STATISTICAL INFERENCE

MScSTAT 102(N)/ MASTAT 102(N)

BLOCK 1: ESTIMATION THEORY

UNIT 1: POINT AND INTERVAL ESTIMATION

Definition 5.1.1 The random variables $X_1, \ldots, X_n$ are called a *random sample of size n from the population* $f(x)$ if $X_1, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, \ldots, X_n$ are called *independent and identically distributed random variables with pdf or pmf* $f(x)$. This is commonly abbreviated to iid random variables.

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots\cdots f(x_n) = \prod_{i=1}^{n} f(x_i). \tag{5.1.1}$$

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta), \tag{5.1.2}$$

where the same parameter value θ is used in each of the terms in the product. If, in a statistical setting, we assume that the population we are observing is a member of a specified parametric family but the true parameter value is unknown, then a random sample from this population has a joint pdf or pmf of the above form with the value of θ unknown. By considering different possible values of θ, we can study how a random sample would behave for different populations.

Example 5.1.2 (Sample pdf–exponential) Let $X_1, \ldots, X_n$ be a random sample from an exponential(β) population. Specifically, $X_1, \ldots, X_n$ might correspond to the times until failure (measured in years) for n identical circuit boards that are put on test and used until they fail. The joint pdf of the sample is

$$f(x_1, \ldots, x_n|\beta) = \prod_{i=1}^{n} f(x_i|\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}.$$

This pdf can be used to answer questions about the sample. For example, what is the probability that all the boards last more than 2 years? We can compute

$$\begin{aligned}
&P(X_1 > 2, \ldots, X_n > 2) \\
&\quad = \int_2^\infty \cdots \int_2^\infty \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n \\
&\quad = e^{-2/\beta} \int_2^\infty \cdots \int_2^\infty \prod_{i=2}^{n} \frac{1}{\beta} e^{-x_i/\beta} \, dx_2 \cdots dx_n \quad \text{(integrate out } x_1\text{)} \\
&\quad \vdots \qquad\qquad \text{(integrate out the remaining } x_i\text{s successively)} \\
&\quad = (e^{-2/\beta})^n \\
&\quad = e^{-2n/\beta}.
\end{aligned}$$

If β, the average lifelength of a circuit board, is large relative to n, we see that this probability is near 1.

$$
\begin{aligned}
&P(X_1 > 2, \ldots, X_n > 2) \\
&\quad = P(X_1 > 2) \cdots P(X_n > 2) && \text{(independence)} \\
&\quad = [P(X_1 > 2)]^n && \text{(identical distributions)} \\
&\quad = (e^{-2/\beta})^n && \text{(exponential calculation)} \\
&\quad = e^{-2n/\beta}. && \|
\end{aligned}
$$

The random sampling model in Definition 5.1.1 is sometimes called sampling from an *infinite* population. Think of obtaining the values of $X_1, \ldots, X_n$ sequentially. First, the experiment is performed and $X_1 = x_1$ is observed. Then, the experiment is repeated and $X_2 = x_2$ is observed. The assumption of independence in random sampling implies that the probability distribution for X_2 is unaffected by the fact that $X_1 = x_1$ was observed first. "Removing" x_1 from the infinite population does not change the population, so $X_2 = x_2$ is still a random observation from the same population.

When sampling is from a *finite* population, Definition 5.1.1 may or may not be relevant depending on how the data collection is done. A finite population is a finite set of numbers, $\{x_1, \ldots, x_N\}$. A sample $X_1, \ldots, X_n$ is to be drawn from this population.

Suppose a value is chosen from the population in such a way that each of the N values is equally likely (probability $= 1/N$) to be chosen. (Think of drawing numbers from a hat.) This value is recorded as $X_1 = x_1$. Then the process is repeated. Again, each of the N values is equally likely to be chosen. The second value chosen is recorded as $X_2 = x_2$. (If the same number is chosen, then $x_1 = x_2$.) This process of drawing from the N values is repeated n times, yielding the sample $X_1, \ldots, X_n$. This kind of sampling is called *with replacement* because the value chosen at any stage is "replaced" in the population and is available for choice again at the next stage. For this kind of sampling, the conditions of Definition 5.1.1 are met. Each X_i is a discrete random variable that takes on each of the values $x_1, \ldots, x_N$ with equal probability. The random variables $X_1, \ldots, X_n$ are independent because the process of choosing any X_i is the same, regardless of the values that are chosen for any of the other variables.

A second method for drawing a random sample from a finite population is called sampling *without replacement.* Sampling without replacement is done as follows. A value is chosen from $\{x_1, \ldots, x_N\}$ in such a way that each of the N values has probability $1/N$ of being chosen. This value is recorded as $X_1 = x_1$. Now a second value is chosen from the remaining $N - 1$ values. Each of the $N - 1$ values has probability $1/(N - 1)$ of being chosen. The second chosen value is recorded as $X_2 = x_2$. Choice of the remaining values continues in this way, yielding the sample $X_1, \ldots, X_n$. But once a value is chosen, it is unavailable for choice at any later stage.

A sample drawn from a finite population without replacement does not satisfy all the conditions of Definition 5.1.1. The random variables $X_1, \ldots, X_n$ are not mutually independent. To see this, let x and y be distinct elements of $\{x_1, \ldots, x_N\}$. Then $P(X_2 = y | X_1 = y) = 0$, since the value y cannot be chosen at the second stage if it was already chosen at the first. However, $P(X_2 = y | X_1 = x) = 1/(N - 1)$. The

probability distribution for X_2 depends on the value of X_1 that is observed and, hence, X_1 and X_2 are not independent. However, it is interesting to note that $X_1, \ldots, X_n$ are identically distributed. That is, the marginal distribution of X_i is the same for each $i = 1, \ldots, n$. For X_1 it is clear that the marginal distribution is $P(X_1 = x) = 1/N$ for each $x \in \{x_1, \ldots, x_N\}$. To compute the marginal distribution for X_2, use Theorem 1.2.11(a) and the definition of conditional probability to write

$$P(X_2 = x) = \sum_{i=1}^{N} P(X_2 = x | X_1 = x_i) P(X_1 = x_i).$$

For one value of the index, say k, $x = x_k$ and $P(X_2 = x | X_1 = x_k) = 0$. For all other $j \neq k$, $P(X_2 = x | X_1 = x_j) = 1/(N-1)$. Thus,

$$P(X_2 = x) = (N-1)\left(\frac{1}{N-1}\,\frac{1}{N}\right) = \frac{1}{N}. \tag{5.1.3}$$

Similar arguments can be used to show that each of the X_is has the same marginal distribution.

Sampling without replacement from a finite population is sometimes called *simple random sampling*. It is important to realize that this is not the same sampling situation as that described in Definition 5.1.1. However, if the population size N is large compared to the sample size n, $X_1, \ldots, X_n$ are nearly independent and some approximate probability calculations can be made assuming they are independent. By saying they are "nearly independent" we simply mean that the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ is not too different from the marginal distribution of X_i. For example, the conditional distribution of X_2 given X_1 is

$$P(X_2 = x_1 | X_1 = x_1) = 0 \quad \text{and} \quad P(X_2 = x | X_1 = x_1) = \frac{1}{N-1} \quad \text{for } x \neq x_1.$$

This is not too different from the marginal distribution of X_2 given in (5.1.3) if N is large. The nonzero probabilities in the conditional distribution of X_i given $X_1, \ldots, X_{i-1}$ are $1/(N-i+1)$, which are close to $1/N$ if $i \leq n$ is small compared with N.

When a sample $X_1, \ldots, X_n$ is drawn, some summary of the values is usually computed. Any well-defined summary may be expressed mathematically as a function $T(x_1, \ldots, x_n)$ whose domain includes the sample space of the random vector $(X_1, \ldots, X_n)$. The function T may be real-valued or vector-valued; thus the summary is a random variable (or vector), $Y = T(X_1, \ldots, X_n)$.

This be used to describe the distribution of Y in terms of the distribution of the population from which the sample was obtained. Since the random sample $X_1, \ldots, X_n$ has a simple probabilistic structure (because the X_is are independent and identically distributed), the distribution of Y is particularly tractable. Because this distribution is usually derived from the distribution of the variables in the random sample, it is called the *sampling distribution* of Y. This distinguishes the probability distribution of Y from the distribution of the population, that is, the marginal distribution of each X_i.

Definition 5.2.1 Let $X_1, \ldots, X_n$ be a random sample of size n from a population and let $T(x_1, \ldots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable or random vector $Y = T(X_1, \ldots, X_n)$ is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution of* Y.

The definition of a statistic is very broad, with the only restriction being that a statistic cannot be a function of a parameter. The sample summary given by a statistic can include many types of information. For example, it may give the smallest or largest value in the sample, the average sample value, or a measure of the variability in the sample observations. Three statistics that are often used and provide good summaries of the sample are now defined.

Definition 5.2.2 The *sample mean* is the arithmetic average of the values in a random sample. It is usually denoted by

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 5.2.3 The *sample variance* is the statistic defined by

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

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$.

Theorem 5.2.4 *Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then*
a. $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$,
b. $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Proof: To prove part (a), add and subtract $\bar{x}$ to get

$$\begin{aligned}\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2. \qquad \text{(cross term is 0)}\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$.

To prove part (b), take $a = 0$ in the above. □

Lemma 5.2.5 *Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathrm{E}g(X_1)$ and $\mathrm{Var}\ g(X_1)$ exist. Then*

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{E}g(X_1)\right) \tag{5.2.1}$$

and

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\left(\mathrm{Var}\ g(X_1)\right). \tag{5.2.2}$$

Proof: To prove (5.2.1), note that

$$\mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right) = \sum_{i=1}^{n} \mathrm{E}g(X_i) = n\left(\mathrm{E}g(X_1)\right).$$

Since the X_is are identically distributed, the second equality is true because $\mathrm{E}g(X_i)$ is the same for all i. Note that the independence of $X_1, \ldots, X_n$ is not needed for (5.2.1)

To prove (5.2.2), note that

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) &= \mathrm{E}\left[\sum_{i=1}^{n} g(X_i) - \mathrm{E}\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 && \text{(definition of variance)} \\
&= \mathrm{E}\left[\sum_{i=1}^{n} \left(g(X_i) - \mathrm{E}g(X_i)\right)\right]^2. && \begin{pmatrix}\text{expectation property and} \\ \text{rearrangement of terms}\end{pmatrix}
\end{aligned}$$

In this last expression there are n^2 terms. First, there are n terms $\left(g(X_i) - \mathrm{E}g(X_i)\right)^2$, $i = 1, \ldots, n$, and for each, we have

$$\begin{aligned}
\mathrm{E}\left(g(X_i) - \mathrm{E}g(X_i)\right)^2 &= \mathrm{Var}\ g(X_i) && \text{(definition of variance)} \\
&= \mathrm{Var}\ g(X_1). && \text{(identically distributed)}
\end{aligned}$$

The remaining $n(n-1)$ terms are all of the form $\left(g(X_i) - \mathrm{E}g(X_i)\right)\left(g(X_j) - \mathrm{E}g(X_j)\right)$, with $i \neq j$. For each term,

$$\begin{aligned}
\mathrm{E}\left[\left(g(X_i) - \mathrm{E}g(X_i)\right)\left(g(X_j) - \mathrm{E}g(X_j)\right)\right] &= \mathrm{Cov}\left(g(X_i), g(X_j)\right) && \begin{pmatrix}\text{definition of} \\ \text{covariance}\end{pmatrix} \\
&= 0. && \begin{pmatrix}\text{independence} \\ \text{Theorem 4.5.5}\end{pmatrix}
\end{aligned}$$

Thus, we obtain equation (5.2.2). □

Theorem 5.2.6 *Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Then*

a. $\mathrm{E}\bar{X} = \mu$,

b. $\mathrm{Var}\ \bar{X} = \dfrac{\sigma^2}{n}$,

c. $\mathrm{E}S^2 = \sigma^2$.

Proof: To prove (a), let $g(X_i) = X_i/n$, so $\mathrm{E}g(X_i) = \mu/n$. Then, by Lemma 5.2.5,

$$\mathrm{E}\bar{X} = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}n\mathrm{E}X_1 = \mu.$$

Similarly for (b), we have

$$\mathrm{Var}\,\bar{X} = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}n\,\mathrm{Var}\,X_1 = \frac{\sigma^2}{n}.$$

For the sample variance, using Theorem 5.2.4, we have

$$\begin{aligned}\mathrm{E}S^2 &= \mathrm{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]\right)\\ &= \frac{1}{n-1}\left(n\mathrm{E}X_1^2 - n\mathrm{E}\bar{X}^2\right)\\ &= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2,\end{aligned}$$

establishing part (c) and proving the theorem. □

Theorem 5.2.7 *Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is*

$$M_{\bar{X}}(t) = [M_X(t/n)]^n.$$

Example 5.2.8 (Distribution of the mean) Let $X_1, \ldots, X_n$ be a random sample from a $\mathrm{n}(\mu, \sigma^2)$ population. Then the mgf of the sample mean is

$$\begin{aligned}M_{\bar{X}}(t) &= \left[\exp\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right]^n\\ &= \exp\left(n\left(\mu\frac{t}{n} + \frac{\sigma^2(t/n)^2}{2}\right)\right) = \exp\left(\mu t + \frac{(\sigma^2/n)t^2}{2}\right).\end{aligned}$$

Thus, $\bar{X}$ has a $\mathrm{n}(\mu, \sigma^2/n)$ distribution.

Theorem 5.2.11 *Suppose $X_1, \ldots, X_n$ is a random sample from a pdf or pmf $f(x|\theta)$, where*

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{i=1}^{k} w_i(\theta)t_i(x)\right)$$

is a member of an exponential family. Define statistics $T_1, \ldots, T_k$ by

$$T_i(X_1, \ldots, X_n) = \sum_{j=1}^{n} t_i(X_j), \quad i = 1, \ldots, k.$$

If the set $\{(w_1(\theta), w_2(\theta), \ldots, w_k(\theta)), \theta \in \Theta\}$ contains an open subset of $\Re^k$, then the distribution of $(T_1, \ldots, T_k)$ is an exponential family of the form

$$f_T(u_1, \ldots, u_k|\theta) = H(u_1, \ldots, u_k)[c(\theta)]^n \exp\left(\sum_{i=1}^{k} w_i(\theta)u_i\right). \tag{5.2.6}$$

The open set condition eliminates a density such as the $n(\theta, \theta^2)$ and, in general, eliminates curved exponential families from Theorem 5.2.11.

Example 5.2.12 (Sum of Bernoulli random variables) Suppose $X_1, \ldots, X_n$ is a random sample from a Bernoulli(p) distribution. From Example 3.4.1 (with $n = 1$) we see that a Bernoulli(p) distribution is an exponential family with $k = 1, c(p) = (1-p), w_1(p) = \log(p/(1-p))$, and $t_1(x) = x$. Thus, in the previous theorem, $T_1 = T_1(X_1, \ldots, X_n) = X_1 + \cdots + X_n$. From the definition of the binomial distribution in Section 3.2, we know that T_1 has a binomial(n, p) distribution. From Example 3.4.1 we also see that a binomial(n, p) distribution is an exponential family with the same $w_1(p)$ and $c(p) = (1-p)^n$. Thus expression (5.2.6) is verified for this example. ||

Theorem 4.4.3 *If X and Y are any two random variables, then*

$$\mathrm{E}X = \mathrm{E}\left(\mathrm{E}(X|Y)\right), \tag{4.4.1}$$

provided that the expectations exist.

Proof: Let $f(x, y)$ denote the joint pdf of X and Y. By definition, we have

$$\mathrm{E}X = \int\int x f(x, y)\, dx\, dy = \int \left[\int x f(x|y)\, dx\right] f_Y(y)\, dy, \tag{4.4.2}$$

where $f(x|y)$ and $f_Y(y)$ are the conditional pdf of X given $Y = y$ and the marginal pdf of Y, respectively. But now notice that the inner integral in (4.4.2) is the conditional expectation $\mathrm{E}(X|y)$, and we have

$$\mathrm{E}X = \int \mathrm{E}(X|y) f_Y(y)\, dy = \mathrm{E}\left(\mathrm{E}(X|Y)\right),$$

as desired. Replace integrals by sums to prove the discrete case. □

Theorem 4.4.7 (Conditional variance identity) *For any two random variables X and Y,*

$$\mathrm{Var}\, X = \mathrm{E}\left(\mathrm{Var}(X|Y)\right) + \mathrm{Var}\left(\mathrm{E}(X|Y)\right), \tag{4.4.4}$$

provided that the expectations exist.

Proof: By definition, we have

$$\mathrm{Var}\, X = \mathrm{E}\left([X - \mathrm{E}X]^2\right) = \mathrm{E}\left([X - \mathrm{E}(X|Y) + \mathrm{E}(X|Y) - \mathrm{E}X]^2\right),$$

where in the last step we have added and subtracted $\mathrm{E}(X|Y)$. Expanding the square in this last expectation now gives

$$\begin{aligned} \mathrm{Var}\, X = {} & \mathrm{E}\left([X - \mathrm{E}(X|Y)]^2\right) + \mathrm{E}\left([\mathrm{E}(X|Y) - \mathrm{E}X]^2\right) \\ & + 2\mathrm{E}\left([X - \mathrm{E}(X|Y)][\mathrm{E}(X|Y) - \mathrm{E}X]\right). \end{aligned} \tag{4.4.5}$$

The last term in this expression is equal to 0, however, which can easily be seen by iterating the expectation:

(4.4.6) $\mathrm{E}\left([X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]\right)=\mathrm{E}\left(\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\}\right).$

In the conditional distribution $X|Y$, X is the random variable. So in the expression

$$\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\},$$

$\mathrm{E}(X|Y)$ and $\mathrm{E}X$ are constants. Thus,

$$\begin{aligned}\mathrm{E}\left\{[X-\mathrm{E}(X|Y)][\mathrm{E}(X|Y)-\mathrm{E}X]|Y\right\}&=(\mathrm{E}(X|Y)-\mathrm{E}X)\left(\mathrm{E}\left\{[X-\mathrm{E}(X|Y)]|Y\right\}\right)\\&=(\mathrm{E}(X|Y)-\mathrm{E}X)\left(\mathrm{E}(X|Y)-\mathrm{E}(X|Y)\right)\\&=(\mathrm{E}(X|Y)-\mathrm{E}X)\,(0)\\&=0.\end{aligned}$$

Thus, from (4.4.6), we have that $\mathrm{E}((X-\mathrm{E}(X|Y))(\mathrm{E}(X|Y)-\mathrm{E}X))=\mathrm{E}(0)=0$. Referring back to equation (4.4.5), we see that

$$\begin{aligned}\mathrm{E}\left([X-\mathrm{E}(X|Y)]^2\right)&=\mathrm{E}\left(\mathrm{E}\left\{[X-\mathrm{E}(X|Y)]^2|Y\right\}\right)\\&=\mathrm{E}\left(\mathrm{Var}(X|Y)\right)\end{aligned}$$

and

$$\mathrm{E}\left([\mathrm{E}(X|Y)-\mathrm{E}X]^2\right)=\mathrm{Var}\left(\mathrm{E}(X|Y)\right),$$

establishing

$$\mathrm{Var}\,X=\mathrm{E}\left(\mathrm{Var}(X|Y)\right)+\mathrm{Var}\left(\mathrm{E}(X|Y)\right)$$ □

Definition 4.6.5 Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be random vectors with joint pdf or pmf $f(\mathbf{x}_1,\ldots,\mathbf{x}_n)$. Let $f_{\mathbf{X}_i}(\mathbf{x}_i)$ denote the marginal pdf or pmf of $\mathbf{X}_i$. Then $\mathbf{X}_1,\ldots,\mathbf{X}_n$ are called *mutually independent random vectors* if, for every $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$,

$$f(\mathbf{x}_1,\ldots,\mathbf{x}_n)=f_{\mathbf{X}_1}(\mathbf{x}_1)\cdot\cdots\cdot f_{\mathbf{X}_n}(\mathbf{x}_n)=\prod_{i=1}^{n}f_{\mathbf{X}_i}(\mathbf{x}_i).$$

If the X_is are all one-dimensional, then $X_1,\ldots,X_n$ are called *mutually independent random variables.*

Theorem 4.6.6 *Let $X_1,\ldots,X_n$ be mutually independent random variables. Let $g_1,\ldots,g_n$ be real-valued functions such that $g_i(x_i)$ is a function only of x_i, $i=1,\ldots,n$. Then*

$$\mathrm{E}(g_1(X_1)\cdot\cdots\cdot g_n(X_n))=(\mathrm{E}g_1(X_1))\cdot\cdots\cdot(\mathrm{E}g_n(X_n)).$$

Theorem 4.6.7 *Let $X_1,\ldots,X_n$ be mutually independent random variables with mgfs $M_{X_1}(t),\ldots,M_{X_n}(t)$. Let $Z=X_1+\cdots+X_n$. Then the mgf of Z is*

$$M_Z(t)=M_{X_1}(t)\cdot\cdots\cdot M_{X_n}(t).$$

In particular, if $X_1,\ldots,X_n$ all have the same distribution with mgf $M_X(t)$, then

$$M_Z(t)=(M_X(t))^n.$$

Lemma 4.2.7 *Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$. Then X and Y are independent random variables if and only if there exist functions $g(x)$ and $h(y)$ such that, for every $x \in \Re$ and $y \in \Re$,*

Proof: The "only if" part is proved by defining $g(x) = f_X(x)$ and $h(y) = f_Y(y)$ and using (4.2.1). To prove the "if" part for continuous random variables, suppose that $f(x, y) = g(x)h(y)$. Define

$$\int_{-\infty}^{\infty} g(x)\,dx = c \quad \text{and} \quad \int_{-\infty}^{\infty} h(y)\,dy = d,$$

where the constants c and d satisfy

$$\begin{aligned} cd &= \left(\int_{-\infty}^{\infty} g(x)\,dx\right)\left(\int_{-\infty}^{\infty} h(y)\,dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)\,dx\,dy \qquad (4.2.2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy \\ &= 1. \qquad (f(x, y) \text{ is a joint pdf}) \end{aligned}$$

Furthermore, the marginal pdfs are given by

(4.2.3)

$$f_X(x) = \int_{-\infty}^{\infty} g(x)h(y)\,dy = g(x)d \quad \text{and} \quad f_Y(y) = \int_{-\infty}^{\infty} g(x)h(y)\,dx = h(y)c.$$

Thus, using (4.2.2) and (4.2.3), we have

$$f(x, y) = g(x)h(y) = g(x)h(y)cd = f_X(x)f_Y(y),$$

showing that X and Y are independent. Replacing integrals with sums proves the lemma for discrete random vectors. □

Theorem 4.6.11 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be random vectors. Then $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are mutually independent random vectors if and only if there exist functions $g_i(\mathbf{x}_i), i = 1, \ldots, n$, such that the joint pdf or pmf of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ can be written as*

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) = g_1(\mathbf{x}_1)\cdot\cdots\cdot g_n(\mathbf{x}_n).$$

Theorem 4.6.12 *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors. Let $g_i(\mathbf{x}_i)$ be a function only of $\mathbf{x}_i, i = 1, \ldots, n$. Then the random variables $U_i = g_i(\mathbf{X}_i), i = 1, \ldots, n$, are mutually independent.*

Central Limit Theorem *Let $X_1, X_2, \ldots$ be a sequence of iid random variables with $\mathrm{E}X_i = \mu$ and $0 < \mathrm{Var}\, X_i = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy;$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ has a limiting standard normal distribution.

Slutsky's Theorem *If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then*

a. *$Y_n X_n \to aX$ in distribution.*

b. *$X_n + Y_n \to X + a$ in distribution.*

Example Normal approximation with estimated variance Suppose that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathrm{n}(0,1), \quad \text{but the value of } \sigma \text{ is unknown.}$$

$$P(|S_n^2 - \sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}(S_n^2 - \sigma^2)^2}{\epsilon^2} = \frac{\mathrm{Var}\, S_n^2}{\epsilon^2}$$

if $\lim_{n\to\infty} \mathrm{Var}\, S_n^2 = 0$, then $S_n^2 \to \sigma^2$ in probability. $\sigma/S_n \to 1$ in probability. Hence, Slutsky's Theorem tells us

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} = \frac{\sigma}{S_n}\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \to \mathrm{n}(0,1). \qquad \|$$

The Delta Method

Example 5.5.19 (Estimating the odds) Suppose we observe $X_1, X_2, \ldots, X_n$ independent Bernoulli(p) random variables. The typical parameter of interest is p, the success probability, but another popular parameter is $\frac{p}{1-p}$, the *odds*. For example, if the data represent the outcomes of a medical treatment with $p = 2/3$, then a person has odds $2:1$ of getting better. Moreover, if there were another treatment with success probability r, biostatisticians often estimate the *odds ratio* $\frac{p}{1-p}/\frac{r}{1-r}$, giving the relative odds of one treatment over another.

As we would typically estimate the success probability p with the observed success probability $\hat{p} = \sum_i X_i/n$, we might consider using $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$. But what are the properties of this estimator? How might we estimate the variance of $\frac{\hat{p}}{1-\hat{p}}$? Moreover, how can we approximate its sampling distribution?

Intuition abandons us, and exact calculation is relatively hopeless, so we have to rely on an approximation. The Delta Method will allow us to obtain reasonable, approximate answers to our questions. $\|$

Definition 5.5.20 If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{d^r}{dx^r}g(x)$ exists, then for any constant a, the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

Taylor's major theorem, which we will not prove here, is that the *remainder* from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term.

Theorem 5.5.21 (Taylor) *If* $g^{(r)}(a) = \frac{d^r}{dx^r}g(x)\big|_{x=a}$ *exists, then*

$$\lim_{x\to a}\frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

In general, we will not be concerned with the explicit form of the remainder. one useful one being

$$g(x) - T_r(x) = \int_a^x \frac{g^{(r+1)}(t)}{r!}(x-t)^r dt.$$

For the statistical application of Taylor's Theorem, we are most concerned with the *first-order* Taylor series, that is, an approximation using just the first derivative

Let $T_1, \ldots, T_k$ be random variables with means $\theta_1, \ldots, \theta_k$, and define $\mathbf{T} = (T_1, \ldots, T_k)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Suppose there is a differentiable function $g(\mathbf{T})$ (an estimator of some parameter) for which we want an approximate estimate of variance. Define

$$g_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial t_i}g(\mathbf{t})\big|_{t_1=\theta_1,\ldots,t_k=\theta_k}.$$

The first-order Taylor series expansion of g about $\boldsymbol{\theta}$ is

$$g(\mathbf{t}) = g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i) + \text{Remainder}.$$

For our statistical approximation we forget about the remainder and write

$$g(\mathbf{t}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})(t_i - \theta_i). \tag{5.5.7}$$

Now, take expectations on both sides of (5.5.7) to get

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta}} g(\mathbf{T}) &\approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\mathrm{E}_{\boldsymbol{\theta}}(T_i - \theta_i) \\ &= g(\boldsymbol{\theta}). \qquad (T_i \text{ has mean } \theta_i) \end{aligned} \tag{5.5.8}$$

We can now approximate the variance of $g(\mathbf{T})$ by

$$\begin{aligned} \mathrm{Var}_{\boldsymbol{\theta}}\, g(\mathbf{T}) &\approx \mathrm{E}_{\boldsymbol{\theta}}\left([g(\mathbf{T}) - g(\boldsymbol{\theta})]^2\right) && \text{(using (5.5.8))} \\ &\approx \mathrm{E}_{\boldsymbol{\theta}}\left(\left(\sum\nolimits_{i=1}^{k} g_i'(\boldsymbol{\theta})(T_i - \theta_i)\right)^2\right) && \text{(using (5.5.7))} \end{aligned}$$

$$= \sum_{i=1}^{k} [g_i'(\boldsymbol{\theta})]^2 \mathrm{Var}_{\boldsymbol{\theta}} T_i + 2 \sum_{i>j} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) \mathrm{Cov}_{\boldsymbol{\theta}}(T_i, T_j), \tag{5.5.9}$$

useful because it gives us a variance formula for a general function, using only simple variances and covariances.

Example 5.5.22 (Continuation of Example 5.5.19) Recall that we are interested in the properties of $\frac{\hat{p}}{1-\hat{p}}$ as an estimate of $\frac{p}{1-p}$, where p is a binomial success probability. In our above notation, take $g(p) = \frac{p}{1-p}$ so $g'(p) = \frac{1}{(1-p)^2}$ and

$$\begin{aligned} \mathrm{Var}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx [g'(p)]^2 \mathrm{Var}(\hat{p}) \\ &= \left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}, \end{aligned}$$

giving us an approximation for the variance of our estimator. ‖

Example 5.5.23 (Approximate mean and variance) Suppose X is a random variable with $\mathrm{E}_\mu X = \mu \neq 0$. If we want to estimate a function $g(\mu)$, a first-order approximation would give us

$$g(X) = g(\mu) + g'(\mu)(X - \mu).$$

If we use $g(X)$ as an estimator of $g(\mu)$, we can say that approximately

$$\begin{aligned} \mathrm{E}_\mu g(X) &\approx g(\mu), \\ \mathrm{Var}_\mu\, g(X) &\approx [g'(\mu)]^2 \mathrm{Var}_\mu\, X. \end{aligned}$$

For a specific example, take $g(\mu) = 1/\mu$. We estimate $1/\mu$ with $1/X$, and we can say

$$\mathrm{E}_\mu\left(\frac{1}{X}\right) \approx \frac{1}{\mu},$$

$$\mathrm{Var}_\mu\left(\frac{1}{X}\right) \approx \left(\frac{1}{\mu}\right)^4 \mathrm{Var}_\mu X. \qquad \|$$

Using these Taylor series approximations for the mean and variance, we get the following useful generalization of the Central Limit Theorem, known as the *Delta Method*.

Theorem 5.5.24 (Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \to \mathrm{n}(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.} \tag{5.5.10}$$

Proof: The Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \text{ Remainder}, \tag{5.5.11}$$

where the remainder $\to 0$ as $Y_n \to \theta$. Since $Y_n \to \theta$ in probability it follows that the remainder $\to 0$ in probability. By applying Slutsky's Theorem

$$\sqrt{n}[g(Y_n) - g(\theta)] = g'(\theta)\sqrt{n}(Y_n - \theta)$$

the result now follows. □

Example 5.5.25 (Continuation of Example 5.5.23) Suppose now that we have the mean of a random sample $\bar{X}$. For $\mu \neq 0$, we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \to \text{n}\left(0, \left(\frac{1}{\mu}\right)^4 \text{Var}_\mu X_1\right)$$

in distribution.

If we do not know the variance of X_1, to use the above approximation requires an estimate, say S^2. Moreover, there is the question of what to do with the $1/\mu$ term, as we also do not know μ. We can estimate everything, which gives us the approximate variance

$$\widehat{\text{Var}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 S^2.$$

Furthermore, as both $\bar{X}$ and S^2 are consistent estimators, we can again apply Slutsky's Theorem to conclude that for $\mu \neq 0$,

$$\frac{\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right)}{\left(\frac{1}{\bar{X}}\right)^2 S} \to \text{n}(0,1)$$

in distribution.

There are two extensions of the basic Delta Method that we need to deal with to complete our treatment. The first concerns the possibility that $g'(\mu) = 0$.

If $g'(\theta) = 0$, we take one more term in the Taylor expansion to get

$$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder}.$$

If we do some rearranging (setting $g' = 0$), we have

$$g(Y_n) - g(\theta) = \frac{g''(\theta)}{2}(Y_n - \theta)^2 + \text{ Remainder}. \tag{5.5.12}$$

Now recall that the square of a $\text{n}(0,1)$ is a χ^2_1, which implies that

$$\frac{n(Y_n - \theta)^2}{\sigma^2} \to \chi^2_1$$

Theorem 5.5.26 (Second-order Delta Method) *Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta) \to \mathrm{n}(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then*

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 \text{ in distribution.} \tag{5.5.13}$$

Example 5.5.27 (Moments of a ratio estimator) Suppose X and Y are random variables with nonzero means μ_X and μ_Y, respectively. The parametric function to be estimated is $g(\mu_X, \mu_Y) = \mu_X/\mu_Y$. It is straightforward to calculate

$$\frac{\partial}{\partial \mu_X} g(\mu_X, \mu_Y) = \frac{1}{\mu_Y}$$

and

$$\frac{\partial}{\partial \mu_Y} g(\mu_X, \mu_Y) = \frac{-\mu_X}{\mu_Y^2}.$$

The first-order Taylor approximations (5.5.8) and (5.5.9) give

$$\mathrm{E}\left(\frac{X}{Y}\right) \approx \frac{\mu_X}{\mu_Y}$$

and

$$\begin{aligned}\mathrm{Var}\left(\frac{X}{Y}\right) &\approx \frac{1}{\mu_Y^2}\mathrm{Var}\, X + \frac{\mu_X^2}{\mu_Y^4}\mathrm{Var}\, Y - 2\frac{\mu_X}{\mu_Y^3}\mathrm{Cov}(X, Y) \\ &= \left(\frac{\mu_X}{\mu_Y}\right)^2 \left(\frac{\mathrm{Var}\, X}{\mu_X^2} + \frac{\mathrm{Var}\, Y}{\mu_Y^2} - 2\frac{\mathrm{Cov}(X, Y)}{\mu_X \mu_Y}\right).\end{aligned}$$

Thus, we have an approximation for the mean and variance of the ratio estimator, and the approximations use only the means, variances, and covariance of X and Y. Exact calculations would be quite hopeless, with closed-form expressions being unattainable.

present a CLT to cover an estimator such as the ratio estimator.

Theorem 5.5.28 (Multivariate Delta Method) *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample with $\mathrm{E}(X_{ij}) = \mu_i$ and $\mathrm{Cov}(X_{ik}, X_{jk}) = \sigma_{ij}$. For a given function g with continuous first partial derivatives and a specific value of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ for which $\tau^2 = \Sigma\Sigma\sigma_{ij} \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_i} \cdot \frac{\partial g(\boldsymbol{\mu})}{\partial \mu_j} > 0$,*

$$\sqrt{n}[g(\bar{X}_1, \ldots, \bar{X}_s) - g(\mu_1, \ldots, \mu_p)] \to \mathrm{n}(0, \tau^2) \text{ in distribution .}$$

Suppose the vector-valued random variable $\mathbf{X} = (X_1, \ldots, X_p)$ has mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and covariances $\mathrm{Cov}(X_i, X_j) = \sigma_{ij}$, and we observe an independent random sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$ and calculate the means $\bar{X}_i = \sum_{k=1}^n X_{ik}$, $i = 1, \ldots, p$. For a function $g(\mathbf{x}) = g(x_1, \ldots, x_p)$ we can write

$$g(\bar{x}_1, \ldots, \bar{x}_p) = g(\mu_1, \ldots, \mu_p) + \sum_{k=1}^{p} g_k'(\mathbf{x})(\bar{x}_k - \mu_k),$$

UNIT 2: SUFFICIENCY

SUFFICIENCY

Any statistic, $T(\mathbf{X})$, defines a form of data reduction or data summary. An experimenter who uses only the observed value of the statistic, $T(\mathbf{x})$, rather than the entire observed sample, $\mathbf{x}$, will treat as equal two samples, $\mathbf{x}$ and $\mathbf{y}$, that satisfy $T(\mathbf{x}) = T(\mathbf{y})$ even though the actual sample values may be different in some ways.

Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Then $T(\mathbf{x})$ partitions the sample space into sets $A_t, t \in \mathcal{T}$, defined by $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. The statistic summarizes the data in that, rather than reporting the entire sample $\mathbf{x}$, it reports only that $T(\mathbf{x}) = t$ or, equivalently, $\mathbf{x} \in A_t$. For example, if $T(\mathbf{x}) = x_1 + \cdots + x_n$, then $T(\mathbf{x})$ does not report the actual sample values but only the sum. There may be many different sample points that have the same sum.

The Sufficiency

A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

SUFFICIENCY PRINCIPLE: If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

The Sufficiency

A *sufficient statistic* for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

SUFFICIENCY PRINCIPLE: If $T(\mathbf{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two sample points such that $T(\mathbf{x}) = T(\mathbf{y})$, then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.

Definition A statistic $T(\mathbf{X})$ is a *sufficient statistic for* θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.

Theorem 6.2.2 *If $p(\mathbf{x}|\theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of θ.*

Proof: since $\{\mathbf{X} = \mathbf{x}\}$ is a subset of $\{T(\mathbf{X}) = T(\mathbf{x})\}$,

$$\begin{aligned} P_\theta(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) &= \frac{P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x}))}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{P_\theta(\mathbf{X} = \mathbf{x})}{P_\theta(T(\mathbf{X}) = T(\mathbf{x}))} \\ &= \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)}, \end{aligned}$$

where $p(\mathbf{x}|\theta)$ is the joint pmf of the sample $\mathbf{X}$ and $q(t|\theta)$ is the pmf of $T(\mathbf{X})$. Thus, $T(\mathbf{X})$ is a sufficient statistic for θ if and only if, for every $\mathbf{x}$, the above ratio of pmfs is constant as a function of θ.

Example 6.2.3 (Binomial sufficient statistic) Let $X_1, \ldots, X_n$ be iid Bernoulli random variables with parameter θ, $0 < \theta < 1$. We will show that $T(\mathbf{X}) = X_1 + \cdots + X_n$ is a sufficient statistic for θ. Note that $T(\mathbf{X})$ counts the number of X_is that equal 1, so $T(\mathbf{X})$ has a binomial(n, θ) distribution. The ratio of pmfs is thus

$$\begin{aligned} \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{\Pi\theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{\theta^{\Sigma x_i}(1-\theta)^{\Sigma(1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} && \text{(define } t = \Sigma x_i\text{)} \\ &= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}} = \frac{1}{\binom{n}{\Sigma x_i}}. && (\Pi\theta^{x_i} = \theta^{\Sigma x_i}) \end{aligned}$$

Since this ratio does not depend on θ, by Theorem 6.2.2, $T(\mathbf{X})$ is a sufficient statistic for θ. The interpretation is this: The total number of 1s in this Bernoulli sample contains all the information about θ that is in the data. Other features of the data, such as the exact value of X_3, contain no additional information. ||

Example 6.2.4 (Normal sufficient statistic) Let $X_1, \ldots, X_n$ be iid n(μ, σ^2), where σ^2 is known. We wish to show that the sample mean, $T(\mathbf{X}) = \bar{X} = (X_1 + \cdots + X_n)/n$, is a sufficient statistic for μ. The joint pdf of the sample $\mathbf{X}$ is

$$\begin{aligned} f(\mathbf{x}|\mu) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-(x_i-\mu)^2/(2\sigma^2)\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)\right) \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2/(2\sigma^2)\right) && \text{(add and subtract } \bar{x}\text{)} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)/(2\sigma^2)\right). && (6.2.1) \end{aligned}$$

The last equality is true because the cross-product term $\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - \mu)$ may be rewritten as $(\bar{x} - \mu)\sum_{i=1}^n (x_i - \bar{x})$, and $\sum_{i=1}^n (x_i - \bar{x}) = 0$. Recall that the sample mean $\bar{X}$ has a $n(\mu, \sigma^2/n)$ distribution. Thus, the ratio of pdfs is

$$\begin{aligned}\frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)/(2\sigma^2)\right)}{(2\pi\sigma^2/n)^{-1/2}\exp(-n(\bar{x} - \mu)^2/(2\sigma^2))} \\ &= n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp\left(-\sum_{i=1}^n (x_i - \bar{x})^2/(2\sigma^2)\right),\end{aligned}$$

which does not depend on μ. By Theorem 6.2.2, the sample mean is a sufficient statistic for μ. ||

Example 6.2.5 (Sufficient order statistics) Let $X_1, \ldots, X_n$ be iid from a pdf f, where we are unable to specify any more information about the pdf (as is the case in *nonparametric* estimation). It then follows that the sample density is given by

$$f(\mathbf{x}) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n f(x_{(i)}), \tag{6.2.2}$$

where $x_{(i)} \le x_{(2)} \le \cdots \le x_{(n)}$ are the order statistics. By Theorem 6.2.2, we can show that the order statistics are a sufficient statistic. Of course, this is not much of a reduction, but we shouldn't expect more with so little information about the density f.

However, even if we do specify more about the density, we still may not be able to get much of a sufficiency reduction. For example, suppose that f is the Cauchy pdf $f(x|\theta) = \frac{1}{\pi(x-\theta)^2}$ or the logistic pdf $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}$. We then have the same reduction as in (6.2.2), and no more. So reduction to the order statistics is the most we can get in these families

It turns out that outside of the exponential family of distributions, it is rare to have a sufficient statistic of smaller dimension than the size of the sample, so in many cases it will turn out that the order statistics are the best that we can do.

It may be unwieldy to use the definition of a sufficient statistic to find a sufficient statistic for a particular model. To use the definition, we must guess a statistic $T(\mathbf{X})$ to be sufficient, find the pmf or pdf of $T(\mathbf{X})$, and check that the ratio of pdfs or pmfs does not depend on θ. The first step requires a good deal of intuition and the second sometimes requires some tedious analysis. Fortunately, the next theorem, due to Halmos and Savage (1949), allows us to find a sufficient statistic by simple inspection of the pdf or pmf of the sample.[1]

Theorem 6.2.6 (Factorization Theorem) *Let $f(\mathbf{x}|\theta)$ denote the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that, for all sample points $\mathbf{x}$ and all parameter points θ,*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}). \tag{6.2.3}$$

Proof: We give the proof only for discrete distributions.

Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(T(\mathbf{X}) = t)$ and $h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$. Because $T(\mathbf{X})$ is sufficient, the conditional probability defining $h(\mathbf{x})$ does not depend on θ. Thus this choice of $h(\mathbf{x})$ and $g(t|\theta)$ is legitimate, and for this choice we have

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta(\mathbf{X} = \mathbf{x}) \\ &= P_\theta(\mathbf{X} = \mathbf{x} \text{ and } T(\mathbf{X}) = T(\mathbf{x})) \\ &= P_\theta(T(\mathbf{X}) = T(\mathbf{x}))P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})) \qquad \text{(sufficiency)} \\ &= g(T(\mathbf{x})|\theta)h(\mathbf{x}). \end{aligned}$$

So factorization (6.2.3) has been exhibited. We also see from the last two lines above that

$$P_\theta(T(\mathbf{X}) = T(\mathbf{x})) = g(T(\mathbf{x})|\theta),$$

so $g(T(\mathbf{x})|\theta)$ is the pmf of $T(\mathbf{X})$.

Now assume the factorization (6.2.3) exists. Let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. To show that $T(\mathbf{X})$ is sufficient we examine the ratio $f(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$. Define $A_{T(\mathbf{x})} = \{\mathbf{y}: T(\mathbf{y}) = T(\mathbf{x})\}$. Then

$$\begin{aligned} \frac{f(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{q(T(\mathbf{x})|\theta)} && \text{(since (6.2.3) is satisfied)} \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}} g(T(\mathbf{y})|\theta)h(\mathbf{y})} && \text{(definition of the pmf of } T) \\ &= \frac{g(T(\mathbf{x})|\theta)h(\mathbf{x})}{g(T(\mathbf{x})|\theta)\Sigma_{A_{T(\mathbf{x})}} h(\mathbf{y})} && \text{(since } T \text{ is constant on } A_{T(\mathbf{x})}) \\ &= \frac{h(\mathbf{x})}{\Sigma_{A_{T(\mathbf{x})}} h(\mathbf{y})}. \end{aligned}$$

Since the ratio does not depend on θ, by Theorem 6.2.2, $T(\mathbf{X})$ is a sufficient statistic for θ. □

Example 6.2.7 (Continuation of Example 6.2.4) For the normal model described earlier, we saw that the pdf could be factored as

$$f(\mathbf{x}|\mu) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\right) \exp(-n(\bar{x} - \mu)^2/(2\sigma^2)). \tag{6.2.4}$$

We can define

$$h(\mathbf{x}) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^{n}(x_i - \bar{x})^2/(2\sigma^2)\right),$$

which does not depend on the unknown parameter μ. The factor in (6.2.4) that contains μ depends on the sample $\mathbf{x}$ only through the function $T(\mathbf{x}) = \bar{x}$, the sample mean. So we have

$$g(t|\mu) = \exp\left(-n(t - \mu)^2/(2\sigma^2)\right)$$

and note that

$$f(\mathbf{x}|\mu) = h(\mathbf{x})g(T(\mathbf{x})|\mu).$$

Thus, by the Factorization Theorem, $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. ‖

Example 6.2.8 (Uniform sufficient statistic) Let $X_1, \ldots, X_n$ be iid observations from the discrete uniform distribution on $1, \ldots, \theta$. That is, the unknown parameter, θ, is a positive integer and the pmf of X_i is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & x = 1, 2, \ldots, \theta \\ 0 & \text{otherwise.} \end{cases}$$

Thus the joint pmf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x}|\theta) = \begin{cases} \theta^{-n} & x_i \in \{1, \ldots, \theta\} \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

The restriction "$x_i \in \{1, \ldots, \theta\}$ for $i = 1, \ldots, n$" can be re-expressed as "$x_i \in \{1, 2, \ldots\}$ for $i = 1, \ldots, n$ (note that there is no θ in this restriction) and $\max_i x_i \leq \theta$." If we define $T(\mathbf{x}) = \max_i x_i$,

$$h(x) = \begin{cases} 1 & x_i \in \{1, 2, \ldots\} \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise,} \end{cases}$$

and

$$g(t|\theta) = \begin{cases} \theta^{-n} & t \leq \theta \\ 0 & \text{otherwise,} \end{cases}$$

it is easily verified that $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$ for all $\mathbf{x}$ and θ. Thus, the largest order statistic, $T(\mathbf{X}) = \max_i X_i$, is a sufficient statistic in this problem.

This type of analysis can sometimes be carried out more clearly and concisely using indicator functions. Recall that $I_A(x)$ is the indicator function of the set A; that is, it is equal to 1 if $x \in A$ and equal to 0 otherwise. Let $\mathcal{N} = \{1, 2, \ldots\}$ be the set of positive integers and let $\mathcal{N}_\theta = \{1, 2, \ldots, \theta\}$. Then the joint pmf of $X_1, \ldots, X_n$ is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{-1} I_{\mathcal{N}_\theta}(x_i) = \theta^{-n} \prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i).$$

Defining $T(\mathbf{x}) = \max_i x_i$, we see that

$$\prod_{i=1}^{n} I_{\mathcal{N}_\theta}(x_i) = \left(\prod_{i=1}^{n} I_{\mathcal{N}}(x_i)\right) I_{\mathcal{N}_\theta}(T(\mathbf{x})).$$

Thus we have the factorization

$$f(\mathbf{x}|\theta) = \theta^{-n} I_{\mathcal{N}_\theta}(T(\mathbf{x})) \left(\prod_{i=1}^{n} I_{\mathcal{N}}(x_i)\right)$$

The first factor depends on $x_1, \ldots, x_n$ only through the value of $T(\mathbf{x}) = \max_i x_i$

It is easy to find a sufficient statistic for an exponential family of distributions using the Factorization Theorem. The proof of the following important result is left as Exercise

Theorem 6.2.10 *Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f(x|\boldsymbol{\theta})$ that belongs to an exponential family given by*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$, $d \leq k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is a sufficient statistic for $\boldsymbol{\theta}$.

Minimal Sufficient Statistics

In any problem there are, in fact, many sufficient statistics.

It is always true that the complete sample, $\mathbf{X}$, is a sufficient statistic. We can factor the pdf or pmf of $\mathbf{X}$ as $f(\mathbf{x}|\theta) = f(T(\mathbf{x})|\theta)h(\mathbf{x})$, where $T(\mathbf{x}) = \mathbf{x}$ and $h(\mathbf{x}) = 1$ for all $\mathbf{x}$. By the Factorization Theorem, $T(\mathbf{X}) = \mathbf{X}$ is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose $T(\mathbf{X})$ is a sufficient statistic and define $T^*(\mathbf{x}) = r(T(\mathbf{x}))$ for all $\mathbf{x}$, where r is a one-to-one function with inverse r^{-1}. Then by the Factorization Theorem there exist g and h such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) = g(r^{-1}(T^*(\mathbf{x}))|\theta)h(\mathbf{x}).$$

Defining $g^*(t|\theta) = g(r^{-1}(t)|\theta)$, we see that

$$f(\mathbf{x}|\theta) = g^*(T^*(\mathbf{x})|\theta)h(\mathbf{x}).$$

So, by the Factorization Theorem, $T^*(\mathbf{X})$ is a sufficient statistic.

Recall that the purpose of a sufficient
statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all the information about θ might be considered preferable.

Definition 6.2.11 A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

To say that $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ simply means that if $T'(\mathbf{x}) = T'(\mathbf{y})$, then $T(\mathbf{x}) = T(\mathbf{y})$. In terms of the partition sets
if $\{B_{t'}: t' \in \mathcal{T}'\}$ are the partition sets for $T'(\mathbf{x})$ and $\{A_t: t \in \mathcal{T}\}$ are the partition sets for $T(\mathbf{x})$, then Definition 6.2.11 states that every $B_{t'}$ is a subset of some A_t. Thus, the partition associated with a minimal sufficient statistic, is the *coarsest* possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Example 6.2.12 (Two normal sufficient statistics) The model considered in Example 6.2.4 has $X_1, \ldots, X_n$ iid $n(\mu, \sigma^2)$ with σ^2 known. Using factorization (6.2.4), we concluded that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. Instead, we could write down factorization (6.2.5) for this problem (σ^2 is a known value now) and correctly conclude that $T'(\mathbf{X}) = (\bar{X}, S^2)$ is a sufficient statistic for μ in this problem. Clearly $T(\mathbf{X})$ achieves a greater data reduction than $T'(\mathbf{X})$ since we do not know the sample variance if we know only $T(\mathbf{X})$. We can write $T(\mathbf{x})$ as a function of $T'(\mathbf{x})$ by defining the function $r(a, b) = a$. Then $T(\mathbf{x}) = \bar{x} = r(\bar{x}, s^2) = r(T'(\mathbf{x}))$. Since $T(\mathbf{X})$ and $T'(\mathbf{X})$ are both sufficient statistics, they both contain the same information about μ. Thus, the additional information about the value of S^2, the sample variance, does not add to our knowledge of μ since the population variance σ^2 is known.

Of course, if σ^2 is unknown,
as in Example 6.2.9, $T(\mathbf{X}) = \bar{X}$ is not a sufficient statistic and $T'(\mathbf{X})$
contains more information about the parameter (μ, σ^2) than does $T(\mathbf{X})$. ||

Theorem 6.2.13 *Let $f(\mathbf{x}|\theta)$ be the pmf or pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.*

Proof:

First we show that $T(\mathbf{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set, A_t, as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set $A_t, T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$ and, hence, $f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then it can be seen that

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta) h(\mathbf{x})$$

and, by the Factorization Theorem, $T(\mathbf{X})$ is a sufficient statistic for θ.

Now to show that $T(\mathbf{X})$ is minimal, let $T'(\mathbf{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta) h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal. □

Example 6.2.14 (Normal minimal sufficient statistic) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, both μ and σ^2 unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $(\bar{x}, s_{\mathbf{x}}^2)$ and $(\bar{y}, s_{\mathbf{y}}^2)$ be the sample means and variances corresponding to the $\mathbf{x}$ and $\mathbf{y}$ samples, respectively. Then, using (6.2.5), we see that the ratio of densities is

$$\begin{aligned}\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{x}-\mu)^2+(n-1)s_{\mathbf{x}}^2\right]/(2\sigma^2)\right)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{y}-\mu)^2+(n-1)s_{\mathbf{y}}^2\right]/(2\sigma^2)\right)} \\ &= \exp\left(\left[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_{\mathbf{x}}^2-s_{\mathbf{y}}^2)\right]/(2\sigma^2)\right).\end{aligned}$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x}=\bar{y}$ and $s_{\mathbf{x}}^2 = s_{\mathbf{y}}^2$. Thus, by Theorem 6.2.13, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2). ||

Example 6.2.15 (Uniform minimal sufficient statistic) Suppose $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Then the joint pdf of $\mathbf{X}$ is

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \theta < x_i < \theta+1,\ i=1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases}$$

which can be written as

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \max_i x_i - 1 < \theta < \min_i x_i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for two sample points $\mathbf{x}$ and $\mathbf{y}$, the numerator and denominator of the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ will be positive for the same values of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1. Thus, letting $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$, we have that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter. ||

A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

$(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is also a minimal sufficient statistic in Example 6.2.15

$(\Sigma_{i=1}^n X_i, \Sigma_{i=1}^n X_i^2)$ is also a minimal sufficient statistic in Example 6.2.14.

sufficient statistics contain all the information about θ that is available in the sample we introduce a different sort of statistic, one that has a complementary purpose.

Definition 6.2.16 A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic.*

Alone, an ancillary statistic contains no information about θ. An ancillary statistic is an observation on a random variable whose distribution is fixed and known, unrelated to θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about θ.

Example 6.2.17 (Uniform ancillary statistic) As in Example 6.2.15, let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta+1), -\infty < \theta < \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics from the sample. We show below that the range statistic, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic by showing that the pdf of R does not depend on θ. Recall that the cdf of each X_i is

$$F(x|\theta) = \begin{cases} 0 & x \le \theta \\ x - \theta & \theta < x < \theta + 1 \\ 1 & \theta + 1 \le x. \end{cases}$$

Thus, the joint pdf of $X_{(1)}$ and $X_{(n)}$, is

$$g(x_{(1)}, x_{(n)}|\theta) = \begin{cases} n(n-1)(x_{(n)} - x_{(1)})^{n-2} & \theta < x_{(1)} < x_{(n)} < \theta + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Making the transformation $R = X_{(n)} - X_{(1)}$ and $M = (X_{(1)} + X_{(n)})/2$, which has the inverse transformation $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$ with Jacobian 1, we see that the joint pdf of R and M is

$$h(r, m|\theta) = \begin{cases} n(n-1)r^{n-2} & 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2) \\ 0 & \text{otherwise.} \end{cases}$$

(Notice the rather involved region of positivity for $h(r, m|\theta)$.) Thus, the pdf for R is

$$\begin{aligned} h(r|\theta) &= \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2}\,dm \\ &= n(n-1)r^{n-2}(1-r), \quad 0 < r < 1. \end{aligned}$$

This is a beta pdf with $\alpha = n - 1$ and $\beta = 2$. More important, the pdf is the same for all θ. Thus, the distribution of R does not depend on θ, and R is ancillary. ‖

The ancillarity of R does not depend on the uniformity of the X_is, but rather on the parameter of the distribution being a location parameter.

Example 6.2.18 (Location family ancillary statistic) Let $X_1, \ldots, X_n$ be iid observations from a location parameter family with cdf $F(x-\theta), -\infty < \theta < \infty$. We will show that the range, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic. We work with $Z_1, \ldots, Z_n$ iid observations from $F(x)$ (corresponding to $\theta = 0$) with $X_1 = Z_1 + \theta, \ldots, X_n = Z_n + \theta$. Thus the cdf of the range statistic, R, is

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \le r) \\ &= P_\theta(\max_i X_i - \min_i X_i \le r) \\ &= P_\theta(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i + \theta - \theta \le r) \end{aligned}$$

$$F_R(r|\theta) = P_\theta(\max_i Z_i - \min_i Z_i \le r).$$

The last probability does not depend on θ because the distribution of $Z_1, \ldots, Z_n$ does not depend on θ. Thus, the cdf of R does not depend on θ and, hence, R is an ancillary statistic. ||

Example 6.2.19 (Scale family ancillary statistic) Scale parameter families also have certain kinds of ancillary statistics. Let $X_1, \ldots, X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma), \sigma > 0$. Then any statistic that depends on the sample only through the $n-1$ values $X_1/X_n, \ldots, X_{n-1}/X_n$ is an ancillary statistic. For example,

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

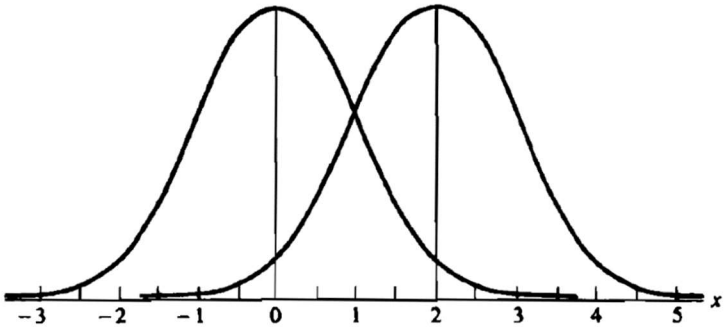
is an ancillary statistic. To see this fact, let $Z_1, \ldots, Z_n$ be iid observations from $F(x)$ (corresponding to $\sigma = 1$) with $X_i = \sigma Z_i$. The joint cdf of $X_1/X_n, \ldots, X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\sigma) &= P_\sigma(X_1/X_n \le y_1, \ldots, X_{n-1}/X_n \le y_{n-1}) \\ &= P_\sigma(\sigma Z_1/(\sigma Z_n) \le y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \le y_{n-1}) \\ &= P_\sigma(Z_1/Z_n \le y_1, \ldots, Z_{n-1}/Z_n \le y_{n-1}). \end{aligned}$$

The last probability does not depend on σ because the distribution of $Z_1, \ldots, Z_n$ does not depend on σ. So the distribution of $X_1/X_n, \ldots, X_{n-1}/X_n$ is independent of σ, as is the distribution of any function of these quantities.

In particular, let X_1 and X_2 be iid n$(0, \sigma^2)$ observations. From the above result, we see that X_1/X_2 has a distribution that is the same for every value of σ. But, in we saw that, if $\sigma = 1, X_1/X_2$ has a Cauchy$(0, 1)$ distribution. Thus, for any $\sigma > 0$, the distribution of X_1/X_2 is this same Cauchy distribution. ‖

Definition 3.5.2 Let $f(x)$ be any pdf. Then the family of pdfs $f(x - \mu)$, indexed by the parameter μ, $-\infty < \mu < \infty$, is called the *location family with standard pdf* $f(x)$ and μ is called the *location parameter* for the family.



Two members of the same location family: means at 0 and 2

It is clear from Figure that the area under the graph of $f(x)$ between $x = -1$ and $x = 2$ is the same as the area under the graph of $f(x - \mu)$ between $x = \mu - 1$ and $x = \mu + 2$. Thus if X is a random variable with pdf $f(x - \mu)$, we can write

$$P(-1 \leq X \leq 2|0) = P(\mu - 1 \leq X \leq \mu + 2|\mu),$$

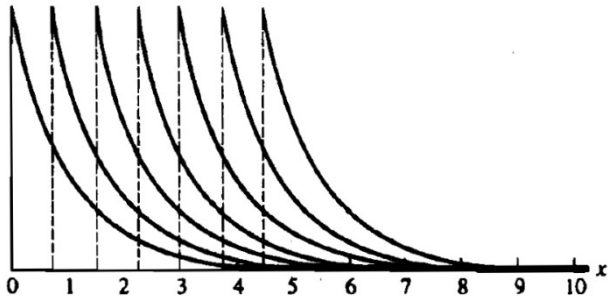
where the random variable X has pdf $f(x - 0) = f(x)$ on the left of the equality and pdf $f(x - \mu)$ on the right.

If X is a random variable with pdf $f(x-\mu)$, then X may be represented as $X = Z+\mu$, where Z is a random variable with pdf $f(z)$. This representation is a consequence of Theorem 3.5.6 (with $\sigma = 1$), which will be proved later.

Example 3.5.3 (Exponential location family) Let $f(x) = e^{-x}$, $x \geq 0$, and $f(x) = 0$, $x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x - \mu \geq 0 \\ 0 & x - \mu < 0 \end{cases}$$
$$= \begin{cases} e^{-(x-\mu)} & x \geq \mu \\ 0 & x < \mu. \end{cases}$$

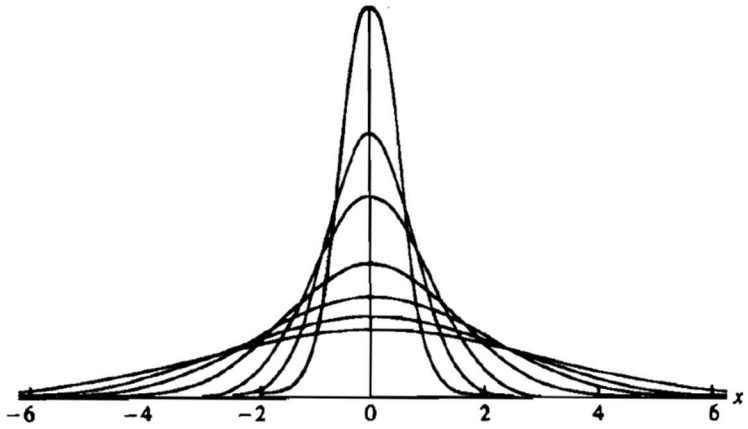
Graphs of $f(x|\mu)$ for various values of μ are shown in Figure
the graph has been shifted. Now the positive part of the graph starts at μ rather than at 0. If X measures time, then μ might be restricted to be nonnegative so that X will be positive with probability 1 for every value of μ. In this type of model, where μ denotes a bound on the range of X, μ is sometimes called a *threshold parameter.* ||



Exponential location densities

Definition 3.5.4 Let $f(x)$ be any pdf. Then for any $\sigma > 0$, the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the *scale family with standard pdf* $f(x)$ and σ is called the *scale parameter* of the family.

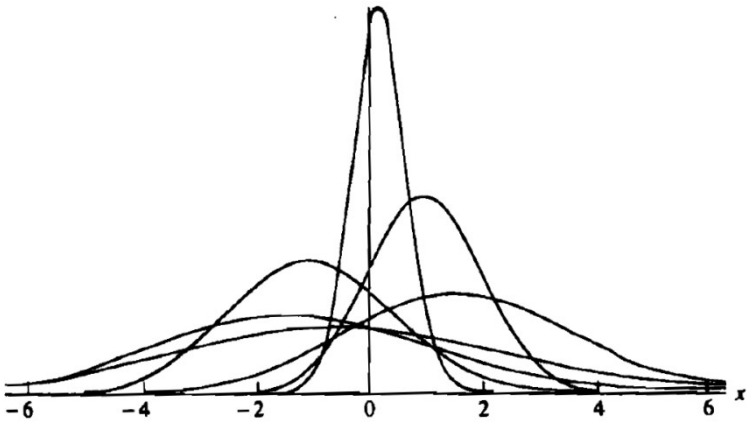
The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. This is illustrated in Figure Most often when scale parameters are used, $f(x)$ is either symmetric about 0 or positive only for $x > 0$. In these cases the stretching is either symmetric about 0 or only in the positive direction. But, in the definition, any pdf may be used as the standard.



Members of the same scale family

Definition 3.5.5 Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ, σ), is called the *location–scale family with standard pdf* $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter.*

The effect of introducing both the location and scale parameters is to stretch ($\sigma > 1$) or contract ($\sigma < 1$) the graph with the scale parameter and then shift the graph so that the point that was above 0 is now above μ. Figure illustrates this transformation of $f(x)$. The normal and double exponential families are examples of location–scale families. the Cauchy as a location–scale family.



Members of the same location–scale family

Theorem 3.5.6 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$. Then $X = g(Z)$, g is a monotone function, $g^{-1}(x) = (x-\mu)/\sigma$, and $\left|(d/dx)g^{-1}(x)\right| = 1/\sigma$. Thus the pdf of X is

$$f_X(x) = f_Z\left(g^{-1}(x)\right)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)$. $g^{-1}(z) = \sigma z + \mu$, $\left|(d/dz)g^{-1}(z)\right| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X. \qquad \square$$

Theorem 3.5.7 *Let Z be a random variable with pdf $f(z)$. Suppose* $\mathrm{E}Z$ *and* $\mathrm{Var}\,Z$ *exist. If X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$, then*

$$\mathrm{E}X = \sigma\mathrm{E}Z + \mu \quad \text{and} \quad \mathrm{Var}\,X = \sigma^2\mathrm{Var}\,Z.$$

In particular, if $\mathrm{E}Z = 0$ *and* $\mathrm{Var}\,Z = 1$, *then* $\mathrm{E}X = \mu$ *and* $\mathrm{Var}\,X = \sigma^2$.

Proof: By Theorem 3.5.6, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So $\mathrm{E}X = \sigma\mathrm{E}Z^* + \mu = \sigma\mathrm{E}Z + \mu$ and $\mathrm{Var}\,X = \sigma^2\mathrm{Var}\,Z^* = \sigma^2\mathrm{Var}\,Z$. $\square$

Probabilities for any member of a location–scale family may be computed in terms of the standard variable Z because

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right).$$

Thus, if $P(Z \le z)$ is tabulated or easily calculable for the standard variable Z, then probabilities for X may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

Definition 6.2.21 Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called *complete* if $\mathrm{E}_\theta g(T) = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a *complete statistic.*

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. For example, if X has a $\mathrm{n}(0,1)$ distribution, then defining $g(x) = x$, we have that $\mathrm{E}g(X) = \mathrm{E}X = 0$. But the function $g(x) = x$ satisfies $P(g(X) = 0) = P(X = 0) = 0$, not 1. However, this is a particular distribution, not a family of distributions. If X has a $\mathrm{n}(\theta, 1)$ distribution, $-\infty < \theta < \infty$, we shall see that no function of X, except one that is 0 with probability 1 for all θ, satisfies $\mathrm{E}_\theta g(X) = 0$ for all θ. Thus, the family of $\mathrm{n}(\theta, 1)$ distributions, $-\infty < \theta < \infty$, is complete.

Example 6.2.22 (Binomial complete sufficient statistic) Suppose that T has a binomial(n,p) distribution, $0 < p < 1$. Let g be a function such that $\mathrm{E}_p g(T) = 0$. Then

$$0 = \mathrm{E}_p g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t}$$

$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

for all p, $0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all r, $0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0. Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \ldots, n$. Since T takes on the values $0, 1, \ldots, n$ with probability 1, this yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic. ||

Example 6.2.23 (Uniform complete sufficient statistic) Let $X_1, \ldots, X_n$ be iid uniform$(0,\theta)$ observations, $0 < \theta < \infty$. Using an argument similar to that in Example 6.2.8, we can see that $T(\mathbf{X}) = \max_i X_i$ is a sufficient statistic and, by Theorem 5.4.4, the pdf of $T(\mathbf{X})$ is

$$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n} & 0 < t < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $g(t)$ is a function satisfying $\mathrm{E}_\theta g(T) = 0$ for all θ. Since $\mathrm{E}_\theta g(T)$ is constant as a function of θ, its derivative with respect to θ is 0. Thus we have that

$$0 = \frac{d}{d\theta}\mathrm{E}_\theta g(T) = \frac{d}{d\theta}\int_0^\theta g(t) n t^{n-1}\theta^{-n} dt$$

$$= (\theta^{-n})\frac{d}{d\theta}\int_0^\theta n g(t) t^{n-1} dt + \left(\frac{d}{d\theta}\theta^{-n}\right)\int_0^\theta n g(t) t^{n-1} dt$$

$$= \theta^{-n} n g(\theta)\theta^{n-1} + 0 \qquad \begin{pmatrix}\text{applying the product}\\ \text{rule for differentiation}\end{pmatrix}$$

$$= \theta^{-1} n g(\theta).$$

The first term in the next to last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to $\mathrm{E}_\theta g(T)$, which is 0. Since $\theta^{-1} n g(\theta) = 0$ and $\theta^{-1} n \neq 0$, it must be that $g(\theta) = 0$. This is true for every $\theta > 0$; hence, T is a complete statistic.

note that the Fundamental Theorem of Calculus does
not apply to all functions, but only to functions that are *Riemann-integrable.* The equation

$$\frac{d}{d\theta} \int_0^\theta g(t)dt = g(\theta)$$

is valid only at points of continuity of Riemann-integrable g. Thus, strictly speaking, the above argument does not show that T is a complete statistic, since the condition of completeness applies to all functions, not just Riemann-integrable ones. From a more practical view, however, this distinction is not of concern since the condition of Riemann-integrability is so general that it includes virtually any function we could think of.) ‖

Theorem 6.2.24 (Basu's Theorem) *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof: We give the proof only for discrete distributions.

Let $S(\mathbf{X})$ be any ancillary statistic. Then $P(S(\mathbf{X}) = s)$ does not depend on θ since $S(\mathbf{X})$ is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x}\colon S(\mathbf{x}) = s\} | T(\mathbf{X}) = t),$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic (recall the definition!). Thus, to show that $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that

$$P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s) \tag{6.2.6}$$

for all possible values $t \in \mathcal{T}$. Now,

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) P_\theta(T(\mathbf{X}) = t).$$

Furthermore, since $\sum_{t \in \mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s) P_\theta(T(\mathbf{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s),$$

the above two equations show that

$$\mathrm{E}_\theta g(T) = \sum_{t \in \mathcal{T}} g(t) P_\theta(T(\mathbf{X}) = t) = 0 \quad \text{for all } \theta.$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence (6.2.6) is verified. □

Basu's Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

Theorem 6.2.25 (Complete statistics in the exponential family) *Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right), \tag{6.2.7}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in $\Re^k$.

The condition that the parameter space contain an open set is needed to avoid a situation like the following. The $\mathrm{n}(\theta, \theta^2)$ distribution can be written in the form (6.2.7); however, the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistic $T(\mathbf{X})$ that is an unbiased estimator of 0
(Recall that exponential families such as the $\mathrm{n}(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called *curved exponential families*)

Example 6.2.26 Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n}.$$

We first note that the exponential distributions form a scale parameter family and thus, by Example 6.2.19, $g(\mathbf{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so, by Theorem 6.2.25,

$$T(\mathbf{X}) = \sum_{n} X_i$$

is a complete statistic and, by Theorem 6.2.10, $T(\mathbf{X})$ is a sufficient statistic. (As noted below, we need not verify that $T(\mathbf{X})$ is minimal, although it could easily be verified using Theorem 6.2.13.) Hence, by Basu's Theorem, $T(\mathbf{X})$ and $g(\mathbf{X})$ are independent. Thus we have

$$\theta = \mathrm{E}_\theta X_n = \mathrm{E}_\theta T(\mathbf{X})g(\mathbf{X}) = (\mathrm{E}_\theta T(\mathbf{X}))(\mathrm{E}_\theta g(\mathbf{X})) = n\theta \mathrm{E}_\theta g(\mathbf{X}).$$

Hence, for any $\theta, \mathrm{E}_\theta g(\mathbf{X}) = n^{-1}$. ‖

Theorem 6.2.28 *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.*

UNIT 3: COMPLETENESS

Minimal Sufficient Statistics

In any problem there are, in fact, many sufficient statistics.

It is always true that the complete sample, $\mathbf{X}$, is a sufficient statistic. We can factor the pdf or pmf of $\mathbf{X}$ as $f(\mathbf{x}|\theta) = f(T(\mathbf{x})|\theta)h(\mathbf{x})$, where $T(\mathbf{x}) = \mathbf{x}$ and $h(\mathbf{x}) = 1$ for all $\mathbf{x}$. By the Factorization Theorem, $T(\mathbf{X}) = \mathbf{X}$ is a sufficient statistic.

Also, it follows that any one-to-one function of a sufficient statistic is a sufficient statistic. Suppose $T(\mathbf{X})$ is a sufficient statistic and define $T^*(\mathbf{x}) = r(T(\mathbf{x}))$ for all $\mathbf{x}$, where r is a one-to-one function with inverse r^{-1}. Then by the Factorization Theorem there exist g and h such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x}) = g(r^{-1}(T^*(\mathbf{x}))|\theta)h(\mathbf{x}).$$

Defining $g^*(t|\theta) = g(r^{-1}(t)|\theta)$, we see that

$$f(\mathbf{x}|\theta) = g^*(T^*(\mathbf{x})|\theta)h(\mathbf{x}).$$

So, by the Factorization Theorem, $T^*(\mathbf{X})$ is a sufficient statistic.

Recall that the purpose of a sufficient
statistic is to achieve data reduction without loss of information about the parameter θ; thus, a statistic that achieves the most data reduction while still retaining all the information about θ might be considered preferable.

Definition 6.2.11 A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$.

To say that $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ simply means that if $T'(\mathbf{x}) = T'(\mathbf{y})$, then $T(\mathbf{x}) = T(\mathbf{y})$. In terms of the partition sets
if $\{B_{t'}: t' \in \mathcal{T}'\}$ are the partition sets for $T'(\mathbf{x})$ and $\{A_t: t \in \mathcal{T}\}$ are the partition sets for $T(\mathbf{x})$, then Definition 6.2.11 states that every $B_{t'}$ is a subset of some A_t. Thus, the partition associated with a minimal sufficient statistic, is the *coarsest* possible partition for a sufficient statistic, and a minimal sufficient statistic achieves the greatest possible data reduction for a sufficient statistic.

Example 6.2.12 (Two normal sufficient statistics) The model considered in Example 6.2.4 has $X_1, \ldots, X_n$ iid $n(\mu, \sigma^2)$ with σ^2 known. Using factorization (6.2.4), we concluded that $T(\mathbf{X}) = \bar{X}$ is a sufficient statistic for μ. Instead, we could write down factorization (6.2.5) for this problem (σ^2 is a known value now) and correctly conclude that $T'(\mathbf{X}) = (\bar{X}, S^2)$ is a sufficient statistic for μ in this problem. Clearly $T(\mathbf{X})$ achieves a greater data reduction than $T'(\mathbf{X})$ since we do not know the sample variance if we know only $T(\mathbf{X})$. We can write $T(\mathbf{x})$ as a function of $T'(\mathbf{x})$ by defining the function $r(a, b) = a$. Then $T(\mathbf{x}) = \bar{x} = r(\bar{x}, s^2) = r(T'(\mathbf{x}))$. Since $T(\mathbf{X})$ and $T'(\mathbf{X})$ are both sufficient statistics, they both contain the same information about μ. Thus, the additional information about the value of S^2, the sample variance, does not add to our knowledge of μ since the population variance σ^2 is known.

Of course, if σ^2 is unknown,
as in Example 6.2.9, $T(\mathbf{X}) = \bar{X}$ is not a sufficient statistic and $T'(\mathbf{X})$
contains more information about the parameter (μ, σ^2) than does $T(\mathbf{X})$. ||

Theorem 6.2.13 *Let $f(\mathbf{x}|\theta)$ be the pmf or pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ.*

Proof:

First we show that $T(\mathbf{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set, A_t, as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set $A_t, T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})})$ and, hence, $f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then it can be seen that

$$f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}_{T(\mathbf{x})}|\theta) f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} = g(T(\mathbf{x})|\theta) h(\mathbf{x})$$

and, by the Factorization Theorem, $T(\mathbf{X})$ is a sufficient statistic for θ.

Now to show that $T(\mathbf{X})$ is minimal, let $T'(\mathbf{X})$ be any other sufficient statistic. By the Factorization Theorem, there exist functions g' and h' such that $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta) h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta) h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal. □

Example 6.2.14 (Normal minimal sufficient statistic) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, both μ and σ^2 unknown. Let $\mathbf{x}$ and $\mathbf{y}$ denote two sample points, and let $(\bar{x}, s_{\mathbf{x}}^2)$ and $(\bar{y}, s_{\mathbf{y}}^2)$ be the sample means and variances corresponding to the $\mathbf{x}$ and $\mathbf{y}$ samples, respectively. Then, using (6.2.5), we see that the ratio of densities is

$$\begin{aligned}\frac{f(\mathbf{x}|\mu,\sigma^2)}{f(\mathbf{y}|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{x}-\mu)^2+(n-1)s_{\mathbf{x}}^2\right]/(2\sigma^2)\right)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\left[n(\bar{y}-\mu)^2+(n-1)s_{\mathbf{y}}^2\right]/(2\sigma^2)\right)} \\ &= \exp\left(\left[-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_{\mathbf{x}}^2-s_{\mathbf{y}}^2)\right]/(2\sigma^2)\right).\end{aligned}$$

This ratio will be constant as a function of μ and σ^2 if and only if $\bar{x}=\bar{y}$ and $s_{\mathbf{x}}^2 = s_{\mathbf{y}}^2$. Thus, by Theorem 6.2.13, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2). ||

Example 6.2.15 (Uniform minimal sufficient statistic) Suppose $X_1, \ldots, X_n$ are iid uniform observations on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Then the joint pdf of $\mathbf{X}$ is

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \theta < x_i < \theta+1,\ i=1,\ldots,n, \\ 0 & \text{otherwise,} \end{cases}$$

which can be written as

$$f(\mathbf{x}|\theta) = \begin{cases} 1 & \max_i x_i - 1 < \theta < \min_i x_i \\ 0 & \text{otherwise.} \end{cases}$$

Thus, for two sample points $\mathbf{x}$ and $\mathbf{y}$, the numerator and denominator of the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ will be positive for the same values of θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. And, if the minima and maxima are equal, then the ratio is constant and, in fact, equals 1. Thus, letting $X_{(1)} = \min_i X_i$ and $X_{(n)} = \max_i X_i$, we have that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic. This is a case in which the dimension of a minimal sufficient statistic does not match the dimension of the parameter. ||

A minimal sufficient statistic is not unique. Any one-to-one function of a minimal sufficient statistic is also a minimal sufficient statistic.

$(X_{(n)} - X_{(1)}, (X_{(n)} + X_{(1)})/2)$ is also a minimal sufficient statistic in Example 6.2.15

$(\Sigma_{i=1}^n X_i, \Sigma_{i=1}^n X_i^2)$ is also a minimal sufficient statistic in Example 6.2.14.

sufficient statistics contain all the information about θ that is available in the sample we introduce a different sort of statistic, one that has a complementary purpose.

Definition 6.2.16 A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic.*

Alone, an ancillary statistic contains no information about θ. An ancillary statistic is an observation on a random variable whose distribution is fixed and known, unrelated to θ. Paradoxically, an ancillary statistic, when used in conjunction with other statistics, sometimes does contain valuable information for inferences about θ.

Example 6.2.17 (Uniform ancillary statistic) As in Example 6.2.15, let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta+1), -\infty < \theta < \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics from the sample. We show below that the range statistic, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic by showing that the pdf of R does not depend on θ. Recall that the cdf of each X_i is

$$F(x|\theta) = \begin{cases} 0 & x \le \theta \\ x - \theta & \theta < x < \theta + 1 \\ 1 & \theta + 1 \le x. \end{cases}$$

Thus, the joint pdf of $X_{(1)}$ and $X_{(n)}$, is

$$g(x_{(1)}, x_{(n)}|\theta) = \begin{cases} n(n-1)(x_{(n)} - x_{(1)})^{n-2} & \theta < x_{(1)} < x_{(n)} < \theta + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Making the transformation $R = X_{(n)} - X_{(1)}$ and $M = (X_{(1)} + X_{(n)})/2$, which has the inverse transformation $X_{(1)} = (2M - R)/2$ and $X_{(n)} = (2M + R)/2$ with Jacobian 1, we see that the joint pdf of R and M is

$$h(r, m|\theta) = \begin{cases} n(n-1)r^{n-2} & 0 < r < 1, \theta + (r/2) < m < \theta + 1 - (r/2) \\ 0 & \text{otherwise.} \end{cases}$$

(Notice the rather involved region of positivity for $h(r, m|\theta)$.) Thus, the pdf for R is

$$\begin{aligned} h(r|\theta) &= \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2}\,dm \\ &= n(n-1)r^{n-2}(1-r), \quad 0 < r < 1. \end{aligned}$$

This is a beta pdf with $\alpha = n - 1$ and $\beta = 2$. More important, the pdf is the same for all θ. Thus, the distribution of R does not depend on θ, and R is ancillary. ‖

The ancillarity of R does not depend on the uniformity of the X_is, but rather on the parameter of the distribution being a location parameter.

Example 6.2.18 (Location family ancillary statistic) Let $X_1, \ldots, X_n$ be iid observations from a location parameter family with cdf $F(x-\theta), -\infty < \theta < \infty$. We will show that the range, $R = X_{(n)} - X_{(1)}$, is an ancillary statistic. We work with $Z_1, \ldots, Z_n$ iid observations from $F(x)$ (corresponding to $\theta = 0$) with $X_1 = Z_1 + \theta, \ldots, X_n = Z_n + \theta$. Thus the cdf of the range statistic, R, is

$$\begin{aligned} F_R(r|\theta) &= P_\theta(R \le r) \\ &= P_\theta(\max_i X_i - \min_i X_i \le r) \\ &= P_\theta(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r) \\ &= P_\theta(\max_i Z_i - \min_i Z_i + \theta - \theta \le r) \end{aligned}$$

$$F_R(r|\theta) = P_\theta(\max_i Z_i - \min_i Z_i \le r).$$

The last probability does not depend on θ because the distribution of $Z_1, \ldots, Z_n$ does not depend on θ. Thus, the cdf of R does not depend on θ and, hence, R is an ancillary statistic. ||

Example 6.2.19 (Scale family ancillary statistic) Scale parameter families also have certain kinds of ancillary statistics. Let $X_1, \ldots, X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma), \sigma > 0$. Then any statistic that depends on the sample only through the $n-1$ values $X_1/X_n, \ldots, X_{n-1}/X_n$ is an ancillary statistic. For example,

$$\frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$$

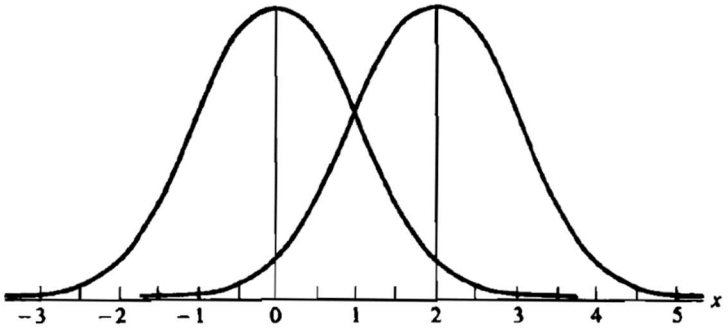
is an ancillary statistic. To see this fact, let $Z_1, \ldots, Z_n$ be iid observations from $F(x)$ (corresponding to $\sigma = 1$) with $X_i = \sigma Z_i$. The joint cdf of $X_1/X_n, \ldots, X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1, \ldots, y_{n-1}|\sigma) &= P_\sigma(X_1/X_n \le y_1, \ldots, X_{n-1}/X_n \le y_{n-1}) \\ &= P_\sigma(\sigma Z_1/(\sigma Z_n) \le y_1, \ldots, \sigma Z_{n-1}/(\sigma Z_n) \le y_{n-1}) \\ &= P_\sigma(Z_1/Z_n \le y_1, \ldots, Z_{n-1}/Z_n \le y_{n-1}). \end{aligned}$$

The last probability does not depend on σ because the distribution of $Z_1, \ldots, Z_n$ does not depend on σ. So the distribution of $X_1/X_n, \ldots, X_{n-1}/X_n$ is independent of σ, as is the distribution of any function of these quantities.

In particular, let X_1 and X_2 be iid n$(0, \sigma^2)$ observations. From the above result, we see that X_1/X_2 has a distribution that is the same for every value of σ. But, in we saw that, if $\sigma = 1, X_1/X_2$ has a Cauchy$(0, 1)$ distribution. Thus, for any $\sigma > 0$, the distribution of X_1/X_2 is this same Cauchy distribution. ‖

Definition 3.5.2 Let $f(x)$ be any pdf. Then the family of pdfs $f(x - \mu)$, indexed by the parameter μ, $-\infty < \mu < \infty$, is called the *location family with standard pdf* $f(x)$ and μ is called the *location parameter* for the family.



Two members of the same location family: means at 0 and 2

It is clear from Figure that the area under the graph of $f(x)$ between $x = -1$ and $x = 2$ is the same as the area under the graph of $f(x - \mu)$ between $x = \mu - 1$ and $x = \mu + 2$. Thus if X is a random variable with pdf $f(x - \mu)$, we can write

$$P(-1 \leq X \leq 2|0) = P(\mu - 1 \leq X \leq \mu + 2|\mu),$$

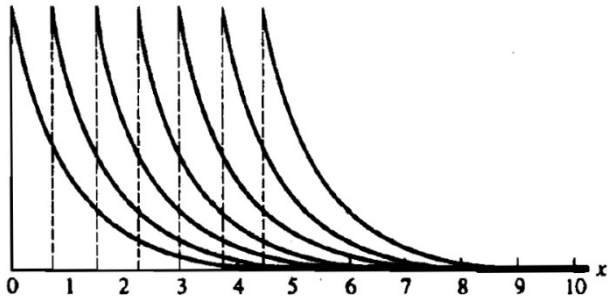
where the random variable X has pdf $f(x - 0) = f(x)$ on the left of the equality and pdf $f(x - \mu)$ on the right.

If X is a random variable with pdf $f(x-\mu)$, then X may be represented as $X = Z+\mu$, where Z is a random variable with pdf $f(z)$. This representation is a consequence of Theorem 3.5.6 (with $\sigma = 1$), which will be proved later.

Example 3.5.3 (Exponential location family) Let $f(x) = e^{-x}$, $x \geq 0$, and $f(x) = 0$, $x < 0$. To form a location family we replace x with $x - \mu$ to obtain

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & x - \mu \geq 0 \\ 0 & x - \mu < 0 \end{cases}$$
$$= \begin{cases} e^{-(x-\mu)} & x \geq \mu \\ 0 & x < \mu. \end{cases}$$

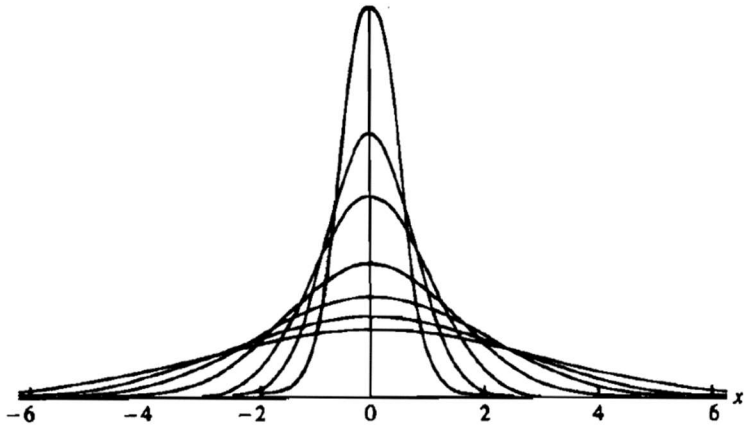
Graphs of $f(x|\mu)$ for various values of μ are shown in Figure
the graph has been shifted. Now the positive part of the graph starts at μ rather than at 0. If X measures time, then μ might be restricted to be nonnegative so that X will be positive with probability 1 for every value of μ. In this type of model, where μ denotes a bound on the range of X, μ is sometimes called a *threshold parameter.* ||



Exponential location densities

Definition 3.5.4 Let $f(x)$ be any pdf. Then for any $\sigma > 0$, the family of pdfs $(1/\sigma)f(x/\sigma)$, indexed by the parameter σ, is called the *scale family with standard pdf* $f(x)$ and σ is called the *scale parameter* of the family.

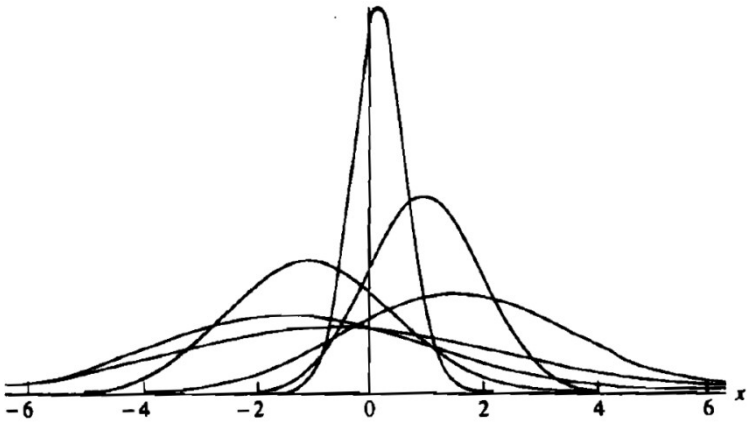
The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. This is illustrated in Figure Most often when scale parameters are used, $f(x)$ is either symmetric about 0 or positive only for $x > 0$. In these cases the stretching is either symmetric about 0 or only in the positive direction. But, in the definition, any pdf may be used as the standard.



Members of the same scale family

Definition 3.5.5 Let $f(x)$ be any pdf. Then for any μ, $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of pdfs $(1/\sigma)f((x-\mu)/\sigma)$, indexed by the parameter (μ, σ), is called the *location–scale family with standard pdf* $f(x)$; μ is called the *location parameter* and σ is called the *scale parameter.*

The effect of introducing both the location and scale parameters is to stretch ($\sigma > 1$) or contract ($\sigma < 1$) the graph with the scale parameter and then shift the graph so that the point that was above 0 is now above μ. Figure illustrates this transformation of $f(x)$. The normal and double exponential families are examples of location–scale families. the Cauchy as a location–scale family.



Members of the same location–scale family

Theorem 3.5.6 *Let $f(\cdot)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$.*

Proof: To prove the "if" part, define $g(z) = \sigma z + \mu$. Then $X = g(Z)$, g is a monotone function, $g^{-1}(x) = (x-\mu)/\sigma$, and $\left|(d/dx)g^{-1}(x)\right| = 1/\sigma$. Thus the pdf of X is

$$f_X(x) = f_Z\left(g^{-1}(x)\right)\left|\frac{d}{dx}g^{-1}(x)\right| = f\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

To prove the "only if" part, define $g(x) = (x-\mu)/\sigma$ and let $Z = g(X)$. $g^{-1}(z) = \sigma z + \mu$, $\left|(d/dz)g^{-1}(z)\right| = \sigma$, and the pdf of Z is

$$f_Z(z) = f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz}g^{-1}(z)\right| = \frac{1}{\sigma}f\left(\frac{(\sigma z+\mu)-\mu}{\sigma}\right)\sigma = f(z).$$

Also,

$$\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X. \qquad \square$$

Theorem 3.5.7 *Let Z be a random variable with pdf $f(z)$. Suppose* $\mathrm{E}Z$ *and* $\mathrm{Var}\,Z$ *exist. If X is a random variable with pdf $(1/\sigma)f((x-\mu)/\sigma)$, then*

$$\mathrm{E}X = \sigma\mathrm{E}Z + \mu \quad \text{and} \quad \mathrm{Var}\,X = \sigma^2\mathrm{Var}\,Z.$$

In particular, if $\mathrm{E}Z = 0$ *and* $\mathrm{Var}\,Z = 1$, *then* $\mathrm{E}X = \mu$ *and* $\mathrm{Var}\,X = \sigma^2$.

Proof: By Theorem 3.5.6, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So $\mathrm{E}X = \sigma\mathrm{E}Z^* + \mu = \sigma\mathrm{E}Z + \mu$ and $\mathrm{Var}\,X = \sigma^2\mathrm{Var}\,Z^* = \sigma^2\mathrm{Var}\,Z$. $\square$

Probabilities for any member of a location–scale family may be computed in terms of the standard variable Z because

$$P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = P\left(Z \le \frac{x-\mu}{\sigma}\right).$$

Thus, if $P(Z \le z)$ is tabulated or easily calculable for the standard variable Z, then probabilities for X may be obtained. Calculations of normal probabilities using the standard normal table are examples of this.

Definition 6.2.21 Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called *complete* if $\mathrm{E}_\theta g(T) = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a *complete statistic.*

Notice that completeness is a property of a family of probability distributions, not of a particular distribution. For example, if X has a $\mathrm{n}(0,1)$ distribution, then defining $g(x) = x$, we have that $\mathrm{E}g(X) = \mathrm{E}X = 0$. But the function $g(x) = x$ satisfies $P(g(X) = 0) = P(X = 0) = 0$, not 1. However, this is a particular distribution, not a family of distributions. If X has a $\mathrm{n}(\theta, 1)$ distribution, $-\infty < \theta < \infty$, we shall see that no function of X, except one that is 0 with probability 1 for all θ, satisfies $\mathrm{E}_\theta g(X) = 0$ for all θ. Thus, the family of $\mathrm{n}(\theta, 1)$ distributions, $-\infty < \theta < \infty$, is complete.

Example 6.2.22 (Binomial complete sufficient statistic) Suppose that T has a binomial(n,p) distribution, $0 < p < 1$. Let g be a function such that $\mathrm{E}_p g(T) = 0$. Then

$$0 = \mathrm{E}_p g(T) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t}$$

$$= (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t$$

for all p, $0 < p < 1$. The factor $(1-p)^n$ is not 0 for any p in this range. Thus it must be that

$$0 = \sum_{t=0}^{n} g(t) \binom{n}{t} \left(\frac{p}{1-p}\right)^t = \sum_{t=0}^{n} g(t) \binom{n}{t} r^t$$

for all r, $0 < r < \infty$. But the last expression is a polynomial of degree n in r, where the coefficient of r^t is $g(t)\binom{n}{t}$. For the polynomial to be 0 for all r, each coefficient must be 0. Since none of the $\binom{n}{t}$ terms is 0, this implies that $g(t) = 0$ for $t = 0, 1, \ldots, n$. Since T takes on the values $0, 1, \ldots, n$ with probability 1, this yields that $P_p(g(T) = 0) = 1$ for all p, the desired conclusion. Hence, T is a complete statistic. ||

Example 6.2.23 (Uniform complete sufficient statistic) Let $X_1, \ldots, X_n$ be iid uniform$(0,\theta)$ observations, $0 < \theta < \infty$. Using an argument similar to that in Example 6.2.8, we can see that $T(\mathbf{X}) = \max_i X_i$ is a sufficient statistic and, by Theorem 5.4.4, the pdf of $T(\mathbf{X})$ is

$$f(t|\theta) = \begin{cases} nt^{n-1}\theta^{-n} & 0 < t < \theta \\ 0 & \text{otherwise.} \end{cases}$$

Suppose $g(t)$ is a function satisfying $\mathrm{E}_\theta g(T) = 0$ for all θ. Since $\mathrm{E}_\theta g(T)$ is constant as a function of θ, its derivative with respect to θ is 0. Thus we have that

$$0 = \frac{d}{d\theta}\mathrm{E}_\theta g(T) = \frac{d}{d\theta}\int_0^\theta g(t) n t^{n-1}\theta^{-n} dt$$

$$= (\theta^{-n})\frac{d}{d\theta}\int_0^\theta n g(t) t^{n-1} dt + \left(\frac{d}{d\theta}\theta^{-n}\right)\int_0^\theta n g(t) t^{n-1} dt$$

$$= \theta^{-n} n g(\theta)\theta^{n-1} + 0 \qquad \begin{pmatrix}\text{applying the product}\\ \text{rule for differentiation}\end{pmatrix}$$

$$= \theta^{-1} n g(\theta).$$

The first term in the next to last line is the result of an application of the Fundamental Theorem of Calculus. The second term is 0 because the integral is, except for a constant, equal to $\mathrm{E}_\theta g(T)$, which is 0. Since $\theta^{-1} n g(\theta) = 0$ and $\theta^{-1} n \neq 0$, it must be that $g(\theta) = 0$. This is true for every $\theta > 0$; hence, T is a complete statistic.

note that the Fundamental Theorem of Calculus does
not apply to all functions, but only to functions that are *Riemann-integrable.* The equation

$$\frac{d}{d\theta} \int_0^\theta g(t)dt = g(\theta)$$

is valid only at points of continuity of Riemann-integrable g. Thus, strictly speaking, the above argument does not show that T is a complete statistic, since the condition of completeness applies to all functions, not just Riemann-integrable ones. From a more practical view, however, this distinction is not of concern since the condition of Riemann-integrability is so general that it includes virtually any function we could think of.) ‖

Theorem 6.2.24 (Basu's Theorem) *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof: We give the proof only for discrete distributions.

Let $S(\mathbf{X})$ be any ancillary statistic. Then $P(S(\mathbf{X}) = s)$ does not depend on θ since $S(\mathbf{X})$ is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x}\colon S(\mathbf{x}) = s\} | T(\mathbf{X}) = t),$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic (recall the definition!). Thus, to show that $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that

$$P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s) \tag{6.2.6}$$

for all possible values $t \in \mathcal{T}$. Now,

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) P_\theta(T(\mathbf{X}) = t).$$

Furthermore, since $\sum_{t \in \mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = \sum_{t \in \mathcal{T}} P(S(\mathbf{X}) = s) P_\theta(T(\mathbf{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\mathbf{X}) = s | T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s),$$

the above two equations show that

$$\mathrm{E}_\theta g(T) = \sum_{t \in \mathcal{T}} g(t) P_\theta(T(\mathbf{X}) = t) = 0 \quad \text{for all } \theta.$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence (6.2.6) is verified. □

Basu's Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

Theorem 6.2.25 (Complete statistics in the exponential family) *Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right), \tag{6.2.7}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in $\Re^k$.

The condition that the parameter space contain an open set is needed to avoid a situation like the following. The $\mathrm{n}(\theta, \theta^2)$ distribution can be written in the form (6.2.7); however, the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistic $T(\mathbf{X})$ that is an unbiased estimator of 0
(Recall that exponential families such as the $\mathrm{n}(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called *curved exponential families*)

Example 6.2.26 Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n}.$$

We first note that the exponential distributions form a scale parameter family and thus, by Example 6.2.19, $g(\mathbf{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so, by Theorem 6.2.25,

$$T(\mathbf{X}) = \sum_{n} X_i$$

is a complete statistic and, by Theorem 6.2.10, $T(\mathbf{X})$ is a sufficient statistic. (As noted below, we need not verify that $T(\mathbf{X})$ is minimal, although it could easily be verified using Theorem 6.2.13.) Hence, by Basu's Theorem, $T(\mathbf{X})$ and $g(\mathbf{X})$ are independent. Thus we have

$$\theta = \mathrm{E}_\theta X_n = \mathrm{E}_\theta T(\mathbf{X})g(\mathbf{X}) = (\mathrm{E}_\theta T(\mathbf{X}))(\mathrm{E}_\theta g(\mathbf{X})) = n\theta \mathrm{E}_\theta g(\mathbf{X}).$$

Hence, for any $\theta, \mathrm{E}_\theta g(\mathbf{X}) = n^{-1}$. ‖

Theorem 6.2.28 *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.*

Definition 2.1.1 A statistics $T(X)$ is called an unbiased estimator for a function of the parameter $g(\theta)$, provided that for every choice of θ,

$$\mathrm{E}T(X) = g(\theta) \tag{2.1.1}$$

Any estimator that is not unbiased is called biased. The bias is denoted by $b(\theta)$.

$$b(\theta) = \mathrm{E}T(X) - g(\theta) \tag{2.1.2}$$

We will now define mean square error (mse)

$$\begin{aligned} MSE[T(X)] &= \mathrm{E}[T(X) - g(\theta)]^2 \\ &= \mathrm{E}[T(X) - \mathrm{E}T(X) + b(\theta)]^2 \\ &= \mathrm{E}[T(X) - \mathrm{E}T(X)]^2 + 2b(\theta)\mathrm{E}[T(X) - \mathrm{E}T(X)] + b^2(\theta) \\ &= \mathrm{V}[T(X)] + b^2(\theta) \\ &= \text{Variance of } [T(X)] + [\text{bias of } T(X)]^2 \end{aligned}$$

Example 2.1.1 Let $(X_1, X_2, \ldots, X_n)$ be Bernoulli rvs with parameter θ, where θ is unknown. $\bar{X}$ is an estimator for θ. Is it unbiased ?

$$\mathrm{E}\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{n\theta}{n} = \theta$$

Thus, $\bar{X}$ is an unbiased estimator for θ.
We denote it as $\hat{\theta} = \bar{X}$.

$$\mathrm{Var}(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{V}(X_i) = \frac{n\theta(1-\theta)}{n^2} = \frac{\theta(1-\theta)}{n}$$

Example 2.1.2 Let $X_i (i = 1, 2, \ldots, n)$ be iid rvs from $N(\mu, \sigma^2)$, where μ and σ^2 are unknown.

Define $nS^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$ and $n\sigma^2 = \sum_{i=1}^{n}(X_i - \mu)^2$
Consider

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}(X_i - \bar{X} + \bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + 2\sum_{i=1}^{n}(X_i - \mu)(\bar{X} - \mu) + n(\bar{X} - \mu)^2 \\ &= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2\end{aligned}$$

Therefore, $$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

$$\begin{aligned}\mathrm{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= \mathrm{E}\left[\sum_{i=1}^{n}(X_i - \mu)^2\right] - n\mathrm{E}[(\bar{X} - \mu)^2] \\ &= n\sigma^2 - \frac{n\sigma^2}{n} = n\sigma^2 - \sigma^2\end{aligned}$$

Hence, $$\mathrm{E}(S^2) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\left(\frac{n-1}{n}\right)$$

Thus, S^2 is a biased estimator of σ^2.

Hence $$b(\sigma^2) = \sigma^2 - \frac{\sigma^2}{n} - \sigma^2 = -\frac{\sigma^2}{n}$$

Further, $\frac{nS^2}{n-1}$ is an unbiased estimator of σ^2.

Best Unbiased Estimators a comparison of estimators based on MSE considerations may not yield a clear favorite. Indeed, there is no one "best MSE" estimator. Many find this troublesome or annoying, and rather than doing MSE comparisons of candidate estimators, they would rather have a "recommended" one.

If W_1 and W_2 are both unbiased estimators of a parameter θ, that is, $\mathrm{E}_\theta W_1 = \mathrm{E}_\theta W_2 = \theta$, then their mean squared errors are equal to their variances, so we should choose the estimator with the smaller variance. If we can find an unbiased estimator with uniformly smallest variance—a best unbiased estimator—then our task is done.

Suppose W^* is an estimator of θ with $\mathrm{E}_\theta W^* = \tau(\theta) \neq \theta$, and we are interested in investigating the worth of W^*. Consider the class of estimators

$$\mathcal{C}_\tau = \{W\colon \mathrm{E}_\theta W = \tau(\theta)\}.$$

For any $W_1, W_2 \in \mathcal{C}_\tau$, $\mathrm{Bias}_\theta W_1 = \mathrm{Bias}_\theta W_2$, so

$$\mathrm{E}_\theta(W_1 - \theta)^2 - \mathrm{E}_\theta(W_2 - \theta)^2 = \mathrm{Var}_\theta W_1 - \mathrm{Var}_\theta W_2,$$

and MSE comparisons, within the class $\mathcal{C}_\tau$, can be based on variance alone. Thus, although we speak in terms of unbiased estimators, we really are comparing estimators that have the same expected value, $\tau(\theta)$.

Definition 7.3.7 An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if it satisfies $\mathrm{E}_\theta W^* = \tau(\theta)$ for all θ and, for any other estimator W with $\mathrm{E}_\theta W = \tau(\theta)$, we have $\mathrm{Var}_\theta W^* \leq \mathrm{Var}_\theta W$ for all θ. W^* is also called a *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

Example 7.3.8 (Poisson unbiased estimation) Let $X_1, \ldots, X_n$ be iid Poisson(λ), and let $\bar{X}$ and S^2 be the sample mean and variance, respectively. Recall that for the Poisson pmf both the mean and variance are equal to λ. Therefore, we have $\mathrm{E}_\lambda \bar{X} = \lambda$, for all λ,

and $\mathrm{E}_\lambda S^2 = \lambda$, for all λ,

so both $\bar{X}$ and S^2 are unbiased estimators of λ.

Even if we can establish that X is better than S^2, consider the class of estimators

$$W_a(\bar{X}, S^2) = a\bar{X} + (1-a)S^2.$$

For every constant a, $\mathrm{E}_\lambda W_a(\bar{X}, S^2) = \lambda$, so we now have infinitely many unbiased estimators of λ. Even if $\bar{X}$ is better than S^2, is it better than every $W_a(\bar{X}, S^2)$? Furthermore, how can we be sure that there are not other, better, unbiased estimators lurking about? ‖

This example shows some of the problems that might be encountered in trying to find a best unbiased estimator, and perhaps that a more comprehensive approach is desirable. Suppose that, for estimating a parameter $\tau(\theta)$ of a distribution $f(x|\theta)$, we can specify a lower bound, say $B(\theta)$, on the variance of *any* unbiased estimator of $\tau(\theta)$. If we can then find an unbiased estimator W^* satisfying $\mathrm{Var}_\theta W^* = B(\theta)$, we have found a best unbiased estimator. This is the approach taken with the use of the Cramér–Rao Lower Bound.

Theorem 7.3.9 (Cramér–Rao Inequality) *Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying*

$$\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left[W(\mathbf{x})f(\mathbf{x}|\theta)\right]\, d\mathbf{x}$$

(7.3.4) *and*

$$\mathrm{Var}_\theta W(\mathbf{X}) < \infty.$$

Then

$$\mathrm{Var}_\theta\,(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)}. \tag{7.3.5}$$

Proof: The proof of this theorem is elegantly simple and is a clever application of the Cauchy–Schwarz Inequality or, stated statistically, the fact that for any two random variables X and Y,

$$[\mathrm{Cov}(X,Y)]^2 \leq (\mathrm{Var}\,X)(\mathrm{Var}\,Y). \tag{7.3.6}$$

If we rearrange (7.3.6) we can get a lower bound on the variance of X,

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

The cleverness in this theorem follows from choosing X to be the estimator $W(\mathbf{X})$ and Y to be the quantity $\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$ and applying the Cauchy–Schwarz Inequality.

First note that

$$\begin{aligned}
\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}) &= \int_{\mathcal{X}} W(\mathbf{x})\left[\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right]\, dx \\
&= \mathrm{E}_\theta\left[W(\mathbf{X})\frac{\frac{\partial}{\partial\theta}f(\mathbf{X}|\theta)}{f(\mathbf{X}|\theta)}\right] \qquad \text{(multiply by } f(\mathbf{X}|\theta)/f(\mathbf{X}|\theta)) \\
&= \mathrm{E}_\theta\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right], \qquad \text{(property of logs)}
\end{aligned} \tag{7.3.7}$$

which suggests a covariance between $W(\mathbf{X})$ and $\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)$. For it to be a covariance, we need to subtract the product of the expected values, so we calculate $\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)$. But if we apply (7.3.7) with $W(\mathbf{x}) = 1$, we have

$$\text{(7.3.8)} \qquad \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta[1] = 0.$$

Therefore $\mathrm{Cov}_\theta(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta))$ is equal to the expectation of the product, and it follows from (7.3.7) and (7.3.8) that

$$\text{(7.3.9)} \quad \mathrm{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}_\theta\left(W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X}).$$

Also, since $\mathrm{E}_\theta(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)) = 0$ we have

$$\text{(7.3.10)} \qquad \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right) = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right).$$

Using the Cauchy–Schwarz Inequality together with (7.3.9) and (7.3.10), we obtain

$$\mathrm{Var}_\theta\,(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)},$$

proving the theorem. □

Corollary 7.3.10 (Cramér–Rao Inequality, iid case) *If the assumptions of Theorem 7.3.9 are satisfied and, additionally, if $X_1, \ldots, X_n$ are iid with pdf $f(x|\theta)$, then*

$$\mathrm{Var}_\theta\, W(\mathbf{X}) \geq \frac{\left(\frac{d}{d\theta}\mathrm{E}_\theta W(\mathbf{X})\right)^2}{n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right)}.$$

Proof: We only need to show that

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right) = n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right).$$

Since $X_1, \ldots, X_n$ are independent,

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2 = \mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right)^2\right)$$

$$= \mathrm{E}_\theta\left(\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) \qquad \text{(property of logs)}$$

$$= \sum_{i=1}^{n}\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) \qquad \text{(expand the square)}$$

$$+ \sum_{i\neq j}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right). \tag{7.3.11}$$

For $i \neq j$ we have

$$\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right)$$

$$= \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right) \qquad \text{(independence)}$$

$$= 0. \qquad \text{(from (7.3.8))}$$

Therefore the second sum in (7.3.11) is 0, and the first term is

$$\sum_{i=1}^{n}\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right)^2\right) = n\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right), \text{ (identical distributions)}$$

which establishes the corollary. □

The quantity $\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right)$ is called the *information number*, or *Fisher information* of the sample. This terminology reflects the fact that the information number gives a bound on the variance of the best unbiased estimator of θ. As the information number gets bigger and we have more information about θ, we have a smaller bound on the variance of the best unbiased estimator.

In fact, the term *Information Inequality* is an alternative to *Cramér–Rao Inequality*

Before looking at some examples, we present a computational result that aids in the application of this theorem. Its proof is left to Exercise

Lemma 7.3.11 *If $f(x|\theta)$ satisfies*

$$\frac{d}{d\theta}\mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right) = \int\frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)f(x|\theta)\right]dx$$

(true for an exponential family), then

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

Example 7.3.12 (Conclusion of Example 7.3.8) Here $\tau(\lambda) = \lambda$, so $\tau'(\lambda) = 1$. Also, since we have an exponential family, using Lemma 7.3.11 gives us

$$\begin{aligned}
\mathrm{E}_\lambda\left(\left(\frac{\partial}{\partial\lambda}\log\prod_{i=1}^{n} f(X_i|\lambda)\right)^2\right) &= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right) \\
&= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log\left(\frac{e^{-\lambda}\lambda^X}{X!}\right)\right) \\
&= -n\mathrm{E}_\lambda\left(\frac{\partial^2}{\partial\lambda^2}(-\lambda + X\log\lambda - \log X!)\right) \\
&= -n\mathrm{E}_\lambda\left(-\frac{X}{\lambda^2}\right) = \frac{n}{\lambda}.
\end{aligned}$$

Hence for any unbiased estimator, W, of λ, we must have

$$\mathrm{Var}_\lambda W \geq \frac{\lambda}{n}.$$

Since $\mathrm{Var}_\lambda \bar{X} = \lambda/n$, $\bar{X}$ is a best unbiased estimator of λ. ‖

It is important to remember that a key assumption in the Cramér–Rao Theorem is the ability to differentiate under the integral sign, which, of course, is somewhat restrictive. As we have seen, densities in the exponential class will satisfy the assumptions but, in general, such assumptions need to be checked, or contradictions such as the following will arise.

Example 7.3.13 (Unbiased estimator for the scale uniform) Let $X_1, \ldots, X_n$ be iid with pdf $f(x|\theta) = 1/\theta, 0 < x < \theta$. Since $\frac{\partial}{\partial\theta}\log f(x|\theta) = -1/\theta$, we have

$$\mathrm{E}_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = \frac{1}{\theta^2}.$$

The Cramér–Rao Theorem would seem to indicate that if W is any unbiased estimator of θ,

$$\mathrm{Var}_\theta W \geq \frac{\theta^2}{n}.$$

We would now like to find an unbiased estimator with small variance. As a first guess, consider the sufficient statistic $Y = \max(X_1, \ldots, X_n)$, the largest order statistic. The pdf of Y is $f_Y(y|\theta) = ny^{n-1}/\theta^n$, $0 < y < \theta$, so

$$\mathrm{E}_\theta Y = \int_0^\theta \frac{ny^n}{\theta^n}dy = \frac{n}{n+1}\theta,$$

showing that $\frac{n+1}{n}Y$ is an unbiased estimator of θ. We next calculate

$$\begin{aligned}
\mathrm{Var}_\theta\left(\frac{n+1}{n}Y\right) &= \left(\frac{n+1}{n}\right)^2 \mathrm{Var}_\theta Y \\
&= \left(\frac{n+1}{n}\right)^2\left[\mathrm{E}_\theta Y^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\
&= \left(\frac{n+1}{n}\right)^2\left[\frac{n}{n+2}\theta^2 - \left(\frac{n}{n+1}\theta\right)^2\right] \\
&= \frac{1}{n(n+2)}\theta^2
\end{aligned}$$

which is uniformly smaller than θ^2/n. This indicates that the Cramér–Rao Theorem is not applicable to this pdf. To see that this is so, we can use Leibnitz's Rule to calculate

$$\begin{aligned}\frac{d}{d\theta}\int_0^\theta h(x)f(x|\theta)\,dx &= \frac{d}{d\theta}\int_0^\theta h(x)\frac{1}{\theta}\,dx \\ &= \frac{h(\theta)}{\theta} + \int_0^\theta h(x)\frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)\,dx \\ &\neq \int_0^\theta h(x)\frac{\partial}{\partial\theta}f(x|\theta)\,dx,\end{aligned}$$

unless $h(\theta)/\theta = 0$ for all θ. Hence, the Cramér–Rao Theorem does not apply. In general, if the range of the pdf depends on the parameter, the theorem will not be applicable. ‖

Example 7.3.14 (Normal variance bound) Let $X_1, \ldots, X_n$ be iid $\mathrm{n}(\mu, \sigma^2)$, and consider estimation of σ^2, where μ is unknown. The normal pdf satisfies the assumptions of the Cramér–Rao Theorem and Lemma 7.3.11, so we have

$$\frac{\partial^2}{\partial(\sigma^2)^2}\log\left(\frac{1}{(2\pi\sigma^2)^{1/2}}e^{-(1/2)(x-\mu)^2/\sigma^2}\right) = \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}$$

and

$$\begin{aligned}-\mathrm{E}\left(\frac{\partial^2}{\partial(\sigma^2)^2}\log f(X|\mu,\sigma^2)\middle|\mu,\sigma^2\right) &= -\mathrm{E}\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\middle|\mu,\sigma^2\right) \\ &= \frac{1}{2\sigma^4}.\end{aligned}$$

Thus, any unbiased estimator, W, of σ^2 must satisfy

$$\mathrm{Var}(W|\mu,\sigma^2) \geq \frac{2\sigma^4}{n}.$$

$$\mathrm{Var}(S^2|\mu,\sigma^2) = \frac{2\sigma^4}{n-1},$$

so S^2 does not attain the Cramér–Rao Lower Bound. ‖

In the above example we are left with an incomplete answer; that is, is there a better unbiased estimator of σ^2 than S^2, or is the Cramér–Rao Lower Bound unattainable?

The conditions for attainment of the Cramér–Rao Lower Bound are actually quite simple. Recall that the bound follows from an application of the Cauchy–Schwarz Inequality, so conditions for attainment of the bound are the conditions for equality in the Cauchy–Schwarz Inequality

Corollary 7.3.15 (Attainment) *Let $X_1, \ldots, X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér–Rao Theorem. Let $L(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér–Rao Lower Bound if and only if*

$$a(\theta)[W(\mathbf{x}) - \tau(\theta)] = \frac{\partial}{\partial\theta} \log L(\theta|\mathbf{x}) \tag{7.3.12}$$

for some function $a(\theta)$.

Proof: The Cramér–Rao Inequality, as given in (7.3.6), can be written as

$$\left[\mathrm{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(X_i|\theta)\right)\right]^2 \le \mathrm{Var}_\theta W(\mathbf{X}) \mathrm{Var}_\theta \left(\frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(X_i|\theta)\right),$$

and, recalling that $\mathrm{E}_\theta W = \tau(\theta)$, $\mathrm{E}_\theta(\frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(X_i|\theta)) = 0$, and using the results of Theorem 4.5.7, we can have equality if and only if $W(\mathbf{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta} \log \prod_{i=1}^n f(x_i|\theta)$. That is exactly what is expressed in (7.3.12). □

Theorem 4.5.7 *For any random variables X and Y,*

a. $-1 \le \rho_{XY} \le 1$.

b. $|\rho_{XY}| = 1$ *if and only if there exist numbers $a \ne 0$ and b such that $P(Y = aX + b) = 1$. If $\rho_{XY} = 1$, then $a > 0$, and if $\rho_{XY} = -1$, then $a < 0$.*

Example 7.3.16 (Continuation of Example 7.3.14) Here we have

$$L(\mu, \sigma^2|\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2)\Sigma_{i=1}^n (x_i-\mu)^2/\sigma^2},$$

and hence

$$\frac{\partial}{\partial\sigma^2} \log L(\mu, \sigma^2|\mathbf{x}) = \frac{n}{2\sigma^4}\left(\sum_{i=1}^n \frac{(x_i-\mu)^2}{n} - \sigma^2\right).$$

Thus, taking $a(\sigma^2) = n/(2\sigma^4)$ shows that the best unbiased estimator of σ^2 is $\sum_{i=1}^n (x_i - \mu)^2/n$, which is calculable only if μ is known. If μ is unknown, the bound *cannot be attained.* ‖

Definition 1.5.3 A family of distributions $\{F(t|\theta) : \theta \in \Theta\}$ is boundedly complete if

$$\mathrm{E}[g(T)] = \int g(t) f(t) dt = 0 \ \forall \ \theta$$

and real statistics $g(t)$ satisfying $|g(t)| < M$, then $g(t) = 0$.

Theorem 1.5.2 *If a family of distributions is complete then it is boundedly complete.*

Remark The converse of the theorem is not true.

Example 1.5.10 Let T be a random variable with the following probability distribution:

$$P[T = 0] = q \quad \text{and} \quad P[T = i + 1] = p^2 q^i, \ i = 0, 1, 2, \ldots, \ 0 < p < 1, \ q = 1 - p$$

Let $\mathrm{E}[g(T)] = 0$ then

$$\begin{aligned} g(0)q + g(1)p^2 + g(2)p^2 q + g(3)p^2 q^2 + \cdots &= 0 \\ g(1) + g(2)q + g(3)q^2 + \cdots &= -g(0)qp^{-2} \\ &= -g(0)q(1-q)^{-2} \\ &= -g(0)[q + 2q^2 + 3q^3 + \cdots] \end{aligned}$$

This implies that $g(1) = 0$, $g(2) = -g(0)$, $g(3) = -2g(0)$, etc.
Hence, $g(i) = -(i-1)g(0)$
If $g(0) = 0$ then $g(t) = 0$ at all nonnegative integers. Otherwise, the function $g(t)$ is unbounded.

Therefore, there are nondegenerate unbiased estimates of zero but they are none that are bounded. Hence, we conclude that the family of distributions is boundedly complete but not complete.

Definition 2.1.1 A statistics $T(X)$ is called an unbiased estimator for a function of the parameter $g(\theta)$, provided that for every choice of θ,

$$\mathrm{E}T(X) = g(\theta) \tag{2.1.1}$$

Any estimator that is not unbiased is called biased. The bias is denoted by $b(\theta)$.

$$b(\theta) = \mathrm{E}T(X) - g(\theta) \tag{2.1.2}$$

We will now define mean square error (mse)

$$\begin{aligned} MSE[T(X)] &= \mathrm{E}[T(X) - g(\theta)]^2 \\ &= \mathrm{E}[T(X) - \mathrm{E}T(X) + b(\theta)]^2 \\ &= \mathrm{E}[T(X) - \mathrm{E}T(X)]^2 + 2b(\theta)\mathrm{E}[T(X) - \mathrm{E}T(X)] + b^2(\theta) \\ &= \mathrm{V}[T(X)] + b^2(\theta) \\ &= \text{Variance of } [T(X)] + [\text{bias of } T(X)]^2 \end{aligned}$$

Empirical Distribution Function

Let $X_1, X_2, \ldots, X_n$ be a random sample from a continuous population with df F and pdf f. Then the order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ is a sufficient statistics. Define $\hat{F}(x) = \frac{\text{Number of } X_i's \leq x}{n}$, same thing can be written in terms of order statistics as,

$$\hat{F}(x) = \begin{cases} 0 \ ; \ X_{(1)} > x \\ \frac{k}{n} \ ; \ X_{(k)} \leq x < X_{(k+1)} \\ 1 \ ; \ x \geq X_{(n)} \end{cases} = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}(x - X_{(j)})$$

where $I(y) = \begin{cases} 1; \ y \geq 0 \\ 0; \ \text{otherwise} \end{cases}$

Example 2.1.10 Show that empirical distribution function is an unbiased estimator of $F(x)$

$$\hat{F}(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{I}(x - X_{(j)})$$

$$\begin{aligned} \mathrm{E}\hat{F}(x) &= \frac{1}{n} \sum_{j=1}^{n} P[X_{(j)} \leq x] \\ &= \frac{1}{n} \sum_{j=1}^{n} \sum_{k=j}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} \\ &= \frac{1}{n} \sum_{j=1}^{k} \sum_{k=1}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} \\ &= \frac{1}{n} \sum_{k=1}^{n} \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k} \sum_{j=1}^{k} (1) \end{aligned}$$

$$\begin{aligned} &= \frac{1}{n} \sum_{k=1}^{n} k \binom{n}{k} [F(x)]^k [[1 - F(x)]^{n-k} \\ &= \frac{1}{n} [nF(x)] = F(x) \end{aligned}$$

Note: One can see that $\mathbf{I}(x - X_{(j)})$ is a Bernoulli random variable. Then $\mathrm{E}\mathbf{I}(x - X_{(j)}) = F(x)$, so that $\mathrm{E}\hat{F}(x) = F(x)$. We observe that $\hat{F}(x)$ has a Binomial distribution with mean $F(x)$ and variance $\frac{F(x)[1-F(x)]}{n}$. Using central limit theorem, for iid rvs, we can show that as $n \to \infty$

$$\sqrt{n} \left[\frac{\hat{F}(x) - F(x)}{\sqrt{F(x)[1 - F(x)]}} \right] \quad \to \quad N(0, 1).$$

Recall that if X and Y are any two random variables, then, provided the expectations exist, we have

$$\mathrm{E}X = \mathrm{E}[\mathrm{E}(X|Y)],$$

$$\mathrm{Var}\, X = \mathrm{Var}[\mathrm{E}(X|Y)] + \mathrm{E}[\mathrm{Var}(X|Y)].$$

Theorem 2.2.1 *Let $h(X)$ be an unbiased estimator of $g(\theta)$. Let $T(X)$ be a sufficient statistics for θ. Define $\Psi(T) = \mathrm{E}(h|T)$. Then $\mathrm{E}[\Psi(T)] = g(\theta)$ and $\mathrm{V}[\Psi(T)] \leq \mathrm{V}(h) \ \forall \ \theta$.*

This theorem is known as Rao–Blackwell Theorem.

Proof
$$\mathrm{E}[h(X)] = \mathrm{E}[\mathrm{E}h(X)|T = t] = \mathrm{E}[\Psi(T)] = g(\theta) \tag{2.2.6}$$

Hence $\Psi(T)$ is unbiased estimator of $g(\theta)$

$$\begin{aligned}\mathrm{V}[h(X)] &= \mathrm{V}[\mathrm{E}(h(X)|T(X))] + \mathrm{E}[\mathrm{V}(h(X)|T(X))] \\ &= \mathrm{V}[\Psi(T)] + \mathrm{E}[\mathrm{V}(h(X)|T(X))]\end{aligned}$$

Since $\mathrm{V}[h(X)|T(X)] \geq 0$ and $\mathrm{E}[\mathrm{V}(h(X)|T(X))] > 0$

Therefore,
$$\mathrm{V}[\Psi(T)] \leq \mathrm{V}[h(X)] \tag{2.2.7}$$

From the definition of sufficiency, we can conclude that the distribution of $h(X)$ given $T(X)$ is independent of θ. Hence $\Psi(T)$ is an estimator.

Theorem 2.2.2 (Lehmann–Scheffe Theorem) *If T is a complete sufficient statistic and there exists an unbiased estimate h of $g(\theta)$, there exists a unique UMVUE of θ, which is given by* $\mathrm{E}h|T$.

Proof Let h_1 and h_2 be two unbiased estimators of $g(\theta)$ Rao–Blackwell theorem, $\mathrm{E}(h_1|T)$ and $\mathrm{E}(h_2|T)$ are both UMVUE of $g(\theta)$.
Hence $\mathrm{E}[\mathrm{E}(h_1|T) - \mathrm{E}(h_2|T)] = 0$
But T is complete therefore

$$[\mathrm{E}(h_1|T) - \mathrm{E}(h_2|T)] = 0$$

This implies $\mathrm{E}(h_1|T) = \mathrm{E}(h_2|T)$.
Hence, UMVUE is unique.

Example 2.2.4 Let $X_1, X_2, \ldots, X_n$ are iid rvs with $B(n, p)$, $0 < p < 1$. In this case, we have to find the UMVUE of $p^r q^s$, $q = 1 - p$, r , $s \neq 0$ and $P[X \leq c]$. Assume n is known.

Binomial distribution belongs to exponential family. So that $\sum_{i=1}^{n} X_i$ is sufficient and complete for p.

(i) The distribution of T is $B(mn, p)$.

Let $U(t)$ be unbiased estimator for $p^r q^s$.

$$\sum_{t=0}^{nm} u(t) \binom{nm}{t} p^t q^{nm-t} = p^r q^s \tag{2.2.13}$$

$$\sum_{t=0}^{nm} u(t) \binom{nm}{t} p^{t-r} q^{nm-t-s} = 1$$

$$\sum_{t=r}^{nm-s} u(t) \frac{\binom{nm}{t}}{\binom{nm-s-r}{t-r}} \binom{nm-s-r}{t-r} p^{t-r} q^{nm-t-s} = 1$$

Then

$$u(t) \frac{\binom{nm}{t}}{\binom{nm-s-r}{t-r}} = 1$$

Hence

$$u(t) = \begin{cases} \frac{\binom{nm-s-r}{t-r}}{\binom{nm}{t}} \; ; \; t = r, r+1, r+2, \ldots, nm-s \\ 0 \qquad\qquad ; \text{otherwise} \end{cases} \tag{2.2.14}$$

(ii) To find UMVUE of $P[X \leq c]$

Now

$$P[X \leq c] = \sum_{x=0}^{c} \binom{n}{x} p^x q^{n-x}$$

Then UMVUE of

$$p^x q^{n-x} = \frac{\binom{nm-n}{t-x}}{\binom{nm}{t}}$$

Hence UMVUE of $P[X \leq c]$

$$= \begin{cases} \sum_{x=0}^{c} \binom{n}{x} \frac{\binom{nm-n}{t-x}}{\binom{nm}{t}} \; ; \; t = x, x+1, x+2, \ldots, nm-n+x, \quad c \leq \min(t, n) \\ 1 \qquad\qquad\qquad\qquad ; \text{otherwise} \end{cases} \tag{2.2.15}$$

Note: UMVUE of $P[X = x] = \binom{n}{x} p^x q^{n-x}$ is $\frac{\binom{n}{x}\binom{nm-n}{t-x}}{\binom{nm}{t}}$; $x = 0, 1, 2, \ldots, t$

Particular cases:

(a) $r = 1, s = 0$. From (2.2.14), we will get UMVUE of p,

$$u(t) = \frac{\binom{nm-1}{t-1}}{\binom{nm}{t}} = \frac{t}{nm} \tag{2.2.16}$$

(b) $r = 0, s = 1$. From (2.2.14), we will get UMVUE of q,

$$u(t) = \frac{\binom{nm-1}{t}}{\binom{nm}{t}} = \frac{nm - t}{nm} = 1 - \frac{t}{nm} \tag{2.2.17}$$

(c) $r = 1, s = 1$. From (2.2.14), we will get UMVUE of pq,

$$u(t) = \left(\frac{t}{nm}\right)\left(\frac{nm - t}{nm - 1}\right) \tag{2.2.18}$$

Example 2.2.5 Let $X_1, X_2, \ldots, X_m$ are iid rvs with $P(\lambda)$. In this case we have to find UMVUE of (i) $\lambda^r e^{-s\lambda}$ (ii) $P[X \leq c]$

Poisson distribution belongs to exponential family. So that $T = \sum_{i=1}^{n} X_i$ is sufficient and complete for λ.

(i) The distribution of T is $P(m\lambda)$.

Let $U(t)$ be unbiased estimator for $\lambda^r e^{-s\lambda}$

$$\sum_{t=0}^{\infty} u(t) \frac{e^{-m\lambda}(m\lambda)^t}{t!} = e^{-s\lambda}\lambda^r \tag{2.2.19}$$

$$\sum_{t=0}^{\infty} u(t) \frac{e^{-(m-s)\lambda} m^t \lambda^{t-r}}{t!} = 1$$

$$\sum_{t=r}^{\infty} u(t) \frac{m^t}{(m-s)^{t-r}} \frac{(t-r)!}{t!} \frac{e^{-(m-s)\lambda}[(m-s)\lambda]^{t-r}}{(t-r)!} = 1$$

Then

$$u(t) \frac{m^t}{(m-s)^{t-r}} \frac{(t-r)!}{t!} = 1$$

$$u(t) = \begin{cases} \frac{(m-s)^{t-r}}{m^t} \frac{t!}{(t-r)!} & ; \ t = r, r+1, \ldots, s \leq m \\ 0 & ; \text{otherwise} \end{cases} \tag{2.2.20}$$

(ii) To find UMVUE of $P[X \leq c]$

$$P[X \leq c] = \sum_{x=0}^{c} \frac{e^{-\lambda}\lambda^x}{x!}$$

Now, UMVUE of $e^{-\lambda}\lambda^x$ is $\frac{(m-1)^{(t-x)}}{m^t} \frac{t!}{(t-x)!}$

$$\text{UMVUE of } P[X \leq c] = \sum_{x=0}^{c} \frac{t!}{(t-x)!x!} \left(\frac{m-1}{m}\right)^t \left(\frac{1}{m-1}\right)^x$$

$$= \begin{cases} \sum_{x=0}^{c} \binom{t}{x} \left(\frac{1}{m}\right)^x \left(\frac{m-1}{m}\right)^{t-x} & ; c \leq t \\ 1 & ; \text{ otherwise} \end{cases} \tag{2.2.21}$$

Remark UMVUE of $P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}$ is $\binom{t}{x} \left(\frac{1}{m}\right)^x \left(\frac{m-1}{m}\right)^{t-x}$; $x = 0, 1, \ldots, t$
Particular cases:
(a) $s = 0, r = 1$
From (2.2.20), we will get the UMVUE of λ,

$$u(t) = \frac{m^{t-1}t!}{m^t(t-1)!} = \frac{t}{m} \tag{2.2.22}$$

(b) $s = 1, r = 0$
From (2.2.20), we will get the UMVUE of $e^{-\lambda}$,

$$u(t) = \left(\frac{m-1}{m}\right)^t \tag{2.2.23}$$

(c) $s = 1, r = 1$
From (2.2.20), we will get the UMVUE of $\lambda e^{-\lambda}$

$$u(t) = \frac{(m-1)^{t-1}t!}{m^t(t-1)!} = \left(\frac{m-1}{m}\right)^t \frac{t}{m-1} \tag{2.2.24}$$

Remark UMVUE of $\lambda e^{-\lambda} \neq$ (UMVUE of λ)(UMVUE of $e^{-\lambda}$)

Theorem 2. Let $\{f_\theta : \boldsymbol{\theta} \in \Theta\}$ be a k-parameter exponential family given by

(2) $$f_{\boldsymbol{\theta}}(\mathbf{x}) = \exp\left[\sum_{j=1}^{k} Q_j(\boldsymbol{\theta})T_j(\mathbf{x}) + D(\boldsymbol{\theta}) + S(\mathbf{x})\right],$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \Theta$, an interval in $\mathcal{R}_k$, $T_1, T_2, \ldots, T_k$, and S are defined on $\mathcal{R}_n$, $\mathbf{T} = (T_1, T_2, \ldots, T_k)$, and $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, $k \leq n$. Let $\mathbf{Q} = (Q_1, Q_2, \ldots, Q_k)$, and suppose that the range of $\mathbf{Q}$ contains an open set in $\mathcal{R}_k$. Then

$$\mathbf{T} = (T_1(\mathbf{X}), T_2(\mathbf{X}), \ldots, T_k(\mathbf{X}))$$

is a complete sufficient statistic.

Proof. For a complete proof in a general setting, we refer the reader to Lehmann [63, pp. 142–143]. Essentially, the unicity of the Laplace transform is used on the probability distribution induced by $\mathbf{T}$. We will content ourselves here by proving the result for the $k = 1$ case when f_θ is a PMF.

Let us write $Q(\theta) = \theta$ in (2), and let $(\alpha, \beta) \subseteq \Theta$. We wish to show that

$$E_\theta g(T(\mathbf{X})) = \sum_t g(t)\, P_\theta\{T(\mathbf{X}) = t\}$$

$$= \sum_t g(t) \exp[\theta t + D(\theta) + S^*(t)] = 0 \qquad \text{for all } \theta \tag{3}$$

implies that $g(t) = 0$.

Let us write $x^+ = x$ if $x \geq 0, = 0$ if $x < 0$, and $x^- = -x$ if $x < 0, = 0$ if $x \geq 0$. Then $g(t) = g^+(t) - g^-(t)$, and both g^+ and g^- are nonnegative functions. In terms of g^+ and g^-, (3) is the same as

$$\sum_t g^+(t) e^{\theta t + S^*(t)} = \sum_t g^-(t) e^{\theta t + S^*(t)} \tag{4}$$

for all θ.

Let $\theta_0 \in (\alpha, \beta)$ be fixed, and write

$$p^+(t) = \frac{g^+(t) e^{\theta_0 t + S^*(t)}}{\sum_t g^+(t) e^{\theta_0 t + S^*(t)}} \quad \text{and} \quad p^-(t) = \frac{g^-(t) e^{\theta_0 t + S^*(t)}}{\sum_t g^-(t) e^{\theta_0 t + S^*(t)}}. \tag{5}$$

Then both p^+ and p^- are PMFs, and it follows from (4) that

$$\sum_t e^{\delta t} p^+(t) = \sum_t e^{\delta t} p^-(t) \tag{6}$$

for all $\delta \in (\alpha - \theta_0, \beta - \theta_0)$. By the uniqueness of MGFs (6) implies that

$$p^+(t) = p^-(t) \qquad \text{for all } t$$

and hence that $\quad g^+(t) = g^-(t) \qquad \text{for all } t,$

which is equivalent to $\quad g(t) = 0 \qquad \text{for all } t.$

Since T is clearly sufficient (by the factorization criterion), it is proved that T is a complete sufficient statistic.

Example 15. Let $X_1, X_2, \ldots, X_n$ be iid $\mathcal{N}(\mu, \sigma^2)$ RVs where both μ and σ^2 are unknown. We know that the family of distributions of $\mathbf{X} = (X_1, \ldots, X_n)$ is a two-parameter exponential family with $T(X_1, \ldots, X_n) = (\sum_1^n X_i, \sum_1^n X_i^2)$. From Theorem 2 it follows that T is a complete sufficient statistic. Examples 10 and 11 fall in the domain of Theorem 2.

Example 10. Let $X_1, X_2, \ldots, X_n$ be iid $b(1, p)$ RVs. Then $T = \sum_1^n X_i$ is a sufficient statistic. We show that T is also complete; that is, the family of distributions of T, $\{b(n, p), 0 < p < 1\}$, is complete.

Example 11. Let X be $\mathcal{N}(0, \theta)$. Then the family of PDFs $\{\mathcal{N}(0, \theta), \theta > 0\}$ is not complete since $EX = 0$ and $g(x) = x$ is not identically zero. Note that $T(X) = X^2$ is complete, for the PDF of $X^2 \sim \theta\chi^2(1)$ is given by

$$f(t) = \begin{cases} \dfrac{e^{-t/2\theta}}{\sqrt{2\pi\theta t}}, & t > 0, \\ 0, & \text{otherwise.} \end{cases}$$

$$E_\theta g(T) = \frac{1}{\sqrt{2\pi\theta}} \int_0^\infty g(t) t^{-1/2} e^{-t/2\theta}\, dt = 0 \qquad \text{for all } \theta > 0,$$

which holds if and only if $\int_0^\infty g(t) t^{-1/2} e^{-t/2\theta}\, dt = 0$, and using the uniqueness property of Laplace transforms, it follows that

$$g(t) t^{-1/2} = 0 \qquad \text{for all } t > 0,$$

that is,

$$g(t) = 0.$$

UNIT 4: EXPONENTIAL FAMILY

Theorem 6.2.24 (Basu's Theorem) *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic.*

Proof: We give the proof only for discrete distributions.

Let $S(\mathbf{X})$ be any ancillary statistic. Then $P(S(\mathbf{X}) = s)$ does not depend on θ since $S(\mathbf{X})$ is ancillary. Also the conditional probability,

$$P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x}\colon S(\mathbf{x}) = s\}|T(\mathbf{X}) = t),$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic (recall the definition!). Thus, to show that $S(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that

$$P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) = P(S(\mathbf{X}) = s) \tag{6.2.6}$$

for all possible values $t \in \mathcal{T}$. Now,

$$P(S(\mathbf{X}) = s) = \sum_{t\in\mathcal{T}} P(S(\mathbf{X}) = s|T(\mathbf{X}) = t)P_\theta(T(\mathbf{X}) = t).$$

Furthermore, since $\sum_{t\in\mathcal{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(S(\mathbf{X}) = s) = \sum_{t\in\mathcal{T}} P(S(\mathbf{X}) = s)P_\theta(T(\mathbf{X}) = t).$$

Therefore, if we define the statistic

$$g(t) = P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s),$$

the above two equations show that

$$\mathrm{E}_\theta g(T) = \sum_{t\in\mathcal{T}} g(t)P_\theta(T(\mathbf{X}) = t) = 0 \quad \text{for all } \theta.$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence (6.2.6) is verified. □

Basu's Theorem is useful in that it allows us to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

Theorem 6.2.25 (Complete statistics in the exponential family) *Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp\left(\sum_{j=1}^{k} w(\theta_j)t_j(x)\right), \tag{6.2.7}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. Then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is complete as long as the parameter space Θ contains an open set in $\Re^k$.

The condition that the parameter space contain an open set is needed to avoid a situation like the following. The $n(\theta, \theta^2)$ distribution can be written in the form (6.2.7); however, the parameter space (θ, θ^2) does not contain a two-dimensional open set, as it consists of only the points on a parabola. As a result, we can find a transformation of the statistic $T(\mathbf{X})$ that is an unbiased estimator of 0
(Recall that exponential families such as the $n(\theta, \theta^2)$, where the parameter space is a lower-dimensional curve, are called *curved exponential families*)

Example 6.2.26 Let $X_1, \ldots, X_n$ be iid exponential observations with parameter θ. Consider computing the expected value of

$$g(\mathbf{X}) = \frac{X_n}{X_1 + \cdots + X_n}.$$

We first note that the exponential distributions form a scale parameter family and thus, by Example 6.2.19, $g(\mathbf{X})$ is an ancillary statistic. The exponential distributions also form an exponential family with $t(x) = x$ and so, by Theorem 6.2.25,

$$T(\mathbf{X}) = \sum_{n} X_i$$

is a complete statistic and, by Theorem 6.2.10, $T(\mathbf{X})$ is a sufficient statistic. (As noted below, we need not verify that $T(\mathbf{X})$ is minimal, although it could easily be verified using Theorem 6.2.13.) Hence, by Basu's Theorem, $T(\mathbf{X})$ and $g(\mathbf{X})$ are independent. Thus we have

$$\theta = \mathrm{E}_\theta X_n = \mathrm{E}_\theta T(\mathbf{X}) g(\mathbf{X}) = (\mathrm{E}_\theta T(\mathbf{X}))(\mathrm{E}_\theta g(\mathbf{X})) = n\theta \mathrm{E}_\theta g(\mathbf{X}).$$

Hence, for any $\theta, \mathrm{E}_\theta g(\mathbf{X}) = n^{-1}$. ‖

Theorem 6.2.28 *If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.*

BLOCK 2: ESTIMATION THEORY

UNIT 1: METHODS OF ESTIMATION

1.1 METHOD OF MOMENTS

The method of moments is, perhaps, the oldest method of finding point estimators, dating back at least to Karl Pearson in the late 1800s. The method of moments essentially amounts to equating the sample moments and corresponding population moments and solving the resulting equations for the parameters to be determined.

Let $f(\cdot;\theta_1,\theta_2,\dots,\theta_k)$ be a density of a random variable X having k parameters $\theta_1,\theta_2,\dots,\theta_k$. Further, let μ_r' be the r^{th} moment about origin, where μ_r' is a known function of the k parameters $\theta_1,\theta_2,\dots,\theta_k$, i.e.

$$\mu_r' = \mu_r'(\theta_1,\theta_2,\dots,\theta_k) = E[X^r]$$

Let $X_1, X_2, \dots, X_n$ be a random sample from the density $f(\cdot;\theta_1,\theta_2,\dots,\theta_k)$ and let M_j' be the j[th] sample moment, i.e.

$$M_j' = \frac{1}{n}\sum_{i=1}^{n} X_i^j \qquad ; i = 1,2,\dots,n$$

Form the k equations,

$$M_j' = \mu_j'(\theta_1,\theta_2,\dots,\theta_k) \qquad ; j = 1,2,\dots,k \qquad (1)$$

in the k variables $\theta_1,\theta_2,\dots,\theta_k$, and let $\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_k$ be their *unique* solution, i.e. $\hat{\Theta}_j$ estimates $\theta_j; \forall\, j = 1,2,\dots,k$. Eq. (1) is obtained by using the first k raw moments.

The estimator $(\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_k)$ is the estimator of $(\theta_1,\theta_2,\dots,\theta_k)$ obtained by the *method-of-moments*. The estimators $(\hat{\Theta}_1, \hat{\Theta}_2, \dots, \hat{\Theta}_k)$ were obtained by replacing population moments by sample moments.

The method-of-moments estimators are not uniquely defined.

If instead of estimating $(\theta_1, \theta_2, \ldots, \theta_k)$, method-of-moments estimators of, say $\tau_1(\theta_1, \theta_2, \ldots, \theta_K), \tau_2(\theta_1, \theta_2, \ldots, \theta_K), \ldots, \tau_r(\theta_1, \theta_2, \ldots, \theta_K)$ are desired, they can be obtained in several ways. One way would be to first find method-of-moments estimates, say $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, of $\theta_1, \theta_2, \ldots, \theta_k$ and then use $\tau_j(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ as an estimate of $\tau_j(\theta_1, \theta_2, \ldots, \theta_k)$ for $j = 1, 2, \ldots, r$. Another way would be to form the equations

$$M_j' = \mu_j'(\tau_1, \tau_2, \ldots, \tau_r) \qquad ; j = 1, 2, \ldots, r$$

and solve them for $\tau_1, \tau_2, \ldots, \tau_r$. Estimators obtained using either way are called method-of-moments estimators and may not be the same in both cases.

Example 2.1.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Let $(\theta_1, \theta_2) = (\mu, \sigma)$. Estimate the parameters μ and σ by the method of moments.

Since we know that,

$$\mu = E[X] = \mu_1'$$

and $$\sigma^2 = E[X^2] - \{E[X]\}^2 = \mu_2' - (\mu_1')^2$$

$\Rightarrow$ $$\mu_2' = \sigma^2 + \mu^2.$$

Using method-of-moments, we have the following equations

$$M_1' = \mu_1' = \mu_1'(\mu, \sigma) = \mu$$

and $$M_2' = \mu_2' = \mu_2'(\mu, \sigma) = \sigma^2 + \mu^2.$$

Thus, the method-of-moments estimator of μ is

$$\hat{\theta}_1 = M_1' = \bar{X}$$

and the method-of-moments estimator of σ^2 is

$$\hat{\theta}_2 = \sqrt{M_2' - \hat{\theta}_1^{\,2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sqrt{\frac{n-1}{n}}S.$$

Example 2.1.2 Let $X_1, X_2, \ldots, X_n$ be a random sample from uniform distribution on $(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma)$. Here, the unknown parameters are μ and σ, which are the population mean and standard deviation.

Let $(\theta_1, \theta_2) = (\mu, \sigma)$ and since we know that,

$$\mu_1' = \mu$$

and

$$\mu_2' = \sigma^2 + \mu^2.$$

It follows that the method-of-moments equations are

$$M_1' = \mu_1' = \mu_1'(\mu, \sigma) = \mu$$

and

$$M_2' = \mu_2' = \mu_2'(\mu, \sigma) = \sigma^2 + \mu^2.$$

Hence, the method-of-moments estimators are

$$\hat{\theta}_1 = \bar{X}$$

and

$$\hat{\theta}_2 = \sqrt{\frac{n-1}{n}}S.$$

1.2 METHOD OF MAXIMUM LIKELIHOOD

Consider an estimation problem where we suppose that an urn contains a number of blue balls and number of red balls. Suppose that it is known that the ratio of the

numbers is 3:1 but that it is not known whether the blue or the red balls are more numerous.

Let X be a random variable which denotes the event of drawing a blue ball. If n balls are drawn with replacement from the urn, the distribution of X is given by the binomial distribution

$$f(x;\ p) = \binom{n}{x} p^x q^{n-x} \qquad ;x = 0, 1, 2, \dots, n,$$

where $q = 1 - p$ and $0 \leq p \leq 1$ is the probability of drawing a blue ball,

i.e. $$p = P[X] = \frac{1}{4}\ or\ \frac{3}{4}$$

We shall draw a sample of three balls, i.e. $n = 3$, with replacement and attempt to estimate the unknown parameter p of the distribution. Let us anticipate the results of drawing the sample. The possible outcomes and their probabilities are given below:

Outcome : x	0	1	2	3
$f\left(x;\ \frac{1}{4}\right)$	$\frac{27}{64}$	$\frac{27}{64}$	$\frac{9}{64}$	$\frac{1}{64}$
$f\left(x;\ \frac{3}{4}\right)$	$\frac{1}{64}$	$\frac{9}{64}$	$\frac{27}{64}$	$\frac{27}{64}$

Here, if $x = 0$ in a sample of 3, the estimate 0.25 for p would be preferred over 0.75 since,

$$f\left(0;\ \frac{1}{4}\right) > f\left(0;\ \frac{3}{4}\right).$$

In other words, a sample with $x = 0$ is more likely to arise in a population with $p = 0.25$ than from one with $p = 0.75$. In general, we should estimate p by 0.25 when $x = 0$ or 1 and by 0.75 when $x = 2$ or 3. So, the estimator may be defined as

$$\hat{p} = \hat{p}(x) = \begin{cases} 0.25, & x = 0, 1 \\ 0.75, & x = 2, 3. \end{cases}$$

The estimator thus selects for every possible x, the value of p, say $\hat{p}$, such that

$$f(x; \hat{p}) > f(x; p'),$$

where p' is the complement value of p.

Now, if we found $x = 6$ in a sample of 25 from a binomial population, we should substitute all possible values of p in the expression

$$f(6\,; p) = \binom{25}{6} p^6 (1-p)^{25-6} \qquad ; 0 \leq p \leq 1 \qquad (2)$$

and choose that value of p as our estimate which maximizes $f(6\,; p)$. For the given possible values of p, we should find our estimate to be $\frac{6}{25}$. The position of its maximum value can be found by putting the derivative of the function defined in Eq. (2) with respect to p equal to zero and solving the resulting equation for p. Thus,

$$\frac{d}{dp} f(6; p) = \binom{25}{6} p^5 (1-p)^{18} \, [6(1-p) - 19p],$$

and on substituting this equal to zero and solving for p, we obtain

$$\frac{d}{dp} f(6; p) = 0 \Rightarrow p = 0, 1, \frac{6}{25}.$$

The possible probabilities for the outcome $x = 6$ are given below:

Probability : p	0	$\mathbf{0.24}$	1

$f(6\,;\,p)$	0	0.18	0

Therefore, our estimate is

$$\hat{p} = \frac{6}{25} = 0.24. \qquad (3)$$

Note that the point where the maximum value of $f(6;\,p)$ takes place for $0 \leq p \leq 1$ in Fig. 1 is the same as that given in Eq. (3) when $n = 25$.

This estimate has the property that

$$f(6;\hat{p}) > f(6;p'),$$

where p' is any other value of p in the interval $0 \leq p \leq 1$.

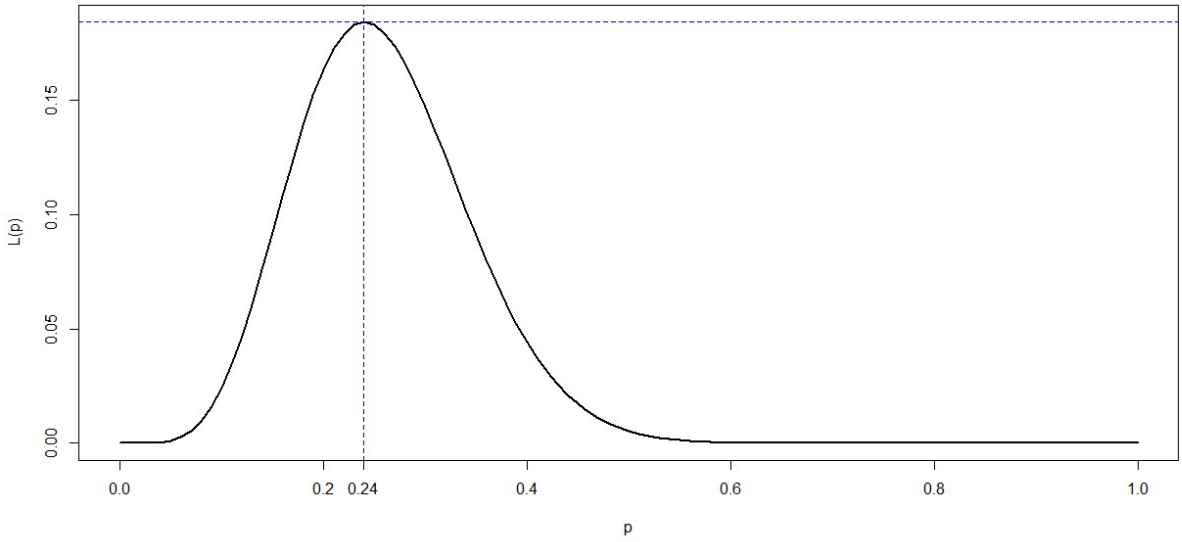


Figure 1. Maximum Likelihood Estimate of p for Bin(25, p) given x = 6

Definition. **Likelihood function**

The *likelihood function* of n random variables $X_1, X_2, \dots, X_n$ is defined to be the joint density of the n random variables, say $f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n)$, which is considered to be a function of θ. In particular, if $X_1, X_2, \dots, X_n$ is a random sample from the density $f(x;\ \theta)$, then the likelihood function is $f(x_1;\ \theta)\ f(x_2;\ \theta)$ ····· $f(x_n;\ \theta)$.

The likelihood function $\mathrm{L}(\theta;\ x_1, x_2, \dots, x_n)$ gives the *likelihood* that the random variables assume a particular value of a density function. The *likelihood* is the value of a density function; so for discrete random variables, it is a probability. Let us suppose that θ is known and denote this known value of θ by θ_0. The particular value of the random variables which is "most likely to occur" is that value $x_1', x_2', \dots, x_n'$ such that $f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n;\ \theta_0)$ is a maximum. For example, let us assume that $n = 1$ and $X_1 \sim N(6,1)$. Then, the value of the random variable which is most likely to occur is $X_1 = 6$. By "most likely to occur", we mean the value x_1' of X_1 such that

$$\Phi_{6,1}(x_1') > \Phi_{6,1}(x_1).$$

Further, suppose that the joint density of n random variables is $f_{X_1,X_2,\dots,X_n}(x_1, x_2, \dots, x_n;\ \theta)$, where θ is unknown. Let the particular values which are observed be represented by $x_1', x_2', \dots, x_n'$. We want to find the value of θ in the parameter space $\boldsymbol{\Theta}$, denoted by $\hat{\theta}$, which maximizes the likelihood function $\mathrm{L}(\theta;\ x_1, x_2, \dots, x_n)$. The value $\hat{\theta}$ which maximizes the likelihood is, in general, a function of $x_1, x_2, \dots, x_n$, say $\hat{\theta} = \hat{\vartheta}(x_1, x_2, \dots, x_n)$. When this is the case, the random variable $\hat{\Theta} = \hat{\vartheta}(X_1, X_2, \dots, X_n)$ is called the *maximum-likelihood estimator* of θ.

Definition. **Maximum-likelihood estimator**

Let $\mathrm{L}(\theta) = \mathrm{L}(\theta;\ x_1, x_2, \dots, x_n)$ be the likelihood function for the random variables $X_1, X_2, \dots, X_n$. If $\hat{\theta}$, a function of $x_1, x_2, \dots, x_n$, is the value of θ in $\boldsymbol{\Theta}$ which

maximizes $\mathrm{L}(\theta)$, then $\hat{\Theta} = \hat{\vartheta}(X_1, X_2, \dots, X_n)$ is the maximum-likelihood estimator of θ and $\hat{\theta} = \hat{\vartheta}(x_1, x_2, \dots, x_n)$ is the maximum-likelihood estimate of θ for the given sample.

Many likelihood functions satisfy regularity conditions. So, the maximum-likelihood estimator is the solution of the equation

$$\frac{dL(\theta)}{d\theta} = 0.$$

Since $L(\theta)$ and $\log L(\theta)$ have their maxima at the same value of θ, it is sometimes easier to find the maximum of the logarithm of the likelihood.

If the likelihood function contains k parameters, i.e. if

$$L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^{n} f(x_i;\ \theta_1, \theta_2, \dots, \theta_k),$$

then the maximum-likelihood estimators of the parameters $\theta_1, \theta_2, \dots, \theta_k$ are the random variables $\hat{\Theta}_1 = \hat{\vartheta}_1(X_1, X_2, \dots, X_n),\ \hat{\Theta}_2 = \hat{\vartheta}_2(X_1, X_2, \dots, X_n), \dots\dots, \hat{\Theta}_k = \hat{\vartheta}_k(X_1, X_2, \dots, X_n)$, where $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$ are the values in $\boldsymbol{\Theta}$ which maximize $L(\theta_1, \theta_2, \dots, \theta_k)$.

If certain regularity conditions are satisfied, the point where the likelihood is maximum is a solution of the k equations

$$\frac{\partial L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_1} = 0$$

$$\frac{\partial L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_2} = 0$$

$$\vdots$$

$$\frac{\partial L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_k} = 0$$

Example 2.2.1 Suppose that a random sample of size n is drawn from the Bernoulli distribution

$$f(x;p) = p^x q^{1-x} I_{(0,1)}(x) \qquad ; 0 \le p \le 1 \; and \; q = 1 - p.$$

Since the random sample is drawn from $B(1,p)$, the sample values $x_1, x_2, \dots, x_n$ will be a sequence of 0's and 1's, and the likelihood function is

$$L(p) = \prod_{i=1}^{n} p^{x_i} q^{1-x_i} = p^{\,y}\,(1-p)^{\,n-y} \qquad ; y = \sum_{i=1}^{n} x_i.$$

We obtain

$$\log L(p) = y \log p - (n-y)\log(1-p).$$

And on differentiating above with respect to the parameter p, we get

$$\frac{d \log L(p)}{dp} = \frac{y}{p} - \frac{n-y}{1-p}.$$

On substituting the last expression equal to zero and solving for p, we find the estimate

$$\hat{p} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{4}$$

which is intuitively what the estimate for this parameter should be. It is also a method-of-moments estimate.

For $n = 3$, let us sketch the likelihood function.

Since the likelihood function depends on the x_i's only through $\sum x_i$, thus the likelihood function can be represented by the following four curves:

$$L_0 = L\left(p; \sum_{i=1}^{3} x_i = 0\right) = (1-p)^3$$

$$L_1 = L\left(p; \sum_{i=1}^{3} x_i = 1\right) = p\,(1-p)^2$$

$$L_2 = L\left(p; \sum_{i=1}^{3} x_i = 2\right) = p^2\,(1-p)$$

$$L_3 = L\left(p; \sum_{i=1}^{3} x_i = 2\right) = p,$$

which are sketched in Fig. 2:

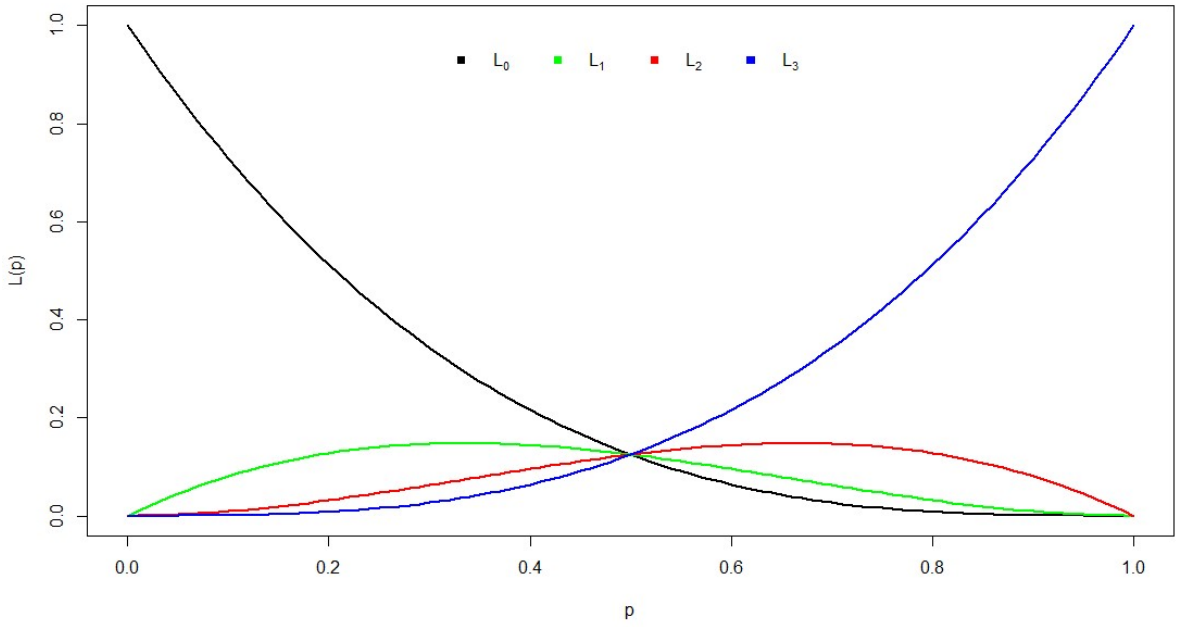


Figure 2. Likelihood Function Plot

Note that the point where the maximum of each of the curves takes place for $0 \leq p \leq 1$ is the same as that given in Eq. (4) when $n = 3$.

Example 2.2.2 Let the random variable X have a uniform density given by

$$f(x;\,\theta) = f(x;\,\mu,\sigma) = \frac{1}{2\sqrt{3}\,\sigma} I_{[\mu-\sqrt{3}\,\sigma,\ \mu+\sqrt{3}\,\sigma]}(x)$$

where $-\infty < \mu < \infty$ and $\sigma > 0$.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. Then, the likelihood function is given by

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\ \mu, \sigma)$$

$$= \left(\frac{1}{2\sqrt{3}\,\sigma}\right)^n \prod_{i=1}^{n} I_{[\mu-\sqrt{3}\,\sigma,\ \mu+\sqrt{3}\,\sigma]}\,(x_i)$$

$$= \left(\frac{1}{2\sqrt{3}\,\sigma}\right)^n I_{[\mu-\sqrt{3}\,\sigma,\ y_1]}(y_n)\, I_{[y_1,\ \mu+\sqrt{3}\,\sigma]}(y_n)$$

$$= \left(\frac{1}{2\sqrt{3}\,\sigma}\right)^n I_{\left[\frac{\mu-y_1}{3},\ \infty\right)}(\sigma)\, I_{\left[\frac{y_n-\mu}{3},\ \infty\right)}(\sigma)\ I_{[y_1,\ \infty)}(y_n),$$

where y_1 is the smallest of the observations and y_n is the largest.

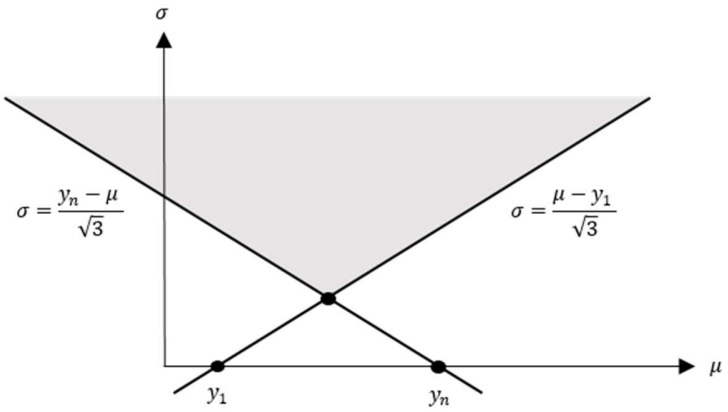


Figure 3

The likelihood function is $\left(2\sqrt{3}\,\sigma\right)^{-n}$ in the shaded area of Fig. 3 and 0 elsewhere. $\left(2\sqrt{3}\,\sigma\right)^{-n}$ within the shaded area is clearly a maximum when σ is smallest, which is the intersection of the lines $\mu - \sqrt{3}\,\sigma = y_1$ and $\mu + \sqrt{3}\,\sigma = y_n$. Hence, the maximum-likelihood estimates of μ and σ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{5}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{6}$$

which are quite different from the method-of-moments estimates given in Example 2.1.2.

Example 2.2.2 shows that one must not always rely on the differentiation process to locate the maximum.

The function $L(\theta)$ may, for example, be represented by the curve in Fig. 4, where the actual maximum is at $\hat{\theta}$, but the derivative set equal to 0 would locate θ' as the maximum.

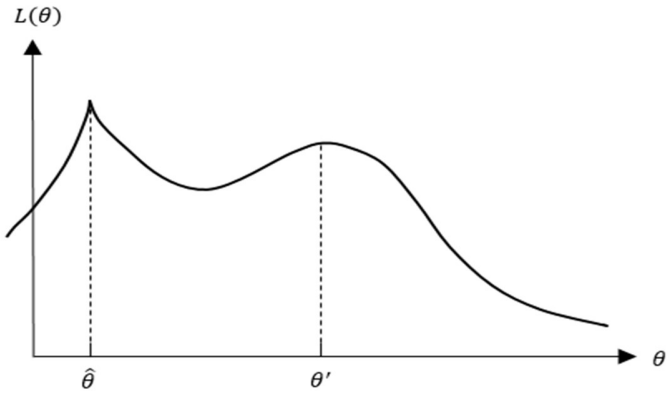


Figure 4

We know that the equation

$$\frac{\partial L}{\partial \theta} = 0$$

locates minima as well as maxima, and hence we must avoid using a root of the equation which actually locates a minimum.

Theorem 2.2.1 Invariance property of maximum-likelihood estimators

Let $\hat{\theta} = \hat{\vartheta}(X_1, X_2, \dots, X_n)$ be the maximum-likelihood estimator of θ in the density $f(x;\ \theta)$, where θ is assumed unidimensional. If $\tau(\cdot)$ is a function with a single-valued inverse, then the maximum-likelihood estimator of $\tau(\theta)$ is $\tau(\hat{\theta})$.

For example, in the normal density with mean μ_0 known, the maximum-likelihood estimator of σ^2 is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

By the invariance property of maximum-likelihood estimators, the maximum-likelihood estimator of σ is

$$\tau_1(\hat{\theta}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2}. \qquad \left[\because \tau_1(\theta) = \sigma = \sqrt{\sigma^2}\right]$$

Similarly, the maximum-likelihood estimator of $\log \sigma^2$ is

$$\tau_2(\hat{\theta}) = \log\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2\right]. \qquad [\because \tau_2(\theta) = \log \sigma^2\ ]$$

The invariance property of maximum-likelihood estimators can be extended in two directions: First, θ can be taken as k-dimensional rather than unidimensional, and, second, the assumption that $\tau(\cdot)$ has a single-valued inverse can be removed. It can be noted that such an extension is necessary by considering an example. Suppose an estimate of the variance, namely $\theta(1-\theta)$, of a Bernoulli distribution is desired. Example 2.2.1 gives the maximum-likelihood estimate of θ to be $\bar{x}$, but since $\theta(1-\theta)$ is not a one-to-one function of θ, Theorem 2.2.1 does not give the maximum-likelihood estimator of $\theta(1-\theta)$. Theorem 2.2.2 below will give such an estimate, and it will be $\bar{x}(1-\bar{x})$.

Let $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ be a k-dimensional parameter, and let $\boldsymbol{\Theta}$ denote the parameter space. Suppose that the maximum-likelihood estimate of $\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_r(\theta))$, where $1 \leq r \leq k$, is desired. Let T denote an r-dimensional range space of the transformation $\tau(\cdot)$. Define

$$M(\tau; x_1, x_2, \ldots, x_n) = \sup_{\{\theta\,:\,\tau(\theta)=\tau\}} L(\theta; x_1, x_2, \ldots, x_n).$$

The function $M(\cdot\,; x_1, x_2, \ldots, x_n)$ is called *the likelihood function induced by* $\tau(\cdot)$. When estimating θ, we maximized the likelihood function $L(\theta; x_1, x_2, \ldots, x_n)$ as a function of θ for fixed $x_1, x_2, \ldots, x_n$. So, in order to estimate $\tau = \tau(\theta)$, we will maximize $M(\tau\,; x_1, x_2, \ldots, x_n)$ as a function of τ for fixed $x_1, x_2, \ldots, x_n$. Thus, the maximum-likelihood estimate of $\tau = \tau(\theta)$, denoted by $\hat{\tau}$, is any value that maximizes the induced likelihood function for fixed $x_1, x_2, \ldots, x_n$, i.e.

$$\{\hat{\tau} : M(\hat{\tau}; x_1, x_2, \ldots, x_n) \geq M(\tau; x_1, x_2, \ldots, x_n), \forall\, \tau \in T\}.$$

The extended form of the *invariance property* of maximum-likelihood estimation is given in the following theorem.

Theorem 2.2.2 Let $\hat{\Theta} = (\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_k)$, where $\hat{\Theta}_j = \hat{\vartheta}_j(\cdot\,; X_1, X_2, \ldots, X_k)$, be a maximum-likelihood estimator of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ in the density function $f(\cdot\,; \theta_1, \theta_2, \ldots, \theta_k)$. If $\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_r(\theta))$ for $1 \leq r \leq k$ is a transformation of the parameter space $\boldsymbol{\Theta}$, then a maximum-likelihood estimator of $\tau(\theta) = (\tau_1(\theta), \tau_2(\theta), \ldots, \tau_r(\theta))$ is $\tau(\hat{\Theta})$, where $\tau(\hat{\Theta}) = (\tau_1(\hat{\Theta}), \tau_2(\hat{\Theta}), \ldots, \tau_r(\hat{\Theta}))$.

PROOF. Let $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)$ be a maximum-likelihood estimate of $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Now, it will be sufficient to show that

$$M(\tau(\hat{\theta}); x_1, x_2, \ldots, x_n) \geq M(\tau(\theta); x_1, x_2, \ldots, x_n)$$

for any $\tau \in T$. We have

$$M(\tau(\theta); x_1, x_2, \dots, x_n) = \sup_{\{\theta\,:\,\tau(\theta)\,=\,\tau\}} L(\theta; x_1, x_2, \dots, x_n)$$

$$\leq \sup_{\{\theta \in \Theta\}} L(\theta; x_1, x_2, \dots, x_n)$$

$$= L\left(\hat{\theta}; x_1, x_2, \dots, x_n\right)$$

$$= \sup_{\{\theta\,:\,\tau(\theta)\,=\,\tau(\hat{\theta})\}} L(\theta; x_1, x_2, \dots, x_n)$$

$$= M\left(\tau(\hat{\theta}); x_1, x_2, \dots, x_n\right)$$

$$\Rightarrow \quad M\left(\tau\left(\hat{\theta}\right); x_1, x_2, \dots, x_n\right) \geq M(\tau(\theta); x_1, x_2, \dots, x_n)$$ ■

Example 2.2.3 Let $X_1, X_2, \dots, X_n$ be a random sample of size n from the normal distribution having density

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

where $-\infty \leq \mu \leq \infty$ and $\sigma > 0$. Then, the likelihood function is given by

$$L(\mu, \sigma^2;\ x_1, x_2, \dots, x_n) = \prod_{i=1}^{n} f(x_i; \mu, \sigma)$$

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

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

The logarithm of the likelihood function is

$$L^* = \log L = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

To find the location of its maximum, we compute

$$\frac{\partial L^*}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

and

$$\frac{\partial L^*}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2,$$

and on putting these derivatives equal to zero and solving the resulting equations for μ and σ^2, we find the estimates

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \tag{7}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{8}$$

which turns out to be the sample moments corresponding to μ and σ^2.

Suppose $\tau(\theta) = \mu + z_q\sigma$, where $\phi(z_q) = q$, is the q^{th} quantile. Here, $\theta = (\mu, \sigma^2)$ and

$$\hat{\Theta}_1 = \bar{X} \text{ and } \hat{\Theta}_2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} = \sqrt{\frac{n-1}{n}}S$$

are the maximum-likelihood estimators of

$$\hat{\theta}_1 = \mu \text{ and } \hat{\theta}_2 = \sigma,$$

respectively. According to Theorem 2.2.2, the maximum-likelihood estimator of $\tau(\theta)$ is given by

$$\tau(\hat{\Theta}) = \tau(\hat{\Theta}_1) + z_q\tau(\hat{\Theta}_2) \qquad [\because \tau(\theta) = \mu + z_q\sigma]$$

$$= \bar{X} + z_q\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

1.3 MINIMUM CHI-SQUARE METHOD

Let $X_1, X_2, \dots, X_n$ be a random sample from a density given by $f_X(x;\, \theta)$, and let $\wp_1, \wp_2, \dots, \wp_k$ be a partition of the range of X.The probability that an observation falls in cell $\wp_j, j = 1, 2, \dots, k$, denoted by $p_j(\theta)$, can be found. For instance, if $f_X(x;\, \theta)$ is the density function of a continuous random variable, then

$$p_j(\theta) = P\left[X\ falls\ in\ cell\ \wp_j\right]$$

$$= \int_{\wp_j} f_X(x;\, \theta)\, dx \qquad ;\ \sum_{j=1}^{k} p_j(\theta) = 1.$$

Let the random variable N_j denote the number of X_i's in the sample which falls in cell $\wp_j, j = 1, 2, \dots, k$. Then, the sample size n is given by

$$n = \sum_{j=1}^{k} n_j\ .$$

Form the following summation:

$$\chi^2 = \sum_{j=1}^{k} \frac{\left[n_j - n\, p_j(\theta)\right]^2}{n\, p_j(\theta)},$$

where n_j is the value of N_j.The numerator of the j^{th} term in the sum is the square of the difference between the observed and the expected number of observations falling in cell $\wp_j$.

The *minimum chi-square* estimate of θ is that $\hat{\theta}$ which minimizes χ^2. It is that θ among all possible θ's which makes the expected number of observations in cell $\wp_j$ "nearest" the observed number. The minimum chi-square estimator depends on the partition $\wp_1, \wp_2, \dots, \wp_k$ selected.

Example 2.3.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution, i.e.

$$f_X(x;\, \theta) = \theta^x (1-\theta)^{1-x} \qquad ;\, x = 0, 1.$$

Further, let N_j be the number of observations equal to j for $j = 0, 1$. Here, the range of the observation X is partitioned into the two sets consisting of the numbers 0 and 1 respectively. We have

$$\chi^2 = \sum_{j=0}^{1} \frac{\left[n_j - n\, p_j(\theta)\right]^2}{n\, p_j(\theta)}$$

$$= \frac{[n_0 - n(1-\theta)]^2}{n(1-\theta)} + \frac{(n_1 - n\theta)^2}{n\theta}$$

$$= \frac{[(n - n_1) - n(1-\theta)]^2}{n(1-\theta)} + \frac{(n_1 - n\theta)^2}{n\theta} \qquad \left[\because \sum_{j=0}^{1} n_j = n\right]$$

$$= \frac{(n_1 - n\theta)^2}{n} \frac{1}{\theta(1-\theta)}.$$

The minimum of χ^2 as a function of θ can be found by inspection by noting that $\chi^2 = 0$ for

$\theta = \frac{n_1}{n}$. Hence,

$$\hat{\theta} = \frac{n_1}{n}.$$

Often it is difficult to locate that $\hat{\theta}$ which minimizes χ^2. Hence, the denominator $np_j(\theta)$ is sometimes changed to n_j and if $n_j = 0$, unity is used. Thus,

$$Modified\ \chi^2 = \sum_{j=1}^{k} \frac{\left[n_j - n\, p_j(\theta)\right]^2}{n_j}.$$

The *modified minimum chi-square estimate* of θ is then that $\hat{\theta}$ which minimizes the modified χ^2.

1.4 MINIMUM DISTANCE METHOD

Let $X_1, X_2, \ldots, X_n$ be a random sample from the distribution given by the cumulative distribution function $F_x(x;\ \theta) = F(x;\ \theta)$, and let $d(F, G)$ be a distance function that measures how "far apart" two cumulative distribution functions F and G are. An example of a distance function is

$$d(F, G) = \sup_x\ |F(x) - G(x)|,$$

which is the largest vertical distance between F and G. This can be easily seen in Fig. 5.

The *minimum-distance estimate* of θ is that $\hat{\theta}$ among all possible θ for which $d(F(x;\ \theta), F_n(x))$ is minimized, where $F_n(x)$ is the *sample cumulative distribution function*. Thus, $\hat{\theta}$ is chosen so that $F(x; \hat{\theta})$ will be "closest" to $F_n(x)$, which is desirable since the Theorem 7.1 of Chap. I states that for a fixed argument x, the sample cumulative distribution function has the same distribution as the mean of the binomial distribution. Hence, by the law of large numbers, $F_n(x)$ converges to $F(x)$.

Example 2.4.1 Let $X_1, X_2, \ldots, X_n$ be a random sample from a Bernoulli distribution; then

$$F(x;\ \theta) = (1 - \theta_0) I_{[0,1)}(x) + I_{[1,\infty)}(x),$$

where $0 \leq \theta \leq 1$.

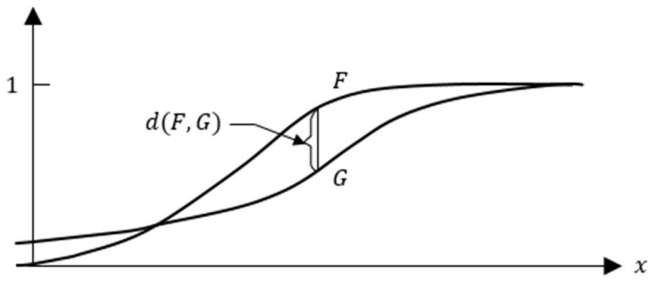


Figure 5

Further, let n_j be the number the observations equal to j ; $j = 0, 1$. Then

$$F_n(x) = \frac{n_0}{n}\, I_{[0,1)}(x) + I_{[1,\infty)}(x).$$

Now, if the distance function

$$d(F, G) = \sup_x |F(x) - G(x)|$$

is used, then $d\big(F(x;\ \theta), F_n(x)\big)$ is minimized if

$$1 - \theta = \frac{n_0}{n}$$

$$\Rightarrow \qquad \theta = \frac{n_1}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \ = \bar{x} \qquad \left[\because \sum_{j=0}^{1} n_j = n\right]$$

Hence, $\hat{\theta} = \bar{x}$.

2. PROPERTIES OF POINT ESTIMATORS

In this section, we will define certain properties, which an estimator may or may not possess, that will help in deciding whether one estimator is better than another.

2.1 CLOSENESS

Let $X_1, X_2, \dots, X_n$ be a random sample from a density, say $f_x(x;\ \theta)$, which is known except for θ. Then, a point estimator of $\tau(\theta)$ is a statistic, say $t(X_1, X_2, \dots, X_n)$, whose value is used as an estimate of $\tau(\theta)$, where $\tau(\theta)$ is a real-valued function of the unknown parameter θ. Ideally, we would like the value of $t(X_1, X_2, \dots, X_n)$ to be the value

of the unknown $\tau(\theta)$, but this is not possible except in trivial cases. For example, assume that one can sample from a density given by

$$f(x;\,\theta) = I_{\left(\theta-\frac{1}{2},\ \theta+\frac{1}{2}\right)}(x),$$

where θ is known to be an integer, i.e. $\boldsymbol{\Theta}$ consists of all integers. Consider estimating θ on the basis of a single observation x_1. If $t(x_1)$ is assigned as its value the nearest integer x_1, then the estimator $t(X_1)$ will always correctly estimate θ. In a sense, this problem is really not statistical since one knows the value of θ after taking one observation.

Not being able to achieve the ultimate of always estimating the unknown $\tau(\theta)$, we look for an estimator that is "close" to $\tau(\theta)$. There are several ways of defining "close". $T = t(X_1, X_2, \ldots, X_n)$ is a statistic and hence has a distribution, or rather a family of distributions, depending on what θ is. The distribution of T tells us how the values t of T are distributed, and we would like to select t so that the values of T distributed near $\tau(\theta)$.

Rather than resorting to characteristics of a distribution, such as its mean and variance, one can define what "concentration" might mean in terms of the distribution itself. Two such definitions follow.

Definition. **More concentrated and most concentrated**

Let $T = t(X_1, X_2, \ldots, X_n)$ and $T' = t'(X_1, X_2, \ldots, X_n)$ be two estimators of $\tau(\theta)$. T' is called a *more concentrated* estimator of $\tau(\theta)$ than T if and only if

$$P_\theta[\tau(\theta) - \lambda < T' \leq \tau(\theta) + \lambda] \geq P_\theta[\tau(\theta) - \lambda < T \leq \tau(\theta) + \lambda]$$

for all $\lambda > 0$ and for each $\theta \in \boldsymbol{\Theta}$. An estimator $T^* = t^*(X_1, X_2, \ldots, X_n)$ is called *most concentrated* if it is *more concentrated* than any other estimator.

The property of most concentrated is highly desirable. Unfortunately, most concentrated estimators seldom exist.

Another criterion for comparing estimators is the following one.

Definition. **Pitman-closer and Pitman-closest**

Let $T = t(X_1, X_2, \dots, X_n)$ and $T' = t'(X_1, X_2, \dots, X_n)$ be two estimators of $\tau(\theta)$. T' is called a *Pitman-closer* estimator of $\tau(\theta)$ than T if and only if

$$P_\theta[\tau(\theta) - \lambda < T' \leq \tau(\theta) + \lambda] \geq P_\theta[\tau(\theta) - \lambda < T \leq \tau(\theta) + \lambda]$$

for each $\theta \in \boldsymbol{\Theta}$. An estimator $T^* = t^*(X_1, X_2, \dots, X_n)$ is called *Pitman-closest* if it is Pitman-closer than any other estimator.

The property of Pitman-closest is, like the property of most concentrated, desirable, yet rarely there will exist a Pitman-closest estimator. Both Pitman-closer and more concentrated are intuitively attractive properties to be used to compare estimators, yet they are not always useful. Given two estimators T and T', one does not have to be more concentrated or Pitman-closer than the other. What often happens is that one, say T, is Pitman-closer or more concentrated for some $\theta \in \boldsymbol{\Theta}$; and since θ is unknown, we cannot say which estimator is preferred.

Competing estimators can be compared by defining a *measure* of the closeness of an estimate to the unknown $\tau(\theta)$. An estimator $T' = t'(X_1, X_2, \dots, X_n)$ of $\tau(\theta)$ will be judged better than an estimator $T = t(X_1, X_2, \dots, X_n)$ if the *measure* of the closeness of T' to $\tau(\theta)$ indicates that T' is closer to $\tau(\theta)$ than T. Here, we assume that n, the sample size, is fixed.

2.2 MEAN-SQUARED ERROR

A useful, though perhaps crude, measure of goodness or closeness of an estimator $t(X_1, X_2, \dots, X_n)$ of $\tau(\theta)$ is what is called the *mean-squared error* of the estimator.

Definition. **Mean-squared error**

Let $T = t(X_1, X_2, \dots, X_n)$ be an estimator of $\tau(\theta)$. $E_\theta[T - \tau(\theta)]^2$ is defined to be the *mean-squared error* of the estimator T.

Let $\mathrm{MSE}_t(\theta)$ denote the mean-squared error of the estimator $\mathrm{T} = t(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n)$ of $\tau(\theta)$.Then,

$$E_\theta[T - \tau(\theta)]^2 = E_\theta[t(X_1, X_2, \dots, X_n) - \tau(\theta)]^2$$

$$= \int \cdots \int [t(x_1, x_2, \dots, x_n) - \tau(\theta)]^2 \, f(x_1;\, \theta) \cdots f(x_n;\, \theta) \, dx_1 \cdots dx_n,$$

where $f(x;\, \theta)$ is the probability density function from which the random sample was selected.

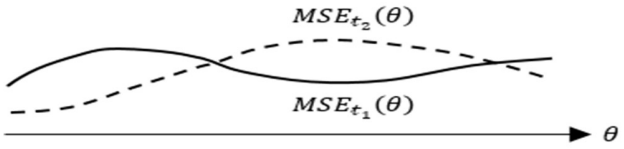


Figure 6

The name “mean-squared error” can be justified if one thinks of the difference $t - \tau(\theta)$, where t is a value of T used to estimate $\tau(\theta)$, as the error made in estimating $\tau(\theta)$, and then interprets the “mean” in “mean-squared error” as expected or average. $E_\theta[T - \tau(\theta)]^2$ is a measure of the spread of T values about $\tau(\theta)$ so that the mean-squared error of an estimator is taken as our standard in assessing the goodness of an estimator. If we were to compare estimators by looking at their respective mean-

squared errors, we could define as best that estimator with small or smallest mean-squared error, but such estimators rarely exist.

For any two estimators $T_1 = t_1(X_1, X_2, \dots, X_n)$ and $T_2 = t_2(X_1, X_2, \dots, X_n)$ of $\tau(\theta)$, their respective men-squared errors $MSE_{t_1}(\theta)$ and $MSE_{t_2}(\theta)$ are likely to cross. So for some θ, t_1 has smaller MSE, and for others t_2 has smaller MSE.We would then have no basis for preferring one of the estimators over the other. This can be easily seen in Fig. 6.

Example 3.2.1 Let $X_1, X_2, \dots, X_n$ be a random sample from the density $f(x;\ \theta)$, where θ is a real number, and consider estimating $\tau(\theta) = \theta$. We seek an estimator, say $T^* = t^*(X_1, X_2, \dots, X_n)$, such that

$$MSE_{t^*}(\theta) \leq MSE_t(\theta)$$

for every θ and for any other estimator $T = t(X_1, X_2, \dots, X_n)$ of θ.

Consider the family of estimators $T_{\theta_0} = t_{\theta_0}(X_1, X_2, \dots, X_n) \equiv \theta_0$ indexed by θ_0 for $\theta_0 \in \boldsymbol{\Theta}$. For each $\theta_0 \in \boldsymbol{\Theta}$, the estimator T_{θ_0} ignores the observations and estimatesθ to be θ_0.

Note that

$$MSE_{t_{\theta_0}}(\theta) = E_\theta\left[t_{\theta_0}(X_1, X_2, \dots, X_n) - \theta\right]^2$$

$$= E_{\theta_0}[\theta_0 - \theta]^2$$

So, $$MSE_{t_{\theta_0}}(\theta_0) = 0. \qquad (9)$$

Hence, if $\exists$ an estimator $T^* = t^*(X_1, X_2, \dots, X_n)$ satisfying $MSE_{t^*}(\theta) \leq MSE_t(\theta), \forall\ \theta$ and for any estimator t,

$$MSE_{t^*}(\theta_0) \leq MSE_{t_{\theta_0}}(\theta_0) = 0 \qquad [Using\ (9)]$$

$$\equiv 0.$$

In order for an estimator t^* to have its mean-squared error identically 0, it must always estimate θ correctly.

Example 3.2.1 shows that except in very rare cases, an estimator with smallest mean-squared error will not exist. One reason for being unable to find an estimator with uniformly smallest mean-squared error is that the class of all possible estimators is too large – it includes some estimators that are extremely prejudiced in favor of particular θ. For instance, in the example above, $t_{\theta_0}(X_1, X_2, \ldots, X_n)$ is highly partial to θ_0 since it always estimates θ to be θ_0. We could restrict the totality of estimators by considering only estimators that satisfy some other property. One such property is that of *unbiasedness*.

Definition. **Unbiased**

An estimator $T = t(X_1, X_2, \ldots, X_n)$ is defined to be an *unbiased* estimator of $\tau(\theta)$ if and only if

$$E_\theta[T] = E_\theta[t(X_1, X_2, \ldots, X_n] = \tau(\theta), \qquad \forall\, \theta \in \boldsymbol{\Theta}.$$

An estimator is unbiased if the mean of its distribution equals $\tau(\theta)$, the function of the parameter being estimated. Consider again the estimator $t_{\theta_0}(X_1, X_2, \ldots, X_n) \equiv \theta_0$ of Example 3.2.1. Since

$$E_\theta[t_{\theta_0}(X_1, X_2, \ldots, X_n) = E_\theta[\theta_0] = \theta_0 \neq \theta,$$

so $t_{\theta_0}(X_1, X_2, \ldots, X_n)$ is not an unbiased estimator of θ. If we restricted the totality of estimators under consideration by considering only unbiased estimators, we could hope to find an estimator with uniformly smallest mean-squared error within the restricted class, i.e. within the class of unbiased estimators.

Remark.

$$MSE_t(\theta) = var[T] + \{\tau(\theta) - E_\theta[T]\}^2. \tag{10}$$

So if T is an unbiased estimator of $\tau(\theta)$, then $MSE_t(\theta) = var[T]$.

PROOF. By definition, we have

$$\begin{aligned} MSE_t[\theta] &= E_\theta[T - \tau(\theta)]^2 \\ &= E_\theta[(T - E_\theta[T]) - \{\tau(\theta) - E_\theta[T]\}]^2 \\ &= E_\theta\big[T - E_\theta[T]\big]^2 - 2\,\{\tau(\theta) - E_\theta[T]\}\,E_\theta\big[T - E_\theta[T]\big] + E_\theta\big[\tau(\theta) - E_\theta[T]\big]^2 \\ &= var[T] + \{\tau(\theta) - E_\theta[T]\}^2. \end{aligned}$$

■

The term $\tau(\theta) - E_\theta[T]$ is called the *bias* of the estimator T and can be either positive, negative, or zero.

Example 3.2.2 Let $X_1, X_2, \ldots, X_n$ be a random sample from density $f(x;\, \theta) = \phi_{\mu,\ \sigma^2}(x)$. In Example 2.2.3, the maximum likelihood estimators of μ and σ^2 are, respectively, $\bar{X}$ and

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

Now, $E_\theta[\bar{X}] = \mu$. So, $\bar{X}$ is an unbiased estimator of μ, and hence

$$MSE_{\bar{X}}(\mu) = E_\theta[\bar{X} - \mu]^2 = var[\bar{X}] = \frac{\sigma^2}{n}.$$

We know that $E_\theta[S^2] = \sigma^2$. So,

$$\begin{aligned} E_\theta\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] &= \left(\frac{n-1}{n}\right) E_\theta\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \\ &= \left(\frac{n-1}{n}\right) E_\theta[S^2] \end{aligned}$$

$$= \left(\frac{n-1}{n}\right)\sigma^2.$$

Hence, the maximum-likelihood estimator of σ^2 is not unbiased. Using Eq. (10), the mean-squared error of the maximum-likelihood estimator of σ^2 is given by

$$MSE_{\left[\frac{1}{n}\sum(X_i-\bar{X})^2\right]}(\sigma^2) = var\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right] + \left\{\sigma^2 - E_\theta\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]\right\}^2$$

$$= \left(\frac{n-1}{n}\right)^2 var[S^2] + \left[\sigma^2 - \left(\frac{n-1}{n}\right)\sigma^2\right]$$

$$= \left(\frac{n-1}{n}\right)^2 \left\{\frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right]\right\} + \frac{\sigma^4}{n^2},$$

using Eq. (5) of Theorem 6.1 in Chap. I.

2.3 CONSISTENCY AND BAN

Properties of point estimators that are defined for a fixed sample size are sometimes referred to as *small-sample* properties, whereas properties that are defined for increasing sample size are sometimes referred to as *large-sample* properties. Consistency and asymptotic efficiency are two properties that are defined in terms of increasing sample size.

When considering a sequence of estimators, it seems that a good sequence of estimators should be one for which the values of the estimators tend to get closer to the quantity being estimated as the sample size increases.

Definition. **Mean-squared-error consistency**

Let $T_1, T_2, \ldots, T_n, \ldots$ be a sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, X_2, \ldots, X_n)$ is based on a sample of size n. This sequence of estimators is defined to be a *mean-squared-error consistent* sequence of estimators of $\tau(\theta)$, if and only if

$$\lim_{n\to\infty} E_\theta[T_n - \tau(\theta)]^2 = 0, \qquad \forall\, \theta \in \boldsymbol{\Theta}.$$

Example 3.3.1 In sampling from any density having mean μ and variance σ^2, let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

be a sequence of estimators of μ and σ^2, respectively. Since

$$E_\theta[\bar{X}_n - \mu]^2 = var[\bar{X}_n] \to 0 \text{ as } n \to \infty.$$

Hence, the sequence $\{\bar{X}_n\}$ is a mean-squared-error consistent sequence of estimators of μ. Again, since

$$E_\theta[S_n^2 - \sigma^2]^2 = var[S_n^2] = \frac{1}{n}\left[\mu_4 - \left(\frac{n-3}{n-1}\right)\sigma^4\right] \to 0 \text{ as } n \to \infty.$$

Hence, the sequence $\{S_n^2\}$ is a mean-squared-error consistent sequence of estimators of σ^2.

Definition. **Simple consistency**

Let $T_1, T_2, \ldots, T_n, \ldots$ be a countable sequence of estimators of $\tau(\theta)$, where $T_n = t_n(X_1, X_2, \ldots, X_n)$. The sequence $\{T_n\}$ is defined to be a *simple* (or *weakly*) *consistent* sequence of estimators of $\tau(\theta)$ if for every $\epsilon > 0$, the following is satisfied:

$$\lim_{n\to\infty} P_\theta[\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon] = 1, \qquad \forall\, \theta \in \boldsymbol{\Theta}.$$

Remark. If an estimator is a mean-squared-error consistent estimator, it is also a simple consistent estimator, but not necessarily vice-versa.

PROOF. We have

$$P_\theta[\tau(\theta) - \epsilon < \mathrm{T_n} < \tau(\theta) + \epsilon] = P_\theta[\,|T_n - \tau(\theta)| < \epsilon]$$

$$= P_\theta[\{T_n - \tau(\theta)\}^2 < \epsilon^2]$$

$$\geq 1 - \frac{E_\theta[T_n - \tau(\theta)]^2}{\epsilon^2}$$

by the Chebyshev inequality. Since

$$E_\theta[T_n - \tau(\theta)]^2 \to 0 \text{ as } n \to \infty.$$

Hence,

$$\lim_{\mathrm{n}\to\infty} P_\theta[\tau(\theta) - \epsilon < T_n < \tau(\theta) + \epsilon] = 1.$$

Definition. **Best asymptotically normal estimators (BAN estimators)**

A sequence of estimators $T_1^*, T_2^*, \ldots, T_n^*, \ldots$of $\tau(\theta)$ is defined to be *best asymptotically normal* (BAN) if and only if the following four conditions are satisfied:

(i) The distribution of $\sqrt{n}\,[T_n^* - \tau(\theta)]$ approaches the normal distribution with mean 0 and variance $\sigma^{*^2}(\theta)$ as n approaches infinity, i.e.

$$\lim_{n\to\infty} \sqrt{n}\,[T_n^* - \tau(\theta)] \to N\left(0, \sigma^{*^2}(\theta)\right) \text{ as } n \to \infty.$$

(ii) For every $\epsilon > 0$,

$$\lim_{n\to\infty} P_\theta[\,|T_n^* - \tau(\theta)| > \epsilon\,]^2 = 0, \quad \forall\, \theta \in \Theta.$$

(iii) Let $\{T_n\}$ be any other sequence of simple consistent estimators for which the distribution of $\sqrt{n}\,[T_n^* - \tau(\theta)]$ approaches the normal distribution with mean 0 and variance $\sigma^2(\theta)$.

(iv) $\sigma^2(\theta)$ is not less than $\sigma^{*^2}(\theta)$ for all θ in any open interval.

The abbreviation BAN is sometimes replaced by CANE, standing for *consistent asymptotically normal efficient*. BAN estimators are necessarily weakly consistent by (ii) of the definition.

Let us consider the maximum-likelihood estimation of the parameter θ, which is to be estimated on the basis of a random sample from a density $f(\cdot\,;\,\theta)$, where θ is assumed to be a real number. For the observed sample $x_1, x_2, \ldots, x_n$, the maximum-likelihood estimate of θ is that value, say $\hat{\theta}$, of θ which maximizes the likelihood function

$$L(\theta; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\ \theta).$$

Let $\widehat{\Theta}_n = \hat{\vartheta}_n(X_1, X_2, \ldots, X_n)$ denote the maximum-likelihood estimator of θ based on a sample of size n.

One property that it seems reasonable to expect of a sequence of estimators is that of *consistency*.

Theorem If the density $f(x;\theta)$ satisfies certain regularity conditions and if $\widehat{\Theta}_n = \hat{\vartheta}_n(X_1, X_2, \ldots, X_n)$ is the maximum-likelihood estimator of θ for a random sample of size n from $f(x;\theta)$, then

(i) $\widehat{\Theta}_n$ is asymptotically normally distributed with mean θ and variance

$$\frac{1}{n\,E_\theta\left[\frac{\partial}{\partial\theta}\log f(X;\ \theta)\right]^2}.$$

(ii) The sequence of maximum-likelihood estimators $\widehat{\Theta}_1, \widehat{\Theta}_2, \ldots, \widehat{\Theta}_n, \ldots$ is best asymptotically normal (BAN).

The theorem says that for large sample size, the maximum-likelihood estimator of θ is as good an estimator as there is. We may point out that the asymptotic normal distribution of the maximum-likelihood estimator is not given in terms of the distribution of the maximum-likelihood estimator. It is given in terms of $f(\cdot;\theta)$, the density sampled. Also, the variance of the asymptotic normal distribution given in the theorem is the Cramér-Rao lower bound.

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from the density

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

We have already derived, in Example 2.2.3, the maximum-likelihood estimators of μ and σ^2, as

$$\hat{\Theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \hat{\Theta}_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\Theta}_1\right)^2.$$

According to the above, the asymptotic large-sample joint distribution of $\hat{\Theta}_1$ and $\hat{\Theta}_2$ is a bivariate normal distribution with means $\mu_1 = \mu$ and $\mu_2 = \sigma^2$. Since

$$\log f(X;\ \theta) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(X-\mu)^2,$$

the required derivatives are

$$\frac{\partial^2}{\partial\mu^2}\log f(X;\ \theta) = -\frac{1}{\sigma^2},$$

$$\frac{\partial^2}{\partial\sigma^2\ \partial\mu}\log f(X;\ \theta) = -\frac{(X-\mu)}{\sigma^4}, \qquad \text{and}$$

$$\frac{\partial^2}{\partial\sigma^4}\log f(X;\ \theta) = \frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}.$$

Since

$$E[X] = \mu \text{ and } E[X-\mu]^2 = \sigma^2;$$

$$E_\theta\left[\frac{\partial^2}{\partial\mu^2}\log f(X;\,\theta)\right] = -\frac{1}{\sigma^2},$$

$$E_\theta\left[\frac{\partial^2}{\partial\sigma^2\,\partial\mu}\log f(X;\,\theta)\right] = 0,$$

$$E_\theta\left[\frac{\partial^2}{\partial\sigma^4}\log f(X;\,\theta)\right] = -\frac{1}{2\sigma^4},$$

which gives

$$\Delta = E_\theta\left[\frac{\partial^2}{\partial\mu^2}\log f(X;\,\theta)\right] E_\theta\left[\frac{\partial^2}{\partial\sigma^4}\log f(X;\,\theta)\right] - \left(E_\theta\left[\frac{\partial^2}{\partial\sigma^2\,\partial\mu}\log f(X;\,\theta)\right]\right)^2$$

$$= \left(-\frac{1}{\sigma^2}\right)\left(-\frac{1}{2\sigma^4}\right) - (0)^2$$

$$= \frac{1}{2\sigma^6}.$$

Finally, then

$$\sigma_1^2 = \frac{-E_\theta\left[\frac{\partial^2}{\partial\sigma^4}\log f(X;\,\theta)\right]}{n\,\Delta} = \frac{\frac{1}{2\sigma^4}}{n\cdot\frac{1}{2\sigma^6}} = \frac{\sigma^2}{n},$$

$$\sigma_2^2 = \frac{-E_\theta\left[\frac{\partial^2}{\partial\mu^2}\log f(X;\,\theta)\right]}{n\,\Delta} = \frac{\frac{1}{\sigma^2}}{n\cdot\frac{1}{2\sigma^6}} = \frac{2\sigma^4}{n}, \qquad \text{and}$$

$$\rho = \frac{-E_\theta\left[\frac{\partial^2}{\partial\sigma^2\,\partial\mu}\log f(X;\,\theta)\right]}{n\,\Delta\,\sigma_1\,\sigma_2} = 0.$$

Theorem 3.2.1 *Let T be a sufficient statistics for the family of pdf(pmf) $f(x|\theta, \theta \in \Theta)$. If an MLE of θ exists and it is unique then it is a function of T.*

Proof It is given that T is sufficient, from the factorization theorem,

$$f(x|\theta) = h(x)g(T|\theta)$$

Maximization of the likelihood function with respect to θ is therefore equivalent to the maximization of $g(T|\theta)$, which is a function of T alone. □

Remark: This theorem does not say that a MLE is itself a sufficient statistics.
In Example 3.2.8, we have shown that MLE need not be a function of sufficient statistics (see Remark 1).

Example 3.2.8 Let $X_1, X_2, \ldots, X_n$ be iid rvs with the following uniform pdf

1. $\cup(0, \theta)$
2. $\cup(\theta, 2\theta)$
3. $\cup(\theta - 1, \theta + 1)$
4. $\cup(\theta, \theta + 1)$

(i) The pdf of X is given by

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & ; \ 0 < x < \theta, \\ 0 & ; \ \text{otherwise} \end{cases} \tag{3.2.51}$$

and the corresponding likelihood function is

$$L(\theta|x) = \theta^{-n} \ ; \ 0 < x_i < \theta, \ i = 1, 2, \ldots, n$$

Consider the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Hence $0 < X_{(1)} < X_{(2)} < \cdots < X_{(n)} < \theta < \infty$. Note that the support of θ is $X_{(n)} < \theta < \infty$
We have to maximize $L(\theta|x)$ which is equivalent to finding the minimum value o θ, and it is given by $\hat{\theta} = X_{(n)}$. Thus,

$$\text{MLE of } \theta \ \text{ is } \ X_{(n)} \tag{3.2.52}$$

(ii) The pdf is given by

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & ; \ \theta < x < 2\theta, \\ 0 & ; \ \text{otherwise} \end{cases}$$

and the corresponding likelihood function is given by

$$L(x|\theta) = \begin{cases} \theta^{-n} & ; \ \theta < X_{(1)} < X_{(n)} < 2\theta, \ i = 1, 2, \ldots, n \\ 0 & ; \ \text{otherwise} \end{cases}$$

$\theta < X_{(1)}$ and $\frac{X_{(n)}}{2} < \theta$

$\Rightarrow \frac{X_{(n)}}{2} < \theta < X_{(1)}$

Maximizing $L(\theta|x)$ occurs at minimum value of θ

That is,
$$\hat{\theta} = \frac{X_{(n)}}{2} \tag{3.2.53}$$

(iii) The pdf and its corresponding likelihood functions are given by

$$f(x|\theta) = \begin{cases} \frac{1}{2} & ; \ \theta - 1 < x < \theta + 1, \\ 0 & ; \ \text{otherwise} \end{cases}$$

$$L(\theta|x) = \begin{cases} \frac{1}{2^n} & ; \ \theta - 1 < X_{(1)} < X_{(n)} < \theta + 1, \\ 0 & ; \ \text{otherwise} \end{cases}$$

The support of θ is $X_{(n)} - 1 \leq \theta \leq X_{(1)} + 1$. Here any value of θ is MLE.

Therefore,
$$\hat{\theta} = \alpha(X_{(n)} - 1) + (1 - \alpha)(X_{(1)} + 1), \tag{3.2.54}$$
where $\alpha \in [0, 1]$

(iv) In this case

$$L(\theta|x) = \begin{cases} 1 & ; \ \theta < x < \theta + 1, \\ 0 & ; \ \text{otherwise} \end{cases}$$

and

$$L(\theta|x) = \begin{cases} 1 & ; \ \theta < X_{(1)} < X_{(n)} < \theta + 1, \\ 0 & ; \ \text{otherwise} \end{cases}$$

The support of θ is $X_{(n)} - 1 < \theta < X_{(1)}$. Here also, any value of θ is MLE,

$$\hat{\theta} = \alpha(X_{(n)} - 1) + (1 - \alpha)X_{(1)} \tag{3.2.55}$$

Remark:

1. In (iii) and (iv), from (3.2.54) and (3.2.55), we can conclude that MLE is not a function of sufficient statistics, if $\alpha = 0$ or 1.
2. From (3.2.54) and (3.2.55), we can say that MLE is not unique.

Example 3.2.15 Find the MLE of the parameter p and σ of the following pdf

$$f(x|p, \sigma) = \frac{1}{\Gamma p} \left(\frac{p}{\sigma}\right)^p e^{-\frac{px}{\sigma}} x^{p-1}; \ x > 0, \ p, \ \sigma > 0$$

For large value of p, one should use $\Psi(p)$,

$$\Psi(p) = \log p - \frac{1}{2p} \quad \text{and} \quad \Psi'(p) = \frac{1}{p} + \frac{1}{2p^2},$$

where $\Psi(p)$ and $\Psi'(p)$ are known as digamma and trigamma functions,

$$\frac{d \log \Gamma p}{dp} = \Psi(p) \text{ and } \frac{d\Psi(p)}{dp} = \Psi'(p) \tag{3.2.64}$$

The corresponding likelihood function is given by,

$$L(p, \sigma|x) = \left(\frac{1}{\Gamma p}\right)^n \left(\frac{p}{\sigma}\right)^n p e^{-\frac{p}{\sigma}\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} x_i^{p-1}$$

$$\log L = -n \log \Gamma p + np[\log p - \log \sigma] - \frac{p}{\sigma} \sum_{i=1}^{n} x_i + (p-1) \sum_{i=1}^{n} \log x_i$$

Let G be the geometric mean of $x_1, x_2, \ldots, x_n$, then

$$\log G = \frac{1}{n} \sum_{i=1}^{n} \log x_i \Rightarrow n \log G = \sum_{i=1}^{n} \log x_i$$

$$\log L = -n \log \Gamma p + np[\log p - \log \sigma] - \frac{pn\bar{x}}{\sigma} + (p-1)n \log G$$

$$\frac{\partial \log L}{\partial \sigma} = -\frac{np}{\sigma} + \frac{np\bar{x}}{\sigma^2} = 0 \Rightarrow \hat{\sigma} = \bar{x}$$

$$\frac{\partial \log L}{\partial p} = -n\Psi(p) + n[\log p - \log \sigma] + \frac{np}{p} - \frac{n\bar{x}}{p} + n \log G$$

$$\Rightarrow \left[-n \log p + \frac{n}{2p}\right] + n[\log p - \log \sigma + 1] - n + n \log G = 0$$

$$\Rightarrow \frac{1}{2p} - \log \bar{x} + 1 - 1 + \log G = 0$$

$$\frac{1}{2p} - \log \frac{\bar{x}}{G} = 0$$

$$\hat{p} = \frac{1}{2 \log \frac{\bar{x}}{G}} \tag{3.2.65}$$

Hence, MLE of p and σ are

Hence, M

$$\hat{p} = \frac{1}{2 \log \frac{\bar{x}}{G}} \quad \text{and} \quad \hat{\sigma} = \bar{x} \tag{3.2.66}$$

Further, we state the following theorems on MLE without proof.

Theorem 3.2.2 (Invariance property of MLE):
If $\hat{\theta}$ is the MLE of θ, then $h(\hat{\theta})$ is the MLE of $h(\theta)$, where $h(\theta)$ is any continuous function of θ.

Theorem 3.2.3 *Let $X_1, X_2, \ldots, X_n$ be iid rvs having common pdf $f(x|\theta), \theta \in \Theta$. Assumption:*

1. *The derivative $\frac{\partial^i \log f(x|\theta)}{\partial \theta^i}$, $i = 1, 2, 3$ exist for almost all x and for every θ belonging to a non-degenerate interval in Θ*
2. *There exists functions $H_1(x)$, $H_2(x)$ and $H_3(x)$ such that $|\frac{\partial f}{\partial \theta}| < H_1(x)$, $|\frac{\partial^2 f}{\partial \theta^2}| < H_2(x)$, $|\frac{\partial^3 f}{\partial \theta^3}| < H_3(x)$, $\forall \theta \in \Theta$, $\int H_1(x)dx < \infty$, $\int H_2(x)dx < \infty$, $\int H_3(x)dx < \infty$,*
3. $$\int \left[\frac{\partial \log f(x|\theta)}{\partial \theta}\right]^2 f(x|\theta)dx$$

   is finite and positive for every $\theta \in \Theta$.

If assumptions (a)–(c) are satisfied and true parameter point θ_0 is an inner point then for sufficiently large n,

1. $$\sum_{j=1}^{n} \frac{\partial \log f(x_j|\theta)}{\partial \theta} = 0$$

   has at least one root $\hat{\theta}_n$ which converges in probability to θ_0.

2. *$\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to $N(0, I^{-1}(\theta))$, where*

$$I(\theta) = \int \left(\frac{\partial \log f(x|\theta)}{\partial \theta}\right)^2 f(x|\theta)dx,$$

   which is the Fisher information contained in the sample size n.

Theorem 3.2.4 *Huzurbazar (1948): The consistent root is unique.*

Theorem 3.2.5 *Wald (1949): The estimate which maximizes the likelihood absolutely is a consistent estimate.*

Newton–Raphson Method

The Newton–Raphson method is a powerful technique for solving equations numerically. Like so much of the differential calculus, it is based on the simple idea of linear approximation.

Let $f(x)$ be a well-behaved function. Let x^* be a root of the equation $f(x) = 0$ which we want to find. To find let us start with an initial estimate x_0. From x_0, we produce to an improved estimate x_1 (if possible) then from x_1 to x_2 and so on. Continue the procedure until two consecutive values x_i and x_{i+1} in ith and $(i + 1)$th steps are very close or it is clear that two consecutive values are away from each other. This style of proceeding is called 'iterative procedure'.

Consider the equation $f(x) = 0$ with root x^*. Let x_0 be a initial estimate. Let $x^* = x_0 + h$ then $h = x^* - x_0$, the number h measures how far the estimate x_0 is from the truth. Since h is small, we can use linear approximation to conclude that

$$0 = f(x^*) = f(x_0 + h) \simeq f(x_0) + hf'(x_0)$$

and therefore, unless $f'(x_0)$ is close to 0,

$$h \simeq -\frac{f(x_0)}{f'(x_0)}$$

This implies,

$$x^* = x_0 + h \simeq x_0 - \frac{f(x_0)}{f'(x_0)}$$

Our new improved estimate x_1 of x^* is given by

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

Next estimate x_2 is obtained from x_1 in exactly the same way as x_1 was obtained from x_0

$$x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$$

Continuing in this way, if x_n is the current estimate,then next estimate x_{n+1} is given by

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)},$$

Example 3.4.1 Consider the Example 3.2.15

$$\log L = -n \log \Gamma p + np[\log p - \log \sigma] - \frac{np\bar{x}}{\sigma} + n(p-1) \log G$$

$$\frac{\partial \log L}{\partial \sigma} = \frac{np}{\sigma} + \frac{np\bar{x}}{\sigma^2} \Rightarrow \hat{\sigma} = \bar{x}$$

$$\frac{\partial \log L}{\partial p} = -n\Psi(p) + n[\log p - \log \sigma] + \frac{np}{p} - \frac{n\bar{x}}{\sigma} + n \log G$$

$$\Rightarrow \log p - \Psi(p) = \log \frac{\bar{x}}{G}$$

Let $\log \frac{\bar{x}}{G} = C$
Hence $\log p - \Psi(p) = C$. By Newton–Raphson iteration method gives

$$\hat{p}_k = \hat{p}_{k-1} - \frac{\log(\hat{p}_{k-1}) - \Psi(\hat{p}_{k-1}) - C}{(\hat{p}_{k-1})^{-1} - \Psi'(\hat{p}_{k-1})}$$

$\hat{p}_k$ denotes the kth iterate starting with initial trial value $\hat{p}_0$ and $\Psi'(p) = \frac{d\Psi}{d}$
The function $\Psi(p)$ and $\Psi'(p)$ are tabulated in Abramowitz and Stegun

UNIT 2: CRITERION FOR GOOD ESTIMATORS

Definition 5.1.1 A sequence of rvs $\{X_n\}$ is said to converge to X in probability, denoted as $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, as $n \to \infty$

$$P[|X_n - X| \geq \epsilon] \to 0. \tag{5.1.1}$$

Equivalently, $X_n \xrightarrow{P} X$, if for every $\epsilon > 0$, as $n \to \infty$

$$P[|X_n - X| < \epsilon] \to 1. \tag{5.1.2}$$

Definition 5.1.2 The sequence of rvs $\{X_n\}$ is said to converge to X almost surely (a.s.) or almost certainly, denoted as $X_n \xrightarrow{a.s.} X$ iff $X_n(w) \to X(w)$ for all w, except those belonging to a null set N.

Thus,

$$X_n \xrightarrow{a.s.} X \text{ iff } X_n(w) \to X(w) < \infty, \text{ for } w \epsilon N^c,$$

where $P(N) = 0$. Hence we can write as

$$P\left[\lim_n X_n = X\right] \to 1 \tag{5.1.3}$$

Definition 5.1.3 Let $F_n(x)$ be the df of a rv X_n and $F(x)$, the df of X. Let $C(F)$ be the set of points of continuity of F. Then $\{X_n\}$ is said to converge to X in distribution or in law or weakly, denoted as $X_n \xrightarrow{L} X$ and/or $F_n \xrightarrow{W} F$, for every $x \in C(F)$.

It may be written as $X_n \xrightarrow{d} X$ or $F_n \xrightarrow{d} F$.

Theorem 5.1.5 *Let k be a constant, $X_n \xrightarrow{L} k \Leftrightarrow X_n \xrightarrow{P} k$.*

Definition 5.1.4 A sequence of rvs $\{X_n\}$ is said to converge to X in the rth mean if $\mathrm{E}|X_n - X|^r \to 0$ as $n \to \infty$. It is denoted as $X_n \xrightarrow{r} X$.

For $r = 2$, it is called convergence in quadratic mean or mean square.

Definition 5.2.1 Let $X_1, X_2, \ldots, X_n$ be a sequence of iid rvs with pdf(pmf) $f(x|\theta)$. A sequence of point estimates T is called consistent estimator of θ, where $T = T(X_1, X_2, \ldots, X_n)$ if for a given $\epsilon, \delta > 0$, there exists $n_0(\epsilon, \delta, \theta)$ such that $\forall \theta \in \Theta$

$$P[|T - \theta| < \epsilon] \geq 1 - \delta, \ \forall \ n > n_0 \tag{5.2.2}$$

or, using the Definition 5.1.1, we can say that $T \xrightarrow{P} \theta$ as $n \to \infty$.

Moreover, we can say that

$$P[|T - \theta| < \epsilon] \to 1 \tag{5.2.3}$$

Note: Some authors (5.2.3) define as a weak consistency and if we use a Definition 5.1.2 then they define it as a strong consistency.

Chebychev's Inequality

Theorem 5.1.12 *Let X be a rv with $\mathrm{E}X = \mu$ and $\mathrm{V}arX = \sigma^2 < \infty$, for any $k > 0$*

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2} \tag{5.1.8}$$

Theorem 5.1.7 $X_n \xrightarrow{r} X \Rightarrow X_n \xrightarrow{P} X$.

$$\lim_{n\to\infty} \mathrm{MSE}(T(X)) = 0 \tag{5.2.1}$$

which means that as the number of observations increase, the mse decreases to zero. For example, if $X_1, X_2, \ldots, X_n \sim N(\theta, 1)$, then $MSE(\bar{X}) = \frac{1}{n}$. Hence $\lim_{n\to\infty} MSE(\bar{X}) = 0$, $\bar{X}$ is consistent estimator of θ.

Example 5.2.1 Let $\{X_i\}_1^m$ be iid $B(n, p)$, then $\dfrac{\bar{X}}{n}$ is consistent estimator for p, where $\bar{X} = \dfrac{\sum_{i=1}^m X_i}{m}$.

Now, $MSE\left(\frac{\bar{X}}{n}\right) = \frac{pq}{mn}, q = 1 - p$. As $m \to \infty \Rightarrow MSE\left(\frac{\bar{X}}{n}\right) \to 0$

Example 5.2.2 Let $\{X_i\}_1^n$ be iid rvs with $P(\lambda)$ $\lambda > 0$ then $\bar{X}$ is a consistent estimator of λ, $\bar{X} = n^{-1} \sum_{i=1}^n X_i$.

Now, $\mathrm{E}\bar{X} = \lambda$ and $MSE(\bar{X}) = \frac{\lambda}{n} \to 0$ as $n \to \infty$.

Example 5.2.3 Let $\{X_i\}_1^n$ be iid rvs with $\cup(0, \theta)$, $\theta > 0$.
$\bar{X}$ is not an consistent estimator of θ.

$MSE(\bar{X}) = \frac{(3n+1)\theta^2}{12n}$ and $\lim_{n\to\infty} \frac{(3n+1)\theta^2}{12n} = \frac{\theta^2}{4} \neq 0$
But $X_{(n)}$ is an consistent estimator.

(i) $EX_{(n)} = \frac{n\theta}{n+1}$ and $MSE(X_{(n)}) = \dfrac{2\theta^2}{(n+1)(n+2)} \to 0$ as $n \to \infty$

(ii) Use the Definition (5.1.1) and assume $\epsilon \leq \theta$, from (5.1.2)

$$P[|X_{(n)} - \theta| < \epsilon] = P[\theta - \epsilon < X_{(n)} < \theta + \epsilon]$$
$$= \int_{\theta-\epsilon}^{\theta} \frac{nx^{n-1}}{\theta^n} dx = 1 - \left(\frac{\theta - \epsilon}{\theta}\right)^n \to 1 \text{ as } n \to \infty$$

(iii) Consider a df of $X_{(n)}$, let it be $H_n(x, \theta)$

$$H_n = P[X_{(n)} \leq x] = \begin{cases} 0 & ; \ x < 0 \\ (\frac{x}{\theta})^n & ; \ 0 \leq x < \theta \\ 1 & ; \ x \geq \theta \end{cases}$$

$$\lim_{n\to\infty} H_n(x, \theta) = H(x, \theta),$$

where

$$H(x, \theta) = \begin{cases} 0 & ; \ x < 0 \\ 1 & ; \ x \geq 0 \end{cases}$$

we have

$$X_{(n)} \xrightarrow{P} \theta \Leftrightarrow H_n \xrightarrow{d} H$$

In this case $H(x, \theta)$ is a df of a singular random variable, i.e., $P[X = \theta] = 1$, then

$$X_{(n)} \xrightarrow{d} X \Rightarrow X_{(n)} \xrightarrow{P} \theta.$$

Example 5.2.4 Consider $\{X_i\}_1^n$ are iid rvs as Cauchy distribution with location parameter θ.

$$f(x|\theta) = \frac{1}{\pi}\left[\frac{1}{1 + (x - \theta)^2}\right]; \ x \in R, \ \theta \in R$$

then $\bar{X}$ is not a consistent estimator for θ.
The distribution of $\bar{X}$ is Cauchy with parameter θ.
Using the Definition 5.1.1,

$$P[|\bar{X} - \theta| < \epsilon] = P[\theta - \epsilon < \bar{X} < \theta + \epsilon] \tag{5.2.4}$$
$$= \int_{\theta-\epsilon}^{\theta+\epsilon} \frac{1}{\pi}\left[\frac{dx}{1 + (x - \theta)^2}\right] = \frac{2}{\pi}\tan^{-1}\epsilon$$

This does not tends to 1.

Hence $\bar{X}$ is not a consistent estimator.

Example 5.2.5 Let $X_1, X_2, \ldots, X_n$ be iid $N\left(\mu, \sigma^2\right)$ rvs. We have to find the consistent estimator for σ^2.

We know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$, where $S^2 = \frac{\sum(x_i-\bar{x})^2}{n-1}$.

From Chebychev's Inequality,
$k\sigma = \epsilon \Rightarrow k = \frac{\epsilon}{\sigma}$

$$P[|S^2 - \sigma^2| > \epsilon] \leq \frac{\mathrm{V}ar(S^2)}{\epsilon^2} = \frac{2\sigma^4}{(n-1)\epsilon^2} \to 0 \text{ as } n \to \infty$$

Hence S^2 is consistent estimator for σ^2.

Theorem 5.2.1 *Let T be a consistent estimator for θ and let g be a continuous function then $g(t)$ is consistent for $g(\theta)$.*

Proof Given any $\epsilon > 0$, there exist a $\delta > 0$, such that, $|g(t) - g(\theta)| < \epsilon$ whenever $|T - \theta| < \delta$

Therefore, $\{x|\ |T - \theta| < \delta\} \subseteq \{x||g(t) - g(\theta)| < \epsilon\}$

Then $P\{x|\ |g(t) - g(\theta)| < \epsilon\} \geq P\{x||T - \theta| < \delta\}$,

Hence,

$$P\{x|\ |g(t) - g(\theta)| < \epsilon\} \to 1$$

Because

$$P\{x|\ |T - \theta| < \delta\} \to 1$$

$g(t)$ is consistent for $g(\theta)$.

Example 5.2.6 Let $X_1, X_2, \ldots, X_n$ be iid $p(\lambda)$ rvs. To find the consistent estimator for $g(\lambda) = e^{-s\lambda}\lambda^r$. We know that $\bar{X}$ is consistent for λ.

Using the Theorem 5.2.1, $g\left(\bar{X}\right) = e^{-s\bar{X}}\left(\bar{X}\right)^r$ is consistent for $g(\lambda) = e^{-s\lambda}\lambda^r$.

Example 5.2.7 Let $X_1, X_2, \ldots, X_m$ be iid $B(n, p)$ rvs. We know that $\dfrac{\bar{X}}{n} = (mn)^{-1}$ $\sum_{i=1}^{m} X_i$ is consistent for p.

Now, using Theorem 5.2.1, $\binom{n}{x}\bar{X}^x\left(1-\bar{X}\right)^{n-x}$ is consistent for $\binom{n}{x}p^x q^{n-x}$, when $m \to \infty$.

Example 5.2.8 Let $X_1, X_2, \ldots, X_m$ be iid $B(n, p)$ rvs, where p is a function of θ, in Bioassay problem, $p(\theta) = \frac{\exp(\theta y)}{1+\exp(\theta y)}$, where $y > 0$ is a given dose level.

Now $\frac{\bar{X}}{n}$ is consistent for p.

$$\frac{\bar{X}}{n} = \frac{\exp(\theta y)}{1+\exp(\theta y)} \Rightarrow \hat{\theta} = \frac{1}{y}\log\frac{\frac{\bar{X}}{n}}{1-\frac{\bar{X}}{n}} \quad , \quad \bar{X} = \frac{\sum_{i=1}^{m} X_i}{n}$$

Example 5.2.9 Let $X_1, X_2, \ldots, X_n$ be iid with $f(x|\theta)$,

$$f(x|\theta) = \theta x^{\theta-1} \;\; ; \;\; 0 < x < 1, \;\; \theta > 0$$

Let $y = -\log x$

$$g(y|\theta) = \theta e^{-\theta y} \;\; ; \;\; y > 0, \;\; \theta > 0$$

One can easily see that $\frac{-n}{\sum \log x_i}$ is consistent for θ.

Definition 5.2.2 Let X be a rv with its df $F(x|\theta)$, $\theta\epsilon\Theta$ then population quantile q_p is defined as

$$P[X \leq q_p] = p, \quad 0 < p < 1$$

If $p = \frac{1}{2}$ then $q_{\frac{1}{2}}$ is median.

If $p = \frac{i}{4}$ $(i = 1, 2, 3)$, then $q_{\frac{i}{4}}$ is called as ith Quartile. In many textbooks, Quartiles such as Q_1, Q_2 and Q_3 are defined.

If $p = \frac{i}{10}$ $(i = 1, 2, \ldots, 9)$, then $q_{\frac{i}{10}}$ is called as ith Decile. In many textbooks, it is defined as $(D_1, D_2, \ldots, D_9)$.

Let the rv X have exponential distribution with mean θ, then to find Q_1, Q_2, Q_3, D_1, D_3, and D_8:

$$f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}} \;\; ; \;\; x > 0, \;\; \theta > 0$$

By Definition 5.2.2,

$$P[X \leq Q_1] = \frac{1}{4}$$

$$1 - e^{-\frac{Q_1}{\theta}} = \frac{1}{4} \Rightarrow Q_1 = -\theta \log \frac{3}{4}$$

Similarly,

$$Q_2 = -\theta \log \frac{1}{2} \text{ and } Q_3 = -\theta \log \frac{1}{4}$$

$$D_1 = -\theta \log \frac{9}{10}, \; D_3 = -\theta \log \frac{7}{10} \text{ and } D_8 = -\theta \log \frac{2}{10}.$$

Lemma 5.2.1 *Let X be a random variable with its df $F(x)$. The distribution of $F(x)$ is $\cup(0, 1)$*

Proof Then

$$P[F(X) \leq z] = P\left[X \leq F^{-1}(z)\right] = F\left[F^{-1}(z)\right] = z$$

Hence $F(x)$ is $\cup(0, 1)$. □

Example 3. Let $X_1, X_2, \ldots$ be iid $b(1, p)$ RVs. Then $EX_1 = p$ and it follows by the WLLN that

$$\frac{\sum_1^n X_i}{n} \xrightarrow{P} p.$$

Thus $\overline{X}$ is consistent for p. Also, $(\sum_1^n X_i + 1)/(n + 2) \xrightarrow{P} p$, so that a consistent estimator need not be unique. Indeed, if $T_n \xrightarrow{P} p$ and $c_n \to 0$ as $n \to \infty$, then $T_n + c_n \xrightarrow{P} p$.

Theorem 1. If $X_1, X_2 \ldots$ are iid RVs with common law $\mathcal{L}(X)$, and $E|X|^p < \infty$ for some positive integer p, then

$$\frac{\sum_1^n X_i^k}{n} \xrightarrow{P} EX^k \qquad \text{for } 1 \leq k \leq p,$$

and $n^{-1}\sum_1^n X_i^k$ is consistent for EX^k, $1 \le k \le p$. Moreover, if c_n is any sequence of constants such that $c_n \to 0$ as $n \to \infty$, then $(n^{-1}\sum X_i^k + c_n)$ is also consistent for EX^k, $1 \le k \le p$. Also, if $c_n \to 1$ as $n \to \infty$, then $\left(c_n n^{-1}\sum X_i^k\right)$ is consistent for EX^k. This is simply a restatement of the WLLN for iid RVs.

Theorem 2. If T_n is a sequence of estimators such that $ET_n \to \psi(\boldsymbol{\theta})$ and $\operatorname{var}(T_n) \to 0$ as $n \to \infty$, then T_n is consistent for $\psi(\boldsymbol{\theta})$.

Proof. We have

$$P\{|T_n - \psi(\boldsymbol{\theta})| > \varepsilon\} \le \varepsilon^{-2}E[T_n - ET_n + ET_n - \psi(\boldsymbol{\theta})]^2$$
$$= \varepsilon^{-2}\{\operatorname{var}(T_n) + [ET_n - \psi(\boldsymbol{\theta})]^2\} \to 0 \qquad \text{as } n \to \infty.$$

Other large-sample properties of estimators are asymptotic unbiasedness, asymptotic normality, and asymptotic efficiency. A sequence of estimators $\{T_n\}$ is *asymptotically unbiased* for $\psi(\boldsymbol{\theta})$ if

$$\lim_{n\to\infty} E_{\boldsymbol{\theta}} T_n(\mathbf{X}) = \psi(\boldsymbol{\theta})$$

for all $\boldsymbol{\theta}$. A consistent sequence of estimators $\{T_n\}$ is said to be *consistent asymptotically normal* (CAN) for $\psi(\boldsymbol{\theta})$ if $T_n \sim AN(\psi(\boldsymbol{\theta}), v(\boldsymbol{\theta})/n)$ for all $\boldsymbol{\theta} \in \Theta$. If $v(\theta) = 1/I(\theta)$, where $I(\theta)$ is the Fisher information , then $\{T_n\}$ is known as a *best asymptotically normal* (BAN) estimator.

Bhattacharya's Bounds

Theorem 4.2.1 *Let $S_1, S_2, \ldots, S_k$ and $T_1, T_2, \ldots, T_k$ be the two sets of random variables such that with probability one S_i's are linearly independent, i.e.,*

$$P[a_1 S_1 + a_2 S_2 + \cdots + a_k S_k = 0] = 1 \tag{4.2.1}$$

Further,

$$\Lambda = Covariance\ matrix\ of\ S_i, i = 1, 2, \ldots, k$$

$$M = Covariance\ matrix\ of\ T_j, j = 1, 2, \ldots, k$$

$$N = Covariance\ matrix\ of\ S_i\ and\ T_j, i \neq j$$

Then the matrix $(M - N'\Lambda^{-1}N) \geq 0$ is positive semi-definite, i.e.,

$$v'(M - N'\Lambda^{-1}N)v \geq 0$$

This is also known as Hodge's Lemma.

Proof Without loss of generality, assume that $\mathrm{E}S_i = 0, \mathrm{E}T_j = 0$ and if $\mathrm{E}S_i$ and $\mathrm{E}T_j \neq 0$, then let $S_i^* = S_i - \mathrm{E}S_i$ and $T_j^* = T_j - \mathrm{E}T_j$, then $\mathrm{V}ar(S_i) = \mathrm{V}ar(S_i^*)$ and $\mathrm{V}ar(T_j) = \mathrm{V}ar(T_j^*)$.

Using Cauchy–Schwarz inequality,

$$\mathrm{C}ov^2(u'S, v'T) \leq \mathrm{V}ar(u's)\mathrm{V}ar(v'T) \tag{4.2.2}$$

$$[u'\mathrm{C}ov(S, T)v)]^2 \leq [u'\mathrm{V}ar(S)u][v'\mathrm{V}ar(T)v]$$

$$(u'Nv)^2 \leq (u'\Lambda u)(v'Mv)$$

Suppose $\Lambda u = Nv \Rightarrow u = \Lambda^{-1}Nv$

$$[(\Lambda^{-1}Nv)'Nv]^2 \leq [(\Lambda^{-1}Nv)'\Lambda\Lambda^{-1}Nv][v'Mv]$$

$$[v'N'\Lambda^{-1}Nv]^2 \leq [v'N'\Lambda^{-1}\Lambda\Lambda^{-1}Nv][v'Mv]$$

$$[v'N'\Lambda^{-1}Nv]^2 \leq [v'N'\Lambda^{-1}Nv][v'Mv]$$

$$[v'N'\Lambda^{-1}Nv] \leq [v'Mv]$$

Therefore,

$$v'(M - N'\Lambda^{-1}N)v \geq 0 \tag{4.2.3}$$

Theorem 4.2.2 *Let $X_1, X_2, \ldots, X_n$ be iid rvs with joint pdf $f(x_1, x_2, \ldots, x_n|\theta)$ satisfying the regularity conditions.*

Let

$$S_i = \frac{1}{f(x|\theta)} \frac{\partial^i f(x|\theta)}{\partial \theta^i}$$

then $\mathrm{E}S_i = 0, i = 1, 2, \ldots, k$

$$\Lambda = Covariance\ matrix\ of\ S_i, i = 1, 2, \ldots, k$$

$$N' = [g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(k)}(\theta)], \quad where \quad g^{(i)}(\theta) = \frac{\partial^i g(\theta)}{\partial \theta^i}; i = 1, 2, \ldots, k$$

$u(x_1, x_2, \ldots, x_n)$ is an unbiased estimator of $g(\theta)$, then

$$\mathrm{V}(u(x)) \geq L_k \quad where \quad L_k = N'\Lambda^{-1}N \tag{4.2.4}$$

(4.2.4) is called Bhattacharya bound.

Proof Let u(x) be an unbiased estimator of $g(\theta)$

Hence,

$$\int \int \cdots \int u(x) f(x|\theta) dx = g(\theta)$$

$$\frac{\partial}{\partial \theta} \int \int \cdots \int u(x) f(x|\theta) dx = \frac{\partial g(\theta)}{\partial \theta}$$

$$\int \int \cdots \int \frac{u(x)}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} f(x|\theta) dx = g^{(1)}(\theta)$$

$$\int \int \cdots \int u(x) S_1 f(x|\theta) dx = g^{(1)}(\theta)$$

$$\mathrm{E}[u(x)S_1] = g^{(1)}(\theta)$$

In general,

$$\int \int \cdots \int \frac{u(x)}{f(x|\theta)} \frac{\partial^i f(x|\theta)}{\partial \theta^i} f(x|\theta) dx = g^{(i)}(\theta)$$

$$\mathrm{E}[u(x)S_i] = g^{(i)}(\theta)$$

We know that $\mathrm{E}S_i = 0 \Rightarrow Cov[u(x), S_i] = g^{(i)}(\theta)$

By using Hodge's Lemma (Theorem 4.2.1),

$$M - N'\Lambda^{-1}N \geq 0$$

In this case $M = \mathrm{V}ar[u(x)]$

$$\mathrm{V}ar[u(x)] - N'\Lambda^{-1}N \geq 0$$

Let $L_k = N'\Lambda^{-1}N$ then

$$\mathrm{V}ar[U(x)] \geq L_k$$

Hence

$$L_k \geq L_{k-1} \geq \cdots \geq L_1. \tag{4.2.5}$$

We know that

By using Hod

In this case $M =$

Let $L_k = N'\Lambda^{-}$

Hence

Note

For $k = 1$, $Var[u(x)] \geq L_1$

$$L_1 = \frac{[g^{(1)}(\theta)]^2}{Var(S_1)},$$

$$S_1 = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} = \frac{\partial \log f(x|\theta)}{\partial \theta}$$

$$Var(S_1) = Var\left[\frac{\partial \log f(x|\theta)}{\partial \theta}\right]$$

CR bound becomes a particular case of Bhattacharya bound for $k = 1$.

Steps to find Bhattacharya bound:

1. To get N', differentiate the given parametric function $g(\theta)$.

   i.e., $N' = \left[\frac{\partial g(\theta)}{\partial \theta}, \frac{\partial^2 g(\theta)}{\partial \theta^2}, \ldots, \frac{\partial^k g(\theta)}{\partial \theta^k}\right]$

2. Find $S_i = \frac{1}{f(x|\theta)} \frac{\partial^i f(x|\theta)}{\partial \theta^i}$; $i = 1, 2, \ldots, k$ and verify $ES_i = 0$

3. Find $Var(S_i) = E(S_i)^2$ and $Cov(S_i, S_j) = E(S_i S_j)(i \neq j)$. Then obtain the covariance matrix of (S_i, S_j), $(i \neq j)$, i.e., Λ.

4. Calculate $N' \Lambda^{-1} N$.

Example 4.2.1 Let $X_1, X_2, \ldots, X_n$ be iid rvs with $N(\theta, 1)$. We will obtain the Bhattacharya bound for $g(\theta) = \theta^2$

$$N' = [g^{(1)}(\theta), g^{(2)}(\theta), \ldots, g^{(k)}(\theta)] = [2\theta, 2, 0, \ldots, 0] \tag{4.2.6}$$

Here, we can take $N' = [2\theta, 2]$.

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

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

$$S_1 = \frac{1}{f(x|\theta)} \frac{\partial f(x|\theta)}{\partial \theta} = n(\bar{x} - \theta) \tag{4.2.7}$$

Then $ES_1 = 0$

$$\frac{\partial^2 f(x|\theta)}{\partial\theta^2} = (2\pi)^{-\frac{n}{2}}\left[\left\{-n\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\right\} + n^2(\bar{x}-\theta)^2\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right]\right]$$

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

$$S_2 = \frac{1}{f(x|\theta)}\frac{\partial^2 f(x|\theta)}{\partial\theta^2} = -n + n^2(\bar{x}-\theta)^2 \tag{4.2.8}$$

Similarly, one can find $S_3, S_4, \ldots, S_k$.

$$\mathrm{E}S_2 = -n + n = 0$$

$$\mathrm{V}ar(S_1) = \mathrm{E}{S_1}^2 = n^2\mathrm{E}(\bar{x}-\theta)^2 = n \tag{4.2.9}$$

$$\mathrm{V}ar(S_2) = \mathrm{E}[n^2(\bar{x}-\theta)^2 - n]^2$$

$$= n^2\mathrm{E}[n(\bar{x}-\theta)^2 - 1]^2 \tag{4.2.10}$$

$$= n^2\mathrm{E}[n^2(\bar{x}-\theta)^4 - 2n(\bar{x}-\theta)^2 + 1]$$

Now, $\mathrm{E}(\bar{x}-\theta)^4 = \dfrac{3}{n^2}$,

$$= n^2[n^2\frac{3}{n^2} - 2n\frac{1}{n} + 1]$$

$$= n^2[3-2+1] = 2n^2$$

$$\mathrm{C}ov(S_1, S_2) = \mathrm{E}S_1S_2 = \mathrm{E}[\{n(\bar{x}-\theta)\}\{n^2(\bar{x}-\theta)^2 - n\}]$$

$$= \mathrm{E}[n^3(\bar{x}-\theta)^3] - \mathrm{E}[n^2(\bar{x}-\theta)] = 0 \tag{4.2.11}$$

Hence

$$\Lambda = \begin{pmatrix} n & 0 \\ 0 & 2n^2 \end{pmatrix}, \Lambda^{-1} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{2n^2} \end{pmatrix},$$

$$L_2 = N'\Lambda^{-1}N = \begin{pmatrix} 2\theta & 2 \end{pmatrix}\begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{1}{2n^2} \end{pmatrix}\begin{pmatrix} 2\theta \\ 2 \end{pmatrix}$$

$$= \frac{4\theta^2}{n} + \frac{2}{n^2}, \tag{4.2.12}$$

and $L_1 = \frac{4\theta^2}{n} =$ CR lower bound

Therefore $L_1 < L_2$.

Chapman – Robbin – Kiefer Bound

Theorem 4.3.1 *Let the random vector X have a pdf(pmf) $f(x|\theta)$. Let $T(X)$ be an unbiased estimator $g(\theta)$, where $g(\theta)$ defined on Θ. Further, assume that $\mathrm{E}T^2 < \infty$ for all $\theta \epsilon \Theta$. If $\theta \neq \alpha$, then assume that $f(x|\theta)$ and $f(x|\alpha)$ are different. Assume that $\mathrm{S}(\theta) = \{f(x|\theta) > 0\}$, $\mathrm{S}(\alpha) = \{f(x|\alpha) > 0\}$ and $\mathrm{S}(\alpha) \subset \mathrm{S}(\theta)$.*

Then,

$$\mathrm{V}ar[T(X)] \geq \sup_{\mathrm{S}(\alpha)\subset\mathrm{S}(\theta),\alpha\neq 0} \frac{[g(\alpha) - g(\theta)]^2}{\mathrm{V}ar\{\frac{f(x|\alpha)}{f(x|\theta)}\}} \; \forall\, \theta \epsilon \Theta \tag{4.3.1}$$

Proof Under $f(x|\theta)$ and $f(x|\alpha)$

$$\mathrm{E}T(X) = \int T(X) f(x|\theta) dx = g(\theta)$$

$$\mathrm{E}T(X) = \int T(X) f(x|\alpha) dx = g(\alpha)$$

$$g(\alpha) - g(\theta) = \int T(X) \frac{f(x|\alpha) - f(x|\theta)}{f(x|\theta)} f(x|\theta) dx$$

$$= \int T(X) \left[\frac{f(x|\alpha)}{f(x|\theta)} - 1\right] f(x|\theta) dx$$

$$\mathrm{C}ov\left[T(X), \frac{f(x|\alpha)}{f(x|\theta)} - 1\right] = g(\alpha) - g(\theta)$$

Using Cauchy-Schwarz inequality,

$$\mathrm{C}ov^2\left[T(X), \frac{f(x|\alpha)}{f(x|\theta)} - 1\right] \leq \mathrm{V}ar[T(X)]\mathrm{V}ar\left[\frac{f(x|\alpha)}{f(x|\theta)} - 1\right]$$

$$= \mathrm{V}arT(X)\mathrm{V}ar\left[\frac{f(x|\alpha)}{f(x|\theta)}\right]$$

Therefore,

$$[g(\alpha) - g(\theta)]^2 \leq \mathrm{V}arT(X)\mathrm{V}ar\left[\frac{f(x|\alpha)}{f(x|\theta)}\right]$$

$$\mathrm{V}ar[T(X)] \geq \frac{[g(\alpha) - g(\theta)]^2}{\mathrm{V}ar\{\frac{f(x|\alpha)}{f(x|\theta)}\}} \; \forall\, \theta \epsilon \Theta \tag{4.3.2}$$

Then, (4.3.1) follows immediately.

Chapman and Robbins (1951) had given the same above-mentioned theorem in different form.