Chapter 9. Properties of Point Estimators and Methods of Estimation

9.1 Introduction

- Estimator $\widehat{\theta} = \widehat{\theta}_n = \widehat{\theta}(Y_1, \ldots, Y_n)$ for θ : a function of n random samples, $Y_1, \ldots, Y_n$.
- Sampling distribution of $\widehat{\theta}$: a probability distribution of $\widehat{\theta}$
- Unbiased estimator, $\widehat{\theta}$: $E(\widehat{\theta}) - \theta = 0$.
- Properties of $\widehat{\theta}$: efficiency, consistency, sufficiency
- Rao-Blackwell theorem : an unbiased estimator with small variance is a function of a sufficient statistic
- Estimation method
  - Minimum-Variance Unbiased Estimation
  - Method of Moments
  - Method of Maximum Likelihood

9.2 Relative Efficiency

- We would like to have an estimator with smaller bias and smaller variance : if one can find several unbiased estimators, we want to use an estimator with smaller variance.

- Relative efficiency

  (Def 9.1) Suppose $\hat{\theta}_1$ and $\hat{\theta}_2$ are two unbiased estimators for θ, with variances, $V(\hat{\theta}_1)$ and $V(\hat{\theta}_2)$, respectively.
  Then *relative efficiency* of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted $\text{eff}(\hat{\theta}_1, \hat{\theta}_2)$, is defined to be the ratio $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$

  - $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) > 1 : V(\hat{\theta}_2) > V(\hat{\theta}_1)$, and $\hat{\theta}_1$ is *relatively more efficient* than $\hat{\theta}_2$.

  - $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) < 1 : V(\hat{\theta}_2) < V(\hat{\theta}_1)$, and $\hat{\theta}_2$ is *relatively more efficient* than $\hat{\theta}_1$.

(Example) want to estimate the mean of a normal distribution. Let $\widehat{\theta}_{med}$ be the sample median and $\widehat{\theta}_{mean}$ be the sample mean. Then $\widehat{\theta}_{mean}$ is better than $\widehat{\theta}_{med}$. Why?
$eff(\widehat{\theta}_{med}, \widehat{\theta}_{mean}) = \frac{V(\widehat{\theta}_{mean})}{V(\widehat{\theta}_{med})} = \frac{\sigma^2/n}{(1.2533)^2\sigma^2/n} = .6366.$

(Exercise) Let $Y_1, \ldots, Y_n$ denote a random sample from a population with mean μ and σ^2. Consider the following three estimates for μ :
$\widehat{\mu}_1 = \frac{1}{2}(Y_1 + Y_2)$, $\widehat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2 + \ldots + Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n$, $\widehat{\mu}_3 = \bar{Y}$.
(a) Show that they are unbiased.
(b) Find the efficiency of $\widehat{\mu}_3$ relative to $\widehat{\mu}_2$ and $\widehat{\mu}_1$, respectively.

9.3 Consistency

- Motivation - (Tossing a coin)

  A coin is tossed n times independently, and p is the (unknown) probability of resulting in heads. Suppose we are interested in the number of heads among n tosses, Y.

  (Q1) what is the distribution of Y?

  Since p is unknown, consider $\widehat{p} = \widehat{p}_n = Y/n$. As n increases, the amount of information in the sample(here, the quality of $\widehat{p}_n$) also increases in the sense that $\widehat{p}_n = Y/n$ should be very close to p as $n \to \infty$.

  (Q2) How one can express "closeness" of $\widehat{p}_n$ to p?

  Since $\widehat{p}_n = Y/n$ is a statistic, consider the probability that $\mid \widehat{p}_n - p \mid$ will be less than some arbitrary positive number as n increases : what is the value of $P(\mid Y/n - p \mid \leq \epsilon)$ as $n \to \infty$?

- Consistency and related theorems

  (Def 9.2) $\widehat{\theta}_n = \widehat{\theta}(Y_1, \ldots, Y_n)$ is said to be a *consistent estimator* of θ(i.e., $\widehat{\theta}_n \xrightarrow{p} \theta$)

  if, for any $\epsilon > 0$,

  $\lim_{n\to\infty} P(\mid \widehat{\theta}_n - \theta \mid \leq \epsilon) = 1$ or $\lim_{n\to\infty} P(\mid \widehat{\theta}_n - \theta \mid > \epsilon) = 0$

  (note) "$\widehat{\theta}_n$ is a *consistent estimator* of θ" means "$\widehat{\theta}_n$ converges in probability to θ"

  (Thm 9.1) An unbiased $\widehat{\theta}_n$ for θ is a consistent estimator of θ if $\lim_{n\to\infty} V(\widehat{\theta}_n) = 0$.

  (Example 9.2) Let $Y_1, \ldots, Y_n$ denote a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Show that $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is a consistent estimator of μ.

How about consistency of $\bar{Y}_1 - \bar{Y}_2$ for $\mu_1 - \mu_2$?

(Thm 9.2) Suppose that $\widehat{\theta}_n$ and $\widehat{\theta}_n^{'}$ are consistent estimators of θ and $\theta^{'}$, respectively. Then,
(a) $\widehat{\theta}_n + \widehat{\theta}_n^{'}$ is a consistent estimator of $\theta + \theta^{'}$
(b) $\widehat{\theta}_n \times \widehat{\theta}_n^{'}$ is a consistent estimator of $\theta \times \theta^{'}$
(c) If $\theta^{'} \neq 0$, $\widehat{\theta}_n / \widehat{\theta}_n^{'}$ is a consistent estimator of $\theta / \theta^{'}$
(d) If $g(\cdot)$ is a real-valued function that is continuous at θ, then $g(\widehat{\theta}_n)$ is a consistent estimator of $g(\theta)$.

(Example 9.3) Let $Y_1, \ldots, Y_n$ denote a random sample from a distribution with finite $E(Y_i) = \mu$, $E(Y_i^2) = \mu_2^{'}$, and $E(Y_i^4) = \mu_4^{'}$. Show that $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y}_n^2)$ is a consistent estimator of $\sigma^2 = V(Y_i)$.

(Question) Suppose $Y_1, \ldots, Y_n$ is a random sample from any distribution with mean μ and known variance σ^2. Then, in Section 8.6, we derived a large-sample confidence interval for μ with confidence coefficient approximately equal to $1 - \alpha$,
$[\bar{Y} - z_{\alpha/2}(\sigma/\sqrt{n}),\ \bar{Y} + z_{\alpha/2}(\sigma/\sqrt{n})]$.
If σ^2 is unknown and n is large, can one replace σ with S_n? Why?

(Thm 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.

(Example 9.4)

(Note)

· Example 9.4 implies 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.

· We know from Chapter 7(p.359) that if the sample is taken from a normal distribution and n is finite, $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ has a t distribution with $n-1$ degrees of freedom

· This implies that if a large sample is taken from a normal distribution, the distribution of $\sqrt{n}(\bar{Y}_n - \mu)/S_n$ can be approximated by a standard normal distribution.

In other words, if n gets large, then the number of degrees of freedom also gets large, and the t-distribution can be approximated by a standard normal distribution (see Table 4 and 5 in pp.848-849).

(Exercise) Suppose that $Y \sim b(n, p)$. Show that

1) $\hat{p}_n = Y/n$ is an unbiased estimator of p

2) $(\hat{p}_n - p)/\sqrt{\hat{p}_n \hat{q}_n / n}$ converges to a standard normal distribution.

3) Derive a large-sampel confidence interval for p with confidence coefficient $1 - \alpha$

9.4 Sufficiency

- We summarize the information in the sample by using a statistic, $g(Y_1, \ldots, Y_n)$ for the parameters of interest.

  (Example) $\bar{Y}$ and S^2 as the unbiased estimator for the population mean μ and variance σ^2

  (Question) Does the process of summarizing the original set of n sample observations to a few statistics, $g_1(Y_1, \ldots, Y_n), \ldots,$ $g_m(Y_1, \ldots, Y_n)$ retain all the information about the parameters of interest in n sample observations?

  (Answer) There are methods to find statistics $g_1(Y_1, \ldots, Y_n), \ldots,$ $g_m(Y_1, \ldots, Y_n)$ summarizing all the information in a sample about the parameters of interest : we call such statistics *sufficient* statistics.

- Sufficient statistic

  (Def 9.3) Let $Y_1, \ldots, Y_n$ denote a random sample from a probability distribution with unknown parameter θ. Then the statistic $U = g(Y_1, \ldots, Y_n)$ is said to be *sufficient* for θ if the conditional distribution of $Y_1, \ldots, Y_n$, given U, does not depend on θ.

  (Example) Consider the outcomes of n independent trials of a binomial experiment, $X_1, \ldots, X_n$ where $X_i \sim b(1, p)$ and $p = P$(the i-th trial is a success). Suppose we are given a value of $Y = \sum_{i=1}^{n} X_i$(i.e., # of successes among n trials). If we know the value of Y, do we have to look at $X_1, \ldots, X_n$ or other functions of $X_1, \ldots, X_n$ in order to gain further information about p?

  Why? $P(X_1 = x_1, \ldots, X_n = x_n \mid Y = y)$ does not depend on p (i.e., as long as Y is known, there is no further information about p from other functions of $X_1, \ldots, X_n$). We call Y a sufficient statistic for p.

- How to find a sufficient statistic

  · (Def 9.3) tells us how to check if a statistic is sufficient or not.

  · *likelihood of the sample*(Def 9.4) and *factorization criterion*(Thm 9.4) help find a sufficient statistic for θ

  (Def 9.4) Let $y_1, \ldots, y_n$ be sample observation taken on corresponding random variables $Y_1, \ldots, Y_n$ whose distribution depends on θ. Then the *likelihood of obtaining the sample* $y_1, \ldots, y_n$ when the parameter is θ, $L(\theta) = L(y_1, \ldots, y_n \mid \theta)$, is defined to be

  $i)$ $p(y_1, \ldots, y_n \mid \theta)$ for discrete $Y_1, \ldots, Y_n$
  $ii)$ $f(y_1, \ldots, y_n \mid \theta)$ for continuous $Y_1, \ldots, Y_n$

  · what $L(\theta_1) > L(\theta_2)$ means for given $Y_1 = y_1, \ldots, Y_n = y_n$?

(Def 9.4 *continued*) If the set of random variables $Y_1, \ldots, Y_n$ is a random sample from $p(y \mid \theta)$ or $f(y \mid \theta)$(i.e., $Y_1, \ldots, Y_n \sim^{iid} p(y \mid \theta)$ or $f(y \mid \theta)$), then the *likelihood of the sample* $L(\theta) = L(y_1, \ldots, y_n \mid \theta)$ is

i) for discrete $Y_1, \ldots, Y_n$,

$$p(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} p(y_i \mid \theta) = p(y_1 \mid \theta) \times \cdots \times p(y_n \mid \theta)$$

ii) for continuous $Y_1, \ldots, Y_n$,

$$f(y_1, \ldots, y_n \mid \theta) = \prod_{i=1}^{n} f(y_i \mid \theta) = f(y_1 \mid \theta) \times \cdots \times f(y_n \mid \theta)$$

(Thm 9.4: factorization criterion) Let U be a statistic based on the random sample $Y_1, \ldots, Y_n$. Then U is a *sufficient statistic* for the estimation of θ iff $L(\theta) = L(y_1, \ldots, y_n \mid \theta)$ can be factored into two nonnegative functions :

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

where $h(y_1, \ldots, y_n)$ is either a constant or a function of $y_1, \ldots, y_n$(not a function of θ).

(Example 9.5)

(Exercises) Let $Y_1, \ldots, Y_n$ be IID samples from $N(\mu, \sigma^2)$. Then,
1) If μ is unknown and σ^2 is known, show $\bar{Y}$ is sufficient for μ.

2) If μ is known and σ^2 is unknown, show $\sum_{i=1}^{n}(Y_i - \mu)^2$ is sufficient for σ^2.

3) If μ and σ^2 are unknown, show $\sum_{i=1}^{n} Y_i$ and $\sum_{i=1}^{n} Y_i^2$ is jointly sufficient for μ and σ^2

- Note for sufficient statistics

  i) there are many possible sufficient statistics for one parameter.

  · the random sample itself, $Y_1, \ldots, Y_n$

  · $Y_{(1)} \leq \ldots \leq Y_{(n)}$: the set of order statistics from $Y_1, \ldots, Y_n$

  · Example 9.5: a one-to one function of $\bar{Y}$

  ii) we want to find a sufficient statistic that reduces the data in the sample as much as possible by using (Thm 9.4: factorization criterion)

  · sufficient statistics are useful to develop unbiased estimators with minimum variance (See 9.5).

9.5 Minimum-Variance Unbiased Estimation

Motivation

Search an estimator with good properties: unbiasedness, sufficiency and minimum variance.

Procedure

S1 Find a sufficient statistic U for θ

S2 Check if $E(U) = \theta$

S3 If yes, U is a *minimum-variance unbiased estimator*(MVUE) for θ.

If no, find a function of U, say $h(U)$ such that $E(h(U)) = \theta$ from $E(U)$. Then $h(U)$ is a *MVUE* for θ.

Why?

(Thm 9.5) **The Rao-Blackwell Theorem** Let $\widehat{\theta}$ be an unbiased estimator for θ such that $V(\widehat{\theta}) < \infty$. If U is a sufficient statistic for θ, define $\widehat{\theta}^{\star} = E(\widehat{\theta} \mid U)$. Then, for all θ,

$$E(\widehat{\theta}^{\star}) = \theta \quad \text{and} \quad V(\widehat{\theta}^{\star}) \leq V(\widehat{\theta}).$$

Why?(continued)

- Given an unbiased estimator $\widehat{\theta}$ for θ and a sufficient statistic U for θ, there is a function of U that is also an unbiased estimator for θ and has no larger variance than $\widehat{\theta}$.

- Which sufficient statistic should we use in this theorem? a sufficient statistic identified from *the factorization criterion(Thm 9.4)*

(Example 9.6)

(Example 9.9)

(Example 9.8)

9.6 The Method of Moments

Motivation

The k-th sample moment $m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$ should provide good estimates of the corresponding k-th population moment $E(Y^k)$ where $E(Y^k)$ are functions of the population parameters.

Procedure

Suppose there are r unknown parameters. Then solve the r equations $E(Y^k) = m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$ for $k = 1, 2, \ldots, r$ with respect to r unknown parameters. We call the solutions to the r equations the *Moment Estimators* for the parameters.

[Note]

1. $m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$ is a consistent estimators of the corresponding $E(Y^k)$.

2. This method is easy to employ and usually provides consistent estimators for respective parameters

3. The estimators derived from this method are often not functions of sufficient statistics, so that their variances are sometimes larger than others(i.e., are sometimes not very efficient)

(Example 9.11, 9.12)

(Example 9.13))

(Exercise) Suppose $Y_1, \ldots, Y_n$ are IID samples from $N(\mu, \sigma^2)$, and μ and σ^2 are unknown. Find the method-of-moments estimators of μ and σ^2.

9.7 The Method of Maximum Likelihood

Motivation

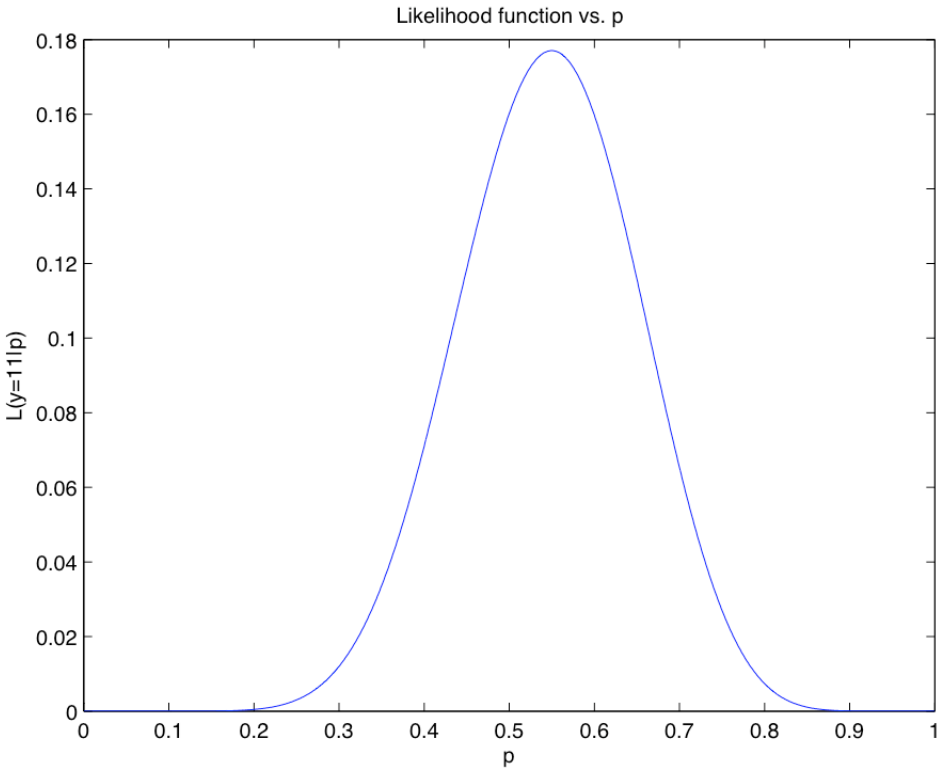
Let $y_1, \ldots, y_n$ be sample observation taken on corresponding random variables $Y_1, \ldots, Y_n$ whose distribution depends on θ. Then, we would like to get the value of θ(i.e., an estimate) which makes the observed data, $y_1, \ldots, y_n$ most likely (i.e., maximize the likelihood of obtaining the observed sample $y_1, \ldots, y_n$).

Example

Suppose one tosses a coin 20 times, and count the number of heads observed. Let p the probability that one observes the head. Suppose we observed 11 heads and 9 tails. Then we would like to get the value of p which maximizes the likelihood of obtaining the observed data.

Example(continued)

Let Y be the number of heads in 20 coin tosses. Then $Y \sim b(20, p)$. Since we observed 11 heads(i.e., $y = 11$), the likelihood of observing $y = 11$ is $L(y = 11 \mid p) = \binom{20}{11} p^{11}(1-p)^9$.



The value of p maximizing $L(y = 11 \mid p)$ is 0.55.

Procedure

S1 Write down the likelihood $L(\theta_1, \ldots, \theta_k) = L(y_1, \ldots, y_n \mid \theta_1, \ldots, \theta_k)$

S2 Find estimates for $\theta_1, \ldots, \theta_k$ that maximize $L(\theta_1, \ldots, \theta_k)$ or $\ell(\theta_1, \ldots, \theta_k) = \log L(\theta_1, \ldots, \theta_k)$. We call them the Maximum Likelihood Estimator(MLE)s for the parameters.

· take the derivative of $\ell(\theta_1, \ldots, \theta_k)$ with respect to $\theta_1, \ldots, \theta_k$, set them to zero, and solve them $\theta_1, \ldots, \theta_k$.

· draw a plot of $L(\theta)$ vs. θ and find $\widehat{\theta}$ maximizing $\ell(\theta)$ (when the range of θ depends on the samples)

(Example 9.14)

(Example 9.15)

(Example 9.16)

[Important notes]

1. This estimation method often leads to MVUEs.

   - The MLE is always some functions of any sufficient statistics U for θ. Why?

   : $L(y_1, \ldots, y_n \mid \theta) = g(u, \theta) \times h(y_1, \ldots, y_n)$

   $\Leftrightarrow \ell(y_1, \ldots, y_n \mid \theta) = \ln g(u, \theta) + \ln h(y_1, \ldots, y_n)$

   Then maximization of $\ell(y_1, \ldots, y_n \mid \theta)$ over θ is equivalent to maximization of maximization of $\ln g(u, \theta)$ over θ, as $\ln h(y_1, \ldots, y_n)$ does not depend on θ.

   Moreover, $\ln g(u, \theta)$ depends on the data only through the values of U.

   - Therefore, if the MLE for θ is adjusted to be unbiased, the resulting estimator often is an MVUE of θ.

2. The MLE for θ has the *invariance property*

   - Suppose $\widehat{\theta}$ is the MLE for θ and one is interested in estimating a function of θ, $h(\theta)$. Then the MLE $\widehat{h(\theta)}$ for $h(\theta)$ is $\widehat{h(\theta)} = h(\widehat{\theta})$.

   In (Example 9.16) the MLE of σ^2 is $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. What is the MLE of σ?