0.1 Confidence Intervals

0.1.1 Confidence interval

- A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, such as a population mean.
- The confidence intervals is calculated from a given set of sample data.
- Confidence intervals are usually calculated so that this percentage is 95%, but we can produce 90%, 99%, 99.9% (or whatever) confidence intervals for the unknown parameter

If a statistic is normally distributed and the standard error of the statistic is known, then a confidence interval for that statistic can be computed as follows:

$$Statistic \pm (Quantile \times Std.Error)$$

The statistic is found in the question. It will be the sample mean, sample proportion or the sample difference.

The quantile depends on the level of confidence desired.

- If the standard error of the statistic is **known**, then the quantile is from the Z distribution.

$$Statistic \pm (Z_{\frac{\alpha}{2}} \times Std.Error)$$

- If the standard error of the statistic is **unknown**, then the quantile is from the $t-$ distribution, with $n-1$ degrees of freedom.

$$Statistic \pm (t_{\frac{\alpha}{2},n-1} \times Std.Error)$$

Generally, we use 1.96 for the 95% confidence interval and 2.58 for the 99% confidence interval.

Confidence Level Important:

- A confidence level for an interval is denoted to $1-\alpha$ (in percentages: $100(1-\alpha)\%$) for some value α.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1-\alpha)\% = 100(1-0.05)\% = 100(0.95)\% = 95\%$
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for α is important when determining quantiles.

Quantiles

- The quantile is a value from a probability distribution that scales the intervals according to the specified confidence level.
- For practical purposes, the quantile can be taken from the standard normal distribution, if the sample is larger than 30, further to the central limit theorem.

- For a specified confidence level $1-\alpha$, the corresponding quantile is the value z_o that satisfies the following identity (when $n > 30$):

$$p(-z_o \leq Z \leq z_o) = 1 - \alpha$$

Quantiles

- When the sample size n is greater than 30, we would normally use compute the quantile using Murdoch Barnes table 3.
- 95% of Random generated Z-scores should be between -1.96 (quantile for 2.5%) and 1.96 (quantile for 97.5%)

Quantiles

- If the confidence level is 95%, then the quantile is 1.96. Recall

$$p(-1.96 \leq Z \leq 1.96) = 0.95$$

- If the confidence level is 90%, then the quantile is 1.645.

$$p(-1.645 \leq Z \leq 1.645) = 0.90$$

- If the confidence level is 99%, then the quantile is 2.576.

$$p(-2.576 \leq Z \leq 2.576) = 0.99$$

Using the $t-$distribution for large samples USEFUL

- The $t-$distribution is used for computing quantiles in the case of small samples (i.e. when sample size $n \leq 30$).
- A key value in the $t-$distribution is the degrees of freedom, denoted df (or sometimes ν). For small samples

$$df = n - 1$$

.

- The $t-$distribution is used for computing quantiles in the case of large samples too, as an alternative to using the Z distribution.
- In this case , use the value ∞ as the degrees of freedom (see bottom row of table 7).

$$df = \infty$$

- This means that we can use the $t-$ distribution for finding the quantiles of all types of confidence intervals.

0.2 Confidence Intervals

Confidence Intervals are a way of taking data from a sample and saying something about the population from which the sample was drawn.

- Confidence intervals can be used both to evaluate and report on the precision of estimates, and the significance of hypothesis tests.
- The level of confidence associated with a confidence interval indicates the long-run percentage of such intervals which would include the parameter being estimated.
- The most frequently used confidence intervals are the 90 percent, 95 percent, and 99 percent confidence intervals.

Confidence Intervals

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.
- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%,99% and 99.9%.

- A confidence level for an interval is denoted to $1-\alpha$ (in percentages: $100(1-\alpha)\%$) for some value α.
- A confidence level of 95% corresponds to $\alpha = 0.05$.
- $100(1-\alpha)\% = 100(1-0.05)\% = 100(0.95)\% = 95\%$.
- For a confidence level of 99%, $\alpha = 0.01$.
- Knowing the correct value for α is important when determining quantiles.

0.2.1 Standard Error

- The standard error measures the dispersion of the sampling distribution.
- For each type of point estimate, there is a corresponding standard error.
- A full list of standard error formulae will be attached in your examination paper.
- The standard error for a mean is
$$S.E(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$
However, we often do not know the value for σ. For practical purposes, we use the sample standard deviation s as an estimate for σ instead.
$$S.E(\bar{x}) = \frac{s}{\sqrt{n}}$$

0.3 Confidence interval for the True Mean Difference

The in above example the estimated average improvement is just over 2 points. Note that although this is statistically significant, it is actually quite a small increase. It would be useful to calculate a confidence interval for the mean difference to tell us within what limits the true difference is likely to lie. A 95% confidence interval for the true mean difference is:

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

or, equivalently

$$\bar{d} \pm (t_{\alpha/2} \times SE(\bar{d}))$$

where $t_{\alpha/2}$ is the 2.5% point of the t-distribution on $n-1$ degrees of freedom.

Using our example: We have a mean difference of 2.05. The 2.5% point of the t-distribution with 19 degrees of freedom is 2.093. The 95% confidence interval for the true mean difference is therefore: $2.05 \pm (2.093 \times 0.634) = 2.05 \pm 1.33 = (0.72, 3.38)$.

This confirms that, although the difference in scores is statistically significant, it is actually relatively small. We can be 95% sure that the true mean increase lies somewhere between just under one point and just over 3 points.

Confidence Intervals (Revision)

- The 95% confidence interval is a range of values which contain the true population parameter (i.e. mean, proportion etc) with a probability of 95%.
- We can expect that a 95% confidence interval will not include the true parameter values 5% of the time.
- A confidence level of 95% is commonly used for computing confidence interval, but we could also have confidence levels of 90%, 99% and 99.9%.

The Central Limit Theorem

- This theorem states that as sample size n is increased, the sampling distribution of the mean (and for other sample statistics as well) approaches the normal distribution in form, regardless of the form of the population distribution from which the sample was taken.
- For practical purposes, the sampling distribution of the mean can be assumed to be approximately normally distributed, even for the most non-normal populations or processes, whenever the sample size is $n > 30$.
- (For populations that are only somewhat non-normal, even a smaller sample size will suffice. A variation of the normal distribution can be used for such circumstances.)

Chapter 1

11. Confidence Intervals

1.1 Confidence Intervals

- Confidence Limits
- Confidence levels
- Margin of Error
- Interpreting Confidence Intervals
- Example 1 : Confidence Interval for a mean
- Example 2 : Confidence Interval for a proportion

1.1.1 Confidence limits

- Confidence limits are the lower and upper boundaries / values of a confidence interval, that is, the values which define the range of a confidence interval.
- The upper and lower bounds of a 95% confidence interval are the 95% confidence limits.
- These limits may be taken for other confidence levels, for example, 90%, 99%, 99.9%.
- The confidence level is the probability value associated with a confidence interval.
- It is often expressed as a percentage. For example, say , then the confidence level is equal to (1-0.05) = 0.95, i.e. a 95% confidence level.
- Density and Distribution Function
- Continuous random variables and probability distributions
  - Uniform Distribution
  - Exponential Distribution

  - The Normal Distribution
  - The Standard Normal Distribution

- Sampling Distributions
- Statistical Inference
- Confidence Interval
- The paired t-test
- Two sample test for independence samples.

Statistical Inference : Confidence Intervals

- Confidence intervals allow us to use sample data to estimate a parameter value, such as a population mean.
- A confidence interval is a range of values for which we can be confident (at a specific level) that parameter value (such as the population mean) lies within.
- A confidence level will have a specified level of confidence, commonly 95%.
- The 95% confidence interval is a range of values which contains the parameter value of interest with a probability of 0.95.
- We can expected that a 95% confidence interval will not contain the parameter value of interest with a probability of 0.05.
- It is natural to interpret a 95% confidence interval on the mean as an interval with a 0.95 probability of containing the population mean.
- However, the proper interpretation is not that simple.
- Consider the case in which 1,000 studies estimating the value of μ in a certain population all resulted in estimates between 30 and 40.
- Suppose one more study was conducted and the 95% confidence interval on μ was computed to be $40 \leq \mu \leq 50$ (based on that one study).
- The probability that μ is between 40 and 50 is very low, the confidence interval not withstanding.

Thirty-eight students took the test. The X-axis shows various intervals of scores (the interval labeled 35 includes any score from 32.5 to 37.5). The Y-axis shows the number of students scoring in the interval or below the interval. ***cumulative frequency distribution***A can show either the actual frequencies at or below each interval (as shown here) or the percentage of the scores at or below each interval. The plot can be a histogram as shown here or a polygon.

Confidence intervals example

A researcher was investigating computer usage among students at a particular university. Three hundred undergraduates and one hundred postgraduates were chosen at random and asked if they owned a laptop. It was found that 150 of the undergraduates and 80 of the postgraduates owned a laptop.

Find a 95% confidence interval for the difference in the proportion of undergraduates and postgraduates who own a laptop. On the basis of this interval, do you believe that postgraduates and undergraduates are equally likely to own a laptop?

1.1.2 Difference of two proportions example

- Two time-sharing systems are compared according to their response time to an editing command. The mean response time of 100 requests submitted to system 1 was measured to be 600 milliseconds with a known standard deviation of 20 milliseconds.
- The mean response time of 100 requests on system 2 was 592 milliseconds with a known standard deviation of 23 milliseconds.
- Using a significance level of 5%,test the hypothesis that system 2 provides a faster response time than system 1.
- Clearly state your null and alternative hypotheses and your conclusion.

Quantiles (1)

- The quantile is a value from a probability distribution that scales the intervals according to the specified confidence level.
- For practical purposes, the quantile can be taken from the standard normal distribution, if the sample is larger than 30, further to the central limit theorem.
- For a specified confidence level $1-\alpha$, the corresponding quantile is the value a that satisfies the following identity (when $n > 30$):

$$p(-a \leq Z \leq a) = 1 - \alpha$$

Quantiles (2) PCs = Percentiles (i.e. Quantiles)

```
>PCs=c(180/200,190:199/200,1998:1999/2000)

>
> PCs
[1] 0.9000 0.9500 0.9550 0.9600 0.9650 0.9700 0.9750
[8] 0.9800 0.9850 0.9900 0.9950 0.9990 0.9995
> qnorm(PCs)
[1] 1.281552 1.644854 1.695398 1.750686 1.811911
[6] 1.880794 1.959964 2.053749 2.170090 2.326348
[11] 2.575829 3.090232 3.290527
>
```

Quantiles (3) Negative values for Qauntiles

```
>PCs
[1] 0.1000 0.0500 0.0450 0.0400 0.0350 0.0300 0.0250
[8] 0.0200 0.0150 0.0100 0.0050 0.0010 0.0005
> qnorm(1-PCs)
[1] -1.281552 -1.644854 -1.695398 -1.750686 -1.811911
[6] -1.880794 -1.959964 -2.053749 -2.170090 -2.326348
[11] -2.575829 -3.090232 -3.290527
```

Margin of Error

- The product of the quantile and the standard error give us the width of the confidence interval
- The width of the confidence interval is known as the ***margin of error***.

$$\text{Margin of Error} = [\text{Quantile} \times \text{Standard Error}]$$

- The margin of error gives us some idea about how uncertain we are about the unknown population parameter.
- A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.
- The only way to control the margin of error is to adjust the sample size accordingly.
- By choosing an appropriate sample size, it is possible to ensure that the margin of error does not excess a certain threshold.

Small samples

- We indicated that use of the normal distribution in estimating a population mean is warranted for any large sample ($n > 30$).
- For a small sample ($n \leq 30$) only if the population is normally distributed **and** σ is known, the standard normal distribution can be used compute quantiles. In practice, this case is unusual.
- Now we consider the situation in which the sample is small and the population is normally distributed, but σ is not known.

1.2 Confidence interval: Worked Example

The intelligence quotient (IQ) of 36 randomly chosen students was measured. Their average IQ was 109.9 with a variance of 324. The average IQ of the population as a whole is 100.

1. Calculate the p-value for the test of the hypothesis that on average students are as intelligent as the population as a whole against the alternative that on average students are more intelligent.

2. Can we conclude at a significance level of 1% that students are on average more intelligent than the population as a whole?

3. Calculate a 95% confidence interval for the mean IQ of all students.

$$Z_{Test} = \frac{X - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{109.9 - 100}{\frac{18}{\sqrt{36}}} = \frac{9.9}{3} = 3.3 \tag{1.1}$$

$$p.value = P(Z \geq Z_{Test}) = P(Z \geq 3.3) = 0.00048 \tag{1.2}$$

$$\bar{X} \pm t_{1-\alpha/2,\nu} S.E.(\bar{X}) \tag{1.3}$$

$\nu = 1.96$

$$t_{1-\alpha/2,\nu} = 1.96 \tag{1.4}$$

$$109.9 \pm (1.96 \times 3) = [104.02, 115.79] \tag{1.5}$$

Confidence Intervals : Worked Example

- In a statistical report on the daily sales of a certain pharmaceutical product the following confidence interval was reported [6.3, 8.1] in hundreds of units per day.
- In the report it was stated that the used confidence level was 99% and the sample size was n = 25.
- The industry standard for that type of analysis recommends the 95% confidence level.

Question: Calculate a 95% confidence interval **Solution:**
The sample mean is 7.2

$$X = \frac{8.1 + 6.3}{2} = 7.2$$

The sample size is n= 25. This is a small sample (i.e. less than 30) We are able to deduce the quantile of the $t-$distribution used to construct the 99% confidence interval
Confidence intervals are always 2 tailed , therefore k=2

- The significance level used is 1%
- The degrees of freedom is 24 (n-1)
- The significance level for the new interval is 5%

Using Murdoch Barnes Table 7

- The quantile used to make the 99% interval was 2.797.
- The quantile used to make the 95% interval is 2.064.

We are now able to work out the standard error.

1.3 Confidence Interval For The Mean

Suppose that you wish to estimate the mean sales amount perretail outlet for a particular consumer product during the past year. The number of retail outlets is large. Determine the 95 percent confidence interval given that the sales amounts are assumed to be normally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.
Ans. $3; 346 : 60 to 3; 503 : 40$

Determine the 95 percent confidence interval given that the population is assumed to benormally distributed, $\bar{X}$ =3,425, s = 200 , and $n = 25$.
Ans. $3; 342 : 44 to 3; 507 : 56$

1.4 Confidence Intervals for Means

The structure of a confidence interval for the mean is as follows:

$$\text{Sample Mean} \pm (\text{Quantile} \times \text{Std. Error})$$

Quantiles:
For large samples (i.e. greater than 30) where a normal distribution can be assumed, the quantiles are as follows

90%	1.645
95%	1.96
99%	2.576

Sample Mean:
The sample mean $\bar{x}$ is usually given in the question.
(Remark: Sample mean is a type of ***point estimate***).

Standard Error :
The standard error is computed using the sample standard deviation (s) and the sample size (n).

$$S.E.(\bar{x}) = \frac{s}{\sqrt{n}}$$

1.4.1 Confidence Intervals

- The length of life of a type of battery is estimated from a sample of 100 test items taken from a large population.
- Sample results show that the mean length of life is 57.4 hours with a standard deviation of 15.1 hours.
- Construct a 95% confidence interval for the mean length of life of all of these batteries.

With a sample standard deviation of 15.1 and a sample size of 100, the standard error is as follows:

$$S.E.(\bar{x}) = \frac{15.1}{\sqrt{100}}$$

With a sample standard deviation of 15.1 and a sample size of 100, the standard error is as follows:

$$57.4 \pm (1.96 \times 1.51)$$

1.5 Confidence Interval of a Mean of the Small Sample

If the data have a normal probability distribution and the sample standard deviation s is used to estimate the population standard deviation σ, the interval estimate is given by:

$$\bar{X} \pm t_{1-\alpha/2,n-1} \frac{s}{\sqrt{n}} \tag{1.6}$$

where $t_{1-\alpha/2,n-1}$ is the value providing an area of $\alpha/2$ in the upper tail of a Students t distribution with n - 1 degrees of freedom.

Confidence Interval for the Difference Between Two Proportions

- A confidence interval gives us some idea of the range of values which an unknown population parameter (such as the mean or variance) is likely to take based on a given set of sample data.
- Many occasions arise where we have to compare the proportions of two different populations. For example, a firm may want to compare the proportions of defective items produced by different machines; medical researchers may want to compare the proportions of men and women who suffer heart attacks etc.
- A confidence interval for the difference between two proportions would specify a range of values within which the difference between the two true population proportions may lie, for such examples.
- The procedure for obtaining such an interval is based on the sample proportions, p1 and p2, from their respective overall populations.