so it must be the Radon-Nikodym derivative over $\sigma(T)$. By the chain rule again,

$$\frac{d\tilde{\nu}}{d\tilde{Q}} \;=\; \frac{d\tilde{\nu}}{d\tilde{P}_\theta} \frac{d\tilde{P}_\theta}{d\tilde{Q}} \;=\; E_\theta[I_A(X)|T]\tilde{g}(T(\cdot), \theta) \quad , \quad \tilde{Q} \; a.s. \tag{4.36}$$

and also (see (4.33))

$$\frac{d\nu}{dQ}(x) \;=\; \frac{d\nu}{dP_\theta} \frac{dP_\theta}{dQ}(x) \;=\; I_A(x)\tilde{g}(T(x), \theta) \,, \quad \tilde{Q} \; a.s.$$

Using the same reasoning as in (4.34),

$$\frac{d\tilde{\nu}}{d\tilde{Q}} \;=\; E_Q[I_A(X)\tilde{g}(T,\theta)|T] \quad , \quad \tilde{Q}\ a.s. \tag{4.37}$$

$$= \tilde{g}(T,\theta)E_Q[I_A(X)|T]\,, \quad \tilde{Q}\ a.s.$$

where Theorem 2.3.7(h) was used at the last step. Here, we use $E_Q[\cdot|T]$ to mean computation of conditional expectation under Law$[X] = Q$. Combining (4.36) and (4.37) gives

$$P_Q[X \in A|T]\tilde{g}(T,\theta) \;=\; P_\theta[X \in A|T]\tilde{g}(T,\theta) \quad , \quad \tilde{Q}\ a.s. \tag{4.38}$$

Since $Q$-a.s. is the same as $P_\theta$-a.s. for all $\theta$, we may replace $\tilde{Q}$-a.s. in (4.38) (which implies $Q$-a.s.) by $P_\theta$-a.s., all $\theta$. Also, for all $\theta$, $\tilde{g}(T,\theta) > 0$, $P_\theta$-a.s. since $dP_\theta/dQ = \tilde{g}(T,\theta)$. Hence, for all $\theta$,

$$P_\theta[X \in A|T] \;=\; P_Q[X \in A|T] \quad , \quad P_\theta\ a.s.$$

Thus, a regular conditional probability distribution $P_Q[X \in \cdot|T = t]$ is also a version of the regular conditional probability distribution $P_\theta[X \in \cdot|T = t]$, and $T$ is sufficient.

Now for the converse, suppose $T$ is sufficient and let $Q$ be as above. Denote

$$p(B,t) \;=\; P[X \in B|T = t]$$

Then, for all $\theta$, all $A \in \sigma(T)$, and all $B \in \mathcal{B}_n$,

$$\int_A p(B,T(x))\,dP_\theta(x) \;=\; P_\theta(A \cap B)\ , \tag{4.39}$$

by the defining property (ii) of conditional expectation. (Here, we are using $(\Xi, \mathcal{G}, P_\theta)$ as the underlying probability space. Note that $p(B,T(x))$ is a version of $P[B|T](x) = E[I_B|T](x)$ on this probability space. Hence, $\int_A P[B|T]dP_\theta = \int_A I_B dP_\theta = P_\theta(A \cap B)$ for all $\sigma(T)$ measurable subsets $A$ of $\Xi$.) Since $Q = \sum a_i P_{\theta_i}$, it follow by taking the infinite convex combination of both sides of (4.39) (and applying Exercise 1.2.35)

$$\int_A p(B,T(x))\,dQ(x) \;=\; Q(A \cap B)\ .$$

This shows that $p(B,t)$ can serve as a version of the conditional distribution for $X$ given $T = t$ under Law$[X] = Q$. Let $\tilde{P}_\theta$ and $\tilde{Q}$ denote the restrictions of $P_\theta$ and $Q$ to $\sigma(T)$ as above. Let

$$h_\theta(x) \;=\; \frac{d\tilde{P}_\theta}{d\tilde{Q}}(x)\ .$$

Since $h_\theta$ is $\sigma(T)$ measurable, by Theorem 2.3.1 it is a function of $T$, so there is a $g(\cdot,\theta)$ defined on the range space of $T$ such that

$$h_\theta(x) \;=\; g(T(x),\theta)\ .$$

We will show that $g(\cdot,\theta)$ is also a version of $dP_\theta/dQ$. To this end, put $A=\Xi$ in (4.39) to obtain

$$P_\theta(B) \;=\; \int_\Xi p(B,T(x))\,dP_\theta(x) \;=\; \int p(B,T(x))\,d\tilde{P}_\theta(x)\ . \tag{4.40}$$

The last equation follows since $p(B,T(\cdot))$ is $\sigma(T)$ measurable, so we may integrate w.r.t. $P_\theta$ restricted to $\sigma(T)$ (see the proof of equation (1.61) in Theorem 1.5.2). Now using the Radon-Nikodym derivative of $\tilde{P}_\theta$ w.r.t. $\tilde{Q}$ and Proposition 1.4.2 (a), the last displayed quantity in (4.40) is equal to

$$\int p(B,T(x))g(T(x),\theta)\,d\tilde{Q}(x) \;=\; \int E_Q[I_B(X)|T=T(x)]g(T(x),\theta)\,d\tilde{Q}(x)\ ,$$

where we have used that $p(B,t)$ is a version of $P_Q[X\in B|T=t]$. By Theorem 1.5.7 (h), this equals

$$\int E_Q[g(T(x),\theta)I_B(X)|T=T(x)]\,d\tilde{Q}(x)$$

$$=\int g(T(x),\theta)I_B(X)\,dQ(x) \;=\; \int_B g(T(x),\theta)\,dQ(x)\ ,$$

where the second to last equation follows from the definition of conditional expectation. Thus, $P_\theta(B)$ the last displayed quantity, which shows that $g(\cdot,\theta)$ is a version of $dP_\theta/dQ$.

From this claim and the chain rule (Proposition 1.4.2 (c)),

$$\frac{dP_\theta}{d\mu}(x) \;=\; \left[\frac{dP_\theta}{dQ}\,\frac{dQ}{d\mu}\right](x) \;=\; g(T(x),\theta)\frac{dQ}{d\mu}(x)$$

which shows that the factorization (4.29) holds with $h=dQ/d\mu$.

□

**Proof of Lemma 7.** Without loss of generality, we can assume the dominating measure $\mu$ is finite rather than just $\sigma$–finite (Exercise 4.2.9). Let $\mathbf{Q}$ be the collection of all probability measures $Q$ on $\Xi$ of the form $\sum a_i P_{\theta_i}$ where the $a_i\ge 0$ and $\sum a_i=1$. Then $\mathbf{Q}$ is also dominated by $\mu$ and for any $Q\in\mathbf{Q}$ given by $Q=\sum a_i P_{\theta_i}$, let $q=\sum a_i(dP_{\theta_i}/d\mu)=dQ/d\mu$ denote the Radon-Nikodym derivative of $Q$ w.r.t. $\mu$. We will show that there is a $Q^*\in\mathbf{Q}$ for which $Q^*(A)=0\Rightarrow Q(A)=0$ for all $Q\in\mathbf{Q}$, which implies the lemma since $\mathbf{P}\subset\mathbf{Q}$.

Let $\mathcal{C}$ be the collection of all subsets $C$ of $\Xi$ for which there is some $Q \in \mathbf{Q}$ such that $q > 0$ $\mu$-a.e. on $C$ and $Q(C) > 0$. Note that $\mathcal{C}$ is closed under countable unions since if $C_1$, $C_2$, ..., are all in $\mathcal{C}$ and $Q_1$, $Q_2$, ..., are the corresponding measures in $\mathbf{Q}$, then $Q = \sum 2^{-i} Q_i$ is in $\mathbf{Q}$ and $Q$ has $\mu$-density $q = \sum 2^{-i} q_i$ which is $> 0$ $\mu$-a.e. on $C = \bigcup_i C_i$. Also, if $C \in \mathcal{C}$ and $C_1 \subset C$, then clearly $C_1 \in \mathcal{C}$.

Let $\mu(C_i)$ tend to $\sup\{\mu(C) : C \in \mathcal{C}\} < \infty$ since $\mu$ is finite, and let $Q_i$ denote elements of $\mathbf{Q}$ for which their densities $q_i > 0$ $\mu$-a.e. on $C_i$ and $Q_i(C_i) > 0$. Put $C^* = \bigcup_i C_i$ and $q^* = \sum 2^{-i} q_i$. Denote by $Q^* = \sum 2^{-i} Q_i$. Then $C^* \in \mathcal{C}$ and $Q^* \in \mathbf{Q}$ as noted above, and clearly $\mu(C^*) = \sup\{\mu(C) : C \in \mathcal{C}\}$ (since $\mu(C^*) \geq \mu(C_i) \uparrow \sup\{\mu(C) : C \in \mathcal{C}\}$ and $\mu(C^*) \leq \sup\{\mu(C) : C \in \mathcal{C}\}$ because $C^* \in \mathcal{C}$). Let $A$ be such that $Q^*(A) = 0$, and let $Q \in \mathbf{Q}$. We will show that $Q(A) > 0$ leads to a contradiction, and hence we must have $Q(A) = 0$, which gives the claim above about $Q^*$.

Let $C = \{x : q(x) > 0\}$. Note that $A$ can be decomposed into three disjoint subsets as in

$$A \;=\; [A \cap C^*] \bigcup [A \cap C^{*c} \cap C^c] \bigcup [A \cap C^{*c} \cap C] \,. \tag{4.41}$$

We will show that the $Q$ measure of the first two subsets of $A$ is 0. Now $A \cap C^* \subset A$ so $Q^*(A \cap C^*) = 0$, and hence $\mu(A \cap C^*) = 0$ (because $0 = Q^*(A \cap C^*) = \int_{A \cap C^*} q^*(x) d\mu(x)$ implies $q^* I_{A \cap C^*} = 0$ $\mu$-a.e. by Proposition 1.2.6 (b), but $A \cap C^* \subset C^*$ and $q^* > 0$ on $C^*$ $\mu$-a.e., so we must have $I_{A \cap C^*} = 0$ $\mu$-a.e., i.e. $\mu(A \cap C^*) = 0$). Since $Q \ll \mu$, we have $Q(A \cap C^*) = 0$. Also $Q(A \cap C^{*c} \cap C^c) = 0$ since $q = 0$ on $C^c$.

Finally, $Q(A \cap C^{*c} \cap C) > 0$ would imply $\mu(A \cap C^{*c} \cap C) > 0$ (by $Q \ll \mu$) and hence that

$$\mu[C^* \bigcup (A \cap C^{*c} \cap C)] \;=\; \mu[C^*] \,+\, \mu[A \cap C^{*c} \cap C] \;>\; \mu[C^*] \,.$$

Now $A \cap C^{*c} \cap C \in \mathcal{C}$ since it is a subset of $C$, and so also $C^* \bigcup (A \cap C^{*c} \cap C) \in \mathcal{C}$ since $\mathcal{C}$ is closed under unions. But the last displayed equation is a contradiction to the definition of $C^*$ since its $\mu$ measure is largest among elements of $\mathcal{C}$.

□

## Exercises for Section 4.2.

**4.2.1** (a) Suppose $\underline{X} \sim N(\underline{\mu}, V)$ is an $n$-dimensional normal random vector. Let $A$ be an $m \times n$ matrix. Show that $\underline{Y} = A\underline{X}$ has a normal distribution on $\mathbb{R}^m$ and determine the parameters.

(b) Let $\underline{X}$ and $\underline{Y}$ be random vectors. Let $\psi_{(\underline{X},\underline{Y})}(\underline{u}, \underline{v}) = E[\exp[\underline{u}'\underline{X} + \underline{v}'\underline{Y}]]$ be the joint moment generating function. Assume $\psi_{(\underline{X},\underline{Y})}$ is finite in a neighborhood of the origin. Show that $\underline{X}$ and $\underline{Y}$ are independent if and only if the joint m.g.f. factors into the product of the marginals, i.e. $\psi_{(\underline{X},\underline{Y})}(\underline{u}, \underline{v}) = \psi_{\underline{X}}(\underline{u})\psi_{\underline{Y}}(\underline{v})$.

(c) Suppose $\underline{X}$ and $\underline{Y}$ are random vectors which have a joint normal distribution. Show that $\underline{X}$ and $\underline{Y}$ are independent if and only if $\mathrm{Cov}[\underline{X}, \underline{Y}] = 0$.

**4.2.2** Show that $\mathrm{Law}[X|Y = y] = \mathrm{Law}[X]$ for $\mathrm{Law}[Y]$-almost all $y$ if and only if $X$ and $Y$ are independent.

**4.2.3** Let $Y_1$, $Y_2$, ..., $Y_n$ be independent Bernoulli random variables with success probabilities

$$p_i \;=\; P_\eta[Y_i = 1] \;=\; \frac{e^{\eta_1 + \eta_2 x_i}}{1 + e^{\eta_1 + \eta_2 x_i}}$$

where $(\eta_1, \eta_2) \in \mathbb{R}^2$ is unknown and $x_1$, $x_2$, ..., $x_n$ are known constants. Show that the joint distribution of the $Y_i$'s can be written as an exponential family with $\eta = (\eta_1, \eta_2)$ as the natural parameter.

Hints: The density of $Y_i$ w.r.t. counting measure on $\{0, 1\}$ is

$$f(y_i) \;=\; p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

Also, one can show

$$\eta_1 + \eta_2 x_i \;=\; \log\left(\frac{p_i}{1 - p_i}\right)$$

**4.2.4** For Examples 2.3.1 and 2.3.2, find sufficient statistics which are as "minimal" as you can make them.

**4.2.5** For each of the families in Exercise 2.3.10, (a) through (d), find a sufficient statistic which is as "minimal" as you can make it.

**4.2.6** For each of the families in Exercises 4.2.4 and 4.2.5, find a minimal sufficient statistic.

**4.2.7** Let $\mathbf{P} = \{P_i : 0 \le i \le k\}$ be a finite family as in Proposition 4.2.4 (b). Let

$$\mu \;=\; \sum_{i=0}^{k} P_i$$

Show that $\mu$ is $\sigma$–finite and $\mathbf{P} \ll \mu$. Thus, we may assume that a finite family of models is dominated, as was done implicitly in Proposition 4.2.4 (b).

**4.2.8** Let $\nu_1$, $\nu_2$, ..., be measures dominated by a $\sigma$–finite measure $\mu$ and suppose $a_1$, $a_2$, ..., are nonnegative real numbers. Let

$$\nu \;=\; \sum_{i=1}^{\infty} a_i \nu_i \;\; .$$

Show that $\nu \ll \mu$ and

$$\frac{d\nu}{d\mu} \;=\; \sum_{i=1}^{\infty} a_i \frac{d\nu_i}{d\mu} \quad , \quad \mu - a.e.$$

**4.2.9** Suppose $\mu$ is $\sigma$–finite. Show that there is a finite measure $\nu$ which is equivalent to $\mu$.

**4.2.10** Let $\underline{X}$ be a random vector with distribution in an exponential family. Show that for the measure $Q$ in Lemma 4.2.6, we can take *any* $Q$ of the form $\sum a_i P_{\theta_i}$ where $a_i \geq 0$ and $\sum a_i = 1$. In particular, we can take $Q = P_{\theta_0}$ for some $\theta_0$.

**4.2.11** (a) Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. with a $Unif[a,b]$ distribution with $\theta = (a,b)$, $-\infty < a < b < \infty$. Show that for the measure $Q$ in Lemma 4.2.6 we may take $Q_0^n$ where

$$Q_0 \;=\; \sum_{i=1}^{\infty} 2^{-i} Unif[-i,i] \;\; .$$

(b) Suppose $X_1$, $X_2$, ..., $X_n$ are i.i.d. with $Law[X_i] = Q_0$. What is $Law_{Q_0}[\underline{X} \; | (X_{(1)}, X_{(n)}) = (x_{(1)}, x_{(n)})]$?

(c) For the setup as in part (a), explain why it would not suffice in the proof of Theorem 4.2.2 to take $Q = P_{\theta_0}$ for some fixed $\theta_0$.

**4.2.12** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. from a truncation family $P_\theta$ where $\theta = (a,b)$, $-\infty < a < b < \infty$, and the Lebesgue density of a single $X_i$ is given by

$$f_\theta(x) \;=\; \frac{g(x)}{\int_a^b \, g(\xi)\, d\xi} \quad , \quad a < x < b \;\; .$$

(See the discussion after Example 2.3.4.)

(a) Show that $(X_{(1)}, X_{(n)})$ are sufficient.

(b) Show that $(X_{(1)}, X_{(n)})$ are minimal sufficient.

**4.2.13** Show that the exponential family in Example 4.2.4 is not full rank.

**4.2.14** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. $N(\sigma^2, \sigma^2)$. Find a minimal sufficient statistic for $\sigma^2$.

**4.2.15** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. $N(\sigma, \sigma^2)$. Find a minimal sufficient statistic for $\sigma > 0$.

**4.2.16** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. $N(\mu_X, \sigma_X^2)$ and $Y_1$, $Y_2$, ..., $Y_m$ be i.i.d. $N(\mu_Y, \sigma_Y^2)$, and assume the $\underline{X}$ and $\underline{Y}$ samples are independent. Find minimal sufficient statistics under each of the following:

**(i)** The parameters $\mu_X$, $\sigma_X^2$ $\mu_Y$, $\sigma_Y^2$ are unrestricted.

**(ii)** $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

**(iii)** $\mu_X = \mu_Y = \mu$.

## 4.3 Complete Statistics and Ancillary Statistics.

Now we introduce a concept which is related to minimal sufficiency but is more useful.

**Definition 4.3.1** *Let $T$ be a random $p$-vector with possible distributions $\{Law_\theta[T] : \theta \in \Theta\}$. We say $T$ is* complete *for $\theta$ iff for all $f : \mathbb{R}^p \to \mathbb{R}$ such that $E_\theta|f(T)| < \infty$ for all $\theta$,*

$$E_\theta[f(T)] = 0 \quad \text{for all } \theta \Rightarrow f(T) = 0,\ Law_\theta[T] - a.s. \text{ for all } \theta.$$

*We say $T$ is* boundedly complete *for $\theta$ iff this condition holds for $f$ bounded.*

□

**Remarks 4.3.1** (a) A complete statistic is always boundedly complete.

(b) If $T$ is complete and $U = h(T)$, then $U$ is complete since $E_\theta[f(U)] = E_\theta[f(h(T))] = 0$ for all $\theta$ implies $f(h(T)) = f(U) = 0$ $P_\theta$-a.s. for all $\theta$.

Suppose $P_\theta$ is a dominated family as in the factorization theorem. Then if $T$ is sufficient and $T = H(V)$ for some $V$ then $V$ is sufficient. To see this note that $f_\theta(x) = g(T(x), \theta)h(x) = g(H(V(x)), \theta)h(x) = \tilde{g}(V(x), \theta)h(x)$.

In terms of $\sigma$-fields:

$T$ complete, $\sigma(U) \subset \sigma(T) \quad \Rightarrow \quad U$ complete;

$T$ sufficient, $\sigma(T) \subset \sigma(V) \quad \Rightarrow \quad V$ sufficient.

In words, if $T$ is complete and $U$ is less informative than $T$, then $U$ is complete. If $T$ is sufficient and $V$ is more informative than $T$, then $V$ is sufficient. The next result shows that under some conditions, if $T$ is complete and sufficient then no less informative statistic is sufficient. One can see that any "strictly" more informative statistic cannot be complete (Exercise 4.3.2).

**Proposition 4.3.1** *Let the random vector $T$ be complete and sufficient for $\{P_\theta : \theta \in \Theta\}$, and assume $E_\theta[\|T\|] < \infty$ for all $\theta$. Suppose there is a random vector $U$ which is a minimal sufficient statistic. Then $T$ is minimal sufficient.*

**Proof.** Since $U$ is minimal sufficient and $T$ is sufficient, $U = h(T)$. By sufficiency of $U$, $E_\theta[T|U] = E[T|U] = g(U) = g(h(T))$ where of course $g(u) = E[T|U = u]$. By the Law of Total Expectation, Theorem 1.5.7 (d),

$$\underline{0} = E_\theta[T] - E_\theta[E[T|U]] = E_\theta[T - E[T|U]] = E_\theta[T - g(h(T))]\,.$$

By completeness of $T$, $T - g(h(T)) = T - g(U) = \underline{0}$, $P_\theta$-a.s., all $\theta$, i.e. $T = g(U)$ $P_\theta$-a.s., all $\theta$. This shows $T$ is a function of $U$ and hence that $T$ is minimal sufficient.

□

**Remarks 4.3.2** (a) One can prove the theorem without assuming $T$ is integrable, i.e. that $E[\|T\|] < \infty$, but the proof is much harder. One can accomplish this by considering $I_A(T)$ for all $A \in \mathcal{B}_n$.

(b) The result holds more generally in that one need not assume the existence of a minimal sufficient statistic. See Lehmann and Scheffe (19??).

□

Determining whether or not a statistic is complete can be a trying task. In exponential families, it is possible to give a simple sufficient condition.

**Theorem 4.3.2** *If $X$ is distributed according to an exponential family*

$$f_\theta(x) \;=\; \exp\left[\,\eta(\theta)'T(x) - B(\theta)\,\right] h(x)$$

*which is of full rank, then $T(X)$ is complete.*

**Proof.** We assume the family is in canonical form. By changing $\eta$ to $\eta - \eta_0$ where $\eta_0$ is an interior point of the natural parameter space, we can obtain a new parameterization where $0 = \eta_0 - \eta_0$ is an interior point of the natural parameter space which we denote $\Lambda$. Assume $h(T)$ is integrable for all $\eta$ and $E_\eta[h(T)] = 0$ for all $\eta$ in some neighborhood of 0, i.e. for some $\epsilon > 0$

$$\int h(T(x))e^{\eta' T(x) - A(\eta)}\, d\mu(x) \;=\; 0 \quad , \quad \text{for all } \|\eta\| < \epsilon$$

where $\mu$ is the dominating measure, but this is the same as

$$\int h_+(T(x))e^{\eta' T(x)}\, d\mu(x) \;=\; \int h_-(T(x))e^{\eta' T(x)}\, d\mu(x)\,, \quad \text{for all } \|\eta\| < \epsilon \;\; (4.42)$$

where $h_+$ and $h_-$ denote the positive and negative parts of $h$. We may assume by renormalizing that $\int h_+ \circ T\, d\mu = \int h_- \circ T\, d\mu = 1$, i.e. $h_\pm \circ T$ is a probability density w.r.t. $\mu$. But (4.42) says the two corresponding probability measures have the same finite m.g.f. in a neighborhood of 0. Hence, by uniqueness of m.g.f.'s (Proposition 2.2.1 (d)) the two p.m.'s are the same and so by uniqueness of Radon-Nikodym derivatives we have

$$h_+(T(x)) \;=\; h_-(T(x)) \quad , \quad \text{for } \mu - \text{ almost all } x\ .$$

But this says that $h(T) = h_+(T) - h_-(T) = 0$ $P_\theta$-a.s. for all $\theta$, since $P_\theta \ll \mu$ for all $\theta$. This shows $T$ is complete as desired.

□

**Example 4.3.1** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. with $Gamma(\alpha, \beta)$ distribution. Then the joint density is

$$f_{(\alpha,\beta)}(\underline{x}) \;=\; \exp\left[\,\alpha\sum_{i=1}^{n}\log(x_i) \;-\; \frac{1}{\beta}\sum_{i=1}^{n}x_i \;-\; n\log(\Gamma(\alpha)\beta^{\alpha})\,\right]\,\prod_{i=1}^{n}x_i^{-1}I_{(0,\infty)}(x_i)\;\;.$$

The natural parameter is $\underline{\eta} = (\alpha, -1/\beta)$ and the natural parameter space is $(0,\infty)\times(-\infty,0)$ which is a subset of $\mathbb{R}^2$ with nonempty interior (it clearly conatains a nonempty open rectangle, in fact it is one). Thus, the family is full rank and $\underline{T} = (\sum_i \log(X_i), \sum_i X_i)$ is complete and sufficient. Of course, by Proposition 4.2.5, $\underline{T}$ is also minimal sufficient.

□

**Example 4.3.2** Suppose $X_1$, $X_2$, ..., $X_n$ are i.i.d. with $Unif(0, b)$ distribution, $b > 0$. We will show that $T = X_{(n)} = \max\{X_i : 1 \le i \le n\}$ is complete and sufficient. For sufficiency, apply the factorization theorem to the joint density

$$f_b(\underline{x}) \;=\; \prod_{i=1}^{n} b^{-1}I_{(0,b)}(x_i) \;=\; b^{-n}I_{(0,b)}(x_{(n)})\;\;.$$

Now we compute $\mathrm{Law}_b[T]$ using the c.d.f.:

$$\begin{aligned} F_b(t) \;&=\; P_b[T \le t] \;=\; P_b[X_1 \le t \,\&\, X_2 \le t \,\&\, \ldots \,\&\, X_n \le t] \\ &= \prod_{i=1}^{n} P_b[X_i \le t] \;=\; (P_b[X_1 \le t])^n \\ &= (t/b)^n \quad , \quad 0 \le t \le b\;. \end{aligned}$$

Hence, $\mathrm{Law}_b[T]$ has Lebesgue density

$$f_b(t) \;=\; \frac{nt^{n-1}}{b^n} \quad , \quad 0 < t < b\;.$$

Now suppose $h$ is such that $E_b[h(T)] = 0$ for all $b > 0$. Then for all $b > 0$,

$$\int_0^b h(t)nt^{n-1}/b^n\,dt \;=\; 0$$

or what is the same

$$\int_0^b h_+(t)t^{n-1}\,dt \;=\; \int_0^b h_-(t)t^{n-1}\,dt\,, \qquad \text{for all } \; b > 0\,, \tag{4.43}$$

where $h_+$ and $h_-$ are the positive and negative parts of $h$. Now suppose $h$ is not identically 0 Lebesgue a.e. and then fix $N$, a positive integer, such that

$$\int_0^N h_+(t)t^{n-1}\,dt \;=\; C \;>\; 0\;\;.$$

The existence of such and $N$ follows from $h_+$ not identically 0, Lebesgue a.e. Note that $C$ is finite since we assume $h(T)$ is integrable for all $b$. Let $H_+$ be the p.m. given by

$$H_+(B) \;=\; C^{-1} \int_{B\cap(0,N)} h_+(t) t^{n-1}\, dt$$

and similarly for $H_-$. Then (4.43) says that $H_+$ and $H_-$ have the same c.d.f., so they are the same p.m., so $h_+ = h_-$ Lebesgue a.e. on $(0, N)$, and hence $h = 0$ Lebesgue a.e. on $(0, N)$, a contradicition. Thus, since each measure in $\{\text{Law}_b[T] : b > 0\}$ is dominated by Lebesgue, it follows that $T$ is complete.

□

**Example 4.3.3** Now we consider a nonparametric family. Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. with Lebesgue density $f$ which is unknown. We will use $f$ to denote the parameter in expectations, etc. We claim $\underline{T} = \mathbf{Sort}(\underline{X})$ is complete and sufficient. We already know $\underline{T}$ is sufficient from Example 4.2.1. We will show that $E_f[h(\underline{T})] = 0$ for all Lebesgue probability densities $f$ implies that $h(t) = 0$ Lebesgue a.e. on $I\!P^n = \{\underline{x} \in I\!R^n : x_1 \le x_2 \le ... \le x_n\}$. This implies that $h(T) = 0$ $\text{Law}_f[T]$-a.s. for all $f$ since $\mathbf{P} = \{\text{Law}_f[T] : f$ is a Lebesgue probability density $\} \ll m^n$. Now if $m^n(I\!P^n \cap \{\underline{x} : h(\underline{x}) \neq 0\}) > 0$ then

$$m^n([-N, N]^n \cap Pn \cap \{\underline{x} : h(\underline{x}) \neq 0\}) \;>\; 0 \quad , \qquad \text{for some } N. \tag{4.44}$$

Fix $N$ at such a value. Given a probability vector $(p_1, p_2, ..., p_n)$ and a $\underline{\eta} \in I\!R^n$, let

$$f(x) \;=\; \sum_{j=1}^{n} p_j \exp\left[\, \eta_j x \;-\; A(\eta_j) \,\right] I_{[-N,N]}(x) \tag{4.45}$$

where

$$A(\eta_j) \;=\; \log \int_{-N}^{N} e^{\eta_j x}\, dx \;.$$

Using $E_f[h(\mathbf{Sort}(\underline{X}))] = 0$ for $f$ in (4.45), we conclude that

$$\int_{[-N,N]^n} h(\mathbf{Sort}(\underline{x})) \left[ \prod_{i=1}^{n} \sum_{j=1}^{n} p_j \exp[\eta_j x_i] \right] d\underline{x} \;=\; 0 \;.$$

Expanding the l.h.s. gives

$$\sum_{j_1=1}^{n} \sum_{j_2=1}^{n} ... \sum_{j_n=1}^{n} \int_{-N}^{N} \int_{-N}^{N} ... \int_{-N}^{N} h(\mathbf{Sort}(x_1, x_2, ..., x_n))$$

$$\exp\left[ \sum_{i=1}^{n} \eta_{j_i} x_i \right] dx_1\, dx_2 \,...\, dx_n \prod_{i=1}^{n} p_{j_i} \;=\; 0.$$

Considering the expression on the l.h.s. of the latter equation as a function of $\underline{p}$ for fixed $\underline{\eta}$, we have a homogeneous polynomial of degree $n$ in the variables $p_1$,

$p_2$, ..., $p_n$ which vanishes identically on the set $\{\underline{p} \in I\!R^n : p_j \geq 0,\ 1 \leq j \leq n, \text{ and } \sum_j p_j = 1\}$. This can only happen if all coefficients vanish. (See Exercise 4.3.3.) Hence, looking at the coefficient with $j_1 = 1$, $j_2 = 2$, ... $j_n = n$ we obtain

$$\int_{-N}^{N} \int_{-N}^{N} ... \int_{-N}^{N} h(\mathbf{Sort}(x_1, x_2, ..., x_n)) \exp\left[\sum_{i=1}^{n} \eta_i x_i\right] dx_1\, dx_2 \ldots dx_n \;=\; 0$$

for all $\eta_1$, $\eta_2$, ..., $\eta_n$. In more compact notation,

$$\int_{[-N,N]^n} h(\mathbf{Sort}(\underline{x})) \exp[\underline{\eta}'\underline{x}]\, d\underline{x} \;=\; 0 \;\text{ for all } \underline{\eta} \in I\!R^n\ .$$

Now as in the proof of Theorem 4.3.2, or similarly to Example 4.3.2, we obtain

$$\int_{[-N,N]^n} h_+(\mathbf{Sort}(\underline{x})) \exp[\underline{\eta}'\underline{x}]\, d\underline{x} \;= \tag{4.46}$$

$$\int_{[-N,N]^n} h_-(\mathbf{Sort}(\underline{x})) \exp[\underline{\eta}'\underline{x}]\, d\underline{x} \;>\; 0\ .$$

The last inequality follows from (4.44). As in the proof of Theorem 4.3.2, we can normalize the nonnegative functions $h_+ \circ \mathbf{Sort}$ and $h_- \circ \mathbf{Sort}$ to be probability densities w.r.t. Lebesgue measure. Since (4.46) holds for all $f$ of the form (4.44), it follows that the p.m.'s corresponding to $h_+ \circ \mathbf{Sort}$ and $h_- \circ \mathbf{Sort}$ have the same finite m.g.f. by (4.46), and hence that $h_+ = h_-$ Lebesgue a.e. on $I\!P^n \cap [-N, N]^n$, i.e. $h = 0$ Lebesgue a.e. on $I\!P^n \cap [-N, N]^n$, a contradiction. Thus, we conclude $h = 0$ Lebesgue a.e. on $I\!P^n$ and hence that $h(\mathbf{Sort}(\underline{X})) = 0$, $P_f$-a.s. for all $f$, i.e. that $\underline{T} = \mathbf{Sort}(\underline{X})$ is complete.

□

**Definition 4.3.2** *A statistic $V = V(X)$ is called* ancillary *to $\theta$ iff $Law_\theta[V] = Law[V]$ doesn't depend on $\theta$.*

□

The notion of an ancillary statistic is in some sense complementary to the notion of a sufficient statistic. A sufficient statistic contains all the information about the unknown parameter, but an ancillary statistic contains no information about the unknown parameter since its distribution is the same no matter what parameter value Nature chooses. The notions of ancillarity and complete sufficiency come together in the following result.

**Theorem 4.3.3 (Basu's Theorem.)** *Suppose $T$ is boundedly complete and sufficient for $I\!P^n = \{P_\theta : \theta \in \Theta\}$ and $V$ is ancillary for $\theta$. Then $T$ and $V$ are independent for all $\theta$.*

**Proof.** By ancillarity, $P_\theta[V \in B] = P[V \in B]$. Also, by sufficiency, $P_\theta[V \in B|T] = P[V \in B|T]$. By the law of total expectation (Theorem 1.5.7 (d)), $E_\theta[P[V \in B|T]] = P[V \in B]$. Thus, for all $\theta$,

$$E_\theta\{\, P[V \in B|T] \,-\, P[V \in B]\,\} \;=\; 0$$

Note that for fixed $B$, the quantity inside braces { ... } is a function of $T$ ($P[V \in B]$ is a constant function of $T$). Hence, by completeness,

$$P[V \in B|T] \;=\; P[V \in B] \quad , \qquad \text{for all } \theta \; .$$

This shows that Law$[V]$ can be used as a version of Law$[V|T = t]$, which implies $V$ and $T$ are independent by Supplementary Exercise for Chapter 2, #1???.

□

**Example 4.3.4** Let $X_1, X_2, ..., X_n$ be i.i.d. with $N(\mu, \sigma^2)$ density. We will show that the sample mean and variance,

$$\bar{X} \;=\; \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 \;=\; \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 \;\; ,$$

respectively, are independent. For this purpose, fix $\sigma^2$ and let $\mu$ be the unknown parameter ranging over $\mathbb{R}$. The joint density w.r.t. $m^n$ is

$$f_\mu(\underline{X}) \;=\; \exp\left[\frac{(n\mu)}{\sigma^2}\bar{X} \,-\, \frac{n\mu^2}{2\sigma^2}\right]\left\{(2\pi\sigma^2)^{-n/2}\exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}X_i^2\right]\right\}$$

which is an exponential family with $T = \bar{X}$ and $\eta = (n\mu)/\sigma^2$. Since $\eta$ ranges over all of $\mathbb{R}$, the family is full rank and it follows that $\bar{X}$ is complete and sufficient for $\mu$. To show that $S^2$ is ancillary, first recall that

$$\text{Law}_\mu[\frac{(n-1)}{\sigma^2}S^2] \;\sim\; \chi^2_{(n-1)}$$

where $\chi^2_r$ denotes the $\chi$-squared distribution with $r$ degrees of freedom, which is the same as $Gamma(r/2, 2)$. This distribution is obviously independent of the unknown parameter $\mu$, so this shows ancillarity, and hence independence of $\bar{X}$ and $S^2$. However, we show ancillarity by an invariance argument that works more generally. Clearly, $S^2$ as a function of $\underline{X}$ is invariant of the location parameter $\mu$, since

$$S^2(\underline{X}) \;=\; \frac{1}{(n-1)}\sum_{i=1}^{n}\Big[\,(X_i - \mu) - (\bar{X} - \mu)\,\Big]^2$$

$$= S^2(\underline{X} - \mu\underline{1})$$

where $\underline{1}$ is an $n$-vector with all entries 1. Now

$$\text{Law}_\mu[\underline{X} - \mu\underline{1}] \;=\; \text{Law}_0[\underline{X}] \;.$$

Hence,

$$\text{Law}_\mu[S^2(\underline{X})] \;=\; \text{Law}_\mu[S^2(\underline{X} - \mu\underline{1})] \;=\; \text{Law}_0[S^2(\underline{X})]$$

and the latter is clearly independent of $\mu$.

It is a source of some confusion that while probably both $\mu$ and $\sigma^2$ are unknown, we treated $\sigma^2$ as known. This is done purely from a mathematical standpoint. Certainly all of the distribution theory stated above is true for any fixed value of $\sigma^2$, whether that value is known to be a specific number or not.

□

**Example 4.3.5** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. with unknown Lebesgue density $f$, as in Example 4.3.3. Let $\mathbf{Rank}(\underline{X})$ denote the ranks of the random sample as in Proposition 2.5.2. According to that Proposition, $\mathbf{Rank}(\underline{X})$ has a distribution which doesn't depend on $f$. Hence, $\mathbf{Rank}(\underline{X})$ is independent of the complete and sufficient statistic $\mathbf{Sort}(\underline{X})$.

□

## Exercises for Section 4.3.

**4.3.1** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with unknown Lebesgue density $f$ which is known to satisfy the condition of existence of the first $k$ moments:

$$E_f[|X_1|^j] \ < \ \infty \quad , \quad 1 \leq j \leq k \ .$$

Show that $\mathbf{Sort}(\underline{X})$ is complete and sufficient for $f$. (Hint: This problem is easy in that you don't need to make any long arguments.)

**4.3.2** Suppose $\{P_\theta : \theta \in \Theta\}$ is a dominated family as in the factorization theorem. Suppose $T$ is a sufficient statistic and $\sigma(T)$ is essentially a proper subset of $\sigma(V)$ in that $\sigma(T) \subset \sigma(V)$ and there is a set $A \in \sigma(V)$ and a $\theta$ such that there is no $B \in \sigma(T)$ with $I_A = I_B$, $P_\theta$-a.s. (This means that $V$ is strictly more informative than $T$.) Show that $V$ is not complete.

**4.3.3** The following result was used in Example 4.3.3. Let

$$g(\underline{p}) \ = \ \sum_{j_1=1}^{n} \sum_{j_2=1}^{n} ... \sum_{j_n=1}^{n} a_{j_1 j_2 ... j_n} \prod_{i=1}^{n} p_{j_i}$$

be a homogeneous polynomial of degree $n$ in $p_1$, ..., $p_n$. Here, the polynomial is called homogeneous since

$$g(b\underline{p}) \ = \ b^n g(\underline{p})$$

for all $b > 0$. Assume

$$g(\underline{p}) \ = \ 0 \quad , \text{ for all } \underline{p} \text{ such that} \tag{4.47}$$

$$p_i \geq 0 \text{ for } 1 \leq i \leq n \text{ and } \sum_i p_i = 1 \ .$$

(a) Using homogeniety, show that $g(\underline{x}) = 0$ for all $\underline{x} \in I\!R^n$ such that $x_i \geq 0$ for $1 \leq i \leq n$.

(b) Show the following by induction on $n$: If $g(\underline{x})$ is a polynomial of degree $\leq k$ in $\underline{x} \in I\!R^n$ which vanishes identically in $\{\underline{x} : x_i \geq 0 \text{ for } 1 \leq i \leq n\}$ then all coefficients of $g$ vanish. (Hint: For $n = 1$ the result is trivial since such a polynomial has at most $k$ roots, unless all coefficients are 0. If $g(x_1, x_2, ..., x_n)$ is a polynomial in $n$ variables, then one can fix $x_n$ and obtain a polynomial in $n-1$ variables.)

(c) Using (a) and (b), show that all coefficients of $g$ in (4.47) must vanish.

**4.3.4** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with unknown Lebesgue density $f$ and let $Y_1$, $Y_2$, ..., $Y_m$ be i.i.d. r.v.'s with unknown Lebesgue density $g$, and suppose the $\underline{X}$ and $\underline{Y}$ samples are independent. Show that $(\mathbf{Sort}(\underline{X}), \mathbf{Sort}(\underline{Y}))$ are complete and sufficient for $(f, g)$.

**4.3.5** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with $Gamma(\alpha, \beta)$ density. Show that $\underline{X}/\sum_i X_i$ is independent of $\sum_i X_i$.

**4.3.6** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with $Unif[0, \theta]$ distribution where $\theta > 0$. Show that $V = \underline{X}/X_{(n)}$ is independent of $X_{(n)}$.

**4.3.7** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with $Unif[\theta_1, \theta_2]$ where $\theta_1 < \theta_2$ are unknown.

(a) Show that $\underline{T} = (X_{(1)}, X_{(n)})$ is complete for $\underline{\theta}$.

(b) Let $\underline{V} = \underline{X}/(X_{(n)} - X_{(1)})$. Show that $\underline{V}$ is independent of $\underline{T}$.

**4.3.8** Let $X_1$, $X_2$, ..., $X_n$ be i.i.d. r.v.'s with $Unif[\theta - 1/2, \theta + 1/2]$ where $\theta \in I\!R$ is unknown. Show that $(X_{(1)}, X_{(n)})$ is not complete.