Chapters 8. Estimation

Basic Setup of Estimation

1. Quantity of interest: *population* parameter (i.e., *target* parameter).

   Population distribution: $f(x;\theta)$, and θ is the target parameter. The form of f is known (parametric), except the knowledge of θ.

2. Random samples: Samples, say $X_1, X_2, \ldots, X_n$ are iid (independent identically distributed) draws from the population. They all have common distribution $f(x;\theta)$.

3. Estimator: An estimator , say $\hat{\theta}$, is an estimate for the population parameter through these random samples. It is a function of the random samples, $\hat{\theta} = T(X_1, X_2, \ldots, X_n)$.

   Estimate $\hat{\theta}$ is a random variable!

4. Analysis of the estimator $\hat{\theta}$: Accuracy, confidence interval, bias, efficiency, consistency, and so on.

A baby example

A coin with $P(H) = p$, with p unknown.

1. Quantity of interest: p (play the role of θ).
2. Random samples: Toss coin n times. Let
$$X_i \doteq \begin{cases} 1 & , \quad \text{if the } i\text{-th toss is a heads} \\ 0 & , \quad \text{if the } i\text{-th toss is a tails} \end{cases}$$
   $\{X_1, X_2, \ldots, X_n\}$ are iid, Bernoulli with parameter p.
3. Estimator:
$$\hat{p} \doteq \frac{X_1 + X_2 + \ldots + X_n}{n}$$
4. Analysis of the estimator $\hat{p}$: $\hat{p}$ (random variable) is "unbiased" since
$$E[\hat{p}] = p.$$
   It is the best estimate for p (most efficient).

Bias and MSE

Definition: An estimate $\hat{\theta}$ is said to be unbiased if

$$E[\hat{\theta}] = \theta.$$

Definition: The bias of an estimate $\hat{\theta}$ is defined as

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta.$$

Definition: The Mean Square Error (MSE) of an estimate $\hat{\theta}$ is defined as

$$\text{MSE}[\hat{\theta}] = E\left[(\hat{\theta} - \theta)^2\right] = \left[\text{Bias}(\hat{\theta})\right]^2 + \text{Var}[\hat{\theta}].$$

"Estimators with smaller MSE are more preferable."

Examples

1. The estimator in the "coin toss" problem. Compare with

$$\hat{\theta}' \doteq w_1 X_1 + w_2 X_2 + \cdots + w_n X_n$$

   where $w_i \geq 0$ and $w_1 + w_2 + \cdots + w_n = 1$.

2. Estimating the size of population: A box contain N balls marked from 1 through N. We make n selections from the box, and let $X_1, X_2, \ldots, X_n$ be the observed numbers. Consider the following two estimators:

   (a)

$$\hat{\theta} \doteq 2\bar{X} - 1 = 2 \cdot \frac{X_1 + X_2 + \cdots + X_n}{n} - 1.$$

   (b)

$$\hat{\theta} \doteq \frac{n+1}{n} \cdot \max\{X_1, X_2, \ldots, X_N\}.$$

Common Unbiased Estimators

Textbook, Table 8.1, page 371.

1. Estimating Population Mean μ: iid random samples $Y_1, Y_2, \ldots, Y_n$ are drawn from the population.

$$\hat{\mu} = \bar{Y} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \text{sample mean.}$$

(a). Unbiased since $E[\bar{Y}] = \mu$.

(b). Standard deviation of $\bar{Y}$.

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation.

Special case – Binomial parameter p. $\{Y_i\}$ iid Bernoulli with parameter p.

$$Y = Y_1 + Y_2 + \cdots + Y_n$$

is the total number of "success".

The estimator is

$$\hat{p} = \frac{Y}{n}.$$

Unbiased,

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

2. Estimating Difference of Population Means $\theta = \mu_1 - \mu_2$: n iid samples $\{X_i\}$ from Population 1, and m iid samples $\{Y_j\}$ from Population 2.

$$\begin{aligned}\hat{\theta} &= \bar{X} - \bar{Y} = \frac{1}{n}(X_1 + \cdots + X_n) - \frac{1}{m}(Y_1 + \cdots + Y_m) \\ &= \text{Difference of Sample Mean.}\end{aligned}$$

Unbiased.

$$\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

where σ_i is the population standard deviation for Population i, $i = 1, 2$.

Special case – Difference of binomial parameter $\theta = p_1 - p_2$. $\{X_i\}$ iid Bernoulli with parameter p_1. $\{Y_j\}$ are iid Bernoulli with parameter p_2.

$$\begin{aligned} X &= X_1 + X_2 + \cdots + X_n, \\ Y &= Y_1 + Y_2 + \cdots + Y_m. \end{aligned}$$

The estimator is

$$\hat{\theta} = \frac{X}{n} - \frac{Y}{m}.$$

Unbiased.

$$\sigma_{\hat{\theta}} = \sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$$

3. Estimating Population Variance σ^2: iid samples $Y_1, Y_2, \ldots, Y_n$ are drawn from the population.

$$\text{Sample variance} = \hat{\sigma}^2 \doteq \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

where $\bar{Y}$ is the sample mean.

Unbiased.

Sample variance is often used as an approximation of the population variance!

Confidence Interval

Confidence Interval is a measurement for the accuracy of the estimator.

Illustration through example

1. Population distribution is $N(\mu, 1)$. Wish to estimate μ.

iid samples $\{X_1, X_2, \ldots, X_n\}$ are drawn. The (unbiased) estimator is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

The 95% confidence interval is an interval of type

$$[\bar{X} - b, \bar{X} + b]$$

such that

$$P(\mu \in [\bar{X} - b, \bar{X} + b]) = 95\%.$$

Remark: It is NOT the parameter μ that is random, it is the confidence interval that is random since $\bar{X}$ is a random variable.

Computation of b.

$$\bar{X} \text{ is } N(\mu, 1/n) \qquad \Rightarrow \qquad Z = \sqrt{n}(\bar{X} - b) \text{ is } N(0,1).$$

$$\begin{aligned} P(\mu \in [\bar{X} - b, \bar{X} + b]) &= P(-b \leq \bar{X} - \mu \leq b) \\ &= P(-b\sqrt{n} \leq Z \leq b\sqrt{n}) \\ &= 0.95. \end{aligned}$$

$$b\sqrt{n} = 1.96 \approx 2, \qquad b = \frac{2}{\sqrt{n}}.$$

The 95% confidence interval is $\left[\bar{X} - \frac{2}{\sqrt{n}}, \bar{X} + \frac{2}{\sqrt{n}}\right]$

What is the meaning of confidence interval?

Suppose in the previous example we simulate 100 samples, and get sample mean $\bar{X} = 2.3$. The 95% confidence interval is

$$[2.3 - 0.2, 2.3 + 0.2] = [2.1, 2.5].$$

Is it true that THIS interval, namely [2.1, 2.5], covers the true population mean μ with probability 95%? [NO]

Remark: When we say 95% confidence interval covers the true value with probability 95%, the true value is regarded as FIXED, while the confidence interval is regarded as RANDOM.

For example, if one repeats the experiment 10 times (independently), each times producing a 95% confidence interval. Then the number of intervals that cover the true parameter has a distribution $B(10, 0.95)$.

Remark: "95%" is called confidence level or confidence coefficient. In general, it can be $1-\alpha$ with $\alpha \in (0,1)$.

Remark: If the population distribution is $N(\mu, \sigma^2)$ with σ known, then the $(1-\alpha)$ confidence interval for μ is

$$[\bar{X} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}]$$

where $z_{\alpha/2}$ is defined such that

$$P(N(0,1) \geq z_{\alpha/2}) = \frac{\alpha}{2}.$$

Note that we can also rewrite the confidence interval as

$$[\bar{X} - \sigma_{\bar{X}} z_{\alpha/2}, \bar{X} + \sigma_{\bar{X}} z_{\alpha/2}].$$

Remark: The tighter the confidence interval, the better the estimate. As n increase, the confidence interval becomes tighter, whence the estimate becomes more accurate.

2. A coin $P(\mathrm{H}) = p$. Wishes to estimate p.

Toss the coin n times (assume n big), X = total number of heads. The estimator is

$$\hat{p} = \frac{X}{n}.$$

What is the 95% confidence interval?

Solution: $\hat{p}$ is unbiased,

$$E[\hat{p}] = p, \quad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

The distribution of $\hat{p}$ is approximately (normal approximation)

$$N\left(p, \sigma_{\hat{p}}^2\right)$$

As before, the 95% confidence interval will be approximately

$$[\hat{p} - 2\sigma_{\hat{p}}, \hat{p} + 2\sigma_{\hat{p}}]$$

But we do not know $\sigma_{\hat{p}}$. In this case, we can approximate

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

For example, suppose we toss coin 100 times and get 58 heads. Then $\hat{p} = 0.58$, and the 95% confidence interval is

$$\left[0.58 - 2\sqrt{\frac{0.58(1-0.58)}{100}}, 0.58 + 2\sqrt{\frac{\hat{0}.58(1-0.58)}{100}}\right]$$

or

$$[0.48, 0.68]$$

Generalization: Large-Sample Confidence Interval

Consider a target parameter θ and an unbiased estimator $\hat{\theta}$. When the sample size are large, the distribution of $\hat{\theta}$ can often be approximated by normal distribution. Example include: $\mu, p, \mu_1 - \mu_2, p_1 - p_2$.

More precisely, the distribution of $\hat{\theta}$ is approximately $N(\theta, \sigma_{\hat{\theta}})$. And

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is approximately $N(0, 1)$.

The $(1 - \alpha)$ confidence interval is just

$$\left[\hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \quad \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}\right]$$

Remark: Occasionally $\sigma_{\hat{\theta}}$ is known. More often it has to be estimated from the sample.

Examples

1. In order to estimate the average television viewing time per family in a large southern city, a sociologist took a random sample of 500 families. The sample yielded a mean of 28.4 hours per week, and the sample standard deviation is 8.3 hours per week. Find the 95% confidence interval for the population mean.

Remark: Let $\{X_1, X_2, \ldots, X_n\}$ be the iid samples. The sample variance is

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

and the sample standard deviation is

$$S = \sqrt{S^2}.$$

2. Estimate the difference in mean life of nonsmokers and smokers.

	sample-size	sample-mean	sample std
Nonsmokers	$n = 36$	$\bar{x} = 72$	$s_1 = 9$
Smokers	$m = 44$	$\bar{y} = 62$	$s_2 = 11$

Find the 95% confidence interval for the difference of population means.

Remark: Note that $\hat{\theta} = \bar{X} - \bar{Y}$, and

$$\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \approx \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}$$

Selecting sample size

Sometimes we have a prescribed length for confidence intervals, and the question is how large the sample size n should be.

In the set up when CLT approximation holds, the length of a $(1-\alpha)$ confidence interval is

$$2z_{\alpha/2}\sigma_{\hat{\theta}}.$$

1. The population distribution is $N(\mu, 1)$. Wish to estimate μ. How many samples do we need so that the 95% confidence interval is within ± 0.1 of μ.

   Solution: The estimator $\hat{\theta}$ is sample mean and the confidence interval length is

$$2z_{\alpha/2}\sigma_{\hat{\theta}} = 2z_{\alpha/2}\frac{1}{\sqrt{n}} = 4\frac{1}{\sqrt{n}}.$$

   So

$$4\frac{1}{\sqrt{n}} \leq 2 \times 0.1 = 0.2$$

   or

$$n \geq 400.$$

2. We wish to estimate the population proportion p of voters in favor of Democratic. And we want the 95% confidence interval to be within $\pm 3\%$ of the true value p. How large should the sample be?

Solution:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

So we need to find n such that

$$2z_{\alpha/2}\sigma_{\hat{\theta}} = 4\sqrt{\frac{p(1-p)}{n}} \leq 0.03 \times 2$$

or

$$n \geq \left(\frac{2\sqrt{p(1-p)}}{0.03}\right)^2$$

(a) If we know p is approximately, say 0.6, then

$$n \geq \left(\frac{2\sqrt{0.6(1-0.6)}}{0.03} \right)^2 = 1067.$$

(b) If we do not know p. We can have a conservation bound using the inequality $p(1-p) \leq 1/4$ to obtain

$$n \geq (1/0.03)^2 = 1111.$$

General confidence intervals

A general $(1-\alpha)$ two-sided confidence interval is $[\hat{\theta}_L, \hat{\theta}_U]$ such that

- $\hat{\theta}_L$ and $\hat{\theta}_U$ are both functions of samples. So they are RANDOM.
- $P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$.

The Pivotal Method

This is a general method to obtain a confidence interval. Let samples be $X_1, X_2, \ldots, X_n$.

1. Find a quantity that is a function of $\{X_i\}$ and θ.
2. The distribution of this quantity is independent of θ.

Examples

1. Let $X_1, X_2, \ldots, X_n$ be iid samples from a uniform distribution on $(0, \theta)$. Wish to estimate θ. Our estimate is

$$\hat{\theta} = X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}.$$

Consider the function

$$Y = \frac{X_{(n)}}{\theta}.$$

Then Y has a density

$$f(y) = \begin{cases} ny^{n-1} & , \quad \text{if } 0 < y < 1 \\ 0 & , \quad \text{otherwise} \end{cases}$$

For a $(1-\alpha)$ confidence interval, consider β_1 and β_2 such that

$$P(\beta_1 \leq Y \leq \beta_2) = 1 - \alpha.$$

There are infinitely many such choices. For each such choice,

$$P\left[\frac{X_{(n)}}{\beta_2} \leq \theta \leq \frac{X_{(n)}}{\beta_1}\right] = 1 - \alpha$$

A special choice is that

$$P(Y < \beta_1) = P(Y > \beta_2) = \alpha/2.$$

2. Confidence Interval for σ^2 for normal random variables. Assume $\{X_i\}$ are iid samples from $N(\mu, \sigma^2)$. Both unknown. An unbiased estimate for σ^2 is

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

Then

$$Y \doteq \frac{(n-1)\hat{\sigma}^2}{\sigma^2}$$

has the so called $\chi^2(n-1)$ distribution.

Remark: A chi-square distribution with degree of freedom k, or $\chi^2(k)$, is the distribution of

$$Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

where $\{Z_1, Z_2, \ldots, Z_k\}$ are iid $N(0, 1)$.