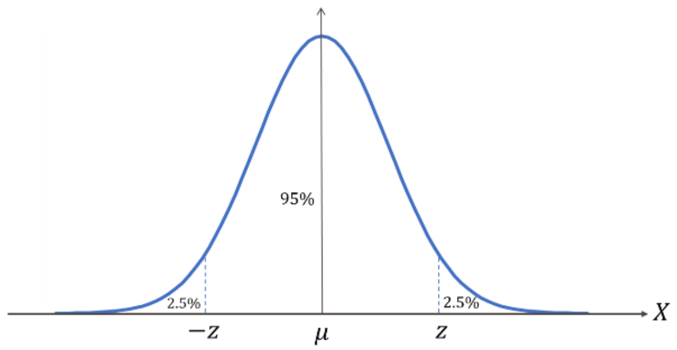
Confidence Intervals I Cheat Sheet

When estimating a population parameter, it is preferable to have a range of values rather than just one value. The single value, known as a **point estimate**, is likely to be inaccurate whereas it is more probable that the true population statistic is held within a selected range of values. This range of estimates is known as a **confidence interval**. When finding the confidence interval, it is important to consider the probability of that range including the wanted value. This is called the **confidence level**.

Symmetrical Confidence Intervals

It is preferable to have a method that allows the confidence interval to have a 95% probability of containing the population parameter. For instance, to estimate the population mean of a sample, the sample mean $\bar{X}$ would be used. Some random variables are taken from the normal distribution such that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ where μ is the mean of the population, n is the size of the sample and σ is the standard deviation for one value of X. The value of interest is μ and hence the confidence interval should be a region which is symmetrical about μ and has a 95% chance of containing $\bar{X}$. This is displayed graphically below by a bell-shaped curve.



The graph has a symmetry around μ. The critical z-score can be determined using the geometric properties of the graph using the relation:

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

where $\frac{\sigma}{\sqrt{n}}$ can be referred to as the standard error. The width of a confidence interval is $2z\frac{\sigma}{\sqrt{n}}$

A sample taken from a normal population with a 95% confidence interval is given by:

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

This is generalised for a c% symmetric confidence interval:

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

This confidence interval can also be used for a sample taken from any distribution where n is large enough.

If the true population variance is known following a normal distribution:

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

where $z = \Phi^{-1}\left(0.5 + \frac{\frac{1}{2}c}{100}\right)$. The value for the z- score can be determined using the tables and also by understanding that it is a measurement determined from the relation $z = \frac{x-\mu}{\sigma}$.

Example 1: The following sample of 10 values is taken from a normally distributed population with a standard deviation of 3:

$$[\,56, 48, 52, 41, 39, 52, 60, 54, 59, 44\,].$$

a) Construct a 98% confidence interval for the population mean.
b) Determine the width of the confidence interval.

a) The mean of the sample is the first value that needs to be determined.	$\bar{x} = \frac{\sum x}{n} = \frac{505}{10} = 50.5$
The z-score can then be determined either using the tables or a calculator.	$z = \Phi^{-1}\left(0.5 + \frac{\frac{1}{2} \times 98}{100}\right)$ $z = 2.326\ (4s.f.)$
The confidence interval is found using the equation for the range and substituting in the calculated values.	$(\bar{x} - z\frac{\sigma}{\sqrt{n}}, \bar{x} + z\frac{\sigma}{\sqrt{n}})$ $\left(50.5 - 2.326\frac{3}{\sqrt{10}}, 50.5 + 2.326\frac{3}{\sqrt{10}}\right)$ The confidence interval is $(48.3, 52.7)$ to $3s.f.$
b) The width of the confidence interval is found using the equation $2z\frac{\sigma}{\sqrt{n}}$.	Width = $2z\frac{\sigma}{\sqrt{n}} = 2(2.326)\frac{3}{\sqrt{10}}$ The width of the interval is 4.41 (3 s. f.).

Symmetrical Confidence Intervals for a Large Sample Size

When it is not certain if a population is normally distributed, the sample size has to be large enough ($n \geq 30$) to use the central limit theorem. This means the data can be approximated to a normal distribution.

Additionally, if the population variance and mean are not known, then sample statistics can be taken to estimate the mean.

Example 2: A palaeontologist is investigating the width of fossilised footsteps ($r\ cm$). They find the data can be summarised by:

$$\sum r = 405, \sum r^2 = 11702, n = 50$$

Find the 90% confidence interval for the mean.

The first step is to determine the sample statistics including the mean and the variance. This can then be used to determine the unbiased estimate of the variance.	$\bar{r} = \frac{405}{50} = 8.1$ $\sigma^2 = \frac{11702}{50} - 8.1^2 = 168.43$ $s^2 = \frac{50}{49} \times \sigma^2 = \frac{50}{49} \times 168.43 \approx 171.9$ $s = 13.1\ (3\ s.f.)$
The z-score can be determined by using the known equation.	$z = \Phi^{-1}\left(0.5 + \frac{\frac{1}{2} \times 90}{100}\right) = 1.65\ (3s.f.)$
By substituting in the determined values into the inequality, a range for μ can be found.	$\bar{r} - z\frac{s}{\sqrt{n}} < \mu < \bar{r} + z\frac{s}{\sqrt{n}}$ $8.1 - 1.65\frac{13.1}{\sqrt{50}} < \mu < 8.1 + 1.65\frac{13.1}{\sqrt{50}}$ $5.04 < \mu < 11.2\ (3s.f.)$

Inferences from Confidence Levels

Once the confidence interval has be found, it is possible that you might be asked to comment on the significance of μ being a certain value. When the value of μ is within the confidence interval range, then there is evidence to support the statement. However, if the value falls outside the interval, then you can state there is not enough evidence to support the claim.

Example 3: A forest is known to contain trees for various heights and the range of heights can be modelled as a normal distribution with a mean height of 18.6m and standard deviation of 2.3m.

a) For a sample of 60 trees, find a 95% confidence interval for the mean height of all the trees in the forest.
b) A biologist suggests that if the trees have a mean height under 20m then the forest may not be experiencing enough rainfall for the trees to grow to a large height. Does the confidence interval suggest that the trees are not experiencing enough rainfall?

a) For a c% confidence interval, the z-score can be determined using the z-score table or using a calculator.	$z = \Phi^{-1}(0.95) = 1.64\ (3s.f.)$
The range of values for the confidence interval can be determined using the symmetric properties of the graph.	$18.6 + 1.64 \times \frac{2.3}{\sqrt{60}} = 19.09$ $18.6 - 1.64 \times \frac{2.3}{\sqrt{60}} = 19.09 = 18.1$ The confidence interval is [18.1, 19.1] to $3s.f.$
b) In this section, it is important to consider whether the given true mean $(20m)$ agrees with your confidence interval. In other words, for it to be suitable, the value should fall within the given range.	A mean height of 20m is not within the confidence interval so the confidence level may suggest that the amount of rainfall is unsuitable for the trees.