Interval estimation:
Estimating the difference in two means

Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\mu_1, \sigma_1^2)$. Let $Y_1, Y_2, \cdots, Y_m$ be a random sample from $N(\mu_2, \sigma_2^2)$ and the two samples are independent. We are interested in the estimation of $\mu_1 - \mu_2$.

From previous results, we have:

$$\bar{X} \sim N\left(\mu_1, \ \frac{\sigma_1^2}{n}\right)$$

$$\bar{Y} \sim N\left(\mu_2, \ \frac{\sigma_2^2}{m}\right)$$

$$\bar{X} - \bar{Y} \sim N\left(\mu_1 - \mu_2, \ \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1)$$

For 95% Confidence Interval (by using normal table):

$$\bar{x} - \bar{y} - 1.96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{x} - \bar{y} + 1.96\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

In general,

$$\bar{x} - \bar{y} - z\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{x} - \bar{y} + z\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Example

Average income of 100 males is 33.7 (thousand), while average income of 64 females is 31.9 (thousand). The two populations are independently normal with known variances 12.5 and 8 respectively. Construct a 90% confidence interval for the difference in means of the two populations.

Solution

Here, $n = 100$, $\bar{x} = 33.7$; $m = 64$, $\bar{y} = 31.9$

$\sigma_1^2 = 12.5$, $\sigma_2^2 = 8$

For 90% C.I.

$z = 1.645$ (from the normal table)

$$\bar{x} - \bar{y} - 1.645\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} < \mu_1 - \mu_2 < \bar{x} - \bar{y} + 1.645\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

$\therefore 0.978 < \mu_1 - \mu_2 < 2.623$

Approximate Confidence Interval for Mean

When we have a large sample from any population with finite mean and variance, we can use Central Limit Theorem (CLT) to use normal distribution for the construction of confidence intervals.

Example

Let $X_1, X_2, \cdots, X_n$ be a random sample from Bernoulli(p). By CLT, for large n,

$$Y = \sum_{i=1}^{n} X_i \sim N\big(np, np(1-p)\big)$$

approximately. That is, for large n, binomial variable is approximately normally distributed. Also, Let

$$\hat{p} = \frac{Y}{n} = \bar{X}$$

denote the sample proportion. Then, for large n,

$$\hat{p} \sim N\left(p, \ \frac{p(1-p)}{n}\right),$$

so that

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1).$$

Since for large samples, $\hat{p}$ is close to p, we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1).$$

Thus, 95% confidence interval for p is given by

$$\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example

A sample of 100 transistors is tested to determine if they meet the current standards. If 80 of them meet the standards, construct a 95% confidence interval for the population proportion.

Solution

$$\hat{p} = \frac{80}{100} = 0.8$$

So, 95% confidence interval for p is given by:

$$0.8 - 1.96\sqrt{\frac{0.8(1-0.8)}{100}} < p < 0.8 + 1.96\sqrt{\frac{0.8(1-0.8)}{100}}$$

$$\therefore 0.7216 < p < 0.8784$$

Confidence Interval for Parameter of Exponential Distribution

Let $X_1, X_2, \cdots, X_n$ be a random sample from exponential(λ) distribution. Then,

$$\sum_{i=1}^{n} X_i \sim \text{gamma}(n, \lambda)$$

It can be shown that

$$2\lambda \sum_{i=1}^{n} X_i \sim \text{gamma}(n, 1/2\ )$$

which is a chi-square distribution with $2n$ degrees of freedom. From χ^2 table, for 95% (say) confidence interval:

$$a < \chi^2 < b$$

$$\therefore a < \ 2\lambda \sum_{i=1}^{n} X_i \ < b$$

$$\therefore \ \frac{a}{2\sum_{i=1}^{n} X_i} < \ \lambda \ < \frac{b}{2\sum_{i=1}^{n} X_i}$$

Bayes' Estimator

Here, we assume that unknown parameter θ is a random variable with 'prior distribution' $g(\theta)$. Random observations have distribution $h(x \mid \theta)$ which may be discrete or continuous. Given the data, we update the distribution of θ, that is, we obtain the 'posterior distribution' of θ denoted by $h(\theta \mid x)$ as follows:

$$h(\theta \mid x) = \frac{h(x, \theta)}{h(x)} = \frac{h(x, \theta)}{\int_{-\infty}^{\infty} h(x, \theta)\, d\theta} = \frac{h(x \mid \theta)\, g(\theta)}{\int_{-\infty}^{\infty} h(x \mid \theta)\, g(\theta)\, d\theta}$$

We use the above posterior distribution for the estimation of θ.

Example

Let $X \sim \text{binomial}(n, \theta)$.

$\theta \sim \text{uniform}(0, 1)$ [prior distribution]

So, $g(\theta) = 1,\ \ 0 < \theta < 1$

$$h(x \mid \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

So, the posterior distribution is:

$$h(\theta \mid x) = \frac{\binom{n}{x} \theta^x (1-\theta)^{n-x}\, 1}{\int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x}\, 1\, d\theta}$$

$$= \frac{\theta^x (1-\theta)^{n-x}\, 1}{\int_0^1 \theta^x (1-\theta)^{n-x}\, 1\, d\theta}$$

$$= \frac{(n+1)!}{x!\,(n-x)!}\, \theta^x (1-\theta)^{n-x}, \quad 0 < \theta < 1$$

For an estimate of θ, we can consider the mean of the posterior distribution:

$$E(\theta \mid x) = \int_0^1 \theta\, \frac{(n+1)!}{x!\,(n-x)!}\, \theta^x (1-\theta)^{n-x}\, d\theta$$

$$= \frac{(n+1)!}{x!\,(n-x)!}\, \frac{(x+1)!\,(n-x)!}{(n+2)!}$$

$$= \frac{x+1}{n+2}$$

Note

It can be shown that the MLE of θ is

$$\hat{\theta} = \frac{x}{n}$$

while the Bayes estimate (with uniform prior) is

$$\tilde{\theta}_B = \frac{x+1}{n+2}$$