Theorem N.4.24 (Chebyshev's Inequality). Suppose X has mean μ and variance σ^2. For all $t > 0$, $P(|X - \mu| < t\sigma) \geq 1 - \frac{1}{t^2}$.

Justification: Apply Markov's Inequality to $Y = \left(\frac{X-\mu}{\sigma}\right)^2$ with $r = t^2$ to get $P\left[\left(\frac{X-\mu}{\sigma}\right)^2 \geq t^2\right] \leq \frac{E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right]}{t^2} = \frac{1}{t^2}$. The event $\left(\frac{X-\mu}{\sigma}\right)^2 \geq t^2$ is the same as the event $|X - \mu| \geq t\sigma$. The complement of this event is $|X - \mu| < t\sigma$ and its probability must be $\geq 1 - \frac{1}{t^2}$. □

For $t = 1, 2$ and 3, Chebyshev's Inequality says that $P(|X - \mu| < \sigma) \geq 0$ (which doesn't say much), $P(|X - \mu| < 2\sigma) \geq 0.75$ and $P(|X - \mu| < 3\sigma) \geq 0.8888$.

CHAPTER 5 – Properties of random samples

In the first four chapters of C&B we have studied the theory of probability and random variables. Now in Chapter 5 we will apply some of this theory in a statistical context. A fundamental activity in statistics is the selection of a random sample from a population in order to discover something about the population. We will now study properties of random samples.

Definition CB.5.1.1. If $X_1, X_2, \ldots, X_n$ are independent random variables having the same distribution, we call them a *random sample* from a population with that distribution. We also call them *independent and identically distributed* (*i.i.d.*) random variables.

Caution: When sampling from a finite population, if a sample is selected at random with replacement, then it is a random sample (according to the definition above). If a sample is selected at random without replacement, it is called a *simple random sample*. Thus our terminology requires us to say, somewhat awkwardly, that a simple random sample is not a random sample. If the population size is much larger than the sample size, then there is very little difference between sampling with replacement and sampling without replacement, and so in that case a simple random sample is almost a random sample.

The sample mean

The *sample mean* of a random sample $X_1, X_2, \ldots, X_n$ is $\bar{X} = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$.

Theorem CB.5.2.2(a,b). Suppose $X_1, X_2, \ldots, X_n$ is a random sample.

(a) Suppose the population has mean μ. Then $\mathrm{E}(\bar{X}) = \mu$.

(b) Suppose the population has variance σ^2. Then $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

Proof: (a) $\mathrm{E}(\bar{X}) = \mathrm{E}[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] = \frac{1}{n}\mathrm{E}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n}[\mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)] = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{1}{n}n\mu = \mu$.

(b) $\mathrm{Var}(\bar{X}) = \mathrm{Var}[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)] = \frac{1}{n^2}\mathrm{Var}(X_1 + X_2 + \cdots + X_n) =$ (by independence) $\frac{1}{n^2}[\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n)] = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$. □

Thus the mean and variance of the sample mean can be easily expressed in terms of the mean and variance of the population. Can we express the entire distribution of the sample mean in terms of the population distribution?

Theorem CB.5.2.3. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population whose distribution has mgf $M_X(t)$. Then $M_{\bar{X}}(t) = [M_X(\frac{t}{n})]^n$.

Proof: Let $Y = X_1 + X_2 + \cdots + X_n$. By Lemma N.4.20, $M_Y(t) = M_{X_1}(t)M_{X_2}(t)\cdots M_{X_n}(t) = [M_X(t)]^n$. Since $\bar{X} = \frac{1}{n}Y$, Theorem CB.2.3.5 implies that $M_{\bar{X}}(t) = M_Y(\frac{1}{n}t) = M_Y(\frac{t}{n}) = [M_X(\frac{t}{n})]^n$. □

Examples. (1) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Normal(μ, σ^2). From the table in the back of C&B, we know $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$. By the preceding theorem, $M_{\bar{X}}(t) = [e^{\mu(\frac{t}{n}) + \frac{1}{2}\sigma^2(\frac{t}{n})^2}]^n = e^{\mu t + \frac{1}{2}\frac{\sigma^2}{n}t^2}$, which is the mgf of the Normal$(\mu, \frac{\sigma^2}{n})$ distribution. Therefore $\bar{X} \sim$ Normal$(\mu, \frac{\sigma^2}{n})$.

(2) Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. Gamma(α, β). From the table in the back of C&B, we know $M_X(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$. By the preceding theorem, $M_{\bar{X}}(t) = \left[\left(\frac{1}{1-\beta(\frac{t}{n})}\right)^{\alpha}\right]^n = \left(\frac{1}{1-(\frac{\beta}{n})t}\right)^{n\alpha}$, which is the mgf of the Gamma$(n\alpha, \frac{\beta}{n})$ distribution. Therefore $\bar{X} \sim$ Gamma$(n\alpha, \frac{\beta}{n})$. ‖

Another theorem that can help in determining the distribution of $\bar{X}$ is the following.

Theorem CB.5.2.4 (Convolution Formula). Suppose X and Y are independent random variables with pdf's $f_X(x)$ and $f_Y(y)$. Let $Z = X + Y$. Its pdf is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\mathrm{d}x .$$

Justification: Let $Z = X + Y$ and $W = X$. Then $X = W$ and $Y = Z - W$, and the absolute value of the determinant of the Jacobian is 1. By (CB.4.3.2), $f_{Z,W}(z,w) = f_{X,Y}(x,y)\cdot 1 = f_X(x)f_Y(y)$ [by independence] $= f_X(w)f_Y(z-w)$. Now $f_Z(z) = \int_{-\infty}^{\infty} f_X(w)f_Y(z-w)\mathrm{d}w = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\mathrm{d}x$, because the dummy variable of integration can be denoted by x just as well as w. □

Example. (a) Suppose X and Y are i.i.d. Cauchy(0, 1).
Let $Z = X + Y$. By the preceding theorem,

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)\mathrm{d}x = \int_{-\infty}^{\infty} \frac{1}{\pi(1+x^2)} \cdot \frac{1}{\pi(1+(z-x)^2)}\mathrm{d}x$$

$$= \frac{1}{\pi^2}\int_{-\infty}^{\infty} \frac{1}{(1+x^2)(1+(z-x)^2)}\mathrm{d}x .$$

A trick for integrating this integrand is to write

$$\frac{1}{(1+x^2)(1+(z-x)^2)} = \frac{a_1 x + a_0}{1+x^2} + \frac{b_1 x + b_0}{1+(z-x)^2}$$

where a_1, a_0, b_1 and b_0 do not involve x but may involve z. Now combine the two terms on the right-hand side using a common denominator $(1+x^2)(1+(z-x)^2)$ and a numerator of the form $c_3x^3 + c_2x^2 + c_1x + c_0$. The c_i's involve a_1, a_0, b_1, b_0 and z but do not involve x. Set $c_3 = 0$, $c_2 = 0$, $c_1 = 0$, $c_0 = 1$. Solve these four equations for the four unknowns a_1, a_0, b_1, b_0 to obtain $a_1 = \frac{2}{4z+z^3}$, $a_0 = \frac{1}{4+z^2}$, $b_1 = \frac{-2}{4z+z^3}$, $b_0 = \frac{3}{4+z^2}$. Therefore,

$$f_Z(z) = \frac{1}{\pi^2}\left\{ \frac{2}{4z+z^3}\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx + \frac{1}{4+z^2}\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx + \frac{-2}{4z+z^3}\int_{-\infty}^{\infty} \frac{x}{1+(z-x)^2}\,dx \right.$$
$$\left. + \frac{3}{4+z^2}\int_{-\infty}^{\infty} \frac{1}{1+(z-x)^2}\,dx \right\}.$$

Recognizing $\frac{1}{1+x^2}$ as a kernel of a Cauchy$(0,1)$ distribution and $\frac{1}{1+(z-x)^2} = \frac{1}{1+(x-z)^2}$ as a kernel of a Cauchy$(z,1)$ distribution, we see that $\int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx = \int_{-\infty}^{\infty} \frac{1}{1+(z-x)^2}\,dx = \pi$. Now

$$f_Z(z) = \frac{1}{\pi^2}\left\{ \frac{4\pi}{4+z^2} + \frac{2}{4z+z^3}\left[\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx - \int_{-\infty}^{\infty} \frac{x}{1+(z-x)^2}\,dx \right] \right\}.$$

In the second integral put $u = x - z$ to obtain

$$\int_{-\infty}^{\infty} \frac{x}{1+(z-x)^2}\,dx = \int_{-\infty}^{\infty} \frac{u+z}{1+u^2}\,du = \int_{-\infty}^{\infty} \frac{x+z}{1+x^2}\,dx = \int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx + \int_{-\infty}^{\infty} \frac{z}{1+x^2}\,dx,$$

because the dummy variable of integration can be denoted by x just as well as u. Now

$$\int_{-\infty}^{\infty} \frac{x}{1+x^2}\,dx - \int_{-\infty}^{\infty} \frac{x}{1+(z-x)^2}\,dx = -\int_{-\infty}^{\infty} \frac{z}{1+x^2}\,dx = -\pi z.$$

Hence

$$f_Z(z) = \frac{1}{\pi^2}\left\{ \frac{4\pi}{4+z^2} + \frac{2}{4z+z^3}(-\pi z) \right\} = \frac{2}{\pi(4+z^2)} = \frac{1}{2\pi[1+(\frac{z}{2})^2]},$$

which is the pdf of the Cauchy$(0,2)$ distribution. Thus we see that if X_1 and X_2 are i.i.d. Cauchy$(0,1)$, then $X_1 + X_2 \sim$ Cauchy$(0,2)$.

(b) If X_1 and X_2 are i.i.d. Cauchy$(0,1)$, what is the distribution of the sample mean? Let $Z = X_1 + X_2$ and $U = \frac{1}{2}Z = \bar{X}$. Then $f_U(u) = f_Z(z)|\frac{dz}{du}| = f_Z(2u)\cdot 2 = \frac{2}{2\pi[1+(\frac{2u}{2})^2]}$ $= \frac{1}{\pi(1+u^2)}$, which is the pdf of the Cauchy$(0,1)$ distribution. Thus we see that $\bar{X} \sim$ Cauchy$(0,1)$. More generally, it can be shown that if $X_1, X_2, \ldots, X_n$ are i.i.d. Cauchy(θ,σ), then, surprisingly, also $\bar{X} \sim$ Cauchy(θ,σ). ||

Birkes, Winter 2002

Textbook by Mukhopadhyay

Notes for ST 562

Overview of the textbook

Chapters 1-5: Theory of probability and random variables. (These chapters, which are covered in ST 561, provide background that is needed to develop the theory of statistical inference.)

Chapters 6-12: Theory of statistical inference.

Ch. 6: Fundamental concepts of statistical inference

Ch. 7: Estimation

(ST 562 covers Chapters 6 and 7)

Ch. 8: Testing

Ch. 9: Confidence intervals

(ST 563 covers Chapter 8, 9, and parts of 10, 11 and 12)

A general formulation of statistical inference

The goal of statistical inference is to analyze a set of data in order to conclude something about the population from which it was taken (or can be imagined to have been taken). To develop a theory of statistical inference, we need a formal formulation. The data are formally represented by a random vector $\boldsymbol{X}$. The population is represented by the unknown probability distribution of $\boldsymbol{X}$. The distribution is typically described by the joint pmf or pdf of $\boldsymbol{X}$, denoted by $f(\boldsymbol{x};\boldsymbol{\theta})$. The parameter vector $\boldsymbol{\theta}$ is unknown and could be any vector in a parameter set Θ. This formulation is quite general and includes the following examples.

(1) $X_1, \ldots, X_n$ i.i.d. Bernoulli(p), $0 < p < 1$.
Here we have $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{\theta} = p$.

(2) $X_{11}, \ldots, X_{1n_1}$ i.i.d. Bernoulli(p_1), $X_{21}, \ldots, X_{2n_2}$ i.i.d. Bernoulli(p_2), two independent samples of Bernoulli random variables. Here $\boldsymbol{X} = (X_{11}, \ldots, X_{1n_1}, X_{21}, \ldots, X_{2n_2})$, $\boldsymbol{\theta} = (p_1, p_2)$.

(3) $X_1, \ldots, X_n$ independent, $X_i \sim \text{Normal}(\beta_0 + \beta_1 w_i, \sigma^2)$.
Here $\boldsymbol{X} = (X_1, \ldots, X_n)$, $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$.

A more formal statement of the general goal of statistical inference is as follows:

> Given a data vector $\boldsymbol{X}$ and a model for its distribution,
> in the form of a family $\{f(\boldsymbol{x}|\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ of possible pmf's or pdf's for $\boldsymbol{X}$,
> we want to say something about what the true parameter vector $\boldsymbol{\theta}$ might be.

Notation. We use X to denote the data when we are regarding it as a random vector (before it was randomly selected from the population) and we use x (or x_{obs}) to denote the actual observed values of the data. The symbol x also appears as a mathematical variable (or dummy variable) in expressions like $f(x)$.

Sufficient statistics

A *statistic* is a function of the data that can be calculated from the data alone (i.e., it cannot involve any unknown parameters). We allow it to be vector-valued and often denote it by $T(X)$ or simply T. Some examples are (1) $T = \overline{X} = \sum_{i=1}^{n} X_i/n$, (2) $T = (\overline{X}, S)$ where $S^2 = \sum_{i=1}^{n}(X_i - X)^2/(n-1)$, (3) $T = (X_{(1)}, X_{(n)})$, (4) $T = X$.

A statistic is a random variable or random vector, but not every random variable or random vector is a statistic. For example, if μ and σ are unknown parameters, then $(\overline{X} - \mu)/(\sigma/\sqrt{n})$ and S^2/σ^2 are not statistics.

Given a set of data, it is often desirable to summarize the data in a few statistics. For example, if a team of engineers measures the resistances of 100 resistors, then, rather than report all 100 resistances, they might report only the mean and standard deviation. That is, given a data vector X, one often wants to reduce it to a summary statistic $T(X)$. A good summary statistic is both concise and informative. To be concise, it should not have very many components. To be informative, $T(X)$ should contain most of the "information" that X contains about θ (that is, about the population). Sometimes we are able to choose $T(X)$ so that it contains all of the information that X contains about θ. This is the idea of a sufficient statistic. In precise mathematical terms, we make the following definition.

Definition 6.2.3. A statistic $T(X)$ is a *sufficient statistic* for θ if the conditional distribution of X given $T(X)$ does not involve θ. (This is short for: the conditional distribution of X given $T = t$ does not involve θ for all t.)

In other words, if we are given a sufficient statistic $T(X)$, there is nothing else in X that can tell us anything about θ.

Example. $X_1 \sim \text{Binomial}(n_1, \theta)$, $X_2 \sim \text{Binomial}(n_2, \theta)$ (same θ), independent of one another.

(a) Is $T = X_1$ sufficient? Intuitively we would say no — let's verify this formally. We need to find the conditional distribution of $(X_1, X_2) \mid X_1 = x_1$. Since $X_1 = x_1$ is given as a constant, then we are concerned only with $X_2 \mid X_1 = x_1$. Since X_1 and X_2 are independent, the conditional distribution of X_2 given X_1 is the same as its unconditional distribution.

(This follows from formulas (3.3.6) and (3.5.1).) The unconditional distribution of X_2 is Binomial(n_2, θ), which involves θ. So X_1 is not sufficient.

(b) Is $T = X_1 + X_2$ sufficient? We need to find the conditional distribution of $(X_1, X_2) \mid X_1 + X_2 = t$.

$$\mathrm{P}\{(X_1, X_2) = (x_1, x_2) \mid X_1 + X_2 = t\} = \mathrm{P}(A \mid B) = \mathrm{P}(A \cap B)/\mathrm{P}(B)$$
$$= \frac{\mathrm{P}\{X_1{=}x_1 \text{ and } X_2{=}x_2 \text{ and } X_1{+}X_2{=}t\}}{\mathrm{P}\{X_1{+}X_2{=}t\}}.$$

Note that the numerator is 0 if $x_1 + x_2 \neq t$. Also note that this case does not involve θ. Now suppose $x_1 + x_2 = t$. Then the conditional probability is

$$= \frac{\mathrm{P}\{X_1{=}x_1 \text{ and } X_2{=}x_2\}}{\mathrm{P}\{X_1{+}X_2{=}t\}} = \frac{\mathrm{P}\{X_1{=}x_1\}\mathrm{P}\{X_2{=}x_2\}}{\mathrm{P}\{X_1{+}X_2{=}t\}}.$$
$$= \frac{\binom{n_1}{x_1}\theta^{x_1}(1{-}\theta)^{n_1-x_1}\binom{n_2}{x_2}\theta^{x_2}(1{-}\theta)^{n_2-x_2}}{\binom{n_1+n_2}{t}\theta^{t}(1{-}\theta)^{n_1+n_2-t}} = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n_1+n_2}{t}},$$

which does not involve θ. So $X_1 + X_2$ is a sufficient statistic.

(c) Is $T = X_1 + 2X_2$ sufficient? If $x_1 + 2x_2 = t$, then

$$\mathrm{P}\{(X_1, X_2) = (x_1, x_2) \mid X_1 + 2X_2 = t\} = \frac{\mathrm{P}\{X_1{=}x_1\}\mathrm{P}\{X_2{=}x_2\}}{\mathrm{P}\{X_1{+}2X_2{=}t\}}.$$

There is no nice formula for $\mathrm{P}\{X_1 + 2X_2 = t\}$, so let's just look at a particular case. Suppose $n_1 = 3$, $n_2 = 2$, $x_1 = 1$, $x_2 = 1$. To figure out $\mathrm{P}\{X_1 + 2X_2 = 3\}$, note that $X_1 \in \{0, 1, 2, 3\}$ and $X_2 \in \{0, 1, 2\}$ and the only pairs (x_1, x_2) that give $x_1 + 2x_2 = 3$ are $(1, 1)$ and $(3, 0)$. Therefore, $\mathrm{P}\{X_1 + 2X_2 = 3\} = \mathrm{P}\{(X_1, X_2) = (1, 2) \text{ or } (3, 0)\}$

$$= \binom{3}{1}\theta^1(1-\theta)^2\binom{2}{1}\theta^1(1-\theta)^1 + \binom{3}{3}\theta^3(1-\theta)^0\binom{2}{0}\theta^0(1-\theta)^2$$
$$= 6\theta^2(1-\theta)^3 + \theta^3(1-\theta)^2.$$

Now we have

$$\frac{\mathrm{P}\{X_1{=}1\}\mathrm{P}\{X_2{=}1\}}{\mathrm{P}\{X_1{+}2X_2{=}3\}} = \frac{6\theta^2(1{-}\theta)^3}{6\theta^2(1{-}\theta)^3 + \theta^3(1{-}\theta)^2} = \frac{6{-}6\theta}{6{-}5\theta},$$

which involves θ. So $X_1 + 2X_2$ is not sufficient.

In general, a way to see whether or not $T(X)$ is a sufficient statistic, that is, to see whether or not the conditional distribution of $X \mid T$ involves the parameter vector θ, is to form the ratio $f_X(x\,;\theta)/f_T(T(x)\,;\theta)$ of the pmf's or pdf's and see whether or not θ cancels out. This is what we did in parts (b) and (c) of the preceding example, using pmf's since the random variables were discrete. Next we look at a continuous example.

Example 6.2.3. Consider a single observation $X \sim$ Laplace with scale parameter θ. That is, the pdf of X is $f(x\,;\theta) = \frac{1}{2\theta} e^{-\frac{1}{\theta}|x|}$ for all $x \in (-\infty, \infty)$ for some $\theta > 0$. (Note that we just have $n = 1$.) The textbook shows that $T = |X|$ is a sufficient statistic. Let's verify this in a different way, by showing that θ cancels out of the ratio $f_X(x\,;\theta)/f_T(|x|\,;\theta)$.

First, we refer back to Chapter 4. Applying formula (4.4.2) to the transformation $T = h(X) = |X|$, we obtain (as in Example 4.4.4):

$$f_T(t) \;=\; f_X(t) + f_X(-t) \text{ for all } t > 0.$$

(To arrive at this, note that for any value $t > 0$, there are two values of x for which $|x| = t$, namely $x = t$ and $x = -t$, and note that $|\frac{\mathrm{d}}{\mathrm{d}t} t| = 1$ and $|\frac{\mathrm{d}}{\mathrm{d}t}(-t)| = 1$.) This formula is true for any continuous random variable X.

In our case, $f_T(t\,;\theta) = f(t\,;\theta) + f(-t\,;\theta) = \frac{1}{2\theta} e^{-\frac{1}{\theta}|t|} + \frac{1}{2\theta} e^{-\frac{1}{\theta}|-t|} = \frac{1}{\theta} e^{-\frac{1}{\theta}t}$. (So $|X|$ has an Exponential distribution.) Now

$$\frac{f_X(x\,;\theta)}{f_T(|x|\,;\theta)} \;=\; \frac{\frac{1}{2\theta} e^{-\frac{1}{\theta}|x|}}{\frac{1}{\theta} e^{-\frac{1}{\theta}|x|}} \;=\; \frac{1}{2},$$

which does not involve θ. $\triangle$

Examples 6.2.4, 6.2.6. X_1, X_2 i.i.d. Normal$(\mu, 1)$.

(a) Is $T = X_1 + X_2$ sufficient? In Example 6.2.4 it is shown that T is sufficient by using results from sections 3.6 and 4.6.1. Let us obtain the same answer by looking at the ratio of pdf's.

We know $T \sim$ Normal$(2\mu, 2)$. The ratio $f_X(x\,;\mu)/f_T(T(x)\,;\mu)$ is

$$\frac{f(x_1;\mu) f(x_2;\mu)}{f_T(x_1+x_2\,;\mu)} \;=\; \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_1-\mu)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_2-\mu)^2}}{\frac{1}{\sqrt{4\pi}} e^{-\frac{1}{4}(x_1+x_2-2\mu)^2}}$$

$$= \frac{1}{\sqrt{\pi}} \exp[-\tfrac{1}{2}(x_1-\mu)^2 - \tfrac{1}{2}(x_2-\mu)^2 + \tfrac{1}{4}(x_1+x_2-2\mu)^2].$$

It is not yet obvious, but μ cancels out. To see this, note that

$$(x_1 + x_2 - 2\mu)^2 \;=\; [(x_1-\mu) + (x_2-\mu)]^2$$

$$= (x_1-\mu)^2 + (x_2-\mu)^2 + 2(x_1-\mu)(x_2-\mu)$$

and hence

$$\begin{aligned} &-\tfrac{1}{2}(x_1-\mu)^2 - \tfrac{1}{2}(x_2-\mu)^2 + \tfrac{1}{4}(x_1+x_2-2\mu)^2 \\ &\quad = -\tfrac{1}{4}(x_1-\mu)^2 - \tfrac{1}{4}(x_2-\mu)^2 + \tfrac{1}{2}(x_1-\mu)(x_2-\mu) \\ &\quad = -\tfrac{1}{4}[(x_1-\mu)-(x_2-\mu)]^2 \\ &\quad = -\tfrac{1}{4}(x_1-x_2)^2 . \end{aligned}$$

Therefore, $X_1 + X_2$ is a sufficient statistic.

(b) Is $T = X_1 + 2X_2$ sufficient? In Example 6.2.6 it is shown that T is not sufficient. Let us obtain the same answer by looking at the ratio of pdf's. Note that $T \sim \text{Normal}(3\mu, 5)$. The ratio $f_{\boldsymbol{X}}(\boldsymbol{x}\,;\mu)/f_T(T(\boldsymbol{x})\,;\mu)$ is

$$\frac{f(x_1;\mu)f(x_2;\mu)}{f_T(x_1+2x_2\,;\mu)} = \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_1-\mu)^2}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_2-\mu)^2}}{\frac{1}{\sqrt{10\pi}}e^{-\frac{1}{10}(x_1+2x_2-3\mu)^2}}$$

$$= \frac{1}{\sqrt{4\pi}}\exp[-\tfrac{1}{2}(x_1-\mu)^2 - \tfrac{1}{2}(x_2-\mu)^2 + \tfrac{1}{10}(x_1+2x_2-3\mu)^2] .$$

We can try the same algebraic manipulations as in part (a):

$$\begin{aligned} (x_1+2x_2-3\mu)^2 &= [(x_1-\mu)+2(x_2-\mu)]^2 \\ &= (x_1-\mu)^2 + 4(x_2-\mu)^2 + 4(x_1-\mu)(x_2-\mu) \end{aligned}$$

and

$$\begin{aligned} &-\tfrac{1}{2}(x_1-\mu)^2 - \tfrac{1}{2}(x_2-\mu)^2 + \tfrac{1}{10}(x_1+2x_2-3\mu)^2 \\ &\quad = -\tfrac{4}{10}(x_1-\mu)^2 - \tfrac{1}{10}(x_2-\mu)^2 + \tfrac{4}{10}(x_1-\mu)(x_2-\mu) \\ &\quad = -\tfrac{1}{10}[2(x_1-\mu)-(x_2-\mu)]^2 \\ &\quad = -\tfrac{1}{4}(2x_1-x_2-\mu)^2 , \end{aligned}$$

which involves μ. So $X_1 + 2X_2$ is not sufficient. $\triangle$

In the preceding examples of sufficient statistics, we had to first guess the statistic T and then verify that the conditional distribution of $X \mid T$ did not involve θ. Next we present the Factorization Theorem, which allows us to determine a sufficient statistic without having to guess. And even if we made the right guess, applying the Factorization Theorem is usually much easier than figuring out the conditional distribution of $X \mid T$ or dealing with the ratio $f_X(x\,;\theta)/f_T(T(x)\,;\theta)$.

(Strictly speaking, it is very easy to determine a sufficient statistic because X itself is sufficient. Of course what we are really looking for is a sufficient statistic that is more concise than X.)

Let $\mathcal{X}$ denote the sample space of all possible data vectors.

Factorization Theorem. Let $f(x\,;\theta)$ be the joint pmf or pdf of X. A statistic $T(X)$ is sufficient if and only if the joint pmf or pdf can be written as $f(x\,;\theta) = g(T(x)\,;\theta)\,h(x)$ for all $x \in \mathcal{X}$ and $\theta \in \Theta$.

(This is item (6.2.9) in the textbook; see Theorem 6.2.1 for the case when T is real-valued.) The factor $g(T(x)\,;\theta)$ can involve x only through $T(x)$ but it can involve θ in any form. The factor $h(x)$ can involve x in any form but it cannot involve θ at all. The factors are not unique. For instance, one could multiply g by 2 and divide h by 2 to get a slightly different factorization.

To apply this theorem, first write down the joint pmf or pdf $f(x\,;\theta)$. Try to factor out everything you can that does not involve θ. This constitutes $h(x)$. Now look at what is left to see how x is involved. If x appears only in the form $T(x)$, then the theorem says that $T(X)$ is a sufficient statistic.

Example. $X_1 \sim \text{Binomial}(n_1, \theta)$, $X_2 \sim \text{Binomial}(n_2, \theta)$ (same θ), independent of one another. To find a sufficient statistic, we first write down the joint pmf:

$$f(x_1, x_2\,;\theta) = \binom{n_1}{x_1}\theta^{x_1}(1-\theta)^{n_1-x_1}\binom{n_2}{x_2}\theta^{x_2}(1-\theta)^{n_2-x_2}.$$

Looking for factors that involve only the x's, we form $h(x_1, x_2) = \binom{n_1}{x_1}\binom{n_2}{x_2}$. This leaves

$$\begin{aligned}
&\theta^{x_1}(1-\theta)^{n_1-x_1}\theta^{x_2}(1-\theta)^{n_2-x_2} \\
&\quad = \theta^{x_1}\theta^{x_2}(1-\theta)^{n_1-x_1}(1-\theta)^{n_2-x_2} \\
&\quad = \theta^{x_1+x_2}(1-\theta)^{n_1+n_2-x_1-x_2} \\
&\quad = \theta^{x_1+x_2}(1-\theta)^{n_1+n_2-(x_1+x_2)},
\end{aligned}$$

which is a function of the x's only through $x_1 + x_2$. Thus we have

$$f(x_1, x_2) = g(x_1 + x_2\,;\theta)\,h(x_1, x_2)$$

where $g(t\,;\theta) = \theta^t(1-\theta)^{n_1+n_2-t}$. $\triangle$

Proof of the Factorization Theorem in the discrete case.

1) Assume that $T(X)$ is a sufficient statistic. Then the conditional distribution of $X \mid T$ does not involve θ. That is, the ratio $f(x\,;\theta)/p(T(x)\,;\theta)$ does not involve θ, where $p(t\,;\theta)$ is the pmf of $T(X)$. Let $h(x)$ denote the ratio and let $g(T(x)\,;\theta) = p(T(x)\,;\theta)$. Check that $f(x\,;\theta) = p(T(x)\,;\theta)\,h(x)$.

2) Conversely, assume that $f(x\,;\theta) = g(T(x)\,;\theta)\,h(x)$. The pmf of $T(X)$ is

$$\begin{aligned} p(t\,;\theta) &= \mathrm{P}_\theta\{T(X) = t\} \\ &= \sum_{x\,:\,T(x)=t} \mathrm{P}_\theta\{X = x\} \\ &= \sum_{x\,:\,T(x)=t} f(x\,;\theta) \\ &= \sum_{x\,:\,T(x)=t} g(T(x)\,;\theta)\,h(x) \\ &= g(t\,;\theta)\Big[\sum_{x\,:\,T(x)=t} h(x)\Big]. \end{aligned}$$

The ratio $f(x\,;\theta)/p(T(x)\,;\theta)$ becomes

$$\frac{f(x\,;\theta)}{p(T(x)\,;\theta)} = \frac{g(T(x)\,;\theta)\,h(x)}{g(T(x)\,;\theta)\Big[\sum_{x'\,:\,T(x')=T(x)} h(x')\Big]} = \frac{h(x)}{\sum_{x'\,:\,T(x')=T(x)} h(x')}.$$

This does not involve θ, and so $T(X)$ is a sufficient statistic. $\square$

The proof for the continuous case is more difficult and is omitted.

Example. As in Example 6.2.3, consider a single observation X with pdf $f(x\,;\theta) = \frac{1}{2\theta}e^{-\frac{1}{\theta}|x|}$. This is a function of x only through $|x|$. We can conclude that $T = |X|$ is a sufficient statistic by putting $h(x) = 1$ in the Factorization Theorem. This example is a special case of the following lemma.

Lemma. If $f(x\,;\theta)$ is a function of x only through $T(x)$, then $T(X)$ is a sufficient statistic.

To say that $f(x\,;\theta)$ is a function of x only through $T(x)$ means that $f(x\,;\theta) = g(T(x)\,;\theta)$ for some function g. So the lemma follows from the Factorization Theorem by putting $h(x) = 1$.

Example. $X_1, \ldots, X_n$ i.i.d. Normal$(\mu, 1)$.
(Note that we are assuming that the variance of the population is known to be 1.)
The pdf of a single Normal$(\mu, 1)$ random variable is

$$f(x\,;\mu) \;=\; \tfrac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\mu)^2}\,.$$

Hence the joint pdf of the data vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ is

$$f(\boldsymbol{x}\,;\mu) \;=\; \left(\tfrac{1}{\sqrt{2\pi}}\right)^n \exp\big[-\tfrac{1}{2}\textstyle\sum (x_i-\mu)^2\big]\,.$$

Since

$$(x_i-\mu)^2 \;=\; x_i^2 + 2\,x_i\mu + \mu^2\,,$$

we have

$$\textstyle\sum (x_i-\mu)^2 \;=\; \sum x_i^2 - 2\mu\sum x_i + n\mu^2\,.$$

Hence

$$\begin{aligned} f(\boldsymbol{x}\,;\mu) \;&=\; \left(\tfrac{1}{\sqrt{2\pi}}\right)^n \exp\Big[-\tfrac{1}{2}\textstyle\sum x_i^2\Big]\, \exp\Big[\mu\textstyle\sum x_i - \tfrac{1}{2}\,n\mu^2\Big] \\ &=\; g(\textstyle\sum x_i\,;\mu)\, h(\boldsymbol{x}) \end{aligned}$$

where

$$g(t\,;\mu) \;=\; \exp\big[\mu\, t - \tfrac{1}{2}\,n\mu^2\big] \quad \text{and} \quad h(\boldsymbol{x}) \;=\; \left(\tfrac{1}{\sqrt{2\pi}}\right)^n \exp\Big[-\tfrac{1}{2}\textstyle\sum x_i^2\Big]\,.$$

Thus we see that $T = \sum X_i$ is a sufficient statistic. $\triangle$

Lemma. A one-to-one function of a sufficient statistic is also sufficient.

This is intuitively clear because a one-to-one function of T must retain all the information that T contains. Since $\overline{X}$ is a one-to-one function of $\sum X_i$, we see that $\overline{X}$ is also a sufficient statistic for an i.i.d. sample from a Normal$(\mu, 1)$ population. .

Not every function of a sufficient statistic is sufficient. If the function is not one-to-one, then it may lose some information. That is, if W is a function of T, and if T is sufficient, then it is not necessarily true that W is sufficient. On the other hand, if W is a function of T, and if W is sufficient, then it can be concluded that T is sufficient. This is because, whatever information can be obtained from W, the same information can certainly be obtained from T, because W can be obtained from T.

Lemma. If W is a sufficient statistic, and if it is a function of T, then T is sufficient.

Next we look at an example with 2 parameters.

Example 6.2.10. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2) random variables with μ and σ^2 both unknown. The parameter vector is $\boldsymbol{\theta} = (\mu, \sigma^2)$. The joint pdf is

$$f(\boldsymbol{x}\,;\mu,\sigma^2) \;=\; \left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\Big[-\tfrac{1}{2\sigma^2}\textstyle\sum(x_i-\mu)^2\Big]\,.$$

As in the preceding example, $\sum(x_i-\mu)^2 = \sum x_i^2 - 2\mu\sum x_i + n\mu^2$, and so

$$f(\boldsymbol{x}\,;\mu\,,\sigma^2) = \left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\Big[-\tfrac{n\mu^2}{2\sigma^2}\Big]\exp\Big[\Big(\tfrac{-1}{2\sigma^2}\Big)\textstyle\sum x_i^2 + \big(\tfrac{\mu}{\sigma^2}\big)\sum x_i\Big]\,.$$

$$= g(\textstyle\sum x_i\,,\sum x_i^2\,;\mu,\sigma^2)\,h(\boldsymbol{x})$$

where $g(t_1,t_2\,;\mu,\sigma^2) = \left(\tfrac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\Big[-\tfrac{n\mu^2}{2\sigma^2}\Big]\exp\Big[\Big(\tfrac{-1}{2\sigma^2}\Big)t_2 + \big(\tfrac{\mu}{\sigma^2}\big)t_1\Big]$ and $h(\boldsymbol{x}) = 1$.

Now the Factorization Theorem implies that $(\sum X_i\,,\sum X_i^2)$ is a sufficient statistic. Since $(\overline{X}\,,S^2) = (\frac{1}{n}\,T_1\,,\frac{1}{n-1}[\,T_2 - \frac{1}{n}\,T_1^2])$ is a one-to-one function of $(T_1\,,T_2) = (\sum X_i\,,\sum X_i^2)$, we see that $(\overline{X}\,,S^2)$ is also sufficient. $\triangle$

Caution. In the preceding example, it is <u>not</u> proper to say that $\overline{X}$ is a sufficient statistic for μ nor to say that S^2 is a sufficient statistic for σ^2. It is only proper to make the joint statement that $(\overline{X}\,,S^2)$ is a sufficient statistic for (μ,σ^2). The concept of a sufficient statistic for $\boldsymbol{\theta}$ is well-defined <u>only</u> when $\boldsymbol{\theta}$ is the <u>entire</u> parameter vector. In the preceding example, why would it be wrong to say that $\overline{X}$ is sufficient for μ? To say that would imply that $\overline{X}$ is all one needs in order to obtain optimal inference procedures for μ. But it turns out that an optimal test of $\mathrm{H}_0 : \mu = \mu_0$ is based on the t-statistic $T = (\overline{X}-\mu_0)\Big/\sqrt{S^2/n}$, which requires S^2 as well as $\overline{X}$.

Next is an example in which the support of the distribution depends on the parameter.

Example 6.2.13. Let $X_1,\ldots,X_n$ be i.i.d. Uniform$(0,\theta)$ random variables where θ is an unknown positive parameter. The pdf is

$$f(x\,;\theta) \;=\; \begin{cases} \frac{1}{\theta} & \text{for } 0<x<\theta \\ 0 & \text{otherwise} \end{cases}$$

or

$$f(x\,;\theta) \;=\; \tfrac{1}{\theta}I\{0<x<\theta\}\,.$$

The joint pdf of the sample is

$$f(x_1,\ldots,x_n\,;\theta) \;=\; \tfrac{1}{\theta^n}\prod_{i=1}^{n} I\{0<x_i<\theta\}\,.$$

Indicators functions can be helpful when dealing with a distribution whose support depends on a parameter. In general, for any statement A,

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false} \end{cases}.$$

Note that $I(A)I(B) = I(A \text{ and } B)$ and $I(A_1)\cdots I(A_n) = I(A_1 \text{ and } \cdots \text{and } A_n)$.

In this example, we have $\prod_{i=1}^{n} I\{0 < x_i < \theta\} = I\{0 < x_1 < \theta \text{ and } \cdots \text{and } 0 < x_n < \theta\} = I\{0 < x_{(1)} < x_{(n)} < \theta\} = I\{0 < x_{(1)}\} I\{x_{(n)} < \theta\}$. Now

$$\begin{aligned} f(x_1, \ldots, x_n ; \theta) &= \frac{1}{\theta^n} I\{0 < x_{(1)}\} I\{x_{(n)} < \theta\} \\ &= g(x_{(n)}; \theta)\, h(\boldsymbol{x}) \end{aligned}$$

where $g(t ; \theta) = \frac{1}{\theta^n} I\{t < \theta\}$ and $h(\boldsymbol{x}) = I\{0 < x_{(1)}\}$. So $X_{(n)}$ is a sufficient statistic. $\triangle$

Exponential families

We will see that exponential families behave nicely with regard to sufficient statistics. Let us first review section 3.8 on exponential families.

Let X be a real-valued random variable with pmf or pdf $f(x ; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ may be a vector. The family $\{f(x ; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ is an *exponential family* of pmf's or pdf's if one can express

$$f(x ; \boldsymbol{\theta}) = a(\boldsymbol{\theta}) h(x) \exp\left\{ \sum_{j=1}^{k} b_j(\boldsymbol{\theta})\, R_j(x) \right\}.$$

To be a pmf or pdf requires $a(\boldsymbol{\theta}) > 0$ and $h(x) \geq 0$. An exponential family is called *regular* if

(a) $k = p$,
(b) Θ contains a p-dimensional rectangle, and
(c) the functions $b_j(\boldsymbol{\theta})$ are differentiable.

Examples. The pmf's or pdf's of the following families of distributions are exponential families:

Binomial(n, θ), $0 < \theta < 1$ (n known)
Normal(μ, σ^2), $-\infty < \mu < \infty$, $\sigma^2 > 0$
Normal(μ, σ_0^2), $-\infty < \mu < \infty$ (σ_0^2 known)
Normal(μ_0, σ^2), $\sigma^2 > 0$ (μ_0 known)
Geometric(θ), $0 < \theta < 1$
Negative binomial(μ, k), $\mu > 0$ (k known)
Poisson(λ), $\lambda > 0$
Gamma(α, β), $\alpha > 0$, $\beta > 0$

$\text{Beta}(\alpha, \beta)$, $\alpha > 0$, $\beta > 0$.

These are all regular exponential families. An exponential family that is not regular is

$\text{Normal}(\mu, \mu^2)$, $\mu > 0$.

Examples. The following families are not exponential families:

$\text{Uniform}(0, \theta)$, $\theta > 0$

$\text{Uniform}(\theta_1, \theta_2)$, $-\infty < \theta_1 < \theta_2 < \infty$

$\{f(x\,;\mu,\sigma) : -\infty < \mu < \infty,\ 0 < \sigma < \infty\}$ where

$$f(x\,;\mu,\sigma) = \tfrac{1}{\sigma}\exp\big[-\tfrac{1}{\sigma}(x-\mu)\big]\, I_{(\mu,\infty)}(x)$$

$\text{Cauchy}(\theta, 1)$, $-\infty < \theta < \infty$

$\text{Laplace}(\theta, 1)$, $-\infty < \theta < \infty$

$\text{Weibull}(\alpha, 1)$, $\alpha > 0$.

The first three examples above are not exponential families because their supports depend on the parameter vector. In an exponential family, note that the support $\{x \in \mathcal{X} : f(x\,;\theta) > 0\}$ $= \{x \in \mathcal{X} : h(x) > 0\}$ is the same for all θ. In the last three examples, the supports do not depend on the parameters, but for other reasons (that are not easy to prove) they are not exponential families.

Theorem 6.2.2. Let $X_1, \ldots, X_n$ be i.i.d. random variables with pmf or pdf in an exponential family. The statistic $\boldsymbol{T} = (\sum_{i=1}^n R_1(X_i), \ldots, \sum_{i=1}^n R_k(X_i))$ is sufficient.

To prove this, write out the joint pmf or pdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$, using the facts

$$\prod_{i=1}^n \exp\Big[\sum_{j=1}^k b_j(\boldsymbol{\theta}) R_j(x_i)\Big] \;=\; \exp\Big[\sum_{i=1}^n \sum_{j=1}^k b_j(\boldsymbol{\theta}) R_j(x_i)\Big]$$

and

$$\sum_{i=1}^n \sum_{j=1}^k b_j(\boldsymbol{\theta}) R_j(x_i) \;=\; \sum_{j=1}^k b_j(\boldsymbol{\theta}) \sum_{i=1}^n R_j(x_i)\,.$$

Then apply the Factorization Theorem.

In this theorem note that the exponential family does <u>not</u> have to be regular.

Minimal sufficient statistics

A good summary statistic should be concise and informative. We have seen that if a statistic is sufficient, then it contains as much information about the parameter vector θ as the whole data vector does. So no information is lost by reducing the data to a sufficient statistic. Besides being informative, we would like the statistic to be concise. That is, if we can take a sample of 100 numbers and summarize all the information in just 2 numbers, this is better than summarizing it in 3 numbers. The more concise, the better. The most concise sufficient statistic is a minimal one.

Definition 6.3.1. A statistic T is called a *minimal sufficient statistic* if (i) T is sufficient, and (ii) it is a function of every other sufficient statistic.

Recall that a function is either one-to-one or (at least partly) many-to-one; a function is never one-to-many. A function $T(X)$ of the data vector X can be regarded, according to whether the function is many-to-one or one-to-one, either as a reduction of the data or as equivalent to the data vector. If one function $T(X)$ is a function of another function $W(X)$, then, according to whether the function is many-to-one or one-to-one, T either is a greater reduction than W or is equivalent to W. Thus, the definition above says that a minimal sufficient statistic achieves the greatest reduction in the data while at the same time retaining sufficiency.

Two facts about minimal sufficiency that we can state right away are:

(1) A one-to-one function of a minimal sufficient statistic is also minimal sufficient.

(2) If $T(X)$ is a minimal sufficient statistic and $U(X)$ is a statistic such that $T(X)$ is not a function of $U(X)$, then $U(X)$ is not sufficient.

Lemma. $T(X)$ is a function of $W(X)$ if and only if whenever $W(x) = W(y)$ then $T(x) = T(y)$.

Proof. ($\Rightarrow$): Suppose $T(X)$ is a function of $W(X)$. This means that there is some function $H(w)$ such that $T(x) = H(W(x))$ for all x. If so, then whenever $W(x) = W(y)$, we must have $T(x) = H(W(x)) = H(W(y)) = T(y)$.
($\Leftarrow$): Conversely, suppose that, whenever $W(x) = W(y)$, then $T(x) = T(y)$. Define a function $H(w)$ as follows. For any w, choose x so that $W(x) = w$. (If no such x exists, then $H(w)$ can be assigned any value — it doesn't matter.) Then define $H(w) = T(x)$. Is this definition well-defined? What if we choose a different point y so that $W(y) = w$? Then $W(x) = w = W(y)$, so $T(x) = T(y)$ and we get the same value

by setting $H(w) = T(y)$. This shows that the definition of H is well-defined. Note that $T(x) = H(W(x))$. □

Theorem. Let $f(x\,;\theta)$ be the joint pmf or pdf of a data vector X. Suppose $T(X)$ is a statistic such that $T(x) = T(y)$ implies that $f(x\,;\theta) = c(x,y)f(y\,;\theta)$ for all θ, where $c(x,y)$ does not involve θ. Then $T(X)$ is a sufficient statistic.

Proof. Choose a vector θ_0; it can be any vector, but once it is chosen, it is known. Now $T(x) = T(y)$ implies

$$\frac{f(x\,;\theta)}{f(x\,;\theta_0)} = \frac{c(x,y)f(y\,;\theta)}{c(x,y)f(y\,;\theta_0)} = \frac{f(y\,;\theta)}{f(y\,;\theta_0)}.$$

By the lemma above, this means that $f(x\,;\theta)/f(x\,;\theta_0)$ is a function of $T(x)$. This statement can be made for each value of θ. In symbols, this means that $f(x\,;\theta)/f(x\,;\theta_0) = g(T(x);\theta)$. Letting $h(x) = f(x\,;\theta_0)$, we can write $f(x\,;\theta) = g(T(x);\theta)h(x)$. From the Factorization Theorem we can now conclude that $T(X)$ is sufficient. □

Theorem 6.3.1. Let $f(x\,;\theta)$ be the joint pmf or pdf of a data vector X. Suppose $T(X)$ is a statistic such that $T(x) = T(y)$ if and only if $f(x\,;\theta) = c(x,y)f(y\,;\theta)$ for all θ, where $c(x,y)$ does not involve θ. Then $T(X)$ is a minimal sufficient statistic.

Let us digress from the textbook for a moment to see how this theorem can be used to show the equivalence of two principles of statistical inference. Suppose that an experiment is performed in which a data vector X is generated according to a probability distribution with joint pmf or pdf $f(x\,;\theta)$ where θ is an unknown parameter vector.

Sufficiency Principle: Suppose $T(X)$ is a sufficient statistic for θ. Statistical inference about θ should depend only on $T(X)$.

Note that a principle is not the same thing as a theorem. The Sufficiency Principle is neither true nor false. It is proposed merely as a sensible approach to statistical inference. A statistician can choose to either follow or not follow the Sufficiency Principle, but most statisticians follow it most of the time, except when they do model-checking.

According to this principle, if we are trying, for example, to find a good estimator of a coordinate θ_1 of the parameter vector, we can restrict our attention to functions $W(T(X))$ of the sufficient statistic. The Sufficiency Principle says that a sufficient statistic is "sufficient" (in the nonstatistical everyday sense of the word) for doing inference about the parameters. Another way to express this principle is as follows.

Sufficiency Principle (restated): Suppose $T(X)$ is a sufficient statistic for θ. Suppose x and y are two possible data vectors in the sample space such that $T(x) = T(y)$. In any statistical inferential procedure, the conclusion about θ from observing $X = x$ should be the same as the conclusion from observing $X = y$.

Next consider the concept of likelihood.

Definition. Let X be a random data vector with joint pmf or pdf $f(x\,;\theta)$. Given an observed data vector $X = x$, the *likelihood function* is $L(\theta\,;x) = f(x\,;\theta)$, regarded as a function of θ.

The pmf or pdf is a function of x for each fixed θ, whereas the likelihood function is a function of θ for each fixed x.

Let θ and θ' be two possible parameter vectors. If we observe $X = x$ and if $L(\theta\,;x) > L(\theta';x)$, then θ is more "likely" to be the true parameter vector than θ' is. The ratio $L(\theta\,;x)/L(\theta';x)$ can be called the *relative likelihood* of θ versus θ'. We see that θ is more likely than θ' if and only if the relative likelihood is greater than 1.

Likelihood Principle: Statistical inference about θ should depend only on the relative likelihoods of the possible parameter vectors.

Suppose x and y are two data points such that the corresponding likelihood functions give the same relative likelihoods for all pairs of parameter vectors. That is, $L(\theta\,;x)/L(\theta';x) = L(\theta\,;y)/L(\theta';y)$ for all θ and θ' in the parameter space. This is equivalent to the condition that $L(\theta\,;x)/L(\theta\,;y) = L(\theta';x)/L(\theta';y)$ for all θ and θ', that is, $L(\theta\,;x)/L(\theta\,;y)$ does not involve θ. Thus we have the restatement:

Likelihood Principle (restated): Suppose x and y are two possible data vectors in the sample space such that $f(x\,;\theta) = c(x,y)f(y\,;\theta)$ for all θ, where c may depend on x and y but not on θ. In any statistical inferential procedure, the conclusion about θ from observing $X = x$ should be the same as the conclusion from observing $X = y$.

Recall Theorem 6.3.1. Let (*) denote the property that $T(x) = T(y)$ if and only if $f(x\,;\theta) = c(x,y)f(y\,;\theta)$ for all θ. Theorem 6.3.1 states that if a statistic $T(X)$ satisfies property (*), then it is minimal sufficient. It can be shown that the converse is also true. Therefore, the Likelihood Principle is equivalent to saying that statistical inference about θ should be based on a minimal sufficient statistic, which is therefore equivalent to the Sufficiency Principle.

Now we return to the textbook. Let us see how Theorem 6.3.1 can be used to find minimal sufficient statistics. Suppose we are given a family of pmf's or pdf's $\{f(\boldsymbol{x}\,;\theta) : \theta \in \Theta\}$ and we want to find a minimal sufficient statistic. The steps are: (a) form the ratio $f(\boldsymbol{x}\,;\theta)/f(\boldsymbol{y}\,;\theta)$, (b) try to simplify the ratio using algebraic manipulations, and (c) identify functions of $\boldsymbol{x}$, say $T_1(\boldsymbol{x}), \ldots, T_r(\boldsymbol{x})$, such that the ratio is constant for all θ if and only if $T_1(\boldsymbol{x}) = T_1(\boldsymbol{y}), \ldots,$ and $T_r(\boldsymbol{x}) = T_r(\boldsymbol{y})$. Then $T(\underset{\sim}{X}) = (T_1(\boldsymbol{X}), \ldots, T_r(\boldsymbol{X}))$ is a minimal sufficient statistic.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Geometric(θ), $0 < \theta < 1$. The joint pmf is

$$f(\boldsymbol{x}\,;\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{\sum x_i - n}.$$

Consider the ratio

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{\theta^n (1-\theta)^{\sum x_i - n}}{\theta^n (1-\theta)^{\sum y_i - n}} = (1-\theta)^{\sum x_i - \sum y_i}.$$

The ratio is constant for all θ if and only if $\sum x_i - \sum y_i = 0$, i.e., $\sum x_i = \sum y_i$. By Theorem 6.3.1, $T = \sum X_i$ is a minimal sufficient statistic. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2) random variables with μ and σ^2 both unknown. The joint pdf of the data vector $\boldsymbol{X}$ is

$$\begin{aligned} f(\boldsymbol{x};\mu,\sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}(x_i - \mu)^2 \Big] \\ &= \Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big] \end{aligned}$$

Consider the ratio

$$\begin{aligned} \frac{f(\boldsymbol{x};\mu,\sigma^2)}{f(\boldsymbol{y};\mu,\sigma^2)} &= \frac{\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \Big]}{\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \Big]} \\ &= \exp\Big[\frac{-1}{2\sigma^2} \big\{ \sum (x_i - \mu)^2 - \sum (y_i - \mu)^2 \big\} \Big] \\ &= \exp\Big[\frac{-1}{2\sigma^2} \Big(\sum x_i^2 - \sum y_i^2 \Big) + \frac{\mu}{\sigma^2} \Big(\sum x_i - \sum y_i \Big) \Big]. \end{aligned}$$

If $\sum x_i^2 - \sum y_i^2 = 0$ and $\sum x_i - \sum y_i = 0$, then the ratio is 1 no matter what μ and σ^2 are. And if one of these two differences is nonzero, then the ratio would not be constant. Therefore, by Theorem 6.3.1, $T = (\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(0, \theta)$, $\theta > 0$. The joint pdf is

$$f(\boldsymbol{x}\,;\theta) \;=\; \prod_{i=1}^{n} \tfrac{1}{\theta} I\{0 < x_i < \theta\} \;=\; \tfrac{1}{\theta^n} I\{0 < x_{(1)}\}\, I\{x_{(n)} < \theta\}\,.$$

For this family of distributions, all observations are positive (or speaking more technically, they are positive with probability 1). Therefore, it is always true that $x_{(1)} > 0$ and so we may write the joint pdf as

$$f(\boldsymbol{x}\,;\theta) \;=\; \tfrac{1}{\theta^n} I\{x_{(n)} < \theta\}\,.$$

Form the ratio

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} \;=\; \frac{\frac{1}{\theta^n} I\{x_{(n)} < \theta\}}{\frac{1}{\theta^n} I\{y_{(n)} < \theta\}} \;=\; \frac{I\{x_{(n)} < \theta\}}{I\{y_{(n)} < \theta\}}\,.$$

First we address the potential problem that in this example the ratio might involve division by 0. Note that the ratio $0/0$, which is not well-defined, can be ignored. This is because, in the statement of Theorem 6.3.1 above, the condition on the pdf's is $f(\boldsymbol{x}\,;\theta) = c(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{y}\,;\theta)$. Note that if $f(\boldsymbol{x}\,;\theta) = f(\boldsymbol{y}\,;\theta) = 0$, then the condition is met.

If $x_{(n)} = y_{(n)}$, then we see that the ratio is either 1 or $0/0$ for all θ, and so the ratio is constant for all values of θ that cannot be ignored. Conversely, suppose the ratio is constant for all values of θ that cannot be ignored. The constant must be 1 because when θ is greater than both $x_{(n)}$ and $y_{(n)}$, the ratio is $1/1 = 1$. We will show that $x_{(n)} = y_{(n)}$ by supposing $x_{(n)} \neq y_{(n)}$ and arriving at a contradiction. Suppose $x_{(n)} < y_{(n)}$. Choose a value of θ between them, $x_{(n)} < \theta < y_{(n)}$. Then the ratio of pdf's is $1/0 = \infty \neq 1$. Similarly, if we suppose $y_{(n)} < x_{(n)}$, we obtain a ratio $0/1 = 0 \neq 1$. Thus we have shown that the ratio is constant if and only if $x_{(n)} = y_{(n)}$. This implies that $T = X_{(n)}$ is a minimal sufficient statistic. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(\theta, 2\theta)$, $\theta > 0$. The joint pdf is

$$f(\boldsymbol{x}\,;\theta) \;=\; \prod_{i=1}^{n} \tfrac{1}{\theta} I\{\theta < x_i < 2\theta\} \;=\; \tfrac{1}{\theta^n} I\{\theta < x_{(1)} < x_{(n)} < 2\theta\}\,.$$

Form the ratio

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} \;=\; \frac{\frac{1}{\theta^n} I\{\theta < x_{(1)} < x_{(n)} < 2\theta\}}{\frac{1}{\theta^n} I\{\theta < y_{(1)} < y_{(n)} < 2\theta\}} \;=\; \frac{I\{\theta < x_{(1)} < x_{(n)} < 2\theta\}}{I\{\theta < y_{(1)} < y_{(n)} < 2\theta\}}\,.$$

If $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$, then we see that the ratio is either 1 or $0/0$ for all θ, and so the ratio is constant for all values of θ that cannot be ignored. Conversely, suppose the ratio is constant for all values of θ that cannot be ignored. First we note that it is all right to ignore data vectors whose joint pdf is 0 for all θ. That is, we may as well eliminate from the sample

space any set of data vectors whose probability is 0 for all distributions in the model. In the last expression for the ratio, the numerator and denominator are indicator functions and so their values are either 0 or 1 for all θ. So the possible ratios are $0/0$ (which can be ignored), $0/1 = 0$, $1/0 = \infty$, and $1/1 = 1$. We have just noted that we can suppose each indicator function takes the value 1 for at least one θ, and so the only possible way these ratios can be constant (for all values θ that cannot be ignored) is for the ratio to always be $1/1 = 1$ (other than in the ignorable case when $0/0$). That is, a constant ratio implies that $\theta < x_{(1)} < x_{(n)} < 2\theta$ if and only if $\theta < y_{(1)} < y_{(n)} < 2\theta$. By an argument similar to the one in the preceding example, this implies $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$. $\triangle$

In these examples we have seen that, when applying Theorem 6.3.1 to a family of pdf's whose supports depend on the parameter vector $\boldsymbol{\theta}$, we can ignore the ratio $f(\boldsymbol{x};\theta)/f(\boldsymbol{y};\theta)$ for values of θ for which both pdf's are 0 (because then both $\boldsymbol{x}$ and $\boldsymbol{y}$ are impossible). In order for the ratio be constant for all $\boldsymbol{\theta}$, it is necessary (but not sufficient) that $\{\theta : f(\boldsymbol{x};\theta) > 0\} = \{\theta : f(\boldsymbol{y};\theta) > 0\}$.

Proof of Theorem 6.3.1. Suppose that $T(X)$ is a statistic having the property that $T(\boldsymbol{x}) = T(\boldsymbol{y})$ if and only if $f(\boldsymbol{x};\theta) = c(\boldsymbol{x},\boldsymbol{y})f(\boldsymbol{y};\theta)$ for all θ. By the preceding theorem, this implies T is sufficient. It remains to prove it is minimal. Let $W(X)$ be another sufficient statistic. We must show that T is a function of W. By the lemma on p. 12, we must show that $W(\boldsymbol{x}) = W(\boldsymbol{y}) \Rightarrow T(\boldsymbol{x}) = T(\boldsymbol{y})$. By assumption, this is equivalent to showing that $W(\boldsymbol{x}) = W(\boldsymbol{y}) \Rightarrow f(\boldsymbol{x};\theta)/f(\boldsymbol{y};\theta)$ is ~~constant~~ (the same) for all θ. The sufficiency of W implies, by the Factorization Theorem, that $f(\boldsymbol{x};\theta) = g(W(\boldsymbol{x});\theta)h(\boldsymbol{x})$. Hence, if $W(\boldsymbol{x}) = W(\boldsymbol{y})$, then

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{g(W(\boldsymbol{x});\theta)h(\boldsymbol{x})}{g(W(\boldsymbol{y});\theta)h(\boldsymbol{y})} = \frac{g(W(\boldsymbol{x});\theta)h(\boldsymbol{x})}{g(W(\boldsymbol{x});\theta)h(\boldsymbol{y})} = \frac{h(\boldsymbol{x})}{h(\boldsymbol{y})},$$

which does not involve θ. $\square$

Suppose $\boldsymbol{X}$ is a data vector with joint pmf or pdf $f(\boldsymbol{x};\theta)$. We let n denote the number of observations in $\boldsymbol{X} = (X_1, \ldots, X_n)$. Let p denote the number of parameters in $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. We say p is the *dimension* of the parameter vector. Suppose $T(X)$ is a sufficient statistic. Let r be its dimension; that is, $\boldsymbol{T} = (T_1, \ldots, T_r)$.

If the model is suitably chosen, then one should "almost always" be able to estimate all p parameters from the data. To estimate p parameters requires at least p pieces of information, and so it is "almost always" true that $n \geq p$. By the definition of sufficiency, if the data allow us to estimate all p parameters, then so does any sufficient statistic. This implies that "typically" $r \geq p$. But atypical counterexamples can be described; see Remark 6.3.1.

Even when T is minimal sufficient, it may be that $r > p$. The case when $r = p$ is a particularly nice one, as we will see later.

If the data are i.i.d. with pmf or pdf in an exponential family, then there is a sufficient statistic of dimension k, no matter how large n is. For some models, however, the dimension of a minimal sufficient statistic is n.

Example. Let $X_1, \ldots, X_n$ be i.i.d. from a Cauchy distribution with median θ, $-\infty < \theta < \infty$. The joint pdf is

$$f(\boldsymbol{x};\theta) = \prod_{i=1}^{n} \frac{1}{\pi[1+(x_i-\theta)^2]} = \frac{1}{\pi^n \prod_{i=1}^{n}[1+(x_i-\theta)^2]}.$$

To apply Theorem 6.3.1, consider the ratio

$$\frac{f(\boldsymbol{x};\theta)}{f(\boldsymbol{y};\theta)} = \frac{\pi^n \prod_{i=1}^{n}[1+(y_i-\theta)^2]}{\pi^n \prod_{i=1}^{n}[1+(x_i-\theta)^2]} = \frac{\prod_{i=1}^{n}[1+(y_i-\theta)^2]}{\prod_{i=1}^{n}[1+(x_i-\theta)^2]}.$$

As $\theta \to \infty$, the ratio approaches 1. So it is constant for all θ if and only if

$$\prod_{i=1}^{n}[1 + (x_i - \theta)^2] = \prod_{i=1}^{n}[1 + (y_i - \theta)^2]$$

for all θ. Each side of this equation can be multiplied out to form a polynomial in θ. It is a consequence of the Fundamental Theorem of Algebra that two polynomials in θ are equal for all θ (or even just for ~~all~~ an infinite number of θ ~~in some interval~~) if and only if the coefficients of the two polynomials are equal for all powers of θ. In the case $n = 2$, this leads to the conclusion that $x_1 + x_2 = y_1 + y_2$ and $x_1 x_2 = y_1 y_2$. This holds if and only if $x_{(1)} = y_{(1)}$ and $x_{(2)} = y_{(2)}$. For general n, it can be shown (but it's not easy) that the ratio is constant if and only if $x_{(1)} = y_{(1)}, \ldots,$ and $x_{(n)} = y_{(n)}$. Therefore, the vector of order statistics $T = (X_{(1)}, \ldots, X_{(n)})$ is a minimal sufficient statistic. $\triangle$

Theorem 6.3.3. Let $X_1, \ldots, X_n$ be i.i.d. random variables with pmf or pdf in a regular exponential family. The statistic $T = (\sum_{i=1}^{n} R_1(X_i), \ldots, \sum_{i=1}^{n} R_k(X_i))$ is minimal sufficient.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2) random variables with μ and σ^2 both unknown. The joint pdf is

$$f(\boldsymbol{x};\mu,\sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\left(\sum x_i^2 - 2\mu\sum x_i + n\mu^2\right)\right]$$

$$= \left\{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} n\mu^2\right)\right\} \exp\left[\left(-\frac{1}{2\sigma^2}\right)\left(\sum x_i^2\right) + \left(\frac{\mu}{\sigma^2}\right)\left(\sum x_i\right)\right]$$

$$= \left\{ a(\mu, \sigma^2) \right\} \exp\left[b_1(\mu, \sigma^2) R_1(\boldsymbol{x}) + b_2(\mu, \sigma^2) R_2(\boldsymbol{x})\right] .$$

So this has the exponential family form (put $h(\boldsymbol{x}) = 1$) with $R_1(\boldsymbol{x}) = \sum x_i^2$ and $R_2(\boldsymbol{x}) = \sum x_i$. (If you prefer, you could choose to change the subscripts in order to match the subscripts of the R's with the powers of the x's: $R_1(\boldsymbol{x}) = \sum x_i$ and $R_2(\boldsymbol{x}) = \sum x_i^2$.) The family is regular as defined on p. 10 above because (a) $k = p = 2$, (b) $\Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$, which is the upper half of the real plane, which certainly contains a 2-dimensional rectangle, and (c) the functions $-1/2\sigma^2$ and μ/σ^2 are differentiable. Therefore, $(\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(θ, θ) random variables with θ positive and unknown. The joint pdf is

$$f(\boldsymbol{x}; \theta) = \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left[-\frac{1}{2\theta}\sum_{i=1}^n (x_i - \theta)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left[-\frac{1}{2\theta}\left(\sum x_i^2 - 2\theta\sum x_i + n\theta^2\right)\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left[-\frac{1}{2\theta}\sum x_i^2 + \sum x_i - \frac{n\theta}{2}\right]$$

$$= \left\{\left(\frac{1}{\sqrt{2\pi\theta}}\right)^n \exp\left(-\frac{n\theta}{2}\right)\right\} \exp\left(\sum x_i\right) \exp\left[\left(-\frac{1}{2\theta}\right)\left(\sum x_i^2\right)\right]$$

$$= a(\theta)\, h(\boldsymbol{x}) \exp\left[b(\theta) R(\boldsymbol{x})\right]$$

where $R(\boldsymbol{x}) = \sum x_i^2$. The family is regular because (a) $k = p = 1$, (b) $\Theta = (0, \infty)$, which contains a 1-dimensional rectangle (a 1-dimensional rectangle is simply an interval), and (c) the function $-1/2\theta$ is differentiable. Therefore, $\sum X_i^2$ is a minimal sufficient statistic. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(θ, θ^2) random variables with θ positive and unknown. The joint pdf is

$$f(\boldsymbol{x}; \theta) = \left(\frac{1}{\sqrt{2\pi\theta^2}}\right)^n \exp\left[-\frac{1}{2\theta^2}\sum_{i=1}^n (x_i - \theta)^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\theta^2}}\right)^n \exp\left[-\frac{1}{2\theta^2}\left(\sum x_i^2 - 2\theta\sum x_i + n\theta^2\right)\right]$$

$$= \left\{ \left(\frac{1}{\sqrt{2\pi\theta^2}} \right)^n \exp\left(-\frac{n}{2} \right) \right\} \exp\left[\left(-\frac{1}{2\theta^2} \right) \left(\sum x_i^2 \right) + \left(\frac{1}{\theta} \right) \left(\sum x_i \right) \right]$$

$$= a(\theta) \exp\left[b_1(\theta) R_1(\boldsymbol{x}) + b_2(\theta) R_2(\boldsymbol{x}) \right] .$$

This has the exponential family form, but $k = 2$ and $p = 1$, so it is not regular. So Theorem 6.3.3 cannot be used. But we can apply Theorem 6.3.1 to show that the minimal sufficient statistic is $(\sum X_i , \sum X_i^2)$. $\triangle$

Next we will generalize Theorem 6.3.3. First we extend the definition of an exponential family.

Let $\boldsymbol{X}$ be a data vector with joint pmf or pdf $f(\boldsymbol{x} ; \boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ may be a vector. The family $\{ f(\boldsymbol{x} ; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \}$ is an *exponential family* of pmf's or pdf's if one can express

$$f(\boldsymbol{x} ; \boldsymbol{\theta}) = a(\boldsymbol{\theta}) h(\boldsymbol{x}) \exp\left\{ \sum_{j=1}^{k} b_j(\boldsymbol{\theta}) R_j(\boldsymbol{x}) \right\} .$$

To be a pmf or pdf requires $a(\boldsymbol{\theta}) > 0$ and $h(\boldsymbol{x}) \geq 0$. An exponential family is called *regular* if

(a) $k = p$,

(b) Θ contains a p-dimensional rectangle, and

(c) the functions $b_j(\boldsymbol{\theta})$ are differentiable.

Check that if $X_1, \ldots, X_n$ be i.i.d. from a distribution with pmf or pdf in an exponential family as on p. 10 above, then the joint pmf or pdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$ has the exponential family form with $a(\boldsymbol{\theta})$ being the n-th power of the $a(\boldsymbol{\theta})$ on p. 10, $h(\boldsymbol{x}) = \prod_{i=1}^{n} h(x_i)$, $b_j(\boldsymbol{\theta})$ being the same as on p. 10, and $R_j(\boldsymbol{x}) = \sum_{i=1}^{n} R_j(x_i)$. Note that if the exponential family for the individual X_i's is regular, then so is the exponential family for the vector $\boldsymbol{X}$.

Generalizing Theorem 6.3.3 (see p. 18), we have:

Theorem. Let $\boldsymbol{X}$ be a random vector with joint pmf or pdf in a regular exponential family. The statistic $\boldsymbol{T} = (R_1(\boldsymbol{X}), \ldots, R_k(\boldsymbol{X}))$ is minimal sufficient.

Information

When discussing the concept of sufficiency, we said that a sufficient statistic contains all of the information about $\boldsymbol{\theta}$ that the whole data vector does. This is just an informal description of what the concept of sufficiency is intended to embody, and so the word "information" is being used informally in its everyday sense. But now we will give a formal quantitative

definition of information. Given a data vector X with joint pmf or pdf $f(x\,;\theta)$, $\theta \in \Theta$, we will come up with a ~~number~~ quantity that measures, in some sense, the information that X contains about θ. First we suppose θ is a real-valued parameter.

The definition will be restricted to situations in which the family of joint pmf's or pdf's satisfy the following three regularity conditions:

(RC1) $f(x\,;\theta)$ has the same support for all θ.

(RC2) $f(x\,;\theta)$ is differentiable with respect to θ.

(RC3) For all statistics $W(X)$ whose expectation $\mathrm{E}_\theta(W)$ exists, the expectation is a differentiable function of θ and the derivative can be calculated by differentiating under the summation or integral sign.

Conditions RC1, RC2, and RC3 are satisfied in a one-parameter exponential family, provided that the functions $b_j(\theta)$ are differentiable, which they typically are.

Definition. (a) The *score* (or *score function*) is $\frac{\partial}{\partial\theta}\log f(X\,;\,\theta)$.

(b) The *information* (or *Fisher information*) is $\mathrm{Var}_\theta\big[\frac{\partial}{\partial\theta}\log f(X\,;\,\theta)\big]$.

Note that the score is a random variable, because it is a function of the random data vector X, but it is not a statistic, because it is also a function of the unknown parameter θ. The Fisher information is regarded as a measure of the information about θ that is contained in the data vector X; it is denoted by $\mathcal{I}_X(\theta)$.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal$(\mu\,,\sigma_0^2)$ random variables with μ unknown and σ_0^2 known. The joint pdf is

$$f(x;\mu) \;=\; \Big(\frac{1}{\sqrt{2\pi\sigma_0^2}}\Big)^n \exp\Big[-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu)^2\Big]\,.$$

So

$$\log f(x;\mu) \;=\; -\,n\log\sqrt{2\pi\sigma_0^2} - \frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu)^2\,,$$

and the score is

$$\frac{\partial}{\partial\mu}\log f(X;\mu) \;=\; \frac{1}{\sigma_0^2}\sum_{i=1}^n (X_i-\mu) \;=\; \frac{n}{\sigma_0^2}(\overline{X}-\mu)\,.$$

The information about μ contained in X is

$$\mathcal{I}_X(\mu) \;=\; \mathrm{Var}[\tfrac{n}{\sigma_0^2}(\overline{X}-\mu)] \;=\; \big(\tfrac{n}{\sigma_0^2}\big)^2\mathrm{Var}(\overline{X}) \;=\; \big(\tfrac{n}{\sigma_0^2}\big)^2\frac{\sigma_0^2}{n} \;=\; \frac{n}{\sigma_0^2}\,.$$

This is a sensible measure of information for two reasons. First, the information is proportional to the size of the sample. If the sample size is doubled, then the information is doubled. This makes sense. Second, if the variance of the distribution is larger, then, for a given sample size, the information in the sample is smaller. This also makes sense because variability interferes with our ability to determine the value of μ. $\triangle$

Theorem 6.4.1. Let $X_1, \ldots, X_n$ be i.i.d. with pmf or pdf $f(x;\theta)$ satisfying RC1, RC2, and RC3. Then $\mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta)$.

That is, the information contained in an i.i.d. sample of size n is n times the amount of information contained in a single observation.

Proof. For an i.i.d. sample,

$$f(\boldsymbol{X};\theta) = \textstyle\prod_{i=1}^n f(X_i;\theta)$$

$$\Rightarrow \log f(\boldsymbol{X};\theta) = \textstyle\sum_{i=1}^n \log f(X_i;\theta)$$

$$\Rightarrow \tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta) = \textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta}\log f(X_i;\theta)$$

$$\Rightarrow \mathrm{Var}\big[\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\big] = \mathrm{Var}\big[\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial\theta}\log f(X_i;\theta)\big]$$

$$= \textstyle\sum_{i=1}^n \mathrm{Var}\big[\tfrac{\partial}{\partial\theta}\log f(X_i;\theta)\big] \quad (\text{the } X_i\text{'s are independent})$$

$$= n\mathrm{Var}\big[\tfrac{\partial}{\partial\theta}\log f(X_1;\theta)\big] \quad (\text{the } X_i\text{'s are identically distributed})$$

$$\Rightarrow \mathcal{I}_X(\theta) = n\mathcal{I}_{X_1}(\theta). \square$$

Next we will find two alternative ways to calculate Fisher information.

Lemma. Assume conditions RC1, RC2, and RC3.

(a) If $\mathrm{E}_\theta(W)$ exists for all θ, then $\mathrm{E}_\theta\Big[W(\boldsymbol{X})\,\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = \tfrac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta[W(\boldsymbol{X})]$.

(b) $\mathrm{E}_\theta\Big[\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = 0$.

(c) $\mathrm{E}_\theta\Big(\Big[\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big]^2\Big) = \mathrm{Var}_\theta\Big[\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big]$.

Proof. Recall that $\tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta) = \tfrac{\frac{\partial}{\partial\theta}f(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)}$. For (a), in the continuous case,

$$\mathrm{E}_\theta\left[W(\boldsymbol{X})\,\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{X};\theta)}{f(\boldsymbol{X};\theta)}\right] = \int\left[W(\boldsymbol{x})\,\frac{\frac{\partial}{\partial\theta}f(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right]f(\boldsymbol{x};\theta)\,d\boldsymbol{x}$$

$$= \int W(\boldsymbol{x})\,\frac{\partial}{\partial\theta}f(\boldsymbol{x};\theta)\,d\boldsymbol{x}$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta}\int W(\boldsymbol{x}) f(\boldsymbol{x};\theta)\,d\boldsymbol{x} = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}_\theta[W(\boldsymbol{X})].$$

Part (b) follows from (a) by taking $W = 1$. Part (c) follows from (b). □

Part (c) gives another way to calculate information.

Example (continued). In the example above concerning an i.i.d. sample from a Normal(μ, σ_0^2) population, we could have calculated $\mathcal{I}_{\boldsymbol{X}}(\mu) = \mathrm{E}[\{\frac{n}{\sigma_0^2}(\overline{X} - \mu)\}^2] = (\frac{n}{\sigma_0^2})^2\mathrm{E}[(\overline{X} - \mu)^2]$, but it seems easier in this example to work in terms of the variance. △

Lemma. Assume conditions RC1, RC2, and RC3. Then

$$\mathrm{E}_\theta\left(\left[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\right]^2\right) = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{X};\theta)\right].$$

See Remark 6.4.1. The proof is based on the preceding lemma and involves differentiation under a summation or integral sign.

Example (continued). In the example concerning an i.i.d. sample from a Normal(μ, σ_0^2) population, we could have calculated $\mathcal{I}_{\boldsymbol{X}}(\mu) = -\mathrm{E}[\frac{\partial^2}{\partial\mu^2}\log f(\boldsymbol{X};\mu)]$. From above we know $\frac{\partial}{\partial\mu}\log f(\boldsymbol{X};\mu) = \frac{n}{\sigma_0^2}(\overline{X} - \mu)$, and so $\frac{\partial^2}{\partial^2\mu}\log f(\boldsymbol{X};\mu) = -\frac{n}{\sigma_0^2}$. Now $\mathcal{I}_{\boldsymbol{X}}(\mu) = -\mathrm{E}[-\frac{n}{\sigma_0^2}] = \frac{n}{\sigma_0^2}$. △

Example. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(θ), $0 < \theta < 1$. This family of distributions is a one-parameter exponential family and so conditions RC1, RC2, and RC3 hold. The pmf of a single observation is $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, so

$$\log f(x;\theta) = x\log\theta + (1-x)\log(1-\theta),$$

$$\frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} = \frac{x-\theta}{\theta(1-\theta)}, \text{ and}$$

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}.$$

The information in X_1 can be obtained as

$$\mathcal{I}_{X_1}(\theta) = \mathrm{Var}\left[\frac{X_1-\theta}{\theta(1-\theta)}\right] = \frac{\mathrm{Var}(X_1)}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

By Theorem 6.4.1, the information in the whole sample is $\mathcal{I}_{\boldsymbol{X}}(\theta) = n\mathcal{I}_{X_1}(\theta) = \frac{n}{\theta(1-\theta)}$.

Alternatively, we can calculate

$$\mathcal{I}_{X_1}(\theta) = -\mathrm{E}\left[-\frac{X_1}{\theta^2} - \frac{1-X_1}{(1-\theta)^2}\right] = \mathrm{E}\left[\frac{X_1}{\theta^2} + \frac{1-X_1}{(1-\theta)^2}\right]$$

$$= \frac{\mathrm{E}[X_1]}{\theta^2} + \frac{1-\mathrm{E}[X_1]}{(1-\theta)^2} = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{(1-\theta)} = \frac{1}{\theta(1-\theta)} \cdot \triangle$$

Another way in which Fisher information behaves like a measure of information should behave is given in the following theorem.

Theorem 6.4.2. Let $\underset{\sim}{X}$ be a data vector with pmf or pdf $f(\boldsymbol{x}\,;\theta)$ satisfying RC1, RC2, and RC3. Let $\boldsymbol{T} = \boldsymbol{T}(\underset{\sim}{X})$ be a statistic.

(a) $\mathcal{I}_{\boldsymbol{T}}(\theta) \le \mathcal{I}_{\boldsymbol{X}}(\theta)$ for all θ.

(b) $\mathcal{I}_{\boldsymbol{T}}(\theta) = \mathcal{I}_{\boldsymbol{X}}(\theta)$ for all θ if and only if $\boldsymbol{T}$ is sufficient.

Part (a) says that a function of $\boldsymbol{X}$ cannot contain more information than $\boldsymbol{X}$ does. A measure of information would not be sensible unless it satisfied this property. Part (b) coincides with the motivating idea behind the definition of a sufficient statistic. The formal definition of sufficiency is in terms of conditional distributions, but the motivation for it was the idea that a statistic is sufficient if it contains all of the information about θ that $\boldsymbol{X}$ does.

To calculate the information in $\boldsymbol{T}$, first find the pmf or pdf of $\boldsymbol{T}$, say, $h(\boldsymbol{t}\,;\theta)$ and then calculate $\mathcal{I}_{\boldsymbol{T}}(\theta) = \mathrm{Var}_\theta[\frac{\partial}{\partial\theta}\log h(\boldsymbol{T}\,;\theta)]$.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(θ, θ), $\theta > 0$.

(a) For an i.i.d. sample we can find the information in a single observation and then multiply by n (Theorem 6.4.1).

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left[-\frac{1}{2\theta}(x-\theta)^2 \right].$$

Before differentiating with respect to θ, it can be helpful to try to simplify the expression relative to θ in order to make differentiation as easy as possible. We can manipulate to get

$$\begin{aligned} -\frac{1}{2\theta}(x-\theta)^2 &= -\frac{1}{2\theta}(x^2 - 2\theta x + \theta^2) \\ &= -\frac{1}{2\theta}x^2 + x - \frac{\theta}{2}. \end{aligned}$$

Now

$$\begin{aligned} \log f(x;\theta) &= -\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\theta - \frac{1}{2\theta}x^2 + x - \frac{\theta}{2} \\ \frac{\partial}{\partial\theta}\log f(x;\theta) &= -\frac{1}{2\theta} + \frac{x^2}{2\theta^2} - \frac{1}{2} \\ \frac{\partial^2}{\partial\theta^2}\log f(x;\theta) &= \frac{1}{2\theta^2} - \frac{x^2}{\theta^3} \\ \mathcal{I}_{X_1}(\theta) &= -\mathrm{E}_\theta\left(\frac{1}{2\theta^2} - \frac{X_1^2}{\theta^3}\right) = -\frac{1}{2\theta^2} + \frac{\theta+\theta^2}{\theta^3} = \frac{1}{2\theta^2} + \frac{1}{\theta}. \end{aligned}$$

Therefore, $\mathcal{I}_{\boldsymbol{X}}(\theta) = n\left(\frac{1}{2\theta^2} + \frac{1}{\theta}\right)$.

(b) Next let us calculate the amount of information about θ that the sample mean $\overline{X}$ contains. We know that $\overline{X} \sim \text{Normal}(\theta, \frac{\theta}{n})$. Let $h(\overline{x}\,;\theta)$ denote the pdf of $\overline{X}$.

$$h(\overline{x}\,;\theta) = \frac{1}{\sqrt{2\pi\frac{\theta}{n}}} \exp\left[-\frac{n}{2\theta}(\overline{x}-\theta)^2 \right]$$

$$\log h(\overline{x}\,;\theta) = -\frac{1}{2}\log\frac{2\pi}{n} - \frac{1}{2}\log\theta - \frac{n}{2\theta}\overline{x}^2 + n\overline{x} - \frac{n\theta}{2}$$

$$\frac{\partial}{\partial\theta}\log h(\overline{x}\,;\theta) = -\frac{1}{2\theta} + \frac{n\overline{x}^2}{2\theta^2} - \frac{n}{2}$$

$$\frac{\partial^2}{\partial\theta^2}\log h(\overline{x}\,;\theta) = \frac{1}{2\theta^2} - \frac{n\overline{x}^2}{\theta^3}$$

$$\mathcal{I}_{\overline{X}}(\theta) = -\,\mathrm{E}_\theta\Big(\frac{1}{2\theta^2} - \frac{n\overline{X}^2}{\theta^3}\Big) = -\frac{1}{2\theta^2} + \frac{n(\frac{\theta}{n}+\theta^2)}{\theta^3} = \frac{1}{2\theta^2} + \frac{n}{\theta}\,.$$

For $n > 1$, we see that $\mathcal{I}_{\overline{X}}(\theta) = \frac{1}{2\theta^2} + \frac{n}{\theta} < \frac{n}{2\theta^2} + \frac{n}{\theta} = \mathcal{I}_{\boldsymbol{X}}(\theta)$. Therefore, by Theorem 6.4.2(b), $\overline{X}$ is not sufficient. However, looking at the ratio $\mathcal{I}_{\overline{X}}(\theta)/\mathcal{I}_{\boldsymbol{X}}(\theta) = \Big(\frac{1}{2\theta^2} + \frac{n}{\theta}\Big)\Big/\Big(\frac{n}{2\theta^2} + \frac{n}{\theta}\Big) = \Big(\theta + \frac{1}{2n}\Big)\Big/\Big(\theta + \frac{1}{2}\Big)$, we could say that $\overline{X}$ is "almost sufficient" if we knew that θ was substantially bigger than $\frac{1}{2}$.

(c) By using the Factorization Theorem, we can see that $T = \sum_{i=1}^n X_i^2$ is a sufficient statistic (see p. 19 above). Another way to show the sufficiency of T would be to use Theorem 6.4.2(b). That is, we could calculate the information in T and show it is equal to the information in $\boldsymbol{X}$. However, this is much more difficult than using the Factorization Theorem. The distribution of T is noncentral chi-squared and its pdf is expressed as an infinite series. $\triangle$ that of a multiple of a random variable

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(θ, θ^2), $\theta > 0$.

(a) Steps in the calculation of the information in the whole sample are:

$$f(x;\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left[-\frac{1}{2\theta^2}(x-\theta)^2 \right].$$

$$\log f(x;\theta) = -\frac{1}{2}\log 2\pi - \log\theta - \frac{x^2}{2\theta^2} + \frac{x}{\theta} - \frac{1}{2}$$

$$\frac{\partial^2}{\partial\theta^2}\log f(x;\theta) = \frac{1}{\theta^2} - \frac{3x^2}{\theta^4} + \frac{2x}{\theta^3}$$

$$\mathcal{I}_{X_1}(\theta) = \frac{3}{\theta^2}\,.$$

Therefore, $\mathcal{I}_{\boldsymbol{X}}(\theta) = \frac{3n}{\theta^2}$.

(b) Next let us find the amount of information in the sample mean $\overline{X}$. We know that $\overline{X} \sim \text{Normal}(\theta, \frac{\theta^2}{n})$. Let $h(\overline{x}\,;\theta)$ denote the pdf of $\overline{X}$.

$$h(\overline{x}\,;\theta) \;=\; \frac{1}{\sqrt{2\pi\frac{\theta^2}{n}}}\exp\Big[-\frac{n}{2\theta^2}(\overline{x}-\theta)^2\Big]$$

$$\log h(\overline{x}\,;\theta) \;=\; -\frac{1}{2}\log\frac{2\pi}{n} - \log\theta - \frac{n}{2\theta^2}\overline{x}^2 + \frac{n}{\theta}\overline{x} - \frac{n}{2}$$

$$\frac{\partial}{\partial\theta}\log h(\overline{x}\,;\theta) \;=\; -\frac{1}{\theta} + \frac{n\overline{x}^2}{\theta^3} - \frac{n\overline{x}}{\theta^2}$$

$$\frac{\partial^2}{\partial\theta^2}\log h(\overline{x}\,;\theta) \;=\; \frac{1}{\theta^2} - \frac{3n\overline{x}^2}{\theta^4} + \frac{2n\overline{x}}{\theta^3}$$

$$\mathcal{I}_{\overline{X}}(\theta) \;=\; \frac{n+2}{\theta^2}\,.$$

For $n > 1$, $\mathcal{I}_{\overline{X}}(\theta) = \frac{n+2}{\theta^2} < \frac{3n}{\theta^2} = \mathcal{I}_X(\theta)$. So $\overline{X}$ is not sufficient. Looking at the ratio $\mathcal{I}_{\overline{X}}(\theta)/\mathcal{I}_X(\theta) = (n+2)/3n$, we see that for large n, $\overline{X}$ has only about $\frac{1}{3}$ the information that the whole sample does. $\triangle$

Information about a parameter vector

Consider a data vector $\boldsymbol{X}$ with joint pmf or pdf $f(\boldsymbol{x}\,;\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. When θ is real-valued, we have defined a measure of the information that $\boldsymbol{X}$ contains about θ. For a vector-valued parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, the information will be defined to be, not a single number, but rather a matrix of numbers.

We require the same three regularity conditions, with θ replaced by $\boldsymbol{\theta}$. In RC3, the derivatives are the partial derivative with respect to the θ_j, $j = 1, \ldots, p$.

Conditions RC1, RC2, and RC3 are satisfied in an exponential family, provided that the functions $b_j(\boldsymbol{\theta})$ are differentiable, which they typically are.

Definition. (a) The *score vector* is $\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})$.

(b) The *information matrix* is $\mathrm{Var}_{\boldsymbol{\theta}}\big[\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})\big]$.

Notation. If $g(\boldsymbol{\theta}) = g(\theta_1, \ldots, \theta_p)$ is a real-valued function of several variables, then

$$\frac{\partial}{\partial\boldsymbol{\theta}}g(\boldsymbol{\theta}) \;=\; \frac{\partial g}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}) \;=\; \Big(\frac{\partial g}{\partial\theta_1}(\boldsymbol{\theta}), \ldots, \frac{\partial g}{\partial\theta_p}(\boldsymbol{\theta})\Big)\,.$$

If $\boldsymbol{U}(\boldsymbol{X}) = (U_1(\boldsymbol{X}), \ldots, U_r(\boldsymbol{X}))$ is a vector-valued statistic, then

$$\mathrm{Var}[\boldsymbol{U}(\boldsymbol{X})] \;=\; \begin{pmatrix} \mathrm{Var}[U_1(\boldsymbol{X})] & \cdots & \mathrm{Cov}[U_1(\boldsymbol{X}), U_r(\boldsymbol{X})] \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}[U_r(\boldsymbol{X}), U_1(\boldsymbol{X})] & \cdots & \mathrm{Var}[U_r(\boldsymbol{X})] \end{pmatrix}.$$

This is called the variance-covariance matrix of the random vector $\boldsymbol{U}$. Its (i,j) entry is $\mathrm{Cov}[U_i(\boldsymbol{X}), U_j(\boldsymbol{X})]$.

The (i,j) entry of the information matrix is $\mathrm{Cov}_{\boldsymbol{\theta}}[\frac{\partial}{\partial\theta_i}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})\,,\,\frac{\partial}{\partial\theta_i}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})]$. Similar to the real-valued parameter case, this can be calculated in two alternative ways, as $\mathrm{E}_{\boldsymbol{\theta}}\big[\big(\frac{\partial}{\partial\theta_i}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})\big)\big(\frac{\partial}{\partial\theta_j}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})\big)\big]$ or as $-\,\mathrm{E}_{\boldsymbol{\theta}}\big[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{X}\,;\,\boldsymbol{\theta})\big]$.

Theorems 6.4.1 and 6.4.2 are still true with $\boldsymbol{\theta}$ in place of θ. (The inequality of matrices in the generalization of Theorem 6.4.2(a) is defined using the concept of a positive semi-definite matrix.)

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal$(\mu\,,\sigma^2)$ random variables with unknown parameters μ and σ^2. We will be differentiating with respect to σ^2, and so it is convenient to get rid of the superscript by using the notation $\psi = \sigma^2$. The pdf for a single observation is

$$f(x\,;\mu,\psi) \;=\; \frac{1}{\sqrt{2\pi\psi}}\,\exp\Big[-\,\frac{1}{2\psi}(x-\mu)^2\Big]\,.$$

Now

$$\log f(x\,;\mu,\psi) \;=\; -\,\tfrac{1}{2}\log 2\pi - \tfrac{1}{2}\log\psi - \frac{1}{2\psi}(x-\mu)^2$$

$$\frac{\partial}{\partial\mu}\log f(x\,;\mu,\psi) \;=\; -\,\frac{1}{2\psi}2(x-\mu)(-1) \;=\; \frac{x-\mu}{\psi}$$

$$\frac{\partial}{\partial\psi}\log f(x\,;\mu,\psi) \;=\; -\,\frac{1}{2\psi} + \frac{1}{2\psi^2}(x-\mu)^2$$

$$\frac{\partial^2}{\partial\mu^2}\log f(x\,;\mu,\psi) \;=\; -\,\frac{1}{\psi}$$

$$\frac{\partial^2}{\partial\psi^2}\log f(x\,;\mu,\psi) \;=\; \frac{1}{2\psi^2} - \frac{1}{\psi^3}(x-\mu)^2$$

$$\frac{\partial^2}{\partial\mu\partial\psi}\log f(x\,;\mu,\psi) \;=\; -\,\frac{x-\mu}{\psi^2}$$

$$-\,\mathrm{E}\Big[\frac{\partial^2}{\partial\mu^2}\log f(x\,;\mu,\psi)\Big] \;=\; \frac{1}{\psi} \;=\; \frac{1}{\sigma^2}$$

$$-\,\mathrm{E}\Big[\frac{\partial^2}{\partial\psi^2}\log f(x\,;\mu,\psi)\Big] \;=\; -\,\frac{1}{2\psi^2} + \frac{1}{\psi^3}\psi \;=\; \frac{1}{2\psi^2} \;=\; \frac{1}{2\sigma^4}$$

$$-\,\mathrm{E}\Big[\frac{\partial^2}{\partial\mu\partial\psi}\log f(x\,;\mu,\psi)\Big] \;=\; 0\,.$$

Thus we obtain

$$\mathcal{I}_{\boldsymbol{X}}(\mu,\sigma^2) \;=\; \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \;\triangle$$

In the next chapter, we will see that information and information matrics are useful for establishing lower bounds on the variances of unbiased estimators. Next term you will see that they are useful for approximating the variances of maximum likelihood estimators.

Ancillary statistics

An ancillary statistic is, in some sense, the "opposite" of a sufficient statistic. Whereas a sufficient statistic contains all the information about θ that the data vector does, an ancillary statistic, by itself, contains no direct information about θ. Nevertheless, we will see that some ancillary statistics can be useful, when used together with other statistics, for making inferences about θ.

Definition 6.5.1. A statistic $T(X)$ is called *ancillary* if the distribution of $T(X)$ does not involve θ.

Therefore, an ancillary statistic by itself cannot provide any information that would help in distinguishing between different values of θ. More formally, if T is ancillary, then the information it contains about θ is 0. This follows from the fact that its pmf or pdf, say $g(t)$, does not involve θ, and so $\frac{\partial}{\partial\theta}\log g(T) = 0$.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal$(\mu, 1)$ random variables.
(a) $X_1 - X_2 \sim$ Normal$(0, 2)$. Since this distribution does not depend on μ, then $X_1 - X_2$ is an ancillary statistic.
(b) $\sum(X_i - \overline{X})^2 \sim \chi^2_{n-1}$ (see Theorem 4.4.2(ii); note that we have $\sigma^2 = 1$) and so it is ancillary.
(c) $X_{(n)} - X_{(1)}$ is ancillary. It would be difficult to find the pdf of $X_{(n)} - X_{(1)}$, but fortunately we can show that the distribution does not involve μ without finding the pdf. We can write $X_i = Z_i + \mu$ where $Z_1, \ldots, Z_n$ are i.i.d. Normal$(0, 1)$ (by letting $Z_i = X_i - \mu$). Since adding μ to the Z_i's does not affect their order, we have $X_{(1)} = Z_{(1)} + \mu$ and $X_{(n)} = Z_{(n)} + \mu$. Hence $X_{(n)} - X_{(1)} = (Z_{(n)} + \mu) - (Z_{(1)} + \mu) = Z_{(n)} - Z_{(1)}$. Since the joint distribution of the Z_i's does not depend on μ, then neither does the distribution of any function of them, such as $Z_{(n)} - Z_{(1)}$. (In the preceding sentence, of course the function itself should not involve μ.) Note that the statistics in (a) and (b) also could be shown to be ancillary by expressing them in terms of the Z_i's.
(d) Although $Z_1 = X_1 - \mu \sim$ Normal$(0, 1)$ has a distribution that does not involve μ, note that we should not call it an ancillary statistic, because it is not a statistic (see p. 2 above). $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ random variables.
(a) The range $R = X_{(n)} - X_{(1)}$ is ancillary. This can be shown by deriving its pdf (see Exercise 4.2.8) and noting that it does not involve θ. An easier way to show R is ancillary is to use an argument similar to the one in (c) in the preceding example, that is, by writing $X_i = Z_i + \theta$ where $Z_1, \ldots, Z_n$ are i.i.d. Uniform$(-\frac{1}{2}, +\frac{1}{2})$.

(b) The statistic (M, R), where $M = (X_{(1)} + X_{(n)})/2$ is the midrange, is minimal sufficient because it is a one-to-one function of $(X_{(1)}, X_{(n)})$. At first sight, it seems strange that part of a minimal sufficient statistic, namely R, could be ancillary. The midrange M clearly provides information about θ and in fact could be used to estimate θ. But what sort of information does R provide? It is useful in determining how accurate M is as an estimator of θ. Without using R, we know that $|M - \theta| < \frac{1}{2}$ (because $\theta - \frac{1}{2} < M < \theta + \frac{1}{2}$). By using the value of R, we can say more: $|M - \theta| < \frac{1}{2}(1 - R)$. For example, if $R = 0.8$, then we know $|M - \theta| < 0.1$. $\triangle$

The examples above can be viewed as special cases of a general result about location families.

Definition. Let $g(x)$ be a known pdf on the real line and define $f(x;\theta) = g(x - \theta)$. The family of pdf's $\{f(x;\theta) : -\infty < \theta < \infty\}$ is called a *location family*, and θ is called a *location parameter*.

Examples. (1) The family of Normal$(\mu, 1)$ pdf's is a location family with location parameter μ. To see this, write down the pdf $f(x;\mu) = (1/\sqrt{2\pi})\exp[-\frac{1}{2}(x-\mu)^2]$ and note that it can be expressed as $f(x;\mu) = g(x - \mu)$ where $g(z) = (1/\sqrt{2\pi})\exp[-\frac{1}{2}z^2]$. Note that $g(z)$ is a pdf, namely, the pdf of the Normal$(0, 1)$ distribution.

(2) The family of Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ pdf's is a location family with location parameter θ. The pdf can be expressed as $f(x;\theta) = I\{\theta - \frac{1}{2} < x < \theta + \frac{1}{2}\} =$
$I\{-\frac{1}{2} < x - \theta < \frac{1}{2}\} = g(x - \theta)$ where $g(z) = I\{-\frac{1}{2} < z < \frac{1}{2}\}$, which is the Uniform$(-\frac{1}{2}, \frac{1}{2})$ pdf.

(3) The location parameter does not have to designate the "center" of the distribution, as it does in example (1) and (2). For example, the family of Uniform$(\theta, \theta + 1)$ pdf's is a location family with location parameter θ.

Lemma. Let $X_1, \ldots, X_n$ be i.i.d. random variables with pdf $f(x;\theta) = g(x - \theta)$ in a location family. The statistic $\boldsymbol{T} = (X_1 - X_n, \ldots, X_{n-1} - X_n)$ is ancillary, and so is any statistic that is a function of $\boldsymbol{T}$.

Proof. (i) To see that the differences are (jointly) ancillary, we consider $Z_i = X_i - \theta$. Then $X_i = Z_i + \theta$ and Z_i has pdf $g(z)$ not involving θ. Let us check that if X has pdf $g(x - \theta)$, then $Z = X - \theta$ has pdf $g(z)$. In general, if $Z = k(X)$ for some one-to-one differentiable function $k(x)$, then Theorem 4.4.1 tells us that $f_Z(z) = f_X(x)\left|\frac{dx}{dz}\right|$ where x is expressed in terms of z by the inverse transformation $x = k^{-1}(z)$. In our case we have $z = x - \theta$, $x = z + \theta$, $\frac{dx}{dz} = 1$, $f_X(x) = g(x - \theta)$. Hence $f_Z(z) = g(x - \theta)|1| = g(z)$.

(ii) $Z_1, \ldots, Z_n$ are i.i.d. with pdf $g(z)$, so their joint distribution does not involve θ. Therefore, any function of them, as long as the function does not involve θ, is a random variable (or random vector if the function is vector-valued) whose distribution does not involve θ. So the lemma will be established if we show that T can be expressed as a function of the Z_i's. Each entry in T can be expressed as $T_i = X_i - X_n = (Z_i + \theta) - (Z_n + \theta) = Z_i - Z_n$. □

Lemma. If T is an ancillary statistic and $W = W(T)$ is a statistic that is a function of T, then W is ancillary.

For example, for a location family we have just seen that $T = (X_1 - X_n, \ldots, X_{n-1} - X_n)$ is ancillary. Hence the statistics $X_1 - X_2 = (X_1 - X_n) - (X_2 - X_n) = T_1 - T_2$ and $X_{(n)} - X_{(1)} = \max\{T_1, \ldots, T_{n-1}, 0\} - \min\{T_1, \ldots, T_{n-1}, 0\}$ are ancillary.

In the proof above, it was important that not only do the individual Z_i's have distributions not involving the parameter, but moreover their joint distribution does not involve the parameter. This point is worth pursuing. Suppose we know that Y_1 has a distribution not involving θ and also Y_2 has a distribution not involving θ. Suppose $W = h(Y_1, Y_2)$ where the function $h(y_1, y_2)$ does not involve θ (e.g., $h(y_1, y_2) = y_1 - y_2$ or $h(y_1, y_2) = y_1 y_2$). Can we say that the distribution of W does not involve θ? This is not necessarily true, because the distribution of $h(Y_1, Y_2)$ depends on the joint distribution of (Y_1, Y_2). The marginal distributions of Y_1 and Y_2 individually do not completely determine their joint distribution, which must also take into account the relationship between them.

In particular, if T_1 and T_2 are two ancillary statistics, it is not necessarily true that (T_1, T_2) is an ancillary statistic. However, if T_1 and T_2 are known to be independent, then (T_1, T_2) is ancillary.

Example. Let (X_1, X_2) have a Bivariate Normal distribution (see section 3.6) with known means $\mu_1 = \mu_2 = 0$, known variances $\sigma_1^2 = \sigma_2^2 = 1$, and an unknown correlation coefficient ρ. Then each of the statistics X_1 and X_2 is distributed as Normal$(0, 1)$, which does not involve the parameter ρ. So, individually, each of the two statistics is ancillary. But jointly their distribution involves ρ. In particular, $\text{Cov}(X_1, X_2) = \rho$. △

Next we look at a different variety of ancillary statistics.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(0, \theta)$ random variables for some $\theta > 0$. The statistics X_1/X_2 and $\overline{X}/X_{(n)}$ are ancillary. The easiest way to see this is to define $Z_i = X_i/\theta$ and write $X_i = \theta Z_i$, noting that $Z_1, \ldots, Z_n$ are i.i.d. Uniform$(0, 1)$. Then

$X_1/X_2 = \theta Z_1/\theta Z_2 = Z_1/Z_2$, whose distribution does not involve θ. Check that $\overline{X} = \theta\overline{Z}$ and $X_{(n)} = \theta Z_{(n)}$. Hence $\overline{X}/X_{(n)} = \theta\overline{Z}/\theta Z_{(n)} = \overline{Z}/Z_{(n)}$. $\triangle$

This example is a special case of a general result about scale families.

Definition. Let $g(x)$ be a known pdf on the real line and define $f(x;\delta) = \frac{1}{\delta}g(\frac{x}{\delta})$. The family of pdf's $\{f(x;\delta) : \delta > 0\}$ is called a *scale family*, and δ is called a *scale parameter*.

Examples. (1) The family of Normal$(0,\sigma^2)$ pdf's is a scale family with scale parameter σ. To see this, write down the pdf $f(x;\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}x^2]$ and rewrite it as $\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}(\frac{x}{\sigma})^2] = \frac{1}{\sigma}g(\frac{x}{\sigma})$ where $g(z) = \frac{1}{\sqrt{2\pi}}\exp[-\frac{1}{2}z^2]$, which is the Normal$(0,1)$ pdf.

(2) The family of Uniform$(0,\theta)$ pdf's is a scale family with scale parameter $\theta > 0$. The pdf can be expressed as $f(x;\theta) = \frac{1}{\theta}I\{0 < x < \theta\} = \frac{1}{\theta}I\{0 < \frac{x}{\theta} < 1\} = \frac{1}{\theta}g(\frac{x}{\theta})$ where $g(z) = I\{0 < z < 1\}$, which is the Uniform$(0,1)$ pdf.

Lemma. Let $X_1, \ldots, X_n$ be i.i.d. random variables with pdf $f(x;\delta) = \frac{1}{\delta}g(\frac{x}{\delta})$ in a scale family. The statistic $T = (X_1/X_n, \ldots, X_{n-1}/X_n)$ is ancillary.

Proof. To see that the ratios are (jointly) ancillary, we consider $Z_i = X_i/\delta$. Then $X_i = \delta Z_i$ and Z_i has pdf $g(z)$ (which can be verified by using Theorem 4.4.1). Hence $Z_1, \ldots, Z_n$ are i.i.d. with pdf $g(z)$, so their joint distribution does not involve δ. Therefore, any function of them, as long as the function does not involve δ, has a distribution not involving δ. So the lemma will be established if we show that T can be expressed as a function of the Z_i's. Each entry in T can be expressed as $T_i = X_i/X_n = \delta Z_i/\delta Z_n = Z_i/Z_n$. $\square$

Consequently, any function of these ratios is an ancillary statistic (provided the function does not involve δ). For example, $X_1/X_2 = (X_1/X_n)/(X_2/X_n) = T_1/T_2$ is ancillary, and so is $\overline{X}/X_{(n)} \neq \text{mean}\{T_1, \ldots, T_{n-1}, 1\}/\max\{T_1, \ldots, T_{n-1}, 1\}$.

Next we consider some 2-parameter families of distributions.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ,σ^2) random variables. The statistics $(X_1 - X_2)/(X_1 - X_3)$ and $(X_{(n)} - X_{(1)})^2/\sum(X_i - \overline{X})^2$ are ancillary. To see this, define $Z_i = (X_i - \mu)/\sigma$ and write $X_i = \sigma Z_i + \mu$, noting that $Z_1, \ldots, Z_n$ are i.i.d. Normal$(0,1)$. Then $(X_1 - X_2)/(X_1 - X_3) = [(\sigma Z_1 + \mu) - (\sigma Z_2 + \mu)]/[(\sigma Z_1 + \mu) - (\sigma Z_3 + \mu)] = (Z_1 - Z_2)/(Z_1 - Z_3)$, whose distribution does not involve μ or σ^2. $\triangle$

This example is a special case of a general result about location-scale families.

Definition. Let $g(x)$ be a known pdf on the real line and define $f(x;\theta,\delta) = \frac{1}{\delta} g(\frac{x-\theta}{\delta})$. The family of pdf's $\{f(x;\theta,\delta) : -\infty < \theta < \infty, \delta > 0\}$ is called a *location-scale* family, θ is called a *location parameter*, and δ is called a *scale parameter*.

Examples. (1) The family of Normal(μ, σ^2) pdf's is a location-scale family with location parameter μ and scale parameter σ. To see this, write down the pdf $f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$ and rewrite it as $\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] = \frac{1}{\sigma} g(\frac{x-\mu}{\sigma})$ where $g(z)$ is the Normal$(0,1)$ pdf.

(2) The family of Uniform$(\theta, \theta+\delta)$ pdf's is a location-scale family with location parameter θ and scale parameter $\delta > 0$. The pdf can be expressed as $f(x;\theta,\delta) = \frac{1}{\delta} I\{\theta < x < \theta+\delta\}$ $= \frac{1}{\delta} I\{0 < \frac{x-\theta}{\delta} < 1\} = \frac{1}{\delta} g(\frac{x-\theta}{\delta})$ where $g(z)$ is the Uniform$(0,1)$ pdf.

Lemma. Let $X_1, \ldots, X_n$ be i.i.d. random variables with pdf $f(x;\delta) = \frac{1}{\delta} g(\frac{x-\theta}{\delta})$ in a location-scale family. The statistic $\boldsymbol{T} =$ $((X_1 - X_n)/(X_{n-1} - X_n), \ldots, (X_{n-2} - X_n)/(X_{n-1} - X_n))$ is ancillary.

Proof. To see that these difference ratios are (jointly) ancillary, we consider $Z_i = (X_i - \theta)/\delta$. Then $X_i = \delta Z_i + \theta$ and Z_i has pdf $g(z)$ (which you can verify by using Theorem 4.4.1). Hence $Z_1, \ldots, Z_n$ are i.i.d. with pdf $g(z)$, so their joint distribution does not involve θ or δ. Therefore, any function of them, as long as the function does not involve θ or δ, has a distribution not involving θ or δ. So the lemma will be established if we show that $\boldsymbol{T}$ can be expressed as a function of the Z_i's. Each entry in $\boldsymbol{T}$ can be expressed as $T_i =$ $(X_i - X_n)/(X_{n-1} - X_n) = [(\delta Z_i + \theta) - (\delta Z_n + \theta)]/[(\delta Z_{n-1} + \theta) - (\delta Z_n + \theta)] =$ $(Z_i - Z_n)/(Z_{n-1} - Z_n)$. □

Consequently, any function of these difference ratios is an ancillary statistic (provided the function does not involve θ or δ). For example, $(X_1 - X_2)/(X_1 - X_3) =$ $(T_1 - T_2)/(T_1 - T_3)$ is ancillary.

Conditional inference

In the example on pp. 28-29 above the minimal sufficient statistic contains a component that is ancillary. In general, suppose $\boldsymbol{X}$ is a data vector with joint pmf or pdf $f(\boldsymbol{x};\boldsymbol{\theta})$ indexed by an unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, and suppose $\boldsymbol{T} = (T_1, \ldots, T_r)$ is a minimal sufficient statistic. When $r > p$, it is typically possible to write $\boldsymbol{T}$ or, if necessary, a one-to-one function of $\boldsymbol{T}$ (which would also be minimal sufficient) as $(\boldsymbol{W}, \boldsymbol{A})$ where $\boldsymbol{W}$ has dimension p and $\boldsymbol{A}$ is ancillary. Consider the example on pp. 28-29:

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ random variables. Here we have $p = 1$. Above it was seen that (M, R) is minimal sufficient, where $M = (X_{(1)} + X_{(n)})/2$ and $R = X_{(n)} - X_{(1)}$, and that R is ancillary. We have $r = 2$. This falls under the description above with $W = M$ and $A = R$.

The Sufficiency Principle says that, in analyzing the data (under the assumptions of the model), we can restrict attention to a minimal sufficient statistic. Another principle, stated below, recommends that, furthermore, we should condition on the ancillary part of the minimal sufficient statistic

A Conditionality Principle: Let $T = T(X)$ be a minimal sufficient statistic. Suppose we can write $T = (W, A)$ where A is ancillary. Statistical inference about θ should depend only on the conditional distribution of W given A.

Example (continued). The principle recommends basing our inference about θ on the conditional distribution of M given R. It can be shown (see Exercise 4.2.8) that $M \mid R = r \sim$ Uniform$(\theta - \frac{1}{2}(1-r), \theta + \frac{1}{2}(1-r))$. $\triangle$

Example. Suppose we plan to select a simple random sample of n Corvallis voters and ask them whether or not they favor a higher gasoline tax. Let $X_i = 1$ if the i-th voter says yes and $X_i = 0$ otherwise. A reasonable model is to assume $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ), $0 < \theta < 1$. Suppose that the sample size n is not fixed in advance but is allowed to be however many interviews can be conducted in 2 days. In such a situation we might regard n as a random variable. Let us denote it by N to emphasize its randomness. It would seem reasonable to assume that the distribution of N does not involve θ. With N regarded as random, we would assume that, conditional on $N = n$, $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ). The data would consist of $Y = (N, X_1, \ldots, X_N)$. By applying Theorem 6.3.1 it can be shown that $T = (N, \sum_{i=1}^{N} X_i)$ is minimal sufficient. The statistic N is ancillary. The total information in the sample can be calculated to be $\mathcal{I}_Y(\theta) = \mathcal{I}_T(\theta) = \frac{\mathrm{E}(N)}{\theta(1-\theta)}$. Typically we would just pretend n is fixed and not worry about its randomness. This amounts to following the conditionality principle and basing our inference on the conditional distribution of $W \mid N = n$ where $W = \sum_{i=1}^{N} X_i$. This conditional distribution is simply Binomial(n, θ). Using the conditional pmf of W we can calculate the conditional information to be $\mathcal{I}_{W|N=n}(\theta) = \frac{n}{\theta(1-\theta)}$. Considering all the possible values of the random variable N, we can write this as $\mathcal{I}_{W|N}(\theta) = \frac{N}{\theta(1-\theta)}$. Note that $\mathrm{E}[\mathcal{I}_{W|N}(\theta)] = \mathcal{I}_{(W,N)}(\theta)$. Thus the expected value of the conditional information is equal to the total (unconditional) information. We

conclude that by conditioning on the sample size (provided it is ancillary), we do not lose any information (on the average). △

This example illustrates equation (6.5.12) in the textbook.

A similar argument can be used in regression to justify regarding the explanatory variables as fixed when they are really random. Regarding the explanatory variables as fixed is equivalent to conditioning on them, and this does not lose any information (on the average) when the joint distribution of the explanatory variables does not involve the regression parameters.

Complete statistics

Definition. A statistic $T(X)$ is called a *complete statistic* for θ if the "only" function $h(T)$ for which $\mathrm{E}_\theta[h(T)] = 0$ for all θ is the zero function (i.e., $h(t) = 0$ for all t).

Instead of saying "complete for θ" (or "sufficient for θ"), we sometimes say "complete for the family of distributions" (or "sufficient for the family of distributions").

Technical note. In this definition the word "only" has been put in quotation marks for technical reasons. Actually the condition $\mathrm{E}_\theta[h(T)] = 0$ holds for any function that is 0 with probability 1, i.e., for any function $h(T)$ such that $\mathrm{P}_\theta\{h(T) = 0\} = 1$ for all θ. For example, suppose $T \sim \text{Normal}(\mu, 1)$ and $h(t) = I_{\{3\}}(t)$. Then $h(t) = 0$ for all $t \neq 3$. Since a single point such as $\{3\}$ has probability 0 with respect to a continuous distribution such as a normal distribution, $\mathrm{E}_\mu[h(T)] = 0$ for all μ. A more technically correct statement of the theorem would say that T is complete for θ if the only functions $h(T)$ for which $\mathrm{E}_\theta[h(T)] = 0$ for all θ are functions that are zero with probability 1.

The concept of completeness is useful in the context of unbiased estimation.

It can be helpful to think of this definition in terms of what it says about noncompleteness.

Lemma. A statistic $T(X)$ is not complete if and only if there exists a nonzero function $h(T)$ such that $\mathrm{E}_\theta[h(T)] = 0$ for all θ.

As is implicitly implied by the notation, the function $h(T)$ must not involve the parameters. (Also, to be technically correct — see the technical note above — the function $h(T)$ is required to be nonzero on a set of positive probability.)

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal$(\mu, 1)$, $-\infty < \mu < \infty$.

(a) X_1 is complete but not sufficient.

(b) $X = (X_1, \ldots, X_n)$ is sufficient but not complete.

(c) $\overline{X}$ is complete and sufficient.

In (a), for X_1 to be complete means that, if $\int_{-\infty}^{\infty} h(x)\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x-\mu)^2}\mathrm{d}x = 0$ for all $-\infty < \mu < \infty$, then $h(x) = 0$ for all x (except possibly for x in a set having probability 0). This is difficult to prove. In (b), to see that X is not complete, consider $h(X) = X_1 - X_2$ and note that $\mathrm{E}_\mu(X_1 - X_2) = \mu - \mu = 0$ for all μ, but of course the function $x_1 - x_2$ is not identically 0.

A complete statistic that is also sufficient can be very useful, as we will see in Section 7.5.2.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. By the Factorization Theorem it is not hard to show that $T = X_1 + \cdots + X_n$ is a sufficient statistic. We will show that it is also complete.

(a) To do this, it is helpful to know the distribution of T. By using Theorem 4.3.1, it is not hard to show that $T \sim \text{Poisson}(n\lambda)$.

(b) Now suppose $h(T)$ is a function of T such that $\mathrm{E}_\lambda[h(T)] = 0$ for all $\lambda > 0$. We must show that $h(t) = 0$ for all $t = 0, 1, 2, \ldots$. Writing out the expectation more explicitly, we have $\mathrm{E}_\lambda[h(T)] = \sum_{t=0}^{\infty} h(t)\frac{e^{-n\lambda}(n\lambda)^t}{t!} = e^{-n\lambda}\sum_{t=0}^{\infty} h(t)\frac{n^t}{t!}\lambda^t = 0$.

We can cancel $e^{-n\lambda}$ and write $\sum_{t=0}^{\infty} c_t\lambda^t = 0$ for all $\lambda > 0$, where $c_t = h(t)\frac{n^t}{t!}$.

(c) If $\sum_{t=0}^{\infty} c_t\lambda^t = 0$ for all $\lambda > 0$, then we can conclude that $c_t = 0$ for all t. To see this, it may help to write out the sum as $c_0 + c_1\lambda + c_2\lambda^2 + c_3\lambda^3 + \cdots = 0$.
Letting $\lambda \to 0$, we find that $c_0 = 0$. Now we have $c_1\lambda + c_2\lambda^2 + c_3\lambda^3 + \cdots = 0$.
Factor out λ and cancel it from the equation to get $c_1 + c_2\lambda + c_3\lambda^2 + \cdots = 0$.
Letting $\lambda \to 0$ as before, we find that $c_1 = 0$. Continuing in this way,
we find that $c_t = 0$ for all t.

(d) Now we can say that $h(t)\frac{n^t}{t!} = 0$ for all t. Since $\frac{n^t}{t!} \neq 0$, we must have $h(t) = 0$. $\triangle$

Recall the following lemma about sufficiency from p. 8 above.

Lemma. Suppose a statistic W is a function of another statistic T.
If W is sufficient, then so is T.

For the property of completeness we have an analogous lemma.

Lemma. Suppose a statistic W is a function of another statistic T.
If T is complete, then so is W.

Proof. Suppose $E_\theta[h(W)] = 0$ for all θ. We must show that $h(W) \equiv 0$. Since W is a function of T, then $h(W)$ can be viewed as a function of T. Diagrammatically, $T \to W \to h(W)$. Now we can invoke the completeness of T to conclude $h(W) \equiv 0$. □

Lemma. Suppose a statistic W is a one-to-one function of another statistic T. If T is complete and sufficient, then so is W.

A statistic $T(X)$ either is equivalent to the data vector X (if the function $T(x)$ is one-to-one) or, more typically, is a reduction of X. To be sufficient, it cannot be too much of a reduction because it must retain all the information that X contains about θ.

Usually the data vector is not complete. In particular, if X is a vector of iid random variables, then $E_\theta[X_1 - X_2] = 0$ (provided the distribution has a finite mean). To be complete, a statistic $T(X)$ must be reduced or "condensed" enough that no function of its components can be constructed to have zero expectation.

To be both complete and sufficient, a statistic must reduce X to just the right degree. So complete sufficient statistics are a very special kind of statistic. They do not always exist. They do exist when the distribution of the data vector is in a regular exponential family. See p. 20 above for the general definition of a regular exponential family. We will concentrate on the i.i.d. case, but it should be mentioned that the following general theorem is true.

Theorem. Let X be a random vector with joint pmf or pdf in a regular exponential family. The statistic $T = (R_1(X), \ldots, R_k(X))$ is complete and sufficient.

The same statistic is minimal sufficient, as seen in the theorem on p. 20. The i.i.d. case is:

Theorem 6.6.2. Let $X_1, \ldots, X_n$ be i.i.d. with pmf or pdf in a exponential family,

$$f(x;\theta) = a(\theta)h(x)\exp\left\{\sum_{j=1}^{k} b_j(\theta)\, R_j(x)\right\}, \quad \theta = (\theta_1, \ldots, \theta_p) \in \Theta.$$

Suppose the exponential family is regular, satisfying (a) $k = p$, (b) Θ contains a p-dimensional rectangle, and (c) the functions $b_j(\theta)$ are differentiable. The statistic $T = (\sum_{i=1}^{n} R_1(X_i), \ldots, \sum_{i=1}^{n} R_p(X_i))$ is complete and sufficient.

This theorem could be used in the two preceding examples.

Examples. (1) Let $X_1, \ldots, X_n$ be i.i.d. Normal$(\mu, 1)$, $-\infty < \mu < \infty$. One write the pdf of Normal$(\mu, 1)$ in the exponential-family form with $k = p = 1$ and $R_1(x) = x$. Regularity conditions (b) and (c) are also met. So $\sum X_i$ is a complete sufficient statistic. Since $\overline{X}$ is a one-to-one function of $\sum X_i$, it is also complete and sufficient.

(2) Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. One write the pmf of Poisson(λ) in the exponential-family form with $k = p = 1$ and $R_1(x) = x$. Regularity conditions (b) and (c) are also met. So $\sum X_i$ is a complete sufficient statistic.

(3) Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2), $-\infty < \mu < \infty$, $\sigma^2 > 0$. As on pp. 18-19 above, the pdf of Normal(μ, σ^2) has the exponential-family form with $k = p = 2$ and $R_1(x) = x^2$ and $R_2(x) = x$. Regularity conditions (b) and (c) are also met. So $(\sum X_i^2, \sum X_i)$ is a complete sufficient statistic. Since $(\overline{X}, S^2)$ is a one-to-one function of $(\sum X_i^2, \sum X_i)$ (see p. 9 above), it is also complete and sufficient. $\triangle$

Next we discuss the relationship between complete sufficiency and miniimal sufficiency. In Theorems 6.3.3 and 6.6.2 (and the more general theorems on p. 20 and p. 36) we see that for regular exponential families, the complete sufficient statistic coincides with the minimal sufficient statistic. In general:

- Every complete sufficient statistic is minimal sufficient.
- Not every minimal sufficient statistic is complete sufficient.
- For any statistical model, a minimal sufficient statistic always exists.
- A complete sufficient statistic may not exist.
- If a complete sufficient statistic exists, then a minimal sufficient statistic will be complete sufficient.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(θ, θ^2), $\theta > 0$. As seen in the example on pp. 19-20 above, the statistic $(\sum X_i^2, \sum X_i)$ is minimal sufficient, and hence so is the one-to-one function $(\overline{X}, S^2)$. However, it is not complete, as seen in Exercise 6.6.3. $\triangle$

In general, suppose $\boldsymbol{X}$ is a data vector with joint pmf or pdf $f(\boldsymbol{x}\,;\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. There always exists a minimal sufficient statistic, say $\boldsymbol{T} = (T_1, \ldots, T_r)$. The following two statements are "typically" true, although artificial counterexamples can be constructed.

1. $r \geq p$.
2. $r = p \iff \boldsymbol{T}$ is complete sufficient

In the preceding example, note that $r = 2$ and $p = 1$. Since $r > p$, statement 2 indicates that the minimal sufficient statistic is "probably" not complete. The statement is only "typically" true, and so it is a good idea to formally verify the noncompleteness.

Let us look at a model that is not an exponential family.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Uniform$(0, \theta)$, $\theta > 0$. On p. 16 above we saw that a minimal sufficient statistic is the sample maximum $X_{(n)}$. Here we have $r = p = 1$, so $X_{(n)}$ is "probably" complete sufficient. We can verify this.

(a) As in Example 4.2.7, the pdf of $T = X_{(n)}$ is $f(t;\theta) = \frac{nt^{n-1}}{\theta^n} I\{0 < t < \theta\}$.

(b) Suppose $h(T)$ is a function of T such that $\mathrm{E}_\theta[h(T)] = 0$ for all $\theta > 0$. We must show that $h(t) \equiv 0$. Writing out the expectation more explicitly, we have

$$\mathrm{E}_\theta[h(T)] = \int_{-\infty}^{\infty} h(t) f(t;\theta)\mathrm{d}t = \int_0^\theta h(t)\frac{nt^{n-1}}{\theta^n}\mathrm{d}t = \frac{n}{\theta^n}\int_0^\theta h(t)t^{n-1}\mathrm{d}t = 0.$$

We can cancel $\frac{n}{\theta^n}$ and write $\int_0^\theta h(t)t^{n-1}\mathrm{d}t = 0$ for all $\theta > 0$.

(c) Now we appeal to the Fundamental Theorem of Calculus, which says that $\frac{\mathrm{d}}{\mathrm{d}y}\int_a^y g(t)\mathrm{d}t = g(y)$. Therefore, $\frac{\mathrm{d}}{\mathrm{d}\theta}\int_0^\theta h(t)t^{n-1}\mathrm{d}t = h(\theta)\theta^{n-1}$,

so $h(\theta)\theta^{n-1} = 0$ for all $\theta > 0$, which implies $h(\theta) = 0$ for all $\theta > 0$.

(d) The role of θ in the statement $h(\theta) = 0$ for all $\theta > 0$ is simply a mathematical variable. An equivalent statement is obtained by using any other symbol, such as $h(t) = 0$ for all $t > 0$ (or $h(u) = 0$ for all $u > 0$).

(e) It is enough to show that $h(t) = 0$ for all $t > 0$, because $\mathrm{P}_\theta\{T > 0\} = 1$ for all θ. We can conclude that T is complete. $\triangle$

A strategy for trying to find a complete sufficient statistic is the following.

1. If you have a random sample from a distribution in a regular exponential family, apply one of the theorems on p. 36 above.
2. Otherwise, find a minimal sufficient statistic $\boldsymbol{T(X)}$ (you can use Theorem 6.3.1). Let r and p denote the dimensions of $\boldsymbol{T}$ and $\boldsymbol{\theta}$ respectively.
3. If $r > p$, you can probably find a nonzero function $h(\boldsymbol{T})$ with zero expectation, thus showing that $\boldsymbol{T}$ is not complete. This implies that no complete sufficient statistic exists.
4. If $r = p$, it is probably true that $\boldsymbol{T}$ is complete. To actually prove that it is complete is often difficult.

Theorem 6.6.3 (Basu's Theorem). If $\boldsymbol{T(X)}$ is complete and sufficient, and if $\boldsymbol{S(X)}$ is ancillary, then $\boldsymbol{T}$ and $\boldsymbol{S}$ are independent for all $\boldsymbol{\theta}$.

Proof (in the discrete case). To show that $\boldsymbol{T}$ and $\boldsymbol{S}$ are independent, we will show $(*)$ $\mathrm{P}_{\boldsymbol{\theta}}\{\boldsymbol{S} = \boldsymbol{s} | \boldsymbol{T} = \boldsymbol{t}\} = \mathrm{P}_{\boldsymbol{\theta}}\{\boldsymbol{S} = \boldsymbol{s}\}$ for all θ.

(a) In equation (*), the left-hand conditional probability does not depend on θ because T is sufficient, and the right-hand probability does not depend on θ because S is ancillary. So we can drop the subscripts θ. Thus we want to show (*) $\mathrm{P}\{S = s | T = t\} = \mathrm{P}\{S = s\}$.
(b) Define $h(t) = \mathrm{P}\{S = s | T = t\} - \mathrm{P}\{S = s\}$ and regard s as being a constant. Part (a) shows that this is a valid statistic. We want to show that $h(t) = 0$. Noting that, for fixed s, $\mathrm{P}\{S = s\}$ is a constant, we have

$$\begin{aligned}
\mathrm{E}_\theta[h(T)] &= \mathrm{E}_\theta[\mathrm{P}\{S = s | T\}] - \mathrm{P}\{S = s\} \\
&= \sum_t \mathrm{P}\{S = s | T = t\}\, \mathrm{P}_\theta\{T = t\} - \mathrm{P}\{S = s\} \\
&= \sum_t \mathrm{P}_\theta\{S = s \text{ and } T = t\} - \mathrm{P}\{S = s\} \\
&= \mathrm{P}_\theta\{S = s\} - \mathrm{P}\{S = s\} = 0 \text{ for all } \theta .
\end{aligned}$$

By the completeness of T, we must have $h(T) = 0$ (with probability 1). □

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2). In Theorem 4.4.2 it was proved that $\overline{X}$ and S^2 are independent. An easier proof can be given by using Basu's Theorem. Fix an arbitrary value of σ^2. Consider the family of distributions $\{\text{Normal}(\mu, \sigma^2) : -\infty < \mu < \infty\}$. For this family, with σ^2 regarded as known, $\overline{X}$ is a complete sufficient statistic (use Theorem 6.6.2) and S^2 is an ancillary statistic (see p. 28 above). By Basu's Theorem, they are independent. △

Example. Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2).
(a) On p. 37 we saw that $(\overline{X}, S^2)$ is a complete sufficient statistic. By Basu's Theorem, it is independent of any ancillary statistic.
(b) We have a location-scale family and so, as seen on p. 32 above, a ratio of differences such as $(X_1 - X_2)/(X_1 - X_3)$ is ancillary.
(c) The statistic $W = (X_1 - X_2)^2/S^2$ is also ancillary. To see this, let $Z_i = (X_i - \mu)/\sigma$ so that $X_i = \sigma Z_i + \mu$ and $Z_1, \ldots, Z_n$ are i.i.d. Normal(0, 1). Check that $X_1 - X_2 = \sigma(Z_1 - Z_2)$ and $S^2 = \sum(X_i - \overline{X})^2/(n-1) = \sigma^2 \sum(Z_i - \overline{Z})^2/(n-1) = \sigma^2 S_Z^2$, so that $W = (X_1 - X_2)^2/S^2 = (Z_1 - Z_2)^2/S_Z^2$, which is a function of the Z_i's. The function does not involve any parameters and the joint distribution of the Z_i's does not involve any parameters, so the distribution of W does not involve any parameters.
(d) We can use the independence of W and S^2 to calculate the expectation of W. (To be continued) △

Suppose we have two random variables U and V.
(a) If we want to calculate the expectation of their product,

we must remember that, in general, $\mathrm{E}(UV) \neq \mathrm{E}(U)\mathrm{E}(V)$.

(b) In the case that U and V are independent,

then $\mathrm{E}(UV) = \mathrm{E}(U)\mathrm{E}(V)$ (see Theorem 3.5.1).

(c) If we want to calculate the expectation of the ratio of U and V,

we must beware that, in general, $\mathrm{E}(\frac{U}{V}) \neq \frac{\mathrm{E}(U)}{\mathrm{E}(V)}$.

Even if U and V are independent, we are not guaranteed that the inequality would become an equality. For example, if X_1, X_2 are i.i.d. Bernoulli($\frac{1}{2}$), then $\mathrm{E}(\frac{1+X_1}{1+X_2}) = \frac{9}{8} \neq 1 = \frac{\mathrm{E}(1+X_1)}{\mathrm{E}(1+X_2)}$.

(d) In the case that U/V is independent of V, then $\mathrm{E}(\frac{U}{V}) = \frac{\mathrm{E}(U)}{\mathrm{E}(V)}$.

To verify this, note that $\mathrm{E}(U) = \mathrm{E}(\frac{U}{V}V) = \mathrm{E}(\frac{U}{V})\mathrm{E}(V)$, using the independence in the second equality. Now divide the equation by $\mathrm{E}(V)$. (Technical requirements: It is assumed that $\mathrm{P}\{V \neq 0\} = 1$, $\mathrm{E}(V) \neq 0$, and all the expectations exist.)

Example (continued). Let $X_1, \ldots, X_n$ be i.i.d. Normal(μ, σ^2). Above we showed that $W = (X_1 - X_2)^2/S^2$ is independent of S^2. Therefore, by item (d) above,

$$\mathrm{E}\Big[\frac{(X_1-X_2)^2}{S^2}\Big] = \frac{\mathrm{E}[(X_1-X_2)^2]}{\mathrm{E}(S^2)} = \frac{2\sigma^2}{\sigma^2} = 2\,. \triangle$$

Further discussion. The purpose of the following informal discussion is to try to get a better "feeling" for what completeness means. We will be using the word "information" imprecisely — not in any formal quantitative sense. Let us say that "relevant" information is information that is useful for making inferences about the parameters.

- A statistic is sufficient if and only if it contains all the relevant information that is available.

Note that a sufficient statistic may also contain some ~~irrelevant~~ redundant information.

- A statistic is minimal sufficient if and only if (1) it contains all the relevant information and (2) it contains ~~only relevant~~ no redundant information.

Relevant information can be directly about the parameters or it can be only indirectly relevant, such as by providing an estimate of the variance of an estimator of a parameter. Recall that a minimal sufficient statistic can have components that are ancillary. An ancillary statistic contains no information that is directly relevant for ~~making inference about~~ estimating the parameters, but it can provide useful information about the precision of estimators.

- A statistic is complete if and only if it contains only directly relevant information.

Consider a statistic $T = (T_1, \ldots, T_r)$. It contains a certain amount of "information". Each component T_i contains a piece of information. In fact, each function $h(T)$ represents a piece of information derived from T. (The components are just special functions, $h(t_1, \ldots, t_r) = t_i$.) If $\mathrm{E}_\theta[h(T)] = 0$ for all θ, then we can regard $h(T)$ as a piece of information that says nothing directly about θ. The existence of such a function shows that T contains some information that is not directly relevant. Correspondingly, the formal definition of completeness says that if such a function $h(T)$ exists, then T is not complete.

Point estimation

On p. 1 above there was a statement of the general goal of statistical inference. In Chapter 7 we study the type of statistical inference called point estimation, in which the general goal can be said to be the following.

> Given a data vector X and a model for its distribution,
> in the form of a family $\{f(x;\theta) : \theta \in \Theta\}$ of possible pmf's or pdf's for X,
> and given a real-valued function $\tau(\theta)$ of the parameter vector,
> we want to estimate what the true value of $\tau(\theta)$ might be.

Our estimate will be computed from the data. Any real-valued function $W(X)$ of the data vector can be regarded as an *estimator* of $\tau(\theta)$. In order to be computable from the data, the function cannot involve the parameters. We begin by considering all possible functions as estimators, but of course many of them will be found to be very bad estimators. We make a distinction between the words *estimator* and *estimate*. An estimator $W(X)$ is a function of the data vector X regarded as a vector of random variables before they are actually observed. After observing the data $X = x$, the numerical value $W(x)$ is called an *estimate*.

Method of moments

This method is based on the idea that a sample mean is a natural estimator for a population mean.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(p) random variables and p is unknown. The sample mean $\overline{X}$ is a natural estimator of the population mean $\mathrm{E}(X_1) = p$. Moreover, for any function $\tau(p)$, it is reasonable to use $\tau(\overline{X})$ to estimate $\tau(p)$. For instance, we might use $\overline{X}(1 - \overline{X})/n$ to estimate $\mathrm{Var}(\overline{X}) = p(1-p)/n$. $\triangle$

More generally, suppose $X_1, \ldots, X_n$ are i.i.d. with pmf or pdf $f(x;\theta)$ for a real-valued parameter θ. Then $\mathrm{E}(X_1)$ is a function of θ, say $\mathrm{E}(X_1) = \mu(\theta)$. The *method of moments*

is to equate $\mu(\tilde{\theta}) = \overline{X}$ and solve for $\tilde{\theta}$. This equation can be expressed as

estimated population mean = sample mean.

That is, the method of moments estimator of θ is the value of θ for which the population mean is equal to the sample mean. Then we estimate $\tau(\theta)$ by $\tau(\tilde{\theta})$. The word 'moment' refers to the fact that $\mu(\theta)$ is the first population moment and $\overline{X}$ is the first sample moment.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Geometric(θ), $0 < \theta < 1$. (In Mukhopadhyay's book there are two slightly different definitions of the Geometric distribution; see formula (1.7.7) and Exercise 7.2.6. We will use (1.7.7).) It can be shown that $\mathrm{E}(X_1) = 1/\theta$. Using the method of moments, we set $1/\tilde{\theta} = \overline{X}$, and then solve to get $\tilde{\theta} = 1/\overline{X}$. In order to estimate $\mathrm{SD}(X_1) = \sqrt{1-\theta}/\theta$, we use $\sqrt{1-\tilde{\theta}}/\tilde{\theta} = \sqrt{\overline{X}(\overline{X}-1)}$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(0, \theta)$, $\theta > 0$. The population mean is $\mathrm{E}(X_1) = \theta/2$. Using the method of moments, we set $\tilde{\theta}/2 = \overline{X}$, and then solve to get $\tilde{\theta} = 2\overline{X}$. To estimate $\mathrm{P}\{X_1 \le c\} = \min\{c/\theta, 1\}$ we can use $\min\{c/\tilde{\theta}, 1\}$. $\triangle$

Now suppose $X_1, \ldots, X_n$ are i.i.d. with pmf or pdf $f(x\,;\boldsymbol{\theta})$ for a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Then the j-th population moment is a function of $\boldsymbol{\theta}$, $\mathrm{E}(X_1^j) = \mu_j(\boldsymbol{\theta}) = \mu_j(\theta_1, \ldots, \theta_p)$. Let $m_j = \sum_{i=1}^{n} X_i^j/n$, the j-th sample moment. (Note that $m_1 = \overline{X}$ and $\mu_1(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta})$.) The *method of moments* is to find the value of $\boldsymbol{\theta}$ for which the first p population moments are equal to the corresponding sample moments. That is, we solve the p equations $\mu_j(\tilde{\theta}_1, \ldots, \tilde{\theta}_p) = m_j$, $j = 1, \ldots, p$, for the p unknowns $\tilde{\theta}_1, \ldots, \tilde{\theta}_p$. Then we estimate $\tau(\boldsymbol{\theta})$ by $\tau(\tilde{\boldsymbol{\theta}})$. For $p = 2$, the equations can be expressed as

estimated population mean = sample mean
estimated population 2nd moment = sample 2nd moment

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) random variables. Then $\boldsymbol{\theta} = (\mu, \sigma^2)$ and $\mu_1(\boldsymbol{\theta}) = \mathrm{E}(X_1) = \mu$, $\mu_2(\boldsymbol{\theta}) = \mathrm{E}(X_1^2) = \mu^2 + \sigma^2$, $m_1 = \overline{X}$, $m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Set $\tilde{\mu} = \overline{X}$ and $\tilde{\mu}^2 + \tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$. Solve to get $\tilde{\mu} = \bar{X}$ and $\tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2 = \frac{1}{n}\sum (X_i - \overline{X})^2$, a modified sample variance (dividing by n rather than $n-1$).
To estimate the coefficient of variation $\tau(\mu, \sigma^2) = \sigma/\mu$, use
$\tau(\tilde{\mu}, \tilde{\sigma}^2) = \tilde{\sigma}/\tilde{\mu} = \sqrt{\frac{1}{n}\sum \left(\frac{X_i}{\overline{X}} - 1\right)^2}$. $\triangle$

For $p = 2$, note that the method of moments equations

$$\mu_1(\tilde{\theta}) = \overline{X} \quad \text{and} \quad \mu_2(\tilde{\theta}) = \frac{1}{n}\sum X_i^2$$

are equivalent to the equations

$$\mu_1(\tilde{\theta}) = \overline{X} \quad \text{and} \quad \mathrm{Var}_{\tilde{\theta}}(X_1) = \frac{1}{n}\sum (X_i - \overline{X})^2$$

because $\mathrm{Var}(X_1) = \mu_2(\theta) - \mu_1(\theta)^2$ and $\frac{1}{n}\sum (X_i - \overline{X})^2 = \frac{1}{n}\sum X_i^2 - \overline{X}^2$.

estimated population mean = sample mean,

estimated population variance = modified sample variance.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Gamma(α, β) random variables. Then $\mathrm{E}(X_1) = \alpha\beta$ and $\mathrm{Var}(X_1) = \alpha\beta^2$. Set $\tilde{\alpha}\tilde{\beta} = \overline{X}$ and $\tilde{\alpha}\tilde{\beta}^2 = \frac{1}{n}\sum (X_i - \overline{X})^2$. Solve to get $\tilde{\alpha} = \overline{X}^2 / \frac{1}{n}\sum (X_i - \overline{X})^2$ and $\tilde{\beta} = \frac{1}{n}\sum (X_i - \overline{X})^2 / \overline{X}$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform(θ_1, θ_2) random variables. Then $\mathrm{E}(X_1) = (\theta_1 + \theta_2)/2$ and $\mathrm{Var}(X_1) = (\theta_2 - \theta_1)^2/12$. Set $(\tilde{\theta}_1 + \tilde{\theta}_2)/2 = \overline{X}$ and $(\tilde{\theta}_2 - \tilde{\theta}_1)^2/12 = \frac{1}{n}\sum (X_i - \overline{X})^2 = \frac{n-1}{n}S^2$. Solve to get $\tilde{\theta}_1 = \overline{X} - \sqrt{\frac{3(n-1)}{n}}S$ and $\tilde{\theta}_2 = \overline{X} + \sqrt{\frac{3(n-1)}{n}}S$. To estimate $\mathrm{P}\{X_1 \le c\} = \max\{\min\{(c - \theta_1)/(\theta_2 - \theta_1), 1\}, 0\}$ we can use $\max\{\min\{(c - \tilde{\theta}_1)/(\tilde{\theta}_2 - \tilde{\theta}_1), 1\}, 0\}$. $\triangle$

If a moment in one of the method-of-moments equations does not involve θ, then the method must be modified.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(-\theta, \theta)$, $\theta > 0$. The population mean is $\mathrm{E}(X_1) = 0$. This cannot help us estimate θ, so we try the next moment: $\mathrm{E}(X_1^2) = \theta^2/3$. Set $\tilde{\theta}^2/3 = \frac{1}{n}\sum X^2$ and solve to get $\tilde{\theta} = \sqrt{\frac{3}{n}\sum X^2}$. $\triangle$

Maximum likelihood estimation

To analyze a set of data $\boldsymbol{x} = (x_1, \ldots, x_n)$, we postulate a model. That is, we postulate that $\boldsymbol{x}$ has occurred as the observed value of a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ with distribution given by a joint pmf or pdf $f(\boldsymbol{x}\,; \theta)$ for some $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$. According to the model, one of the vectors in Θ is the vector of true parameters — the "true" parameters being those that specify the distribution of the random mechanism that generated the data — but we do not know which vector it is. A sample can provide only limited information about a population. In other words, the data $\boldsymbol{x}$ contains only a limited amount of information about the

distribution, so we cannot hope to use the data to obtain the exact value of the true θ. The goal of point estimation is to find a vector $\widehat{\theta}$ that is "close" to the true θ.

The method of moments chooses $\widetilde{\theta}$ so that the first p population moments are equal to the first p sample moments. Intuitively, it seems that such a $\widetilde{\theta}$ should be "close" to the true θ.

The method of maximum likelihood (ML) chooses $\widehat{\theta}$ so that the joint pmf or pdf $f(\boldsymbol{x}\,;\theta)$ is maximized, where $\boldsymbol{x}$ denotes the observed data (rather than being simply a mathematical variable). In the following example we will see that this is a sensible procedure. This is called the method of maximum likelihood because $f(\boldsymbol{x}\,;\theta)$, when viewed as a function of θ for a fixed value of $\boldsymbol{x}$, is called the *likelihood function*. We denote it by $L(\theta;\boldsymbol{x})$ or $L(\theta)$ (so that $L(\theta;\boldsymbol{x}) = f(\boldsymbol{x}\,;\theta)$).

Example. Suppose we toss a bent coin 10 times and see which tosses land head up. We record 1 if the toss is a head and record 0 if the toss is a tail. Suppose we observe $\boldsymbol{x} = (1,0,0,1,1,1,0,1,1,1)$. For the model we suppose $X_1,\ldots,X_{10}$ are i.i.d. Bernoulli(θ), $0<\theta<1$. The joint pmf is $f(\boldsymbol{x};\theta) = \mathrm{P}_\theta\{\boldsymbol{X}=\boldsymbol{x}\} = \prod_{i=1}^{10}\mathrm{P}\{X_i = x_i\} = \prod_{i=1}^{10}\theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{10-\sum x_i} = \theta^7(1-\theta)^3$. So, for example, if $\theta = .5$, then $\mathrm{P}\{\boldsymbol{X} = (1,0,0,1,1,1,0,1,1,1)\} = (.5)^7(.5)^3 = .0009765$. If $\theta = .6$, then $\mathrm{P}\{\boldsymbol{X} = (1,0,0,1,1,1,0,1,1,1)\} = (.6)^7(.4)^3 = .0011943$. Thus $\theta = .6$ seems "more likely" to be true than $\theta = .5$. The ML method is to choose the value of θ for which $\mathrm{P}\{\boldsymbol{X} = (1,0,0,1,1,1,0,1,1,1)\}$ is the largest, that is, for which $\theta^7(1-\theta)^3$ is the largest. To find the maximum of a function, it often helps to calculate its derivative (if the derivative exists). In particular, the sign of the derivative tells where the function is increasing and where it is decreasing.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\theta}\left[\theta^7(1-\theta)^3\right] &= \cdots(\text{can use the product rule of differentiation})\cdots \\ &= \theta^6(1-\theta)^2(7-10\theta)\,.\end{aligned}$$

Since $\theta^6(1-\theta)^2 > 0$ for $0<\theta<1$, the sign of the derivative is the sign of $7-10\theta$. The derivative is 0 when $\theta = .7$, is >0 when $\theta < .7$, and is <0 when $\theta > .7$. When the derivative is positive, the function is strictly increasing, and when the derivative is negative, the function is strictly decreasing. From this we see that the function has its maximum value at $\theta = .7$. We say that the MLE of θ is .7. $\triangle$

In general, if the likelihood function $L(\theta)$ is differentiable, it is often easier to differentiate $\log L(\theta)$. Since log is a strictly increasing function, $\widehat{\theta}$ maximizes $L(\theta)$ if and only if $\widehat{\theta}$ maximizes $\log L(\theta)$. In the preceding example, $L(\theta) = \theta^7(1-\theta)^3$, so $\log L(\theta) =$

$7\log\theta + 3\log(1-\theta)$ and $\frac{d}{d\theta}\log L(\theta) = \frac{7}{\theta} - \frac{3}{1-\theta} = \frac{7-10\theta}{\theta(1-\theta)}$. As before, the sign of the derivative is the sign of $7 - 10\theta$.

Definition. (a) For a given value of x, a *maximum likelihood estimate* (MLE) of θ is a value $\hat{\theta} \in \Theta$ such that $L(\hat{\theta}\,; x) \geq L(\theta\,; x)$ for all $\theta \in \Theta$. We sometimes write $\hat{\theta} = \hat{\theta}(x)$.
(b) A *maximum likelihood estimator* (also denoted MLE) of θ is a statistic $\hat{\theta}(X)$ such that $\hat{\theta}(x)$ is an ML estimate for every value of x in the sample space.
(c) For any function $\tau(\theta)$, the ML method estimates it by $\tau(\hat{\theta})$.

Consider the case when θ is real-valued. Often, an MLE can be obtained as a solution of the equation (*) $\frac{d}{d\theta}L(\theta) = 0$ (or equivalently, the equation $\frac{d}{d\theta}\log L(\theta) = 0$). However, sometimes $L(\theta)$ is not differentiable at all θ. When $L(\theta)$ is differentiable at all θ and we have found a solution $\hat{\theta}$ to (*), we must be aware that the solution is not guaranteed to be a valid MLE. To be valid, the solution must satisfy: (i) $\hat{\theta}$ is a global maximum of $L(\theta)$ and (ii) $\hat{\theta} \in \Theta$.

(i) How can we establish that $\hat{\theta}$ is a global maximum? Note that if $g(t)$ is a differentiable function and if the derivative $\frac{d}{dt}g(t)$ is 0 at $t = \hat{t}$, then $\hat{t}$ could be a global maximum but, in general, it might be only a local maximum or a local minimum or a stationary point.
If we also calculate the second derivative $\frac{d^2}{dt^2}g(t)$ at $t = \hat{t}$ and find it to be negative, then $\hat{t}$ is not a local minimum nor is it a stationary point, but it could still be a local maximum and is not necessarily a global maximum. (Local information at $\hat{t}$ cannot tell us about a global maximum.)
We can establish that $\hat{t}$ is a global maximum by showing that the second derivative $\frac{d^2}{dt^2}g(t)$ is negative or zero for all t.
Another way to establish that $\hat{t}$ is a global maximum is to show that $\frac{d}{dt}g(t)$ is positive for all $t < \hat{t}$ and is negative for all $t > \hat{t}$.
(ii) Sometimes we "fudge" and allow $\hat{\theta}$ to be on the boundary of Θ. Another option is to extend Θ to include its boundary.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ), $0 < \theta < 1$. The likelihood function is

$$L(\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}.$$

It is equivalent, but more convenient, to maximize

$$\log L(\theta) = \sum x_i \log\theta + (n - \sum x_i)\log(1-\theta).$$

Its derivative is

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log L(\theta) = \frac{\sum x_i}{\theta} - \frac{n-\sum x_i}{1-\theta} = \frac{n(\overline{x}-\theta)}{\theta(1-\theta)}.$$

The sign of the derivative is the same as the sign of $\overline{x}-\theta$, which is positive for $0<\theta<\overline{x}$, is equal to 0 for $\theta=\overline{x}$, and is negative for $\overline{x}<\theta<1$. So $\widehat{\theta}=\overline{x}$ is the global maximum of $L(\theta)$. In the case that $x_1=\cdots=x_n=0$, we have $\widehat{\theta}=0$, but $0\notin(0,1)=\Theta$. If we strictly followed the definition of MLE, we would say that no MLE exists in this case. However, the MLE does exist if we redefine the parameter set to be $\Theta^*=[0,1]$, which includes 0. The drawback of this option is that the family is no longer an exponential family. $\triangle$

Example. Suppose $X_1,\ldots,X_n$ are i.i.d. Normal$(\mu,1)$, $-\infty<\mu<\infty$.

$$L(\mu) = \prod_{i=1}^{n} f(x_i;\mu) = \cdots = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\Big[-\tfrac{1}{2}\sum(x_i-\mu)^2\Big]$$

$$\log L(\mu) = \text{constant} - \tfrac{1}{2}\sum(x_i-\mu)^2$$

$$\frac{\mathrm{d}}{\mathrm{d}\mu}\log L(\mu) = -\tfrac{1}{2}\sum 2(x_i-\mu)(-1) = \sum(x_i-\mu)$$

$$= \sum x_i - n\mu = n(\overline{x}-\mu).$$

As in the preceding example, the sign of the derivative shows that $\widehat{\mu}=\overline{x}$ is the MLE of μ. For any function $\tau(\mu)$, the ML method estimates it by $\tau(\widehat{\mu})$. For example, the MLE of μ^2 is $\overline{X}^2$.

Example. Suppose $X_1,\ldots,X_n$ are i.i.d. Uniform$(0,\theta)$, $\theta>0$.

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I\{0<x_i<\theta\} = \frac{1}{\theta^n} I\{0<x_{(1)}<x_{(n)}<\theta\}.$$

It is convenient to let the sample space be $(0,\infty)$ rather than the whole real line. Now

$$L(\theta) = \frac{1}{\theta^n} I\{x_{(n)}<\theta\}$$

$$\log L(\theta) = -n\log\theta + \log(I\{x_{(n)}<\theta\}).$$

Note that we have a problem here, because $I\{x_{(n)}<\theta\}$ can be 0 and $\log(0)=-\infty$. Let us look at $L(\theta)$ directly, without taking the logarithm. Remember that $x_{(n)}$ is regarded as a fixed value. For $\theta<x_{(n)}$, the indicator is 0, so $L(\theta)=0$. For $\theta>x_{(n)}$, the indicator is 1, so $L(\theta)=1/\theta^n$. If we graph $L(\theta)$, we see that it is not differentiable and is not even continuous. For $0<\theta<x_{(n)}$, the likelihood function is 0, but then at $\theta=x_{(n)}$, the function jumps up to a height of $1/x_{(n)}^n$, and then decreases for $\theta>x_{(n)}$. The graph makes it clear that the maximum is at $\widehat{\theta}=x_{(n)}$. $\triangle$

Now suppose $\theta=(\theta_1,\ldots,\theta_p)$ is a vector. If the likelihood function is differentiable with respect to θ_j for all $j=1,\ldots,p$, then the MLE can often be found by solving the equations

Now consider the multiparameter case in which $\theta = (\theta_1, \ldots, \theta_p)$ is a vector. If the likelihood function is differentiable with respect to θ_j for all $j = 1, \ldots, p$, then the MLE can often be found by solving the equations

$$\frac{\partial}{\partial\theta_1} L(\hat{\theta}_1, \ldots, \hat{\theta}_p) = 0\,, \ \ldots, \ \frac{\partial}{\partial\theta_p} L(\hat{\theta}_1, \ldots, \hat{\theta}_p) = 0$$

to obtain $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$. These equations are sometimes called the *likelihood equations*. For $\hat{\theta}$ to be an MLE, it must be checked that $\hat{\theta} \in \Theta$ and that $\hat{\theta}$ is a global maximum — which is often true, but not always. As before, it is equivalent and often easier to maximize $\log L(\theta_1, \ldots, \theta_p)$, which is often achieved by solving

$$\frac{\partial}{\partial\theta_1} \log L(\hat{\theta}_1, \ldots, \hat{\theta}_p) = 0\,, \ \ldots, \ \frac{\partial}{\partial\theta_p} \log L(\hat{\theta}_1, \ldots, \hat{\theta}_p) = 0\,.$$

These equations are also sometimes called the *likelihood equations*. Note that $\log L(\theta) = \log\prod_{i=1}^n f(x_i;\theta) = \sum_{i=1}^n \log f(x_i;\theta)$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2). The likelihood function is

$$L(\mu,\sigma) \ = \ \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \exp\Big[-\tfrac{1}{2\sigma^2}\textstyle\sum(x_i-\mu)^2\Big]\,.$$

We want to find $\hat{\mu}$ and $\hat{\sigma}$ that maximize $L(\mu,\sigma)$, or equivalently, maximize

$$\log L(\mu,\sigma) \ = \ -n\log\sqrt{2\pi} - n\log\sigma - \ \tfrac{1}{2\sigma^2}\textstyle\sum(x_i-\mu)^2\,.$$

The procedure has three steps: (1) calculate the partial derivatives of $\log L(\mu,\sigma)$ with respect to μ and σ, (2) set the two partial derivatives equal to 0 and solve for $\hat{\mu}$ and $\hat{\sigma}$, and (3) verify that $(\hat{\mu}, \hat{\sigma})$ is a global maximum of $\log L(\mu,\sigma)$.

Step 1. The partial derivatives are

$$\frac{\partial}{\partial\mu}\log L(\mu,\sigma) \ = \ \tfrac{1}{\sigma^2}\textstyle\sum(x_i-\mu)$$

and

$$\frac{\partial}{\partial\sigma}\log L(\mu,\sigma) \ = \ -\,\frac{n}{\sigma} + \tfrac{1}{\sigma^3}\textstyle\sum(x_i-\mu)^2\,.$$

Step 2. Note that $(\partial/\partial\mu)\log L = (n/\sigma^2)(\bar{x}-\mu)$. Setting this equal to 0, we obtain $\hat{\mu} = \bar{x}$. (This example is special in that the parameter σ cancels out of this equation. Typically, if the equation $(\partial/\partial\theta_1)\log L(\theta_1,\theta_2) = 0$ is solved for θ_1, the solution is a function of θ_2.) Setting $(\partial/\partial\sigma)\log L = 0$, we obtain $\hat{\sigma} = \sqrt{\frac{1}{n}\sum(x_i-\hat{\mu})^2} \ = \sqrt{\frac{1}{n}\sum(x_i-\bar{x})^2}$.

Step 3. Now we need to verify that $(\hat{\mu}, \hat{\sigma})$ is a global maximum. Verifying a global maximum for a two-variable function can be difficult, but in this example the verification can be broken into two relatively easy steps. First, we can see that $(\partial/\partial\mu)\log L > 0$ for $\mu < \bar{x}$

and < 0 for $\mu > \bar{x}$. So, for every fixed value of σ, $\log L$ achieves a maximum at $\mu = \bar{x}$. Next, we want to find a global maximum of $\log L(\bar{x}, \sigma)$ as a function of σ. We can write $(\partial/\partial\sigma)\log L(\bar{x}, \sigma) = (n/\sigma^3)(\hat{\sigma}^2 - \sigma^2)$. Hence this derivative is positive for $\sigma < \hat{\sigma}$ and is negative for $\sigma > \hat{\sigma}$. This implies that $\hat{\sigma}$ maximizes $\log L(\bar{x}, \sigma)$. Therefore, the solution $(\hat{\mu}, \hat{\sigma})$ to the partial derivative equations maximizes $\log L(\mu, \sigma)$. Thus, the MLEs of μ and σ are given by $\hat{\mu}(X) = \bar{X}$ and $\hat{\sigma}(X) = \sqrt{\frac{1}{n}\sum(X_i - \bar{X})^2}$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Gamma(α, β). The pdf is $f(x\,;\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$ for $x > 0$. Rather than maximize the likelihood function directly, it is easier to maximize the log-likelihood function. We have $\log f(x\,;\alpha,\beta) = -\log\Gamma(\alpha) - \alpha\log\beta + (\alpha - 1)\log x - x/\beta$, so the log-likelihood function is

$$\log L(\alpha,\beta) = -n\log\Gamma(\alpha) - n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i .$$

The general procedure has the same three steps as in the preceding example: (1) calculate the partial derivatives of $\log L(\alpha, \beta)$ with respect to α and β, (2) set the two partial derivatives equal to 0 and solve for $\hat{\alpha}$ and $\hat{\beta}$, and (3) verify that $(\hat{\alpha}, \hat{\beta})$ is a global maximum of $\log L(\alpha, \beta)$.

Step 1. The partial derivatives are

$$\frac{\partial}{\partial\alpha}\log L(\alpha,\beta) = -n\frac{\partial}{\partial\alpha}\log\Gamma(\alpha) - n\log\beta + \sum_{i=1}^{n}\log x_i$$

and

$$\frac{\partial}{\partial\beta}\log L(\alpha,\beta) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} x_i .$$

Step 2. In the preceding example we had the good fortune that we could obtain an explicit formula for the MLE $\hat{\mu}$ in terms of the x_i's by setting $(\partial/\partial\mu)\log L = 0$. Here we are not as fortunate, but we are still able to obtain an explicit solution for $\hat{\beta}$ in terms the x_i's and $\hat{\alpha}$ by setting $(\partial/\partial\beta)\log L = 0$. The solution is $\hat{\beta} = \frac{1}{n\hat{\alpha}}\sum_{i=1}^{n} x_i = \bar{x}/\hat{\alpha}$. Now plug $\hat{\beta}$ into the equation $(\partial/\partial\alpha)\log L = 0$ to obtain

$$(*) \qquad \log\hat{\alpha} - \psi(\hat{\alpha}) = \log\bar{x} - \frac{1}{n}\sum_{i=1}^{n}\log x_i$$

where $\psi(\alpha) = \frac{\partial}{\partial\alpha}\log\Gamma(\alpha)$, called the digamma function. There is no explicit formula for the solution $\hat{\alpha}$ to equation (*). Given the values of the data, however, we can use an iterative numerical procedure to obtain a numerical value for $\hat{\alpha}$. Such a numerical procedure requires the ability to evaluate of the digamma function, which is available in the computer packages

SAS and S-Plus but not in Matlab. Tables are available in the book Handbook of Mathematical Functions edited by Abramowitz and Stegun.

Step 3. Let $\widehat{\alpha}$ be a solution to (*) and let $\widehat{\beta} = \overline{x}/\widehat{\alpha}$. It is difficult to verify that these values give a global maximum of the likelihood function, so we will omit this step. In complicated problems, the "MLEs" of the parameters are often found by solving the likelihood equations and without checking whether or not they give a global maximum. Typically such estimators have good properties, even though they may not always be valid MLEs. $\triangle$

Suppose X is a data vector with joint pmf or pdf $f(x\,;\theta)$ where θ is an unknown real-valued parameter. Let $\widehat{\theta}$ be an MLE of θ. The MLE of a function $\tau(\theta)$ is $\tau(\widehat{\theta})$, according to our definition on p. 45 above. Suppose $\tau(\theta)$ is a one-to-one function. Then we could choose to use τ as the parameter for our model instead of θ. For example, the family of Exponential pdf's can be written as $f(x\,;\beta) = \beta^{-1}e^{-x/\beta}$, $\beta > 0$, or as $f(x\,;\lambda) = \lambda e^{-\lambda x}$, $\lambda > 0$. Here $\theta = \beta$ and $\tau(\theta) = 1/\theta = \lambda$. So, in general, we can write the pmf or pdf as $f(x\,;\theta)$ or $f(x\,;\tau)$. This gives two ways to obtain an MLE for $\tau(\theta)$. The first way is to maximize $f(x\,;\theta)$ to obtain an MLE $\widehat{\theta}$ and plug into $\widehat{\tau} = \tau(\widehat{\theta})$. The second way is to maximize $f(x\,;\tau)$ to directly obtain $\widehat{\tau}$. Fortunately, both ways lead to the same estimator.

Example. Let $X_1, \ldots, X_n$ be i.i.d. Exponential(β). The pdf is $f(x\,;\beta) = \frac{1}{\beta}e^{-x/\beta}$ for $x > 0$ for some unknown parameter $\beta > 0$. Let $\lambda = 1/\beta$. (The parameter β is the mean of the distribution, and the parameter λ is called the rate.) Let us derive the MLE of λ in two ways.

(a) First find the MLE of β.

$$L(\beta) = f(x\,;\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-\frac{x_i}{\beta}} = \frac{1}{\beta^n}\exp\Big[-\frac{1}{\beta}\sum_{i=1}^{n} x_i\Big]$$

$$\log L(\beta) = -n\log\beta - \frac{1}{\beta}\sum x_i$$

$$\frac{\partial}{\partial\beta}\log L(\beta) = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum x_i = -\frac{n}{\beta^2}(\beta - \overline{x}).$$

We see that the MLE of β is $\widehat{\beta} = \overline{x}$. The MLE of $\lambda = 1/\beta$ is $\widehat{\lambda} = 1/\widehat{\beta} = 1/\overline{x}$.

(b) Now we use the parameterization in terms of λ.

$$L(\lambda) = f(x\,;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n \exp\Big[-\lambda\sum_{i=1}^{n} x_i\Big]$$

$$\log L(\lambda) = n\log\lambda - \lambda\sum x_i$$

$$\frac{\partial}{\partial\lambda}\log L(\lambda) = \frac{n}{\lambda} - \sum x_i = \frac{n\overline{x}}{\lambda}\Big(\frac{1}{\overline{x}} - \lambda\Big).$$

We see that the MLE of λ is $\widehat{\lambda} = 1/\overline{x}$, which agrees with the result in part (a). $\triangle$

In general, suppose X is a data vector with joint density (pmf or pdf) $f(x;\theta)$.

Lemma. Let $\hat{\theta}$ be an MLE of θ, maximizing $f(x;\theta)$. Suppose $\gamma = h(\theta)$ is a one-to-one transformation of θ, so that the joint density can be expressed as $f(x;\gamma)$. Suppose $\hat{\gamma}$ is an MLE of γ, maximizing $f(x;\gamma)$. Then $\hat{\gamma} = h(\hat{\theta})$.

For example, suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2). To find the MLE we can differentiate with respect to μ and σ^2 to obtain the MLEs $\hat{\mu} = \bar{x}$ and $\hat{\sigma}^2 = \sum(x_i - \bar{x})^2/n$. Here we are parameterizing the pdf's by the parameter vector (μ, σ^2). Alternatively, we can parameterize by (μ, σ) and differentiate with respect to μ and σ, as we did on p. 47 above, to obtain $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = \sqrt{\sum(x_i - \bar{x})^2/n}$.

To see why the lemma is true, let us regard the parameter vector θ as a label for the distributions in a family of distributions. For instance, the Exponential distributions are often parameterized (or labeled) by their means, but some books parameterize them by their rates (recall that the rate is the reciprocal of the mean). The labels are useful, but the distribution is what ultimately matters. By definition, to find the MLE, we must find the distribution whose pdf gives the greatest value to the observed value of the data vector x. Let's call this the "ML distribution". For a particular parameterization, the MLE of the parameter is the value of the parameter that labels the ML distribution. In the lemma, $\gamma = h(\theta)$ labels the same distribution as θ. Therefore, if $\hat{\theta}$ labels the ML distribution in the θ parameterization, then $\hat{\gamma} = h(\hat{\theta})$ labels it in the γ parameterization.

Comparing estimators

Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) and suppose we want to estimate μ. The method of moments estimator (MOME) is $\bar{X}$ (see p. 42 above). This is also the MLE (see pp. 47-48). The sample mean $\bar{X}$ is a natural estimator of the population mean μ. Since a Normal distribution is symmetric about its mean, μ is the population median. It also seems natural to estimate the population median by the sample median $\tilde{X}$. Which of these two estimators, $\bar{X}$ and $\tilde{X}$, is a better estimator of μ? Later we will argue that the sample mean is better. But first we must ask: what is a good way to compare competing estimators?

For comparing two estimators, three relevant properties are: the bias, the variance, and the mean squared error. Let T be a real-valued statistic that is being considered as an estimator of $\tau(\theta)$. The *bias* of T is $E_\theta(T) - \tau(\theta)$. If the bias is 0, then the estimator is said to be *unbiased*. That is, T is an unbiased estimator of $\tau(\theta)$ if and only if $E_\theta(T) = \tau(\theta)$. The bias can be said to measure the "accuracy" of the estimator. The "stability" of an estimator can be measured by its variance $\mathrm{Var}_\theta(T)$ (or by its standard deviation (SD)). The *mean*

squared error (MSE) of an estimator is $\text{MSE}_\theta(T) = \text{E}_\theta[(T - \tau(\theta))^2]$ (or by its root mean squared error). (The root mean squared error (RMSE) is the square root of the MSE; it has the advantage that its units are the same as those of $\tau(\theta)$.) The MSE (or RMSE) can be said to measure the "reliability" of the estimator.

Theorem 7.3.1. $\text{MSE}(T) = \text{Var}(T) + (\text{Bias}(T))^2$

Proof. Let $Z = T - \tau(\theta)$. The theorem follows from the identity $\text{E}(Z^2) = \text{Var}(Z) + (\text{E}(Z))^2$. □

Corollary. If T is unbiased, then $\text{MSE}(T) = \text{Var}(T)$.

We prefer estimators that have small mean squared error. According to the theorem, such an estimator must have both small bias and small variance.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(0, \theta)$, $\theta > 0$. On p. 42 above we found the method of moments estimator (MOME) to be $\tilde{\theta} = 2\overline{X}$. On p. 46 we found the MLE to be $\hat{\theta} = X_{(n)}$. Let us compare these two estimators with regard to their bias, variance, and MSE.

(a) Bias. MOME: $\text{E}(\tilde{\theta}) = \text{E}(2\overline{X}) = 2(\theta/2) = \theta$. So the MOME is unbiased, with bias $= 0$.

MLE: The pdf of $T = X_{(n)}$ is (see p. 38) $g(t) = nt^{n-1}/\theta^n$ for $0 < t < \theta$. So

$$\text{E}(\hat{\theta}) = \int_0^\theta t n t^{n-1}/\theta^n \, \text{d}t = (n/\theta^n)\int_0^\theta t^n \, \text{d}t = (n/\theta^n)\left\{t^{n+1}/(n+1)\right\}_{t=0}^{t=\theta} = n\theta/(n+1).$$

The bias of the MLE is $n\theta/(n+1) - \theta = -\theta/(n+1)$. So the MOME has less bias.

(b) Variance. MOME: $\text{Var}(\tilde{\theta}) = \text{Var}(2\overline{X}) = 4(\theta^2/12)/n = \theta^2/(3n)$.

MLE: $\text{Var}(\hat{\theta}) = \text{E}(T^2) - (\text{E}(T))^2 = n\theta^2/(n+2) - [n\theta/(n+1)]^2 = n\theta^2/[(n+1)^2(n+2)]$.

The MLE has smaller variance for all $n \geq 1$.

(c) MSE. MOME: By the corollary above, $\text{MSE}(\tilde{\theta}) = \text{Var}(\tilde{\theta}) = \theta^2/(3n)$.

MLE: By Theorem 7.3.1, $\text{MSE}(\hat{\theta}) = n\theta^2/[(n+1)^2(n+2)] + [-\theta/(n+1)]^2 = 2\theta^2/[(n+1)(n+2)]$. The two MSEs are equal when $n = 1$ or 2, and the MLE has smaller MSE when $n \geq 3$. So the MLE is a better estimator, according to the criterion of MSE (or of RMSE). As noted above, RMSE is a more interpretable measure of performance than MSE. For large n, the RMSE of the MLE is quite a bit smaller than the RMSE of the MOME, because the ratio $\text{RMSE}(\tilde{\theta})/\text{RMSE}(\hat{\theta}) = \sqrt{(n+1)(n+2)/(6n)} \to \infty$ as $n \to \infty$. For $n = 20$, the ratio is about 2, which tells us that the MOME tends to be about twice as far from θ as the MLE. △

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) and suppose we want to estimate σ^2. The MOME (see p. 42) and the MLE (see p. 47) are both equal to $\hat{\sigma}^2 = \sum(X_i - \overline{X})^2/n = \frac{n-1}{n}S^2$. Let us compare the two estimators $\hat{\sigma}^2$ and S^2.

(a) Bias. We know S^2 is an unbiased estimator of σ^2.

$E(\widehat{\sigma}^2) = E(\frac{n-1}{n}S^2) = \frac{n-1}{n}\sigma^2$, so the bias of $\widehat{\sigma}^2$ is $\frac{n-1}{n}\sigma^2 - \sigma^2 = -\sigma^2/n$.
So S^2 has less bias.

(b) Variance. We know $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$, so $S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{(n-1)}$, hence
$\text{Var}(S^2) = \left(\frac{\sigma^2}{n-1}\right)^2 \text{Var}(\chi^2_{(n-1)}) = \frac{\sigma^4}{(n-1)^2}2(n-1) = \frac{2\sigma^4}{n-1}$.
$\text{Var}(\widehat{\sigma}^2) = \text{Var}(\frac{n-1}{n}S^2) = \left(\frac{n-1}{n}\right)^2 \text{Var}(S^2) = \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} = \frac{2(n-1)\sigma^4}{n^2}$.
It can be shown that $\widehat{\sigma}^2$ has smaller variance for all $n \geq 2$. (We do not consider $n = 1$ because then S^2 is not well-defined.)

(c) MSE. Since S^2 is unbiased, $\text{MSE}(S^2) = \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$.
$\text{MSE}(\widehat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2} + \left(\frac{-\sigma^2}{n}\right)^2 = \frac{(2n-1)\sigma^4}{n^2}$. The MSE of $\widehat{\sigma}^2$ is smaller for all $n \geq 2$, because $(2n-1)/n^2 < 2/(n-1)$, because $(2n-1)(n-1) = 2n^2 - 3n + 1 < 2n^2$.
So the MLE is a better estimator. By multiplying the sample variance by $(n-1)/n$, we introduce some bias, but this is compensated by a decrease in variance ~~of σ^2~~.
For large n, the RMSEs of the two estimators are almost equal, because the ratio $\text{RMSE}(S^2)/\text{RMSE}(\widehat{\sigma}^2) = \sqrt{2n^2/(2n^2-3n+1)} \to 1$ as $n \to \infty$. For $n = 20$, the ratio is 1.04, which tells us that S^2 tends to be about 4% farther from σ^2 than $\widehat{\sigma}^2$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2) and suppose we want to estimate μ. Let us compare the two estimators $\overline{X}$ and $\frac{1}{2}\overline{X}$. We know $\overline{X}$ is unbiased, from which we deduce that $\frac{1}{2}\overline{X}$ has bias $-\frac{1}{2}\mu$. We find that $\frac{1}{2}\overline{X}$ has smaller variance: $\text{Var}(\overline{X}) = \sigma^2/n$ and $\text{Var}(\frac{1}{2}\overline{X}) = \frac{1}{4}\sigma^2/n$. Now, $\text{MSE}(\overline{X}) = \sigma^2/n$ and $\text{MSE}(\frac{1}{2}\overline{X}) = \frac{1}{4}\sigma^2/n + (-\frac{1}{2}\mu)^2 = \frac{1}{4}(\sigma^2/n + \mu^2)$. The MSE of $\overline{X}$ is smaller if and only if $\sigma^2/n < \frac{1}{4}(\sigma^2/n + \mu^2)$, or $|\mu|/\sigma > \sqrt{3/n}$. If, for instance, $\mu > \sigma$ and $n \geq 3$, then $\overline{X}$ has smaller MSE. But if $\mu = 0$, then $\frac{1}{2}\overline{X}$ has smaller MSE. Since we do not know the values of the parameters, we cannot say which estimator has smaller MSE. $\triangle$

Different people may choose different criteria for judging estimators. Some people prefer to use unbiased estimators, even though a biased estimator may have smaller MSE. Other people do not mind a little bias in an estimator if it allows smaller MSE.

Unbiased estimation

Unbiasedness is an appealing property. We might choose to restrict our attention to unbiased estimators. Of course this will not be a good idea if no unbiased estimators exist.

Example. Suppose $X \sim \text{Binomial}(n, \theta)$, $0 < \theta < 1$ and suppose we want to estimate $\tau(\theta) = 1/\theta$. Can we find an unbiased estimator? Suppose $h(X)$ is unbiased for $1/\theta$. Then

$\mathrm{E}[h(X)] = \sum_{x=0}^{n} h(x)\binom{n}{x}\theta^x(1-\theta)^{n-x} = 1/\theta$ for all $0 < \theta < 1$. But this is impossible, because if we let $\theta \to 0$, all the terms in the sum approach 0 except for the term for $x = 0$, which approaches $h(0)$, which is a finite number. However, $1/\theta$ approaches ∞. $\triangle$

If unbiased estimators exist for estimating $\tau(\theta)$, which one should we use? It would be nice if we could find one having the smallest variance. If T^* is an unbiased estimator for $\tau(\theta)$ and if $\mathrm{Var}_\theta(T^*) \le \mathrm{Var}_\theta(T)$ for all unbiased estimators T and all parameter vectors θ, then we say that T^* is a *uniformly minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$. It is also sometimes called a *best unbiased estimator.*

Whereas an MLE almost always exists, a UMVUE exists only in certain situations. Before investigating UMVUEs, let us consider the related topic of BLUEs.

Best linear unbiased estimation

Suppose $X_1, \ldots, X_n$ are i.i.d. with mean μ and variance σ^2 and suppose we want to estimate μ. Instead of considering all unbiased estimators of μ, let us restrict our attention even further to linear unbiased estimators of the form $T = a_1X_1 + \cdots + a_nX_n$ where the coefficients a_i are nonrandom. If T^* is a linear unbiased estimator for μ and if $\mathrm{Var}(T^*) \le \mathrm{Var}(T)$ for all linear unbiased estimators T and all values of μ and σ^2, then we say that T^* is a *best linear unbiased estimator* (BLUE) of μ.

First, which linear ~~unbiased~~ estimators are unbiased for μ? If T is linear, then $T = a_1X_1 + \cdots + a_nX_n$ and $\mathrm{E}(T) = \mathrm{E}(a_1X_1 + \cdots + a_nX_n) = a_1\mathrm{E}(X_1) + \cdots + a_n\mathrm{E}(X_n) = a_1\mu + \cdots + a_n\mu = (a_1 + \cdots + a_n)\mu$. We see that the requirement to be unbiased for μ is that $a_1 + \cdots + a_n = 1$. Next, $\mathrm{Var}(T) = \mathrm{Var}(a_1X_1 + \cdots + a_nX_n) = a_1^2\mathrm{Var}(X_1) + \cdots + a_n^2\mathrm{Var}(X_n) = a_1^2\sigma^2 + \cdots + a_n^2\sigma^2 = (a_1^2 + \cdots + a_n^2)\sigma^2$.

The BLUE of μ is obtained by minimizing $a_1^2 + \cdots + a_n^2$ subject to $a_1 + \cdots + a_n = 1$.

Recall that $\sum(y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$. Letting $y_i = a_i$ and noting that $\bar{y} = \bar{a} = \sum a_i/n = 1/n$, we find that $\sum a_i^2 = \sum(a_i - \bar{a})^2 + n\bar{a}^2 = \sum(a_i - 1/n)^2 + 1/n$. Therefore, $\sum a_i^2 \ge 1/n$ and the minimum value is attained when $a_i = 1/n$ for all i. The BLUE of μ is $(1/n)X_1 + \cdots + (1/n)X_n = \overline{X}$.

In the derivation of the BLUE, note that we did not need to assume any particular distribution for the population. We only assumed that it had a mean μ and a variance σ^2. If $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2), then the sample mean $\overline{X}$ is the MOME and the MLE of μ. We will see below that it is also the UMVUE of μ. To be able to say that $\overline{X}$ is the MLE or UMVUE of μ, we must assume the distribution is Normal. If we drop the assumption of a Normal distribution, we can still say that $\overline{X}$ is the MOME and BLUE of μ.

Rao-Blackwellization

Next we will prove that a sufficient statistic is "sufficient" for minimum variance unbiased estimation. Let X be a data vector with joint pmf or pdf $f(x\,;\theta)$ for some $\theta \in \Theta$.

Theorem 7.4.1 (Rao-Blackwell Theorem). Suppose $T(X)$ is a sufficient statistic. Let $U(X)$ be an unbiased estimator of $\tau(\theta)$. Define the estimator $W(T) = \mathrm{E}(U \mid T)$. Then $W(T)$ is an unbiased estimator of $\tau(\theta)$, and $\mathrm{Var}_\theta(W) \leq \mathrm{Var}_\theta(U)$ for all θ.

The definition of $W(T)$ may look strange at first sight. Let us review the concept of conditional expectation in sections 3.2 and 3.3 in Mukhopadhyay's book. Let Y be a real-valued random variable (such as $U(X)$) and let Z be a random vector (such as $T(X)$) and suppose they have a joint pmf or pdf $f(y, z)$. The conditional distribution of Y given $Z = z$ is described by the conditional pmf or pdf $f(y|z) = f(y, z)/f_Z(z)$.

As a function of y, with z fixed at any value for which $f_Z(z) > 0$, this is a valid pmf or pdf. That is, $f(y|z) \geq 0$ for all y and $\sum_{\text{all } y} f(y|z) = 1$ (or $\int_{-\infty}^{\infty} f(y|z)\mathrm{d}y = 1$). The mean of this distribution is called the conditional mean or conditional expectation,

$$\mathrm{E}(Y|z) = \sum_{\text{all } y} y f(y|z) \quad (\text{or } \int_{-\infty}^{\infty} y f(y|z)\mathrm{d}y).$$

Note that $\mathrm{E}(Y|z)$ is a function of z. If $h(z)$ is any real-valued function of z and we plug in the random vector Z, then $h(Z)$ is a random variable. Thus $\mathrm{E}(Y|Z)$ is a real-valued random variable. Note that its randomness is due to Z and not Y. In particular, in the Rao-Blackwell Theorem, $\mathrm{E}(W|T)$ is a real-valued random variable in so far as it is a real-valued function of the random vector T.

From Theorem 3.3.1 we know that $\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|Z)]$ and $\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|Z)] + \mathrm{Var}[\mathrm{E}(Y|Z)]$, provided that the expectations exist.

Proof of the Rao-Blackwell Theorem. First note that, since T is sufficient for θ, then $\mathrm{E}_\theta(U|T)$ does not depend on θ, and so $W(T)$ is a valid statistic. To show that W is unbiased, $\mathrm{E}_\theta(W) = \mathrm{E}_\theta[\mathrm{E}(U|T)] = \mathrm{E}_\theta[\mathrm{E}_\theta(U|T)] = \mathrm{E}_\theta(U)$, where the last equality is from Theorem 3.3.1(i). Furthermore, $\mathrm{Var}_\theta(W) = \mathrm{Var}_\theta[\mathrm{E}(U|T)] = \mathrm{Var}_\theta[\mathrm{E}_\theta(U|T)] \leq \mathrm{Var}_\theta[\mathrm{E}_\theta(U|T)] + \mathrm{E}_\theta[\mathrm{Var}_\theta(U|T)] = \mathrm{Var}_\theta(U)$, where the last equality is from Theorem 3.3.1(ii) and the inequality is due to the fact that $\mathrm{Var}_\theta(U|T) \geq 0$. □

It can be shown that the variance of $W = \mathrm{E}(U|T)$ is strictly smaller than that of W, i.e., $\mathrm{Var}_\theta(W) < \mathrm{Var}_\theta(U)$, unless U is a function of the sufficient statistic T. If U is a function of T, then $W = U$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ), $0 < \theta < 1$.

(a) Suppose we want to estimate θ unbiasedly. We know $T = \sum X_i$ is a sufficient statistic. Let us find an unbiased estimator of θ that is a function of T. To do this, we can use the Rao-Blackwell Theorem. Actually, in this case, we would not need to bother with the Rao-Blackwell Theorem, because we know $\mathrm{E}_\theta(T) = n\theta$ and so $\mathrm{E}_\theta(T/n) = \theta$, so the desired estimator is $T/n = \overline{X}$. But just to see how the theorem works, let us apply it to this problem. If we take $U = X_1$, which is unbiased because $\mathrm{E}(X_1) = \theta$, then according to the theorem, $W = \mathrm{E}(X_1|T)$ is a function of T that is unbiased for θ. It remains to calculate $\mathrm{E}(X_1|T)$. (This is called "Rao-Blackwellizing" X_1.)

The only possible values of the random variable X_1 are 0 and 1, whether its distribution is unconditional or conditional. In particular, its conditional distribution is Bernoulli, and so $\mathrm{E}(X_1|T) = \mathrm{P}\{X_1 = 1|T\}$. Now we calculate

$$\begin{aligned}\mathrm{P}\{X_1 = 1|T = t\} &= \frac{\mathrm{P}\{X_1{=}1 \text{ and } T{=}t\}}{\mathrm{P}\{T{=}t\}} \\ &= \frac{\mathrm{P}\{X_1{=}1 \text{ and } \sum_{i=2}^n X_i{=}t{-}1\}}{\mathrm{P}\{T{=}t\}} \\ &= \frac{\mathrm{P}\{X_1{=}1\}\mathrm{P}\{\sum_{i=2}^n X_i{=}t{-}1\}}{\mathrm{P}\{T{=}t\}}.\end{aligned}$$

We know $X_1 \sim$ Bernoulli(θ), $T \sim$ Binomial(n, θ), and $\sum_{i=2}^n X_i \sim$ Binomal($n-1, \theta$), so

$$\begin{aligned}\frac{\mathrm{P}\{X_1{=}1\}\mathrm{P}\{\sum_{i=2}^n X_i{=}t{-}1\}}{\mathrm{P}\{T{=}t\}} &= \frac{\theta\binom{n-1}{t-1}\theta^{t-1}(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\ &= \frac{\binom{n-1}{t-1}}{\binom{n}{t}} = \frac{\frac{(n-1)!}{(t-1)!(n-t)!}}{\frac{n!}{t!(n-t)!}} = \frac{t}{n}.\end{aligned}$$

Thus we find that $\mathrm{P}\{X_1 = 1|T\} = T/n = \overline{X}$. $\triangle$

(b) Suppose we want to estimate θ^3 unbiasedly. Let us find an unbiased estimator of θ^3 that is a function of the sufficient statistic $T = \sum X_i$. It is not obvious what function of T would be suitable, and so the Rao-Blackwell Theorem will be useful here. All we need to do is find any unbiased estimator of θ^3 and then we can Rao-Blackwellize it.

Take $U = X_1X_2X_3$. Since the X_i's are independent, $\mathrm{E}(X_1X_2X_3) = \mathrm{E}(X_1)\mathrm{E}(X_2)\mathrm{E}(X_3) = \theta\theta\theta = \theta^3$. The only possible values of $U = X_1X_2X_3$ are 0 and 1, so its distribution, whether conditional or unconditional, is Bernoulli, which implies that $\mathrm{E}(X_1X_2X_3|T) = \mathrm{P}\{X_1X_2X_3 = 1|T\}$. Now we calculate

$$\mathrm{P}\{X_1X_2X_3 = 1|T = t\} = \frac{\mathrm{P}\{X_1X_2X_3{=}1 \text{ and } T{=}t\}}{\mathrm{P}\{T{=}t\}}$$

$$= \frac{\mathrm{P}\{X_1{=}1, X_2{=}1, X_3{=}1 \text{ and } \sum_{i=4}^n X_i{=}t{-}3\}}{\mathrm{P}\{T{=}t\}}$$

$$= \frac{\mathrm{P}\{X_1{=}1\}\mathrm{P}\{X_2{=}1\}\mathrm{P}\{X_3{=}1\}\mathrm{P}\{\sum_{i=4}^n X_i{=}t{-}3\}}{\mathrm{P}\{T{=}t\}} .$$

We know $X_i \sim \text{Bernoulli}(\theta)$, $T \sim \text{Binomial}(n, \theta)$, and $\sum_{i=4}^n X_i \sim \text{Binomal}(n-3, \theta)$, so

$$\frac{\mathrm{P}\{X_1{=}1\}\mathrm{P}\{X_2{=}1\}\mathrm{P}\{X_3{=}1\}\mathrm{P}\{\sum_{i=4}^n X_i{=}t{-}3\}}{\mathrm{P}\{T{=}t\}}$$

$$= \frac{\theta^3 \binom{n-3}{t-3} \theta^{t-3}(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}}$$

$$= \frac{\binom{n-3}{t-3}}{\binom{n}{t}} = \frac{\frac{(n-3)!}{(t-3)!(n-t)!}}{\frac{n!}{t!(n-t)!}} = \frac{t(t-1)(t-2)}{n(n-1)(n-2)} .$$

Thus we obtain $\dfrac{T(T-1)(T-2)}{n(n-1)(n-2)}$ as an unbiased estimator of θ^3 that is a function of the sufficient statistic $T = \sum X_i$. $\triangle$

Example. Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. Suppose we want to estimate $e^{-\lambda}$ unbiasedly. This function of λ is of interest because $\mathrm{P}\{X_1 = 0\} = e^{-\lambda}$. The statistic $T = \sum X_i$ is sufficient (see p. 35 above), and so it seems like a good idea to try to find an estimator that is a function of T. The MLE of λ is $\widehat{\lambda} = \overline{X} = T/n$, so a reasonable estimator of $e^{-\lambda}$ is $e^{-\overline{X}}$, which is a function of T, but it is not unbiased. To find a function of T that is unbiased for $e^{-\lambda}$, we can Rao-Blackwellize.

First we need an unbiased estimator of $e^{-\lambda}$. Since $e^{-\lambda}$ can be expressed as the probability $\mathrm{P}\{X_1 = 0\}$, we can use the following fact. Suppose X is a random variable and A is an event. Then $Y = I\{X \in A\}$ is a Bernoulli random variable and $\mathrm{E}(Y) = \mathrm{P}\{Y = 1\} = \mathrm{P}\{X \in A\}$. Therefore, letting $U = I\{X_1 = 0\}$, we have $\mathrm{E}(U) = \mathrm{P}\{X_1 = 0\} = e^{-\lambda}$.

The desired estimator is $W = \mathrm{E}(U|T) = \mathrm{P}\{X_1 = 0|T\}$. We can calculate

$$\mathrm{P}\{X_1 = 0|T = t\} = \frac{\mathrm{P}\{X_1{=}0 \text{ and } T{=}t\}}{\mathrm{P}\{T{=}t\}}$$

$$= \frac{\mathrm{P}\{X_1{=}0 \text{ and } \sum_{i=2}^n X_i{=}t\}}{\mathrm{P}\{T{=}t\}}$$

$$= \frac{\mathrm{P}\{X_1{=}0\}\mathrm{P}\{\sum_{i=2}^n X_i{=}t\}}{\mathrm{P}\{T{=}t\}} .$$

We know $X_1 \sim \text{Poisson}(\lambda)$, $T \sim \text{Poisson}(n\lambda)$ (see Exercise 4.2.2), and $\sum_{i=2}^n X_i \sim \text{Poisson}((n-1)\lambda)$, so

$$\frac{\mathrm{P}\{X_1{=}0\}\mathrm{P}\{\sum_{i=2}^n X_i{=}t\}}{\mathrm{P}\{T{=}t\}} = \frac{e^{-\lambda}e^{-(n-1)\lambda}[(n-1)\lambda]^t/t!}{e^{-n\lambda}(n\lambda)^t/t!}$$

$$= \frac{(n-1)^t}{n^t} = (1-\tfrac{1}{n})^t .$$

Thus we obtain $W = (1-\frac{1}{n})^T$ as an unbiased estimator of $e^{-\lambda}$ that is a function of the sufficient statistic $T = \sum X_i$.

It is interesting to compare this estimator with the MLE $e^{-\overline{X}}$. They are very similar for large sample sizes, because $e^{-\overline{X}} = (e^{-1})^{\overline{X}}$, $W = (1-\frac{1}{n})^{n\overline{X}} = \left[(1-\frac{1}{n})^n\right]^{\overline{X}}$, and $\lim_{n\to\infty}(1-\frac{1}{n})^n = e^{-1}$. (Recall from calculus that $\lim_{n\to\infty}(1+\frac{x}{n})^n = e^x$.) $\triangle$

Next we look at a sample from a continuous distribution.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d Normal$(\mu, 1)$ and suppose we want to estimate $\mathrm{P}\{X_1 \le c\} = \Phi(c-\mu)$ unbiasedly, where c is a fixed known real number. We know $\overline{X}$ is a sufficient statistic. Let us find an unbiased estimator of $\Phi(c-\mu)$ that is a function of $\overline{X}$. Since $\Phi(c-\mu)$ is expressed as a probability, we can proceed as in the preceding example. Let $U = I\{X_1 \le c\}$. The desired estimator is obtained by Rao-Blackwellizing U to get $W = \mathrm{E}(U|\overline{X}) = \mathrm{P}\{X_1 \le c|\overline{X}\}$.

To calculate this conditional probability, we can use section 3.6 on the Bivariate Normal distribution. A useful lemma is:

Lemma. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ, σ^2). Let $V = a_1X_1 + \cdots + a_nX_n$ and $W = b_1X_1 + \cdots + b_nX_n$. Then (a) V has a Normal distribution, and (b) (V, W) has a Bivariate Normal distribution.

Proof. Part (a) can be shown using mgf's as in section 4.3. The mgf of each X_i is (see (2.3.16)) $M_{X_i}(t) = \exp(t\mu + \frac{1}{2}t^2\sigma^2)$. The mgf of V is $M_V(t) = \mathrm{E}[\exp(tV)] =$ $\mathrm{E}[\exp(t\sum a_iX_i)] = \mathrm{E}[\exp(\sum ta_iX_i)] = \mathrm{E}[\prod\exp(ta_iX_i)] =$ (by independence) $\prod\mathrm{E}[\exp(ta_iX_i)] = \prod M_{X_i}(ta_i) = \prod\exp[(ta_i)\mu + \frac{1}{2}(ta_i)^2\sigma^2] =$ $\exp(\sum[(ta_i)\mu + \frac{1}{2}(ta_i)^2\sigma^2]) = \exp[t(\sum a_i)\mu + \frac{1}{2}t^2(\sum a_i^2)\sigma^2]$, which is the mgf of the Normal$((\sum a_i)\mu, (\sum a_i^2)\sigma^2)$ distribution. For part (b), two ways to show that (V, W) has a Bivariate Normal distribution are (1) to obtain the joint pdf $f(v, w)$ and show it has the form in (3.6.1), or (2) to show that every linear combination $Y = cV + dW$ has a Normal distribution and apply Definition 4.6.1. For this lemma, approach (2) is easier. Write $Y = c(a_1X_1 + \cdots + a_nX_n) + d(b_1X_1 + \cdots + b_nX_n) =$ $(ca_1 + db_1)X_1 + \cdots + (ca_n + db_n)X_n$. This is a linear combination of the X_i's and so part (a) implies that it has a Normal distribution. $\square$

Now we return to the example. Since X_1 and $\overline{X}$ are both linear combinations of the X_i's, the lemma implies that $(X_1, \overline{X})$ has a Bivariate Normal distribution. Using the notation of section 3.6, the parameters are $\mu_1 = \mathrm{E}(X_1) = \mu$, $\mu_2 = \mathrm{E}(\overline{X}) = \mu$, $\sigma_1 = \mathrm{SD}(X_1) =$,

$\sigma_2 = \text{SD}(\overline{X}) = 1/\sqrt{n}$, and ρ is the correlation coefficient between X_1 and $\overline{X}$, that is, $\rho = \text{Cov}(X_1, \overline{X})/\sigma_1\sigma_2 = (1/n)/(1/\sqrt{n}) = 1/\sqrt{n}$. The conditional distribution of X_1 given $\overline{X} = y$ is shown in Theorem 3.6.1(ii) to be Normal with mean $\mu_1 + (\rho\sigma_1/\sigma_2)(y - \mu_2)$ $= y$ and variance $\sigma_1^2(1 - \rho^2) = 1 - 1/n$. Therefore, $\text{P}\{X_1 \le c|\overline{X} = y\} =$ $\Phi((c - y)/\sqrt{1 - 1/n})$. Thus, an unbiased estimator of $\text{P}\{X_1 \le c\} = \Phi(c - \mu)$ that is a function of $\overline{X}$ is given by $\Phi\Big(\frac{c-\overline{X}}{\sqrt{1-\frac{1}{n}}}\Big)$. $\triangle$

The Cramér-Rao lower bound

An approach that sometimes works for finding a UMVUE is to use the Cramér-Rao lower bound. This bound requires some "regularity conditions" on the joint pmf or pdf of the data vector. We will suppose θ is real-valued and will assume the same three regularity conditions that were assumed for the definition of Fisher information (see p. 21). For easy reference, the conditions are:

(RC1) $f(\boldsymbol{x}; \theta)$ has the same support for all θ.

(RC2) $f(\boldsymbol{x}; \theta)$ is differentiable with respect to θ.

(RC3) For all statistics $W(\boldsymbol{X})$ whose expectation $\text{E}_\theta(W)$ exists, the expectation is a differentiable function of θ and the derivative can be calculated by differentiating under the summation or integral sign.

Recall that these conditions are satisfied in a one-parameter exponential family if the functions $b_j(\theta)$ are differentiable.

Theorem 7.5.1 (Cramér-Rao Lower Bound). Let $\boldsymbol{X}$ be a data vector with joint pmf or pdf $f(\boldsymbol{x}; \theta)$ parameterized by a real-valued parameter θ. Suppose conditions RC1, RC2, and RC3 are satisfied. If $U(\boldsymbol{X})$ is unbiased for $\tau(\theta)$, then

$$\text{Var}_\theta(U) \ge \frac{\left[\frac{\text{d}}{\text{d}\theta}\tau(\theta)\right]^2}{\mathcal{I}_{\boldsymbol{X}}(\theta)} \quad \text{for all } \theta.$$

The Cramér-Rao Lower Bound (CRLB) can be used to find a UMVUE of $\tau(\theta)$, because if an unbiased estimator has a variance that achieves the CRLB, then of course it has the smallest possible variance. Note that the theorem does not assume an i.i.d. sample. It can be applied to regression data or time series data, provided, of course, that the regularity conditions are satisfied.

Proof of the CRLB. For any two random variables Y and Z, the Covariance Inequality (Theorem 3.9.6) says that $[\text{Cov}(Y, Z)]^2 \le \text{Var}(Y)\text{Var}(Z)$, which can be rewritten as

$$\mathrm{Var}(Y) \geq \frac{[\mathrm{Cov}(Y,Z)]^2}{\mathrm{Var}(Z)} .$$

This yields the CRLB by taking $Y = U(\boldsymbol{X})$ and $Z = \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)$. By the definition of information (p. 21 above), $\mathrm{Var}(Z) = \mathcal{I}_{\boldsymbol{X}}(\theta)$. It remains to show that

$$\mathrm{Cov}_\theta\Big[U(\boldsymbol{X}), \tfrac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = \tfrac{\mathrm{d}}{\mathrm{d}\theta}\tau(\theta) .$$

Since $\mathrm{E}_\theta\Big[\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = 0$ (by part (b) of the lemma on p. 22), so $\mathrm{Cov}_\theta\Big[U(\boldsymbol{X}), \frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = \mathrm{E}_\theta\Big[U(\boldsymbol{X})\frac{\partial}{\partial\theta}\log f(\boldsymbol{X};\theta)\Big] = \frac{\partial}{\partial\theta}\mathrm{E}_\theta[U(\boldsymbol{X})]$ (by part (a) of the lemma on p. 22). Since $U(\boldsymbol{X})$ is unbiased, $\mathrm{E}_\theta[U(\boldsymbol{X})] = \tau(\theta)$. □

Example. Let $X_1, \ldots, X_n$ be i.i.d. Poisson(λ), $\lambda > 0$. This family of distributions is a regular one-parameter exponential family and so conditions RC1, RC2, and RC3 hold. The pmf of a single observation is $f(x;\lambda) = e^{-\lambda}\lambda^x/x!$, so

$$\log f(x;\lambda) = -\lambda + x\log\lambda - \log x!$$

$$\frac{\partial}{\partial\lambda}\log f(x;\lambda) = -1 + \frac{x}{\lambda} \quad \text{and} \quad \frac{\partial^2}{\partial\lambda^2}\log f(x;\lambda) = -\frac{x}{\lambda^2} .$$

By the lemma on p. 23, the information in X_1 is $\mathcal{I}_{X_1}(\theta) = -\mathrm{E}_\lambda\Big(-\frac{X_1}{\lambda^2}\Big) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$.

By Theorem 6.4.1, the information in $\boldsymbol{X}$ is $\mathcal{I}_{\boldsymbol{X}}(\theta) = n\mathcal{I}_{X_1}(\theta) = \frac{n}{\lambda}$.

(a) We can now calculate the CRLB for the variance of any unbiased estimator of λ. Since $\frac{\mathrm{d}}{\mathrm{d}\lambda}\lambda = 1$, Theorem 7.5.1 implies that the CRLB is $1/(\frac{n}{\lambda}) = \frac{\lambda}{n}$. For the Poisson distribution, $\mathrm{E}(X_1) = \lambda$ and $\mathrm{Var}(X_1) = \lambda$, and so $\mathrm{E}(\overline{X}) = \lambda$ and $\mathrm{Var}(\overline{X}) = \frac{\lambda}{n}$. Thus we see that $\overline{X}$ is an unbiased estimator of λ and it achieves the CRLB. It must therefore be a UMVUE of λ.

(b) Let us calculate the CRLB for the variance of any unbiased estimator of $e^{-\lambda}$. Since $\frac{\mathrm{d}}{\mathrm{d}\lambda}e^{-\lambda} = -e^{-\lambda}$, the CRLB is $(-e^{-\lambda})^2/(\frac{n}{\lambda}) = \frac{\lambda}{n}e^{-2\lambda}$. On p. 56 above we found an unbiased estimator of $e^{-\lambda}$ that is a function of the minimal sufficient statistic $T = \sum X_i$, namely the estimator $W = (1 - \frac{1}{n})^T$. We suspect that W is a good estimator since it is a function of a minimal sufficient statistic, and so it seems quite possible that $\mathrm{Var}(W)$ achieves the CRLB. We can calculate $\mathrm{Var}(W) = \mathrm{E}(W^2) - (\mathrm{E}(W))^2 = \mathrm{E}(W^2) - e^{-2\lambda}$ and

$$\begin{aligned}\mathrm{E}(W^2) &= \mathrm{E}\Big[(1-\tfrac{1}{n})^{2T}\Big] = \sum_{t=0}^{\infty}\Big[(1-\tfrac{1}{n})^{2t}\Big]\frac{e^{-n\lambda}(n\lambda)^t}{t!} \\ &= e^{-n\lambda}\sum_{t=0}^{\infty}\Big[(1-\tfrac{1}{n})^2(n\lambda)\Big]^t/t! \\ &= e^{-n\lambda}e^{(1-\frac{1}{n})^2(n\lambda)} = e^{-2\lambda+\frac{\lambda}{n}} .\end{aligned}$$

Hence

$$\text{Var}(W) \;=\; e^{-2\lambda+\frac{\lambda}{n}} - e^{-2\lambda} \;=\; e^{-2\lambda}(e^{\frac{\lambda}{n}} - 1)\,.$$

This is greater than the CRLB, because $e^{\frac{\lambda}{n}} - 1 > \frac{\lambda}{n}$, because $e^u - 1 > u$ for all $u > 0$, because $e^u = 1 + u + \frac{1}{2!}u^2 + \frac{1}{3!}u^3 + \cdots > 1 + u$ when $u > 0$. So W does not achieve the CRLB. Nevertheless, as we will see later, W is the UMVUE of $e^{-\lambda}$. If an unbiased estimator achieves the CRLB, then it is a UMVUE. The converse is not true. That is, a UMVUE does not necessarily achieve the CRLB. $\triangle$

Next we consider an example in which the data are i.i.d.

Example. Suppose $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \text{Normal}(\alpha + \beta w_i\,, \sigma^2)$, where w_i is a known covariate. This is a simple regression model. Suppose we want to estimate β. The Cramér-Rao lower bound can be generalized to families of distributions with vector-valued parameters, but to stay within the scope of Theorem 7.5.1, let us suppose that α is known to be 0 and σ^2 is known to be σ_0^2. That is, $Y_i \sim \text{Normal}(\beta w_i\,, \sigma_0^2)$. The joint pdf of the data vector $\boldsymbol{Y}$ is

$$\begin{aligned} f(\boldsymbol{y};\beta) \;&=\; \prod_{i=1}^n f_i(y_i;\beta) \;=\; \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\Big[-\tfrac{1}{2\sigma_0^2}(y_i - \beta w_i)^2\Big] \\ &=\; \Big(\frac{1}{\sqrt{2\pi\sigma_0^2}}\Big)^n \exp\Big[-\tfrac{1}{2\sigma_0^2}\sum_{i=1}^n (y_i - \beta w_i)^2\Big] \\ &=\; \Big(\frac{1}{\sqrt{2\pi\sigma_0^2}}\Big)^n \exp\Big[-\tfrac{1}{2\sigma_0^2}\Big(\textstyle\sum y_i^2 - 2\beta\sum w_i y_i + \beta^2\sum w_i^2\Big)\Big] \\ &=\; a(\beta)h(\boldsymbol{y})\exp\big[b(\beta)R(\boldsymbol{y})\big] \end{aligned}$$

where $a(\beta) = \Big(\frac{1}{\sqrt{2\pi\sigma_0^2}}\Big)^n \exp\Big(-\frac{1}{2\sigma_0^2}\beta^2\sum w_i^2\Big)$, $h(\boldsymbol{y}) = \exp\Big(-\frac{1}{2\sigma_0^2}\sum y_i^2\Big)$, $b(\beta) = \frac{\beta}{\sigma_0^2}$, and $R(\boldsymbol{y}) = \sum w_i y_i$. This is a regular one-parameter exponential family, so conditions RC1, RC2, and RC3 are met. We also note that $\sum w_i Y_i$ is a minimal sufficient statistic.

$$\begin{aligned} \log f(\boldsymbol{y};\beta) \;&=\; -n\log\sqrt{2\pi\sigma_0^2} - \tfrac{1}{2\sigma_0^2}\Big(\textstyle\sum y_i^2 - 2\beta\sum w_i y_i + \beta^2\sum w_i^2\Big) \\ \frac{\partial}{\partial\beta}\log f(\boldsymbol{y};\beta) \;&=\; \tfrac{1}{\sigma_0^2}\textstyle\sum w_i y_i - \frac{\beta}{\sigma_0^2}\sum w_i^2 \\ \frac{\partial^2}{\partial\beta^2}\log f(\boldsymbol{y};\beta) \;&=\; -\tfrac{1}{\sigma_0^2}\textstyle\sum w_i^2\,. \end{aligned}$$

Now

$$\mathcal{I}_X(\beta) \;=\; -\,\mathrm{E}_\beta\left[\frac{\partial^2}{\partial\beta^2}\log f(\boldsymbol{y};\beta)\right] \;=\; \frac{1}{\sigma_0^2}\sum w_i^2\,.$$

The CRLB for the variance of unbiased estimators of β is $1/\mathcal{I}_X(\beta) = \sigma_0^2/\sum w_i^2$. Perhaps if we find an unbiased estimator of β that is a function of the minimal sufficient statistic $\sum w_i Y_i$, maybe it will achieve the CRLB and can thus be shown to be a UMVUE.

$$\mathrm{E}(\textstyle\sum w_i Y_i) \;=\; \sum w_i \mathrm{E}(Y_i) \;=\; \sum w_i(\beta w_i) \;=\; \beta\sum w_i^2\,.$$

Therefore, $W = \sum w_i Y_i/\sum w_i^2$ is an unbiased estimator of β that is a function of the minimal sufficient statistic.

$$\begin{aligned}\mathrm{Var}(W) \;&=\; \mathrm{Var}\Big(\frac{\sum w_i Y_i}{\sum w_i^2}\Big) \;=\; \frac{1}{(\sum w_i^2)^2}\mathrm{Var}(\textstyle\sum w_i Y_i)\\ &=\; \frac{1}{(\sum w_i^2)^2}\sum w_i^2\mathrm{Var}(Y_i) \;=\; \frac{1}{\sum w_i^2}\sigma_0^2 \;=\; \mathrm{CRLB}.\end{aligned}$$

We can conclude that $\sum w_i Y_i/\sum w_i^2$ is a UMVUE of β. $\triangle$

The Lehmann-Scheffé Theorem

In examples in which the CRLB can be used to find a UMVUE, it is often easier to find the UMVUE by using the theorem of Lehmann and Scheffé that will be presented next. Moreover, the theorem works for some examples in which the CRLB is not attained. The theorem is stated for models with vector-valued parameters.

Theorem 7.5.3 (Lehmann-Scheffé Theorem). Suppose $\boldsymbol{T}(\boldsymbol{X})$ is a complete sufficient statistic. Suppose $W(\boldsymbol{T})$ is an estimator that is a function of $\boldsymbol{T}$ and is unbiased for $\tau(\boldsymbol{\theta})$. Then $W(\boldsymbol{T})$ is the UMVUE of $\tau(\boldsymbol{\theta})$.

Proof. Let $U(\boldsymbol{X})$ be any unbiased estimator of $\tau(\boldsymbol{\theta})$. We must show that $\mathrm{Var}_{\boldsymbol{\theta}}[W(\boldsymbol{T})] \le \mathrm{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{X})]$ for all $\boldsymbol{\theta}$. By the Rao-Blackwell Theorem, the estimator $U^*(\boldsymbol{T}) = \mathrm{E}(U|\boldsymbol{T})$ is unbiased for $\tau(\boldsymbol{\theta})$ and satisfies $\mathrm{Var}_{\boldsymbol{\theta}}[U^*(\boldsymbol{T})] \le \mathrm{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{X})]$ for all $\boldsymbol{\theta}$. Since $U^*(\boldsymbol{T})$ and $W(\boldsymbol{T})$ both have the same expectation, then $\mathrm{E}_{\boldsymbol{\theta}}[U^*(\boldsymbol{T}) - W(\boldsymbol{T})] = 0$ for all $\boldsymbol{\theta}$. By the definition of completeness, $U^*(\boldsymbol{T}) - W(\boldsymbol{T}) = 0$ (with probability 1); that is, $U^*(\boldsymbol{T}) = W(\boldsymbol{T})$ (with probability 1). Therefore, $\mathrm{Var}_{\boldsymbol{\theta}}[W(\boldsymbol{T})] = \mathrm{Var}_{\boldsymbol{\theta}}[U^*(\boldsymbol{T})]$ $\le \mathrm{Var}_{\boldsymbol{\theta}}[U(\boldsymbol{X})]$ for all $\boldsymbol{\theta}$. $\square$

From the proof of the theorem, we see that the UMVUE is unique (with probability 1). This is true even when no complete sufficient statistic exists.

Corollary. Suppose $\boldsymbol{T}(\boldsymbol{X})$ is a complete sufficient statistic. If $U(\boldsymbol{X})$ is unbiased for $\tau(\boldsymbol{\theta})$, then $\mathrm{E}(U|\boldsymbol{T})$ is the UMVUE of $\tau(\boldsymbol{\theta})$.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(λ), $\lambda > 0$. A complete sufficient statistic is $T = \sum X_i$ (see p. 35 above or use Theorem 6.6.2).
(a) The sample mean $\overline{X}$ is unbiased for the population mean λ. Since $\overline{X} = T/n$ is a function of T, the Lehmann-Scheffé Theorem implies that it is the UMVUE of λ.
(b) On pp. 56-57 above we used the Rao-Blackwell Theorem to obtain $(1 - \frac{1}{n})^T$ as an unbiased estimator of $e^{-\lambda}$. Since it is a function of T, it is the UMVUE of $e^{-\lambda}$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Bernoulli(θ), $0 < \theta < 1$. A complete sufficient statistic is $T = \sum X_i$ (use Theorem 6.6.2).
(a) The sample mean $\overline{X}$ is unbiased for the population mean θ. Since $\overline{X} = T/n$ is a function of T, the Lehmann-Scheffé Theorem implies that it is the UMVUE of θ.
(b) On pp. 55-56 above we used the Rao-Blackwell Theorem to obtain $\frac{T(T-1)(T-2)}{n(n-1)(n-2)}$ as an unbiased estimator of θ^3. Since it is a function of T, it is the UMVUE of θ^3. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(μ,σ^2), $-\infty < \mu < \infty$, $\sigma^2 > 0$.
A complete sufficient statistic is $\boldsymbol{T} = (\overline{X}, S^2)$ (see p. 37 above).
(a) Since $\overline{X}$ is unbiased for μ and is a function of $\boldsymbol{T}$, it is the UMVUE of μ.
(b) Since S^2 is unbiased for σ^2 and is a function of $\boldsymbol{T}$, it is the UMVUE of σ^2.
(c) Similarly, aS, where $a = \sqrt{\frac{1}{2}(n-1)}\,\Gamma(\frac{1}{2}(n-1))/\Gamma(\frac{1}{2}n)$ is the UMVUE of σ. $\triangle$

Example. Suppose $Y_1, \ldots, Y_n$ are i.i.d. Normal(βw_i, σ_0^2), $-\infty < \beta < \infty$. On p. 60 above we saw that this family of distributions is a regular 1-parameter exponential family. By using a theorem on p. 36, one can show that $\sum w_i Y_i$ is a complete sufficient statistic. On p. 61 we found that $\sum w_i Y_i / \sum w_i^2$ is an unbiased estimator of β. Now the Lehmann-Scheffé Theorem implies that it is the UMVUE of β. $\triangle$

Unbiased estimation when no complete sufficient statistic exists

In the examples of unbiased estimators presented from p. 52 to this page, there has been a complete sufficient statistic. Now we consider an example in which no complete sufficient statistic exists.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Normal(θ,θ^2), $\theta > 0$. A minimal sufficient statistic is $(\overline{X}, S^2)$ but it is not complete (see Exercise 6.6.3). The estimator $\overline{X}$ is unbiased for θ and is a function of the minimal sufficient statistic. Another unbiased estimator of θ that is a function of the minimal sufficient statistic is aS where $a = \sqrt{\frac{1}{2}(n-1)}\,\Gamma(\frac{1}{2}(n-1))/\Gamma(\frac{1}{2}n)$.
Which estimator is preferable? Their variances are: $\mathrm{Var}(\overline{X}) = \theta^2/n$ and
$\mathrm{Var}(aS) = (a^2 - 1)\theta^2$. One can show that $\mathrm{Var}(\overline{X}) < \mathrm{Var}(aS)$ for $n = 2$, and

$\text{Var}(aS) < \text{Var}(\overline{X})$ for $n \geq 3$. For large n, it can be shown that $a^2 - 1 \approx 1/(2n)$, so that $\text{Var}(aS) \approx \frac{1}{2}\text{Var}(\overline{X})$. So aS is preferable for most sample sizes. An even better estimator can be obtained by using the idea of a BLUE. Among all unbiased linear combinations of $\overline{X}$ and aS, let us find the one with the smallest variance. Let $U_1 = \overline{X}$, $U_2 = aS$, and $U = c_1U_1 + c_2U_2$. For U to be unbiased, we want $\text{E}(U) = c_1\text{E}(U_1) + c_2\text{E}(U_2) = c_1\theta + c_2\theta = (c_1 + c_2)\theta = \theta$; that is, we need $c_1 + c_2 = 1$. So consider $U = c\overline{X} + (1 - c)aS$. Since $\overline{X}$ and S are independent, $\text{Var}(U) = c^2\text{Var}(\overline{X}) + (1 - c)^2\text{Var}(aS) = c^2\theta^2/n + (1 - c)^2(a^2 - 1)\theta^2 = [c^2/n + (1 - c)^2(a^2 - 1)]\theta^2 = g(c)\theta^2$. We want to minimize $g(c)$. Its derivative is $g'(c) = 2c/n - 2(1 - c)(a^2 - 1) = 2(1/n + a^2 - 1)c - 2(a^2 - 1)$, which is 0 when $c = c^* = (a^2 - 1)/(1/n + a^2 - 1)$. This is a minimum because, as $c \to \pm\infty$, $\text{Var}(U) \to \infty$. An unbiased estimator that has smaller variance than either $\overline{X}$ or aS is $c^*\overline{X} + (1 - c^*)aS$. For large n, this is approximately $\frac{1}{3}\overline{X} + \frac{2}{3}aS$. Its variance is approximately $\theta^2/(3n)$. $\triangle$

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Uniform$(\theta, \theta + 1)$, $-\infty < \theta < \infty$. A minimal sufficient statistic is $(X_{(1)}, X_{(n)})$ (use Theorem 6.3.1) but it is not complete (because $\text{E}[X_{(n)} - X_{(1)}] \neq 0$). There are several unbiased estimators of θ that are functions of the minimal sufficient statistic. The pdf of $W = X_{(1)}$ is $h(w) = n(\theta + 1 - w)^{n-1}$ for $\theta < w < \theta + 1$ (4.2.6), and the pdf of $Y = X_{(n)}$ is $g(y) = n(y - \theta)^{n-1}$ for $\theta < y < \theta + 1$ (4.2.4). Hence, $\text{E}(W) = \int_\theta^{\theta+1} wh(w)\text{d}w = \theta + 1/(n + 1)$, and $\text{E}(Y) = \int_\theta^{\theta+1} yg(y)\text{d}y = \theta + n/(n + 1)$. Therefore, $T_1 = X_{(1)} - 1/(n+1)$ and $T_2 = X_{(n)} - n/(n + 1)$ are both unbiased for θ. Since a Uniform distribution is symmetric, $\text{Var}[X_{(1)}] = \text{Var}[X_{(n)}]$, and so $\text{Var}(T_1) = \text{Var}(T_2) = \text{Var}(Y - \theta) = \text{E}[(Y - \theta)^2] - [n/(n + 1)]^2 = n/[(n + 1)^2(n + 2)]$. Using the idea of a BLUE as in the preceding example, let us see if we can find a linear combination of the two estimators that has smaller variance. Let $T = c_1T_1 + c_2T_2$. For T to be unbiased, we need $c_1 + c_2 = 1$, and so we can write $T = cT_1 + (1 - c)T_2$. Its variance is $\text{Var}(T) = c^2\text{Var}(T_1) + (1 - c)^2\text{Var}(T_2) + 2c(1 - c)\text{Cov}(T_1, T_2)$. The derivative of the variance with respect to c is $2c\text{Var}(T_1) - 2(1 - c)\text{Var}(T_2) + 2(1 - 2c)\text{Cov}(T_1, T_2)$, which is 0 when $c = c^* = [\text{Var}(T_2) - \text{Cov}(T_1, T_2)]/[\text{Var}(T_1) + \text{Var}(T_2) - 2\text{Cov}(T_1, T_2)] = [\text{Var}(T_1) - \text{Cov}(T_1, T_2)]/[2\text{Var}(T_1) - 2\text{Cov}(T_1, T_2)] = \frac{1}{2}$. This is a minimum because, as $c \to \pm\infty$, $\text{Var}(T) \to \infty$. An unbiased estimator of θ with smaller variance than either T_1 or T_2 is $T^* = \frac{1}{2}(T_1 + T_2) = M - \frac{1}{2}$ where $M = \frac{1}{2}[X_{(1)} + X_{(n)}]$ is the midrange. $\triangle$