Confidence Intervals

An interval estimate for an unknown parameter θ is an interval of the form $\widehat{\theta}_1 < \theta < \widehat{\theta}_2$, where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are appropriate values of the random variables $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$, i.e. values such that

$$P(\widehat{\Theta}_1 < \theta < \widehat{\Theta}_2) = 1 - \alpha,$$

for some specified probability $1 - \alpha$. Typically, $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ will be quantities involving random variables of known distribution (e.g. the sample mean $\bar{X}$) and possibly other known parameters, and can be evaluated once a sample has been observed.

For a specified value $1 - \alpha$ we refer to $\hat{\theta}_1 < \theta < \hat{\theta}_2$ as a $(1 - \alpha)100\%$ confidence interval (CI) for θ. For example, when $\alpha = 0.05$ we refer to a 95% CI.

The interpretation of this $(1 - \alpha)100\%$ (observed) CI is the following: *if we obtain a large number of such intervals (each time using a different, independent sample), we expect the true value of the parameter θ to be contained in $(1 - \alpha)100\%$ of them.*

1. Confidence intervals for means

1.1 Population variance σ^2 known

Let $x_1, x_2, \ldots, x_n$ denote a random sample from a population with unknown mean μ. When the population variance σ^2 is known, a $(1 - \alpha)100\%$ CI for μ will be given by

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{1}$$

where $z_{\frac{\alpha}{2}}$ is the appropriate percentage point of the $N(0, 1)$ distribution, i.e. it is such that

$$P\left(Z > z_{\alpha/2}\right) = 1 - \Phi\left(z_{\alpha/2}\right) = \frac{\alpha}{2}. \tag{2}$$

1.2 Population variance σ^2 unknown

Now assume that the population variance σ^2 is unknown. Using the sample $x_1, x_2, \ldots, x_n$ we can estimate it by the sample variance $s^2 = \frac{1}{(n-1)}\left\{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2\right\}$.

Then, if the population is Normal, a $(1 - \alpha)100\%$ CI for μ will be given by

$$\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \quad \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) \tag{3}$$

where $t_{\frac{\alpha}{2}}$ is the point of the t_{n-1} distribution such that

$$P(T > t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}, \qquad \text{with } T \sim t_{n-1}.$$

Note that (3) also gives an approximate CI for non-normal populations.

2. CIs for the difference in the means of two populations

If we are interested in comparing the means of two populations (to investigate whether or not they are different), we can calculate a CI for the difference $\mu_1 - \mu_2$ where μ_1, μ_2 are the unknown means of the two populations.

We then consider whether the value 0 lies in the CI ($\mu_1 - \mu_2 = 0 \Leftrightarrow$ means are equal).

2.1 Population variances σ_1^2 and σ_2^2 known

Let $\bar{x}_1$ and $\bar{x}_2$ denote the means of two random samples from the fisrt and second population respectively. If the population variances σ_1^2 and σ_2^2 are known, and assuming that the two populations are Normal and that the two samples are independent, a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ will be given by

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right) \tag{4}$$

where $z_{\frac{\alpha}{2}}$ is as in (2).

2.1 Population variances σ_1^2 and σ_2^2 unknown

Case (i):

If the population variances σ_1^2 and σ_2^2 are unknown, but both samples are large (say $n_1, n_2 \geq 30$), independent and come from Normal populations, then a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ will be given by expression (4), with the unknown σ_1^2 and σ_2^2 substituted by the sample variances s_1^2 and s_2^2 respectively.

Case (ii):

If either or both samples are small, but come from Normal populations with equal variances (i.e. if we can assume that $\sigma_1^2 = \sigma_2^2 = \sigma^2$), then a $(1-\alpha)100\%$ CI for $\mu_1 - \mu_2$ will be given by

$$\left(\bar{x}_1 - \bar{x}_2 - t\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad \bar{x}_1 - \bar{x}_2 + t\ s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) \tag{5}$$

where

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled sample variance (estimate of σ^2) and $t = t_{n_1+n_2-2,\ \alpha/2}$ is the percentage point of the $t_{n_1+n_2-2}$ distribution such that

$$P(T > t_{n_1+n_2-2,\ \alpha/2}) = \frac{\alpha}{2}, \qquad \text{with } T \sim t_{n_1+n_2-2}.$$