Confidence Intervals

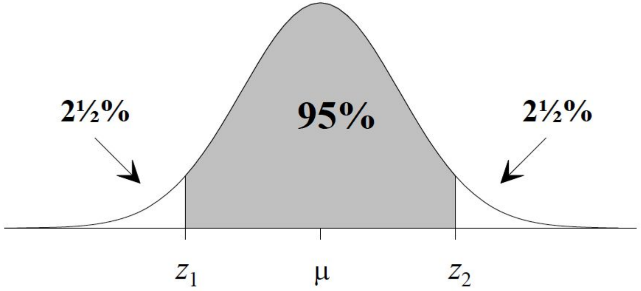
Why confidence intervals?

- A confidence interval provides an “interval estimate” of an unknown parameter (as opposed to a “point estimate”).
- It is designed to contain the parameter’s value with some stated probability.
- The width of the interval provides a measure of the precision accuracy of the estimator involved.
- 95% confidence intervals (α=0.05) are the most common.
- The correct interpretation of a 95% confidence interval (L, U) is that "we are 95% confident that the population parameter is between L and U."

One sample mean (σ Known)

- Confidence intervals will be constructed using the sampling distributions.
- For example, when sampling from a N(μ, σ^2) distribution where σ^2 is known:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \Rightarrow \quad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$



One sample mean (σ known)

- If we require a 95% confidence interval, then we can read off the 2.5% and 97.5% z-values.
- This gives us $z_{0.025}$ = -1.96 and $z_{0.975}$ = +1.96.
- Putting the values of $z_{0.025}$ and $z_{0.975}$ and rearranging we get,

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

This is the 95% confidence interval of the population mean, μ. It is also expressed as:

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

One sample mean (σ known) - Example

The average IQ of a sample of 50 university students was found to be 132. Calculate a symmetrical 95% confidence interval for the average IQ of university students, assuming that IQs are normally distributed. It is known from previous studies that the standard deviation of IQs among students is approximately 20.

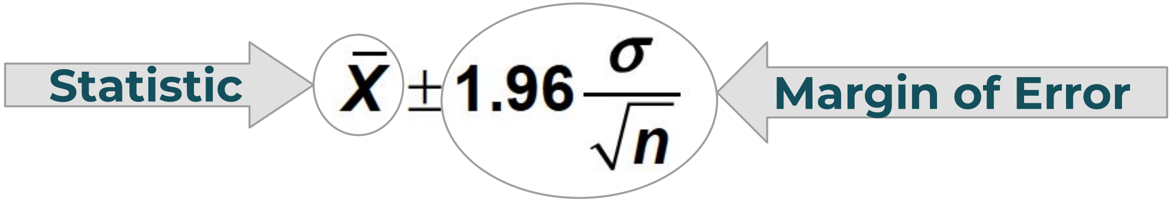
Here, $X \sim N(\mu, 20^2)$.
Given, $n = 50$, $\bar{x} = 132$, $\alpha=0.05$

=> 95% confidence interval for μ is $(132 - 1.96 * 20/\sqrt{50} \ , 132 + 1.96 * 20/\sqrt{50})$, i.e, (126.5, 137.5)
=> We are 95% confident that the average IQ of the population lies between 126.5 and 137.5

Finding sample size

A very common question asked of a statistician is:
“How large a sample is needed?”

Effectively confidence interval consists of two parts:



Statistic

$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

Margin of Error

Width = 2 * MoE

Given, the value of MoE and σ, we can determine n.

One sample mean (σ unknown)

The 95% confidence interval of the population mean, μ is expressed as:

$$\overline{X} \pm t_{0.025,n-1} \frac{S}{\sqrt{n}}$$

$t_{0.025,\, n\text{-}1}$ is the value k such that $P(t_{n\text{-}1} < k) = 0.025$.

Since, t distribution is also symmetric about zero, this confidence interval is also symmetric.

One sample mean (σ unknown) - Example

Calculate a 95% confidence interval for the average height of 10-year-old children, assuming that heights have a $N(\mu, \sigma^2)$ distribution (where μ and σ are unknown), based on a random sample of 5 children whose heights are: 124cm, 122cm, 130cm, 125cm and 132cm.

Here, $n = 5$, $\bar{x} = 126.6$, $s = 4.22$, $\alpha=0.05$

=> $t_{0.025,4} = -2.776$

=> 95% confidence interval for μ is
$(126.6 - 2.776 * 4.22/\sqrt{5}\ ,\ 126.6 + 2.776 * 4.22/\sqrt{5})$, i.e, (121.4, 131.8)

One sample variance

The 95% confidence interval of the population variance, σ^2 is expressed as:

$$\left(\frac{(n-1)S^2}{\chi^2_{0.025,n-1}}, \frac{(n-1)S^2}{\chi^2_{0.975,n-1}} \right)$$

$\chi^2_{0.025,\, n\text{-}1}$ is the value k_1 such that $P(\chi^2_{n\text{-}1} < k_1) = 0.025$
$\chi^2_{0.975,\, n\text{-}1}$ is the value k_2 such that $P(\chi^2_{n\text{-}1} < k_2) = 0.975$

Due to the skewness of the χ^2 distribution these confidence intervals are not symmetrical about the point estimator S^2. So, we can’t write these using the “土” notation.

One sample variance - Example

Calculate a 95% confidence interval for the standard deviation of 10-year-old children, assuming that heights have a N(μ, σ^2) distribution (where μ and σ are unknown), based on a random sample of 5 children whose heights are: 124cm, 122cm, 130cm, 125cm and 132cm.

Here, n = 5, s = 4.22, α=0.05
=> $\chi^2_{0.025,4} = 0.4844$; $\chi^2_{0.975,4} = 11.14$

=> 95% confidence interval for σ^2 is ($4 * 4.22^2/0.4844$, $4 * 4.22^2/11.14$) , i.e, (6.39, 147.0)

=> 95% confidence interval for σ is (√6.39, √147.0), i.e, (2.53, 12.1)

One sample proportion

The 95% confidence interval of the population proportion, θ is expressed as:

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}, \text{ where } \hat{\theta} = \frac{X}{n}$$

This confidence interval is also symmetric.
It is valid only when the binomial distribution can be approximated by normal distribution.

One sample proportion - Example

In a one-year mortality investigation, 45 of the 250 ninety-year-olds present at the start of the investigation died before the end of the year. Assuming that the number of deaths has a binomial distribution with parameters n=250 and p, calculate a symmetrical 90% confidence interval for the unknown mortality rate p.

Here, $n = 250$, $\hat{p} = 45/250 = 0.18$, $\alpha=0.1$

=> $z_{0.05} = -1.6449$; $z_{0.95} = 1.6449$

=> $\sqrt{[\hat{p}(1-\hat{p})/n]} = \sqrt{[0.18 * (1 - 0.18)/250]} = 0.024$

=> 90% confidence interval for p is

$(0.18 - 1.6449 * 0.024, 0.18 + 1.6449 * 0.024)$, i.e, $(0.140, 0.220)$

Two sample means (σ_1,σ_2 known)

The 100(1-α)% confidence interval of the difference in the population means (μ_1 - μ_2) of two normal populations is:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$z_{\alpha/2}$ is the value k such that P(Z < k) = α/2

Two sample means (σ_1,σ_2 unknown)

The 100(1-α)% confidence interval of the difference in the population means (μ_1 - μ_2) of two normal populations is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

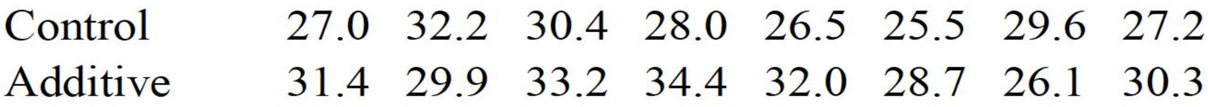
where:

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

Requirement: The two variances should be nearly equal.

Two sample means - Example

A motor company runs tests to investigate the fuel consumption of cars using a newly developed fuel additive. Sixteen cars of the same make and age are used, eight with the new additive and eight as controls. The results, in miles per gallon over a test track under regulated conditions, are as follows:



Control	27.0	32.2	30.4	28.0	26.5	25.5	29.6	27.2
Additive	31.4	29.9	33.2	34.4	32.0	28.7	26.1	30.3

Obtain a 95% confidence interval for the increase in miles per gallon achieved by cars with the additive. State clearly any assumptions required for this analysis.

Two sample variances

The 100(1-α)% confidence interval of the ratio in the population variances $(\sigma^2{}_1 / \sigma^2{}_2)$ of two normal populations is:

$$\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} . F_{n_2-1,n_1-1}$$

Two sample proportions

The 100(1-α)% confidence interval of the difference in the population proportions (θ_1 - θ_2) of two binomial populations (approx. by normal) is:

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}$$

Two sample proportions - Example

In a one-year mortality investigation, 25 of the 100 ninety-year-old males and 20 of the 150 ninety-year-old females present at the start of the investigation died before the end of the year.

Assuming that the numbers of deaths follow binomial distributions, calculate a symmetrical 95% confidence interval for the difference between male and female mortality rates at this age.