Inferential Statistics

Here, we learn how to draw conclusion (inference) about population based on results obtained from the sample. Inference has two parts: (1) estimation and (2) test of hypothesis. (Test of hypothesis will be discussed in the next semester.)

Estimation

Here, we "estimate" the value of the parameter based on the value of the statistic calculated from the sample. We use two types of estimation: point estimation (single-value estimation) and interval estimation (confidence interval).

1. **Point estimation**: Here, we use a single value (a point) as an estimate of the parameter. The statistic used for this purpose is called 'estimator' and its particular value obtained from a particular sample is called 'estimate'.

   Example: Suppose, we want to estimate the population mean (μ) of the monthly income of university graduates. We take a sample of size 100 and calculate the sample mean. Let $\bar{x} = 37.2$ (thousand). This value is a point estimate of μ.

2. **Interval estimation**: Here, we obtain an interval of values such that we are 95% (say) confident that this interval contains the value of the parameter.

   Example: Let $\bar{x} = 37.2$. Using the distribution of $\bar{X}$, we can be 90% (say) confident that μ will be between 36.1 and 38.3 (say).

Point Estimation: *Maximum Likelihood Estimate*

Let $X_1, X_2, \cdots, X_n$ be a random sample from a particular distribution $f(x)$ that has parameter θ (say). When the observations are independent, their joint distribution is

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

which contains the parameter θ. When we want to estimate the unknow parameter θ, we treat the above expression as a function of θ which is then called the 'likelihood function'. We want to obtain the value of θ for which the likelihood function is maximized. The estimate $\hat{\theta}$ that maximizes the likelihood function is called the Maximum Likelihood Estimate (MLE). MLE gives us the value of θ for which the observed sample is most likely.

Note

Since $f(x_1, x_2, \cdots, x_n)$ and $\log f(x_1, x_2, \cdots, x_n)$ have their maximum at the same value of θ, it is often computationally suitable to obtain MLE by maximizing $\log f(x_1, x_2, \cdots, x_n)$ (that is, maximizing the log-likelihood).

Example

Let $X_1, X_2, \cdots, X_n$ be a random sample from exponential(λ) distribution. Obtain the MLE of λ.

Solution

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \lambda \exp(-\lambda x_i) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} x_i\right)$$

$$\therefore \log f(x_1, x_2, \cdots, x_n) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

Maximizing the above log-likelihood with respect to λ, we get

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

Interval estimation: Confidence Intervals for $\boldsymbol{\mu}$

Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu, \sigma^2)$. It can be shown that $\bar{X}$ is the point estimator of μ. For interval estimation, we start with the following theorem:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

From normal table, we are 95% sure that:

$$-1.96 < Z < 1.96$$

$$\therefore -1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

$$\therefore \bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}$$

In general,

$$\bar{X} - z \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z \frac{\sigma}{\sqrt{n}}$$

Confidence Interval (C.I.) when σ is unknown:

$$\bar{X} - t \frac{S}{\sqrt{n}} < \mu < \bar{X} + t \frac{S}{\sqrt{n}}$$

Example

Utilico has developed a new battery. You want to check how long the battery operates continuously. 16 new batteries are randomly selected. The mean operation time of these batteries is 205 minutes with a standard deviation of 15 minutes. Construct a 90% C.I. for the population mean.

Solution

Here, $n = 16, \bar{x} = 205, s = 15.$

For 90% C.I.

$t = 1.753$ (t table, 15 d.f.)

$$\bar{x} - t \frac{s}{\sqrt{n}} = 198.4$$

$$\bar{x} + t \frac{s}{\sqrt{n}} = 211.6$$

$$\therefore 198.4 < \mu < 211.6$$