CHAPTER 9
Properties of Point Estimators and Methods of Estimation

Motivation

- We undertake a more formal and detailed examination of some of the mathematical properties of point estimators—particularly the notions of
  - ✓ ***efficiency***,
  - ✓ ***consistency***, and
  - ✓ ***sufficiency***.

- We present a result, ***the Rao–Blackwell theorem***, that provides a link between sufficient statistics and unbiased estimators for parameters.
  - ✓ demonstrate a method that can sometimes beusedtofindminimum-varianceunbiasedestimatorsforparametersofinterest

- Two other useful methods for deriving estimators:
  - ✓ ***the method of moments*** and
  - ✓ ***the method of maximum likelihood***.

Motivation

- It usually is possible to obtain more than one unbiased estimator for the same target parameter θ
- If $\hat{\theta}_1$ and $\hat{\theta}_2$ denote two unbiased estimators for the same parameter θ, **we prefer to use the estimator with the smaller variance**.
- That is, if both estimators are unbiased, $\hat{\theta}_1$ is relatively more efficient than $\hat{\theta}_2$ if $V(\hat{\theta}_2) > V(\hat{\theta}_1)$

Definition

DEFINITION 9.1

Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ, with variances $V(\hat{\theta}_1)$ and $V(\hat{\theta}_2)$, respectively, then the *efficiency* of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted eff $(\hat{\theta}_1, \hat{\theta}_2)$, is defined to be the ratio

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}.$$

$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) > 1 \Rightarrow V(\hat{\theta}_2) > V(\hat{\theta}_1) \Rightarrow \hat{\theta}_1$ is proffered to $\hat{\theta}_2$

$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) < 1 \Rightarrow V(\hat{\theta}_2) < V(\hat{\theta}_1) \Rightarrow \hat{\theta}_2$ is proffered to $\hat{\theta}_1$

Median vs. Sample Mean

- Let $\hat{\theta}_1$ be **the sample median**, the middle observation when the sample measurements are ordered according to magnitude (n odd) or the average of the two middle observations(n even).
- Let $\hat{\theta}_2$ be **the sample mean**.
- it can be shown that the variance of the sample median, for large n, is $V\left(\hat{\theta}_1\right) = (1.2533)^2(\sigma^2/n) = (1.2533)^2 V\left(\hat{\theta}_2\right)$

$$\text{eff}\left(\hat{\theta}_1, \hat{\theta}_2\right) = \frac{V\left(\hat{\theta}_2\right)}{V\left(\hat{\theta}_1\right)} = \frac{(\sigma^2/n)}{(1.2533)^2(\sigma^2/n)} = \frac{1}{(1.2533)^2} = .6366$$

- Thus, we see that the variance of the sample mean is approximately 64% of the variance of the sample median.
  - Therefore, we would prefer to use the sample mean as the estimator for the population mean.

Definition

EXAMPLE 9.1

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the uniform distribution on the interval $(0, \theta)$. Two unbiased estimators for θ are $\hat{\theta}_1 = 2\bar{Y}$ and $\hat{\theta}_2 = \left(\frac{n+1}{n}\right) Y_{(n)}$,
where $Y_{(n)} = \max(Y_1, Y_2, \ldots, Y_n)$. Find the efficiency of $\hat{\theta}_1$ relative to $\hat{\theta}_2$.

Definition

SOLUTION 9.1

Each Y_i has a uniform distribution on the interval $(0,\theta)$. Thus, $\mu = E(Y_i) = \theta/2$ and $\sigma^2 = V(Y_i) = \theta^2/12$. Therefore,
$E(\hat{\theta}_1) = E(2\bar{Y}) = 2(\mu) = 2\left(\frac{\theta}{2}\right) = \theta$ (unbiased), $V(\hat{\theta}_1) = V(2\bar{Y}) = 4V(\bar{Y}) = 4[\frac{V(Y_i)}{n}] = \left(\frac{4}{n}\right)\left(\frac{\theta^2}{12}\right) = \frac{\theta^2}{3n}$.

To find the mean and variance of $\hat{\theta}_2$, recall (see Exercise 6.74) that the density function of $Y_{(n)}$ is given by

$$g_{(n)}(y) = n[F_Y(y)]^{n-1}f_Y(y) = \begin{cases} n\left(\frac{y}{\theta}\right)^{n-1}\left(\frac{1}{\theta}\right), 0 \le y \le \theta, \\ 0, \qquad\qquad elsewhere \end{cases}$$

Thus, $E(Y_{(n)}) = \frac{n}{\theta^n}\int_0^\theta y^n\,dy = \frac{n}{n+1}\theta \;\to\; E\{\left[\frac{n+1}{n}\right]Y_{(n)}\} = \theta$; that is, $\hat{\theta}_2$ is an unbiased estimator for θ.

Because $E(Y_{(n)}^2) = \frac{n}{\theta^n}\int_0^\theta y^{n+1}dy = (\frac{n}{n+2})\theta^2$,
we obtain

$$V(Y_{(n)}) = E(Y_{(n)}^2) - [E(Y_{(n)})]^2 = \left[\frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2\right]\theta^2$$

$$V(\hat{\theta}_2) = V\left[\left(\frac{n+1}{n}\right)Y_{(n)}\right] = \left(\frac{n+1}{n}\right)^2 V(Y_{(n)}) = \left[\frac{(n+1)^2}{n(n+2)} - 1\right]\theta^2 = \frac{\theta^2}{n(n+2)}.$$

Therefore, eff $(\hat{\theta}_1, \hat{\theta}_2) = V(\hat{\theta}_2)/V(\hat{\theta}_1) = \frac{\theta^2[n(n+2)]}{\theta^2/3n} = \frac{3}{n+2}$.
This efficiency is less than 1 if $n > 1$.
That is, if $n > 1$, $\hat{\theta}_2$has a smaller variance than $\hat{\theta}_1$, and therefore $\hat{\theta}_2$ is generally preferable to $\hat{\theta}_1$ as an estimator of θ.

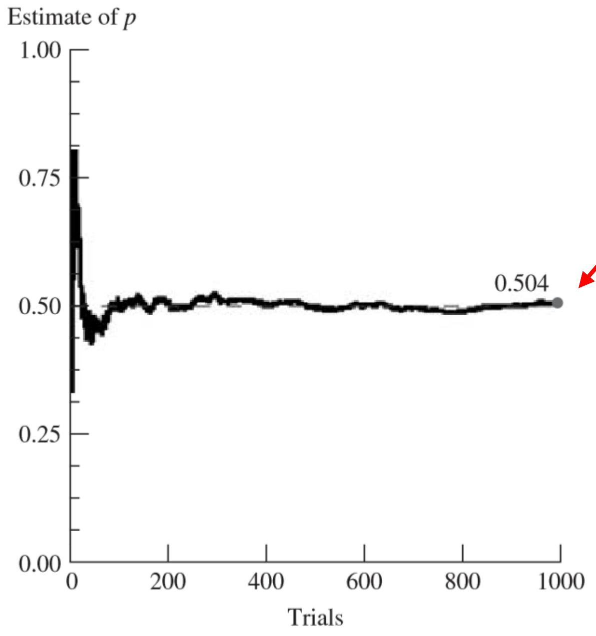
Motivation

- Suppose that a coin, which has probability p of resulting in heads, is tossed n times.
  - If the tosses are independent, then Y, the number of heads among the n tosses, has a binomial distribution.
  - If the true value of p Is unknown, the sample proportion Y/n is an estimator of p.

- What happens to this sample proportion as the number of tosses n increases?
  - Our intuition leads us to believe that as n gets larger, Y/n should get closer to the true value of p.
  - That is, as the amount of information in the sample increases, our estimator should get closer to the quantity being estimated.

Motivation

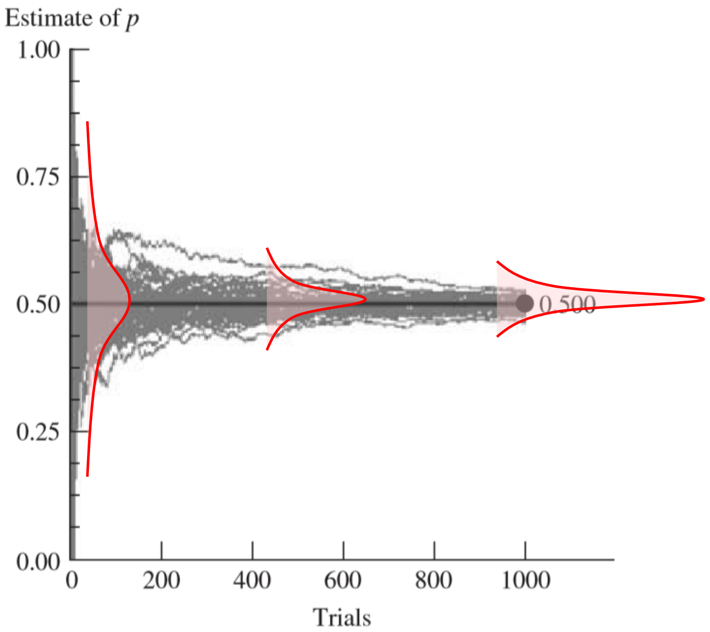
- Figure illustrates the values of $\hat{p} = Y/n$ for a single sequence of 1000 Bernoulli trials when the true value of p is 0.5.
  - ✓ Notice that the values of $\hat{p}$ bounce around 0.5 when the number of trials is small but approach and stay very close to $p = 0.5$ as the number of trials increases.
  - ✓ The single sequence of 1000 trials illustrated in Figure resulted (for larger n) in values for the estimate that were very close to the true value, $p = 0.5$.



This single sequence of result is random!

Motivation

- Would additional sequences yield similar results?
- Figure 9.2 shows the combined results of 50 sequences of 1000 trials.
  - ✓ the 50 distinct sequences were not identical. Rather, Figure shows a "convergence" of sorts to the true value $p = 0.5$.
  - ✓ This is exhibited by a wider spread of the values of the estimates for smaller numbers of trialsbutamuchnarrowerspreadofvaluesoftheestimateswhenthenumberoftrials is larger.



- How can we technically express the type of "convergence" exhibited in Figure?
  - ➢ Because $\hat{p} = Y/n$ is a random variable, we may express this "closeness" to p ***in probabilistic terms***.

Definition

DEFINITION 9.2

The estimator $\hat{\theta}_n$ is said to be a *consistent estimator* of θ if, for any positive number ε,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| \le \varepsilon) = 1$$

or, equivalently,

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

- The notation $\hat{\theta}_n$ expresses that the estimator for θ is calculated by using a sample of size n.
- If this probability in fact does tend to 1 as $n \to \infty$, we then say that
  - (Y/n) is a consistent estimator of p, or
  - (Y/n) "converges in probability to p."

The condition for a Consistent Unbiased Estimator

THEOREM 9.1

An unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator of θ if

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0.$$

Proof:

If Y is any random variable with $E(Y) = \mu$ and $V(Y) = \sigma^2 < \infty$ and if k is any nonnegative constant,

Tchebysheff's theorem (see Theorem 4.13) implies that $P(|Y - \mu| > k\sigma) \leq \frac{1}{k^2}$.

$$P\left(\left|\hat{\theta}_n - \theta\right| > k\sigma_{\hat{\theta}_n}\right) < \frac{1}{k^2}$$

Let n be any fixed sample size. For any positive number ε,

$$k = \varepsilon/\sigma_{\hat{\theta}_n}$$

Is a positive number. For this fixed n and this k shows that

$$P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = P\left(\left|\hat{\theta}_n - \theta\right| > [\frac{\varepsilon}{\sigma_{\hat{\theta}_n}}]\sigma_{\hat{\theta}_n}\right) \leq \frac{1}{\left(\frac{\varepsilon}{\sigma_{\hat{\theta}_n}}\right)^2} = \frac{V\left(\hat{\theta}_n\right)}{\varepsilon^2}$$

The condition for a Consistent Unbiased Estimator

THEOREM 9.1

An unbiased estimator $\hat{\theta}_n$ for θ is a consistent estimator of θ if

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0.$$

Proof:

$$P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) = P\left(\left|\hat{\theta}_n - \theta\right| > [\frac{\varepsilon}{\sigma_{\hat{\theta}_n}}]\sigma_{\hat{\theta}_n}\right) \le \frac{1}{\left(\frac{\varepsilon}{\sigma_{\hat{\theta}_n}}\right)^2} = \frac{V\left(\hat{\theta}_n\right)}{\varepsilon^2}$$

Thus, for any fixed n,

$$0 \le P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \le \frac{V\left(\hat{\theta}_n\right)}{\varepsilon^2}$$

If $\lim_{n\to\infty} V\left(\hat{\theta}_n\right) = 0$ and take the limit as $n \to \infty$ of the preceding sequence of probabilities,

$$\lim_{n\to\infty}(0) \le \lim_{n\to\infty} P\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \le \lim_{n\to\infty} \frac{V\left(\hat{\theta}_n\right)}{\varepsilon^2} = 0$$

Thus, $\hat{\theta}_n$ is a consistent estimator for θ

Overview

THEOREM 9.2

Suppose that $\hat{\theta}_n$ converges in probability to θ and that $\hat{\theta}'_n$ converges in probability to θ'.

a $\hat{\theta}_n + \hat{\theta}'_n$ converges in probability to $\theta + \theta'$.
b $\hat{\theta}_n \times \hat{\theta}'_n$ converges in probability to $\theta \times \theta'$.
c If $\theta' \neq 0$, $\hat{\theta}_n / \hat{\theta}'_n$ converges in probability to θ / θ'.
d If $g(\cdot)$ is a real-valued function that is continuous at θ, then $g(\hat{\theta}_n)$ converges in probability to $g(\theta)$.

Overview

EXAMPLE 9.3

Suppose that $Y_1, Y_2, \ldots, Y_n$ represent a random sample such that $E(Y_i) = \mu$, $E(Y_i^2) = \mu_2'$ and $E(Y_i^4) = \mu_4'$ are all finite. Show that

$$S_n^2 = \frac{1}{(n-1)} \Sigma_{i=1}^{n} (Y_i - \bar{Y}_n)^2$$

is a consistent estimator of $\sigma^2 = V(Y_i)$. (*Note:* We use subscript n on both S^2 and Y to explicitly convey their dependence on the value of the sample size n.)

Overview

SOLUTION 9.3

$$S_n^2 = \frac{1}{(n-1)}\Sigma_{i=1}^{n}(Y_i - \bar{Y}_n)^2 = \frac{1}{(n-1)}\left(\Sigma_{i=1}^{n}Y_i^2 - n\bar{Y}_n^2\right) = \frac{n}{(n-1)}\left(\frac{1}{n}\Sigma_{i=1}^{n}Y_i^2 - \bar{Y}_n^2\right)$$

By the law of large numbers, $\frac{1}{n}\Sigma_{i=1}^{n}Y_i^2$ converges in probability to μ_2'. Also, $\bar{Y}_n$ converges in probability to μ. Because the function $g(x) = x^2$ is continuous for all finite values of x, $\bar{Y}_n^2$ converges in probability to μ^2. Then, $\frac{1}{n}\Sigma_{i=1}^{n}Y_i^2 - \bar{Y}_n^2$ converges in probability to $\mu_2' - \mu^2 = \sigma^2$.

$\rightarrow$ S_n^2, the sample variance, is a consistent estimator for σ^2, the population variance.

Large-Sample Confidence Interval

- We considered **large-sample confidence intervals** for some parameters of practical interest.
- In particular, if $Y_1, Y_2, \dots, Y_n$ is a random sample from **any distribution** with mean μ and variance σ^2, we established that

$$\bar{Y} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Is a valid large-sample confidence interval with confidence coefficient approximately equal to $(1-\alpha)$.

- ✓ If σ^2 is known, this interval can be computed
- ✓ If σ^2 is unknown but the sample size is large we recommended substituting S for σ in the calculation because this entails no significant loss of accuracy

Large-Sample Confidence Interval

THEOREM 9.3

Suppose that U_n has a distribution function that converges to a standard normal distribution function as $n \to \infty$. If W_n converges in probability to 1, then the distribution function of U_n/W_n converges to a standard normal distribution function.

- This result follows from a general result known as Slutsky's theorem (Serfling, 2002).
- The proof of this result is beyond the scope of this text

Overview

EXAMPLE 9.4

Suppose that $Y_1, Y_2, \ldots, Y_n$ is a random sample of size *n* from a distribution with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$. Define S_n^2 as $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2$.

Show that the distribution function of $\sqrt{n}\left(\frac{\bar{Y}_n - \mu}{S_n}\right)$ converges to a standard normal distribution function.

Overview

SOLUTION 9.4

In Example 9.3, we showed that S_n^2 converges in probability to σ^2. Notice that $g(x) = \sqrt{x/c}$ is a continuous function of *x* if both *x* and *c* are positive. Hence, it follows from Theorem 9.2(d) that $S_n/\sigma = \sqrt{S_n^2/\sigma^2}$ converges in probability to 1.

We also know from the central limit theorem (Theorem 7.4) that the distribution function of $U_n = \sqrt{n}\left(\frac{\bar{Y}_n-\mu}{\sigma}\right)$ converges to a standard normal distribution function. Therefore, Theorem 9.3 implies that the distribution function of $\sqrt{n}\left(\frac{\bar{Y}_n-\mu}{\sigma}\right)/\frac{S_n}{\sigma} = \sqrt{n}\left(\frac{\bar{Y}_n-\mu}{S_n}\right)$ converges to a standard normal distribution function.

Motivation

- The result of Example 9.4 tells us that,
  - ✓ when n is large,
  - ✓ $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ has approximately a standard normal distribution **whatever is the form of the distribution from which the sample is taken**

- If the sample is taken from **a normal distribution**,
  - ✓ when n is small or large (does not matter),
  - ✓ $t = \sqrt{n}(\bar{Y}_n - \mu)/S_n$ **has a *t* distribution** with $n - 1$ degrees of freedom (df)

Motivation

- The result of Example 9.4 tells us that,
  - ✓ The sample is taken from **any distribution**
  - ✓ when n is **large**,
  - ✓ $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ has approximately **a standard normal distribution**

- The result of Chapter 7 implies
  - ✓ If the sample is taken from **a normal distribution**,
  - ✓ when $\boldsymbol{n}$ **is small or large (does not matter),**
  - ✓ $t = \sqrt{n}(\bar{Y}_n - \mu)/S_n$ **has a $\boldsymbol{t}$ distribution** with $n - 1$ degrees of freedom (df)

- If a large sample is taken from a normal distribution, the distribution function of $t = \sqrt{n}(\bar{Y}_n - \mu)/S_n$ can be approximated by a standard normal distribution function.
  - ➢ That is, as n gets large and hence as the number of degrees of freedom gets large, the t-distribution function converges to the standard normal distribution function.

Confidence Interval

- If we obtain a large sample from any distribution, we know from Example 9.4 that
  - ✓ $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ has approximately **a standard normal distribution, thus**

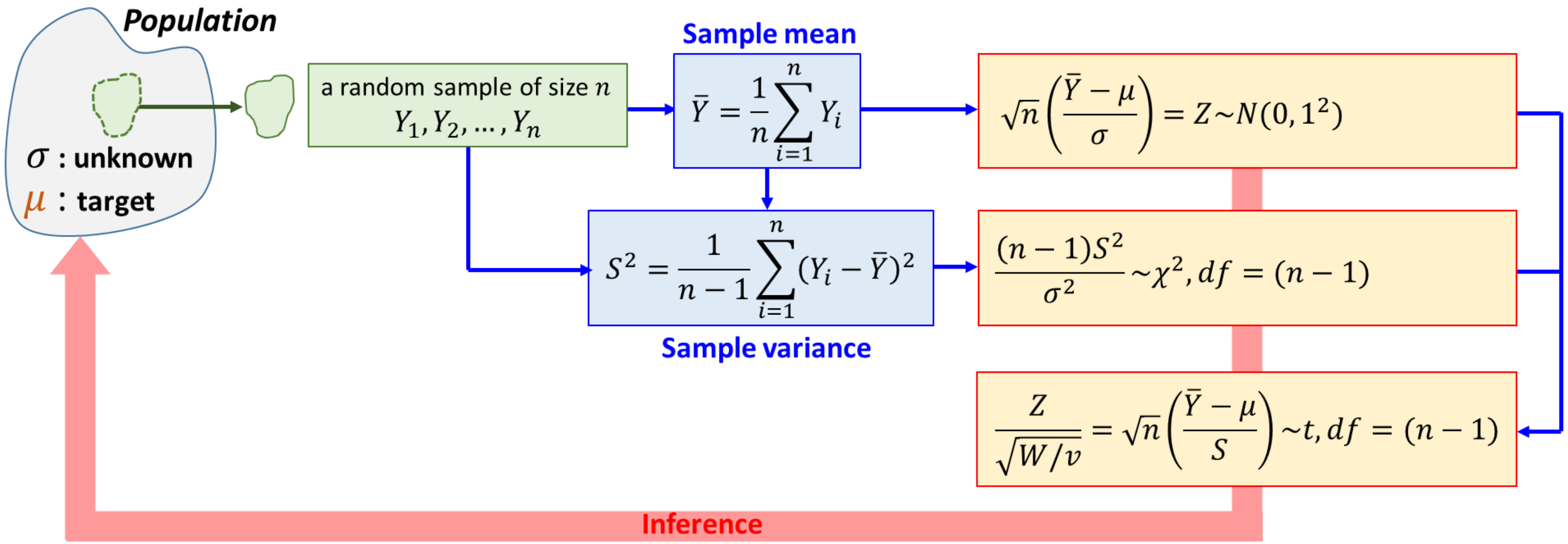
$$P\left[-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{Y}_n - \mu)}{S_n} \leq -z_{\alpha/2}\right] \approx 1 - \alpha$$

$$P\left[\bar{Y}_n - z_{\alpha/2}\left(\frac{S_n}{\sqrt{n}}\right) \leq \mu \leq \bar{Y}_n + z_{\alpha/2}\left(\frac{S_n}{\sqrt{n}}\right)\right] \approx 1 - \alpha$$

- Thus, $\bar{Y}_n \pm z_{\alpha/2}\left(\frac{S_n}{\sqrt{n}}\right)$ forms a valid large-sample confidence interval for μ with confidence coefficient approximately equal to $1 - \alpha$

- Similarly, $\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n \hat{q}_n}{n}}$ is a valid large-sample confidence interval for p

Motivation



- We have used the information in a **sample of size n** to calculate the value of **two statistics** that function as estimators for the parameters of interest.
- At this stage, **the actual sample values** are no longer important; rather, we summarize the information in the sample that relates to the parameters of interest by using the **statistics $\bar{Y}$ and S^2**.
  - Has this process of summarizing or reducing the data to the **two statistics, $\bar{Y}$ and S^2**, retained all the information about **μ and σ^2** in the **original set of n sample observations**?
  - Or has some information about these parameters been lost or obscured through the process of reducing the data?

Motivation

- We present methods for finding statistics that in a sense summarize all the information in a sample about a target parameter.

- Such statistics are said to have the ***property of sufficiency***; or more simply, they are called ***sufficient statistics***.

- As we will see in the next section, "good" estimators are (or can be made to be) functions of any sufficient statistic.
  - Indeed, sufficient statistics often can be used to develop estimators that have the minimum variance among all unbiased estimators.

Illustrative Example

- let us consider the outcomes of n trials of a binomial experiment, $X_1, X_2, \dots, X_n$, where

$$X_1 = \begin{cases} 1, & \text{if the ith trial is a scucess,} \\ 0, & \text{if the ith trial is a failure.} \end{cases}$$

- If p is the probability of success on any trial then, for $i = 1, 2, \dots, n$,

$$X_1 = \begin{cases} 1, & \text{with probability } p, \\ 0, & \text{with probability } q = 1 - p. \end{cases}$$

- Suppose that we are given a value of $Y = \sum_{i=1}^{n} X_i$, the number of successes among the n trials
- If we know the value of Y, can we gain any further information about p by looking at other functions of $X_1, X_2, \dots, X_n$?

Illustrative Example

- One way to answer this question is to look at the conditional distribution of $X_1, X_2, \dots, X_n$, given Y:

$$P(X_1, = x_1, X_2 = x_2, \dots, X_n = x_n | Y = y) = \frac{P(X_1, = x_1, X_2 = x_2, \dots, X_n = x_n, Y = y)}{P(Y = y)}$$

  ✓ The number on the right side of this expression is 0 if $\sum_{i=1}^{n} X_i \neq y$

  ✓ The denominator is the binomial probability of exactly y successes in n trials.

- Therefore, if $y = 0,1,2, \dots, n,$

$$P(X_1, = x_1, X_2 = x_2, \dots, X_n = x_n | Y = y) = \begin{cases} \dfrac{p^y(1-p)^{n-y}}{\binom{n}{y} p^y (1-p)^{n-y}} = \dfrac{1}{\binom{n}{y}}, & \text{if } \sum_{i=1}^{n} X_i = y \\ 0, & \text{otherwise} \end{cases}$$

- It is important to note that the conditional distribution of $X_1, X_2, \dots, X_n$, given Y **does not depend upon** p.

  ✓ That is, once Y is known, no other function of $X_1, X_2, \dots, X_n$ will give information about p.

  ✓ In this sense, Y contains **all the information** about p.

  ✓ Therefore, the statistic Y is said to be **sufficient** for p.

Overview

DEFINITION 9.3

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a probability distribution with unknown parameter θ. Then the statistic $U = g(Y_1, Y_2, \ldots, Y_n)$ is said to be ***sufficient*** for θ if the conditional distribution of $Y_1, Y_2, \ldots, Y_n$, given U, does not depend on θ.

$$\theta \longrightarrow U = g(Y_1, Y_2, \ldots, Y_n) \longrightarrow Y_1, Y_2, \ldots, Y_n$$

- we adopt notation that will permit us to explicitly display the fact that the distribution associated with a random variable Y often depends on the value of a parameter θ
  - If Y is a discrete random variable that has a probability mass function that depends on the value of a parameter θ, we use the notation $p(y|\theta)$ instead of $p(y)$
  - If Y is a continuous random variable that has a density function that depends on the value of a parameter θ, we use the notation $f(y|\theta)$ instead of $f(y)$

Then how to find the sufficient statistics?

Overview

DEFINITION 9.4

Let $y_1, y_2, \ldots, y_n$ be sample observations taken on corresponding random variables $Y_1, Y_2, \ldots, Y_n$ whose distribution depends on a parameter θ.

- If $Y_1, Y_2, \ldots, Y_n$ are discrete random variables, the *likelihood of the sample*, $L(y_1, y_2, \ldots, y_n | \theta)$, is defined to be the joint probability of $y_1, y_2, \ldots, y_n$.
- If $Y_1, Y_2, \ldots, Y_n$ are continuous random variables, the *likelihood of the sample*, $L(y_1, y_2, \ldots, y_n | \theta)$, is defined to be the joint density evaluated at $y_1, y_2, \ldots, y_n$.

- If the set of random variables $Y_1, Y_2, \ldots, Y_n$ denotes a random sample from a discrete distribution with probability function $p(y|\theta)$, then

$$\begin{aligned} L(y_1, y_2, \ldots, y_n | \theta) &= p(y_1, y_2, \ldots, y_n | \theta) \\ &= p(y_1|\theta) \times p(y_2|\theta) \times \cdots \times p(y_n|\theta) \end{aligned}$$

- If the set of random variables $Y_1, Y_2, \ldots, Y_n$ denotes a random sample from a continuous distribution with density function $f(y|\theta)$, then

$$\begin{aligned} L(y_1, y_2, \ldots, y_n | \theta) &= f(y_1, y_2, \ldots, y_n | \theta) \\ &= f(y_1|\theta) \times f(y_2|\theta) \times \cdots \times f(y_n|\theta) \end{aligned}$$

Factorization Criterion

THEOREM 9.4

Let U be a statistic based on the random sample $Y_1, Y_2, \ldots, Y_n$. Then U is a *sufficient statistic* for the estimation of a parameter θ if and only if the likelihood $L(\theta) = L(y_1, y_2, \ldots, y_n|\theta)$ can be factored into two nonnegative functions,

$$L(y_1, y_2, \ldots, y_n|\theta) = g(u, \theta) \times h(y_1, y_2, \ldots, y_n)$$

where $g(u, \theta)$ is a function only of u and θ and $h(y_1, y_2, \ldots, y_n)$ is not a function of θ.

Proof: beyond the scope this book

Example

EXAMPLE 9.5

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample in which Y_i possesses the probability density function

$$f(y_i|\theta) = \begin{cases} \left(\frac{1}{\theta}\right) e^{-y_i/\theta}, & 0 \leq y_i < \infty \\ 0, & \text{elsewhere} \end{cases}$$

where $\theta > 0, \quad i = 1, 2, \ldots, n$. Show that $\bar{Y}$ is a sufficient statistic for the parameter θ.

Example

SOLUTION 9.5

The likelihood $L(\theta)$ of the sample is the joint density

$$L(y_1, y_2, \dots, y_n|\theta) = f(y_1, y_2, \dots, y_n|\theta)$$
$$= f(y_1|\theta) \times f(y_2|\theta) \times \cdots \times f(y_n|\theta)$$
$$= \frac{e^{-y_1/\theta}}{\theta} \times \frac{e^{-y_2/\theta}}{\theta} \times \cdots \times \frac{e^{-y_n/\theta}}{\theta} = \frac{e^{-n\bar{y}/\theta}}{\theta^n}$$

Notice that $L(\theta)$ is a function only of θ and y and that if $g(y,\theta) = \frac{e^{-n\bar{y}/\theta}}{\theta^n}$ and $h(y_1, y_2, \dots, y_n) = 1$,

Then $L(y_1, y_2, \dots, y_n|\theta) = g(y,\theta) \times h(y_1, y_2, \dots, y_n)$.

Hence, Theorem 9.4 implies that Y is a sufficient statistic for the parameter θ.

- Any statistics that is a one-to-one function of $\bar{Y}$ is a sufficient statistics

Comments

- Theorem 9.4 can be used to show that there are many possible sufficient statistics for any one population parameter.
  - First of all, according to Definition 9.3 or the factorization criterion (Theorem9.4), the random sample itself is a sufficient statistic.
  - Second, if $Y_1, Y_2, \dots, Y_n$ denote a random sample from a distribution with a density function with parameter θ, then the set of order statistics $Y(1) \le Y(2) \le \dots \le Y(n)$, which is a function of Y_1, $Y_2, \dots, Y_n$, is sufficient for θ

- Although many statistics are sufficient for the parameter θ associated with a specific distribution, application of **the factorization criterion** typically leads to a statistic that provides **the "best" summary of the information in the data**.

- In the next section, we show how these sufficient statistics can be used to develop unbiased estimators with minimum variance.

Motivation

- Sufficient statistics play an important role in finding good estimators for parameters
- If $\hat{\theta}$ is an unbiased estimator for θ
- If U is a statistic that is sufficient for θ
- Then there is a function $h(U)$ of U that
  - is an unbiased estimator for θ : $E[h(U)] = \theta$
  - has no larger variance than $\hat{\theta}$: $V[h(U)] \leq V(\hat{\theta})$
- If we seek unbiased estimators with small variances, we can restrict our search to **estimators that are functions of sufficient statistics**.

The Rao-Blackwell Theorem

THEOREM 9.5 (The Rao-Blackwell Theorem)

Let $\hat{\theta}$ be an unbiased estimator for θ such that $V(\hat{\theta}) < \infty$. If U is a sufficient statistic for θ, define $\hat{\theta}^* = E(\hat{\theta}|U)$. Then, for all θ,

$$E(\hat{\theta}^*) = \hat{\theta} \text{ and } V(\hat{\theta}^*) \le V(\hat{\theta}).$$

Proof:

- Because U is sufficient for θ, the conditional distribution of any statistic (including $\hat{\theta}$), given U, does not depend on θ.
  - ✓ Thus, $\hat{\theta}^* = E(\hat{\theta}|U)$ is not a function of θ and is therefore a statistic.
- *Theorem* 5.14 implies that $E(\hat{\theta}^*) = E[E(\hat{\theta}|U)] = E(\hat{\theta}) = \theta$. ($\because$ $\hat{\theta}$ is an unbiased estimator for θ)
  - ✓ This concludes that $\hat{\theta}^*$ is **an unbiased estimator**
- *Theorem 5.15* implies that $V(\hat{\theta}) = V[E(\hat{\theta}|U)] + E[V(\hat{\theta}|U)] = V(\hat{\theta}^*) + E[V(\hat{\theta}|U)]$.
  - ✓ because $V(\hat{\theta}|U = u) \ge 0$ for all u, it follows that $E[V(\hat{\theta}|U)] \ge 0$
  - ✓ therefore that $V(\hat{\theta}) \ge V(\hat{\theta}^*)$
  - ✓ This concludes that $\hat{\theta}^*$ is **an estimator with minimum variance**

The Rao-Blackwell Theorem

- Theorem 9.5 implies that ***an unbiased estimator for θ with a small variance*** is or can be made to be a ***function of a sufficient statistic***

- Because many statistics are sufficient for a parameter θ associated with a distribution, **which sufficient statistic should we use when we apply this theorem**?

- For the distributions that we discuss in this text, ***the factorization criterion*** typically identifies a statistic U that best summarizes the information in the data about the parameter θ.
  - ✓ Such statistics are called ***minimal sufficient statistics***.

- These statistics possess another property (***completeness***) that guarantees that, if we apply Theorem 9.5 using U, we not only get an estimator with a smaller variance but also actually obtain ***an unbiased estimator for θ with minimum variance***.

  - ✓ Such an estimator is called ***a minimum-variance unbiased estimator (MVUE).***

The Rao-Blackwell Theorem

- if we start with an unbiased estimator for a parameter θ and the sufficient statistic obtained through the factorization criterion, application of the Rao–Blackwell theorem typically leads to an MVUE for the parameter.

an unbiased estimator for a parameter θ

↓

sufficient obtained through the factorization criterion
(U is the minimal sufficient statistics for θ)

↓

application of the Rao–Blackwell theorem
$\hat{\theta}^* = h(U)$

↓

a minimum-variance unbiased estimator (MVUE)

- **step1**: U is the sufficient statistic that best summarizes the data
- **step2**: Some function of U, $h(U)$, can be found such that $E[h(U)] = \theta$
- **step3**: It follows that $h(U)$ is the MVUE for θ.

Example

EXAMPLE 9.6

Let $Y_1, Y_2, \ldots, Y_n$ denote a random sample from a distribution where $P(Y_i = 1) = p$ and $P(Y_i = 0) = 1 - p$, with p unknown (such random variables are often called *Bernoulli* variables).

Use the factorization criterion to find a sufficient statistic that best summarizes the data.
Give an MVUE for p.

Example

SOLUTION 9.6

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i}, \; y_i = 0, 1$$

Thus, the likelihood $L(p)$ is

$$L(y_1, y_2, \dots, y_n|p) = p(y_1, y_2, \dots, y_n|p) = p^{y_1}(1-p)^{1-y_1} \times \cdots \times p^{y_n}(1-p)^{1-y_n}$$
$$= p^{\Sigma y_i}(1-p)^{n-\Sigma y_i} \times 1 = g(\Sigma y_i, p) \times h(y_1, \dots, y_n)$$

According to the factorization criterion, $U = \Sigma_{i=1}^{n} Y_i$ is sufficient for p. This statistic best summarizes the information about the parameter p. Notice that $E(U) = np$, or equivalently, $E(U/n) = p$. Thus, $U/n = \bar{Y}$ is an unbiased estimator for p. Because this estimator is a function of the sufficient statistic $\Sigma_{i=1}^{n} Y_i$, the estimator $\hat{p} = \bar{Y}$ is the MVUE for p.

Example

EXAMPLE 9.7

Suppose that $Y_1, Y_2, \ldots, Y_n$ denote a random sample from the Weibull density function, given by

$$f(y|\theta) = \begin{cases} \left(\frac{2y}{\theta}\right) e^{-y^2/\theta}, & y > 0 \\ 0, & elsewhere \end{cases}$$

Find an MVUE for θ.

Example

SOLUTION 9.7

We begin by using the factorization criterion to find the sufficient statistic that best summarizes the information about θ.

$$L(y_1, y_2, \dots, y_n|\theta) = f(y_1, y_2, \dots, y_n|\theta)$$
$$= \left(\frac{2}{\theta}\right)^n (y_1 \dots y_n) \exp\left(-\frac{1}{\theta}\Sigma_{i=1}^n y_i^2\right)$$
$$= \left(\frac{2}{\theta}\right)^n \exp\left(-\frac{1}{\theta}\Sigma_{i=1}^n y_i^2\right)(y_1 \dots y_n)$$

Thus, $U = \Sigma_{i=1}^n Y_i^2$ is the minimal sufficient statistic for θ.
We now find a function of this statistic that is unbiased for θ. Letting $W = Y_i^2$,

$$f_W(w) = f(\sqrt{w})\frac{d(\sqrt{w})}{dw} = \left(\frac{2}{\theta}\right)\left(\sqrt{w}e^{-\frac{w}{\theta}}\right)\left(\frac{1}{2\sqrt{w}}\right) = \left(\frac{1}{\theta}\right)e^{-w/\theta}, \qquad w > 0$$

That is, Y_i^2 has an exponential distribution with parameter θ. Because $E\left(Y_i^2\right) = E(W) = \theta$ and $E\left(\Sigma_{i=1}^n Y_i^2\right) = n\theta$, it follows that $\hat{\theta} = \frac{1}{n}\Sigma_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{Y}_{\mathrm{i}}^2$ is an unbiased estimator of ϑ that is a function of the sufficient statistic $\Sigma_{i=1}^n Y_i^2$. Therefore, $\hat{\theta}$ is an MVUE of the Weibull parameter θ.

Example

EXAMPLE 9.8

Suppose $Y_1, Y_2, \dots, Y_n$ denotes a random sample from a normal distribution with unknown mean μ and variance σ^2. Find the MVUEs for μ and σ^2.

Example

SOLUTION 9.8

We first look at the likelihood function,

$$L(y_1, \dots, y_n|\mu, \sigma^2) = f(y_1, \dots, y_n|\mu, \sigma^2)$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\Sigma_{i=1}^n (y_i - \mu)^2\right)$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\left(\Sigma_{i=1}^n y_i^2 - 2\mu\Sigma_{i=1}^n y_i + n\mu^2\right)\right)$$
$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{n\mu^2}{2\sigma^2}\right) \exp\left[-\frac{1}{2\sigma^2}\left(\Sigma_{i=1}^n y_i^2 - 2\mu\Sigma_{i=1}^n y_i\right)\right]$$

Thus, $\Sigma_{i=1}^n Y_i$ and $\Sigma_{i=1}^n Y_i^2$, jointly, are sufficient statistics for μ and σ^2. We know from past work that Y is unbiased for μ and

$$S^2 = \frac{1}{n-1}\Sigma_{i=1}^n (Y_i - \bar{Y})^2 = \frac{1}{n-1}\left[\Sigma_{i=1}^n Y_i^2 - n\bar{Y}^2\right]$$

is unbiased for σ^2. Because these estimators are functions of the statistics that best summarize the information about μ and σ^2, they are MVUEs for μ and σ^2.

Motivation

- The method of moments is a very simple procedure for finding an estimator for one or more population parameters.
- Recall that the kth moment of a random variable, taken about the origin, is

$$\mu_k' = E(Y^k)$$

- The corresponding kth sample moment is the average

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

- The method of moments is based on the intuitively appealing idea that sample moments should provide good estimates of the corresponding population moments.
  - ✓ That is, m_k' should be a good estimator of μ_k' for $k = 1, 2, \ldots$.
  - ✓ Then because the population moments $\mu_1', \mu_2', \ldots, \mu_k'$ are functions of the population parameters, we can equate corresponding population and sample moments and solve for the desired estimators

Method of Moments

Method of Moments

Choose as estimates those values of the parameters that are solutions of the equations, $\mu'_k = m'_k$ for $k = 1,2,\ldots,t$, where t is the number of parameters to be estimated.

Example

EXAMPLE 9.11

A random sample of n observations, $Y_1, Y_2, \ldots, Y_n$, is selected from a population in which Y_i , for $i = 1, 2, \ldots, n$, possesses a uniform probability density function over the interval $(0, \theta)$ where θ is unknown. Use the method of moments to estimate the parameter θ.

Example

SOLUTION 9.11

The value of μ_1' for a uniform random variable is

$$\mu_1' = \mu = \frac{\theta}{2}$$

The corresponding first sample moment is

$$m_1' = \frac{1}{n}\Sigma_{i=1}^{n} Y_i = \bar{Y}$$

Equating the corresponding population and sample moment, we obtain

$$\mu_1' = \frac{\theta}{2} = \bar{Y}$$

The method-of-moments estimator for θ is the solution of the above equation. That is, $\hat{\theta} = 2\bar{Y}$.

Example

EXAMPLE 9.12

Show that the estimator $\hat{\theta} = 2\overline{Y}$, derived in Example 9.11, is a consistent estimator for θ.

Example

SOLUTION 9.12

In Example 9.1, we showed that $\hat{\theta} = 2\bar{Y}$ is an unbiased estimator for θ and that $V(\hat{\theta}) = \theta^2/3n$. Because $\lim_{n\to\infty} V(\hat{\theta}) = 0$, Theorem 9.1 implies that $\hat{\theta} = 2\bar{Y}$ is a consistent estimator for θ.

Note:

- Although the estimator $\hat{\theta}$ derived in Example 9.11 is consistent, it is not necessarily the best estimator for θ.
- Indeed, the factorization criterion yields $Y(n) = \max(Y_1, Y_2, \dots, Y_n)$ to be the best sufficient statistic for θ.
- Thus, according to the Rao–Blackwell theorem, the method-of-moments estimator will have larger variance than an unbiased estimator based on $Y(n)$. This, in fact, was shown to be the case in Example 9.1.

Example

EXAMPLE 9.13

A random sample of n observations, $Y_1, Y_2, \ldots, Y_n$, is selected from a population where Y_i , for $i = 1, 2, \ldots, n$, possesses a gamma probability density function with parameters α and β (see Section 4.6 for the gamma probability density function).

Find method-of-moments estimators for the unknown parameters α and β.

Example

SOLUTION 9.13

Because we seek estimators for two parameters α and β, we must equate two pairs of population and sample moments. The first two moments of the gamma distribution with parameters α and β are

$$\mu_1' = \mu = \alpha\beta$$
$$\mu_2' = \sigma^2 + \mu^2 = \alpha\beta^2 + \alpha^2\beta^2$$

Now equate these quantities to their corresponding sample moments and solve for $\hat{\alpha}$ and $\hat{\beta}$. Thus,

$$\mu_1' = \alpha\beta = m_1' = \bar{Y}$$
$$\mu_2' = \alpha\beta^2 + \alpha^2\beta^2 = m_2' = \frac{1}{n}\Sigma_{i=1}^{n} Y_i^2$$

From the first equation, we obtain $\hat{\beta} = \bar{Y}/\hat{\alpha}$. Substituting into the second equation and solving for $\hat{\alpha}$, we obtain

$$\hat{\alpha} = \frac{\bar{Y}^2}{\left(\frac{\Sigma_{i=1}^{n} Y_i^2}{n}\right) - \bar{Y}^2} = \frac{n\bar{Y}^2}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Substituting $\hat{\alpha}$ into the first equation, we obtain

$$\hat{\beta} = \frac{\bar{Y}}{\hat{\alpha}} = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n\bar{Y}}$$

Remarks

- To summarize, the method of moments finds estimators of unknown parameters by equating corresponding sample and population moments.
- The method is easy to employ and provides ***consistent estimators***.
- However, the estimators derived by this method are often not functions of sufficient statistics.
  - ✓ As a result, method-of-moments estimators are sometimes ***not very efficient***.
- In many cases, the method-of-moments estimators are ***biased***.
- The primary virtues of this method are ***its ease of use*** and that it sometimes yields estimators with reasonable properties.

Motivation

an unbiased estimator for a parameter θ

↓

sufficient obtained through the factorization criterion
(U is the minimal sufficient statistics for θ)

↓

application of the Rao–Blackwell theorem
$\hat{\theta}^* = h(U)$

↓

a minimum-variance unbiased estimator (MVUE)

- **step1**: U is the sufficient statistic that best summarizes the data
- **step2**: Some function of U, $h(U)$, can be found such that $E[h(U)] = \theta$
- **step3**: It follows that $h(U)$ is the MVUE for θ.

Although we have a method for finding a sufficient statistic, the determination of the function of the minimal sufficient statistic that gives us an unbiased estimator can be largely **a matter of hit or miss**.

Motivation

We know that there are a total three balls but do not know the #red balls and #blue balls

We have draw **two red balls**

What would be a good estimate of the total number of red balls in the box?

Motivation

$$P(\text{draw two red balls}|\text{two red balls}) = \frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

$$P(\text{draw two red balls}|\text{three red balls}) = \frac{\binom{2}{3}}{\binom{2}{3}} = 1$$

- It should seem reasonable to choose three as the estimate of the number of red balls in the box **because this estimate maximizes the probability of obtaining the observed sample**.
  - ✓ Of course, it is possible for the box to contain only two red balls, but the observed outcome gives more credence to there being three red balls in the box.

Method of Maximum Likelihood

Method of Moments

Suppose that the likelihood function depends on k parameters $\theta_1, \theta_2, \dots, \theta_k$. Choose as estimates those values of the parameters that maximize the likelihood

$$L(y_1, y_2, \dots, y_n | \theta_1, \theta_2, \dots, \theta_k)$$

- To emphasize the fact that the likelihood function is a function of the parameters $\theta_1, \theta_2, \dots, \theta_k$, we sometimes write the likelihood function as $L(\theta_1, \theta_2, \dots, \theta_k)$.

- It is common to refer to maximum-likelihood estimators as MLEs

Example

EXAMPLE 9.14

A binomial experiment consisting of n trials resulted in observations $y_1, y_2, \ldots, y_n$, where $y_i = 1$ if the ith trial was a success and $y_i = 0$ otherwise. Find the MLE of p, the probability of a success.

Example

SOLUTION 9.14

$$L(p) = L(y_1, \dots, y_n|p) = p^y(1-p)^{n-y}, \qquad where\ y = \Sigma_{i=1}^{n} y_i$$

If $y = 0$, $L(p)$ is maximized when $p = 0$. If $y = n$, $L(p)$ is maximized when $p = 1$. If $y = 1, 2, \dots, n-1$, then $L(p) = p^y(1- \ p)^{n-y}$ is zero when $p = 0$ and $p = 1$ and is continuous for values of p between 0 and 1. Thus, we can find the value of p that maximizes $L(p)$ by setting the derivative $dL(p)/dp$ equal to 0 and solving for p.

Both $\ln[L(p)]$ and $L(p)$ are maximized for the same value of p since $\ln[L(p)]$ is a monotonically increasing function of $L(p)$. We have

$$\ln[L(p)] = \ln[p^y(1-p)^{n-y}] = y\ln p + (n-y)\ln(1-\mathrm{p})$$

If $y = 1, 2, \dots, n-1$, the derivative of $\ln[L(p)]$ with respect to p, is

$$\frac{d\ln[L(p)]}{dp} = y\left(\frac{1}{p}\right) + (n-y)\left(\frac{-1}{1-p}\right)$$

We obtain the estimate $\hat{p} = y/n$. Because $L(p)$ is maximized at $p = 0$ when $y = 0$, at $p = 1$ when $y = n$ and at $p = y/n$ when $y = 1, 2, \dots, n-1$, whatever the observed value of y, $L(p)$ is maximized when $p = y/n$.
The MLE, $\hat{p} = Y/n$, is the fraction of successes in the total number of trials n. Hence, the MLE of p is actually the intuitive estimator for p that we used throughout Chapter 8.

Example

EXAMPLE 9.15

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Find the MLEs of μ and σ^2.

Example

SOLUTION 9.15

$$L(\mu,\sigma^2) = f(y_1,\ldots,y_n|\mu,\sigma^2)$$
$$= f(y_1|\mu,\sigma^2)\cdots f(y_n|\mu,\sigma^2)$$
$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(y_i-\mu)^2\right]$$

Then,

$$\ln[L(\mu,\sigma^2)] = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi - \frac{1}{2\sigma^2}\Sigma_{i=1}^{n}(y_i-\mu)^2$$

The MLEs of μ and σ^2 are the values that make $\ln[L(\mu,\sigma^2)]$ a maximum. Taking derivatives with respect to μ and σ^2, we obtain

$$\frac{\partial\{\ln[L(\mu,\sigma^2)]\}}{\partial\mu} = \frac{1}{\sigma^2}\Sigma_{i=1}^{n}(y_i-\mu)$$
$$\frac{\partial\{\ln[L(\mu,\sigma^2)]\}}{\partial\sigma^2} = -\left(\frac{n}{2}\right)\left(\frac{1}{\sigma^2}\right) + \frac{1}{2\sigma^4}\Sigma_{i=1}^{n}(y_i-\mu)^2$$

Thus,

$$\hat{\mu} = \frac{1}{n}\Sigma_{i=1}^{n}y_i = \bar{y}, \qquad \hat{\sigma}^2 = \frac{1}{n}\Sigma_{i=1}^{n}(y_i-\bar{y})^2$$

And both are the MLEs of μ and σ^2, respectively. Notice that $\bar{Y}$ is unbiased for μ. Although $\hat{\sigma}^2$ is not unbiased for σ^2, it can easily be adjusted to the unbiased estimator S^2. (see Example 8.1)

Example

EXAMPLE 9.16

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample of observations from a uniform distribution with probability density function $f\,(y_i|\theta) = 1/\theta$, for $0 \le y_i \le \theta$ and $i = 1, 2, \ldots, n$. Find the MLE of θ.

Example

SOLUTION 9.16

$$L(\theta) = f(y_1, \dots, y_n|\theta)$$
$$= \begin{cases} \frac{1}{\theta} \times \frac{1}{\theta} \times \cdots \times \frac{1}{\theta} = \left(\frac{1}{\theta}\right)^n, & if\ 0 \le y_i \le \theta, i = 1, \dots, n \\ 0, & otherwise \end{cases}$$

Obviously, $L(\theta)$ is not maximized when $L(\theta) = 0$. You will notice that $1/\theta^n$ is a monotonically decreasing function of θ. Hence, nowhere in the interval $0 < \theta < \infty$ is $d[1/\theta^n]/d\theta$ equal to zero. However, $1/\theta^n$ increases as θ decreases, and $1/\theta^n$ is maximized by selecting θ to be as small as possible, subject to the constraint that all of the y_i values are between zero and θ. The smallest value of θ that satisfies this constraint is the maximum observation in the set $y_1, y_2, \dots, y_n$. That is, $\hat{\theta} = Y_{(n)} = \max(Y_1, Y_2, \dots, Y_n)$ is the MLE for θ. This MLE for θ is not an unbiased estimator of θ, but it can be adjusted to be unbiased, as shown in Example 9.1.

Remarks

- We have seen that sufficient statistics that best summarize the data have desirable properties and often can be used to find an MVUE for parameters of interest.

$$\theta \longrightarrow Y_1, Y_2, \ldots, Y_n \longrightarrow U = f(Y_1, Y_2, \ldots, Y_n) \longrightarrow \text{MLE}$$

- If U is any sufficient statistic for the estimation of a parameter θ, including the sufficient statistic obtained from the factorization criterion, the ***MLE is always some function of U***.
  - ➢ The MLE depends on the sample observations only through the value of a sufficient statistic.
- To show this, we need only observe

$$L(\theta) = L(y_1, y_2, \ldots, y_n | \theta) = g(u, \theta) h(y_1, y_2, \ldots, y_n)$$
$$\Rightarrow \ln[L(\theta)] = \ln[g(u, \theta)] + \ln[h(y_1, y_2, \ldots, y_n)]$$

- Because $\ln[g(u, \theta)]$ depends on the data only through the value of the sufficient statistic U,the MLE for θ is always some function of U.
  - ➢ Consequently, if an MLE for a parameter can be found and then adjusted to be unbiased, the resulting estimator often is an MVUE of the parameter in question.

Property of MLE

- Generally, if θ is the parameter associated with a distribution, we are sometimes interested in estimating some function of θ—say $t(\theta)$—rather than θ itself

- If $t(\theta)$ is a function of θ and $\hat{\theta}$ is the MLE for θ. Then

$$\widehat{t(\theta)} = t(\hat{\theta})$$

The result, sometimes referred to as ***the invariance property of MLEs***

Example

EXAMPLE 9.17

In Example 9.14, we found that the MLE of the binomial proportion p is given by $\hat{p} = Y/n$. What is the MLE for the variance of Y?

Example

SOLUTION 9.17

The variance of a binomial random variable Y is given by $V(Y) = np(1 - p)$. Because $V(Y)$ is a function of the binomial parameter p—namely, $V(Y) = t(p)$ with $t(p) = np(1 - p)$—it follows that the MLE of $V(Y)$ is given by

$$\widehat{V(Y)} = \widehat{t(p)} = t(\hat{p}) = n\left(\frac{Y}{n}\right)\left(1 - \frac{Y}{n}\right)$$

This estimator is not unbiased. However, using the result in Exercise 9.65, we can easily adjust it to make it unbiased. Actually,

$$n\left(\frac{Y}{n}\right)\left(1 - \frac{Y}{n}\right)\left(\frac{n}{n-1}\right) = \left(\frac{n^2}{n-1}\right)\left(\frac{Y}{n}\right)\left(1 - \frac{Y}{n}\right)$$

is the UMVUE for $t(p) = np(1 - p)$.

Summary

- Good estimators are consistent and efficient when compared to other estimators.

- **The most efficient estimators**, those with the smallest variances, are ***functions of the sufficient statistics*** that best summarize all of the information about the parameter of interest.

- Two methods of finding estimators
  - ✓ **The method of moments**
    - ✓ consistent but
    - ✓ generally not very efficient
  - ✓ **The method of maximum likelihood (MLE)**
    - ✓ consistent and,
    - ✓ if adjusted to be unbiased, often lead to minimum-variance unbiased estimators.

Motivation

Population

σ : unknown

μ : target

a random sample of size n
$Y_1, Y_2, \dots, Y_n$

Sample mean

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right) = Z \sim N(0, 1^2)$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2, df = (n-1)$$

$$\frac{Z}{\sqrt{W/v}} = \sqrt{n}\left(\frac{\bar{Y}-\mu}{S}\right) \sim t, df = (n-1)$$

Inference