# Cambridge (CIE) A Level Maths: Probability & Statistics 2

Your notes

## Sampling & Estimation

### Contents

Your notes

# Unbiased Estimates

## What is an unbiased estimator of a population parameter?

- An **estimator** is a **statistic** that is used to estimate a **population parameter**
  - When a sample is used with the estimator, the value that it produces is called an **estimate**
- An estimator is called **unbiased** if the **expected value** of the estimator is **equal** to the **population parameter**
  - An estimate from an unbiased estimator is called an unbiased estimate
  - This means that the **mean** of the **unbiased estimates** will get **closer** to the **population parameter** as more samples are taken

## What are the unbiased estimates for the mean and variance of a population?

- If you had the data for a **whole population** you could find the **actual population mean** and **variance** using
  - $\mu = \frac{\Sigma x}{n}$
  - $\sigma^2 = \frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma x^2}{n} - \mu^2$
- If you are using a **sample to estimate** the **mean** of a population then an **unbiased estimate** is given by
  - $\bar{x} = \frac{\Sigma x}{n}$
  - This is the **same formula** for the population mean
- If you are using a **sample to estimate** the **variance** of a population then an **unbiased estimate** is given by
  - $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$
  - This can be written in different ways
  - $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right) = \frac{n}{n-1}\left(\frac{\Sigma x^2}{n} - \bar{x}^2\right)$
  - This is a **different formula** to the population variance
  - The last formula shows a method for finding an unbiased estimate for the variance

- Find the variance of the sample (treating it as a population)
- Multiply this by $\frac{n}{n-1}$

Your notes

## Is there an unbiased estimate for the standard deviation?

- Unfortunately square rooting an unbiased variance **does not** result in an unbiased standard deviation
- There is **not a formula** for an unbiased estimate for the **standard deviation** that works for all populations
  - Therefore it is better to just work with the variance and not the standard deviation
- If you need an estimate for the standard deviation then you can use:
  - You can use the square root of your unbiased estimate for the population variance
  - $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)}$
  - This **won't be unbiased** but it will be a good estimate

## How do I calculate unbiased estimates?

- If you are given the **summary statistics** $\Sigma x$ and $\Sigma x^2$ then you can simply use the formulae in the **formula booklet**
  - $\bar{x} = \frac{\Sigma x}{n}$
  - $s^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$
- If you are given the **raw** data then you will **first** need to **calculate** $\Sigma x$ **and** $\Sigma x^2$

### Worked Example

The times, $T$ minutes, spent on daily revision of a random sample of 50 A Level students from the UK are summarised as follows.

$$n = 50 \quad \Sigma t = 6174 \quad \Sigma t^2 = 831581$$

Calculate unbiased estimates of the population mean and variance of the times spent on daily revision by A Level students in the UK

Unbiased estimate for the mean

Use formula "$\bar{x} = \frac{\Sigma x}{n}$" ← Replace x with t

$\bar{t} = \frac{6174}{50} = 123.48$

$\bar{t} = 123$ mins (3sf)

Unbiased estimate for the variance

Use formula "$S^2 = \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$" ← Replace x with t

$S_t^2 = \frac{1}{49}\left(831581 - \frac{6174^2}{50}\right) = 1412.56...$

$S_t^2 = 1410$ mins$^2$ (3sf)



## Examiner Tips and Tricks

- Always check whether you need to divide by n or n-1 by looking carefully at the wording in the question.

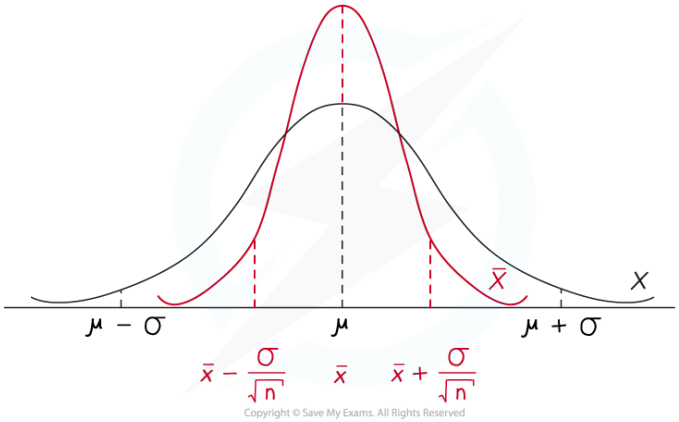
Your notes

Your notes

# Sample Mean Distribution

## What is the distribution of the sample means?

- For any given population it can often be difficult or impractical to find the true value of the **population mean, $\mu$**
  - The population could be too large to collect data using a census or
  - Collecting the data could compromise the individual data values and therefore taking a census could destroy the population
  - Instead, the population mean can be estimated by taking the mean from a sample from within the population
- If a sample of size $n$ is taken from a population, $X$, and the mean of the sample, $\bar{x}$ is calculated then the distribution of the sample means, $\overline{X}$, is the distribution of all values that the sample mean could take
- If the population, $X$, has a normal distribution with mean, $\mu$, and variance, $\sigma^2$, then the mean expected value of the distribution of the sample means, $\overline{X}$ would still be $\mu$ but the variance would be reduced
  - Taking a mean of a sample will reduce the effect of any extreme values
  - The greater the sample size, the less varied the distribution of the sample means would be
- The distribution of the means of the samples of size taken from the population, will have a normal distribution with:
  - Mean, $\overline{x} = \mu$
  - Variance $\frac{\sigma^2}{n}$
  - Standard deviation $\frac{\sigma}{\sqrt{n}}$
- For a random variable $X \sim \mathrm{N}(\mu, \sigma^2)$ the distribution of the sample mean would be $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- The standard deviation of the distribution of the sample means depends on the sample size, $n$
  - It is inversely proportional to the square root of the sample size
  - This means that the greater the sample size, the smaller the value of the standard deviation and the narrower the distribution of the sample means





Your notes

### Worked Example

A random sample of 10 observations is taken from the population of the random variable $X \sim N(30, 25)$ and the sample mean is calculated as $\bar{x}$. Write down the distribution of the sample mean, $\overline{X}$.

10 observations: n = 10

$X \sim N(30, 5^2)$: $\mu = 30$, $\sigma^2 = 25$, $\sigma = \sqrt{25} = 5$

$X \sim N\left(\mu, \frac{\sigma^2}{n}\right)$: $\bar{X} \sim N\left(30, \frac{25}{10}\right)$

$\bar{X} \sim N(30, 2.5)$



### Examiner Tips and Tricks

- Look carefully at the distribution given to determine whether the variance or the standard deviation has been given.

# Central Limit Theorem

## Does the distribution of the sample means always follow a normal distribution?

- If the variable $X$ for the **population follows** a **normal distribution** then the **sample mean** distribution $\overline{X}$ also **follows a normal distribution**
- If the variable $X$ for the population does not follow a normal distribution then the sample mean distribution $\overline{X}$ does not necessarily follow a normal distribution

- If the sample size is **small** then $\overline{X}$ can **not be assumed** to follow a normal distribution
- If the sample size is **big enough** then we can use the **Central Limit Theorem** to approximate $\overline{X}$ using a normal distribution

Your notes

## What is the central limit theorem?

- If a **random sample** of **size n** is taken from a population with **mean** $\mu$ and **variance** $\sigma^2$ then the **Central Limit Theorem** states that $\overline{X}$ can be **approximated** by the **normal distribution** $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$ provided **n is large** enough
  - Notice the variance is still divided by the size of the sample
- We usually say n is large enough if it is **at least 30**
- This is a powerful theorem as it allows us to use the normal distribution for even when the population itself does not follow a normal distribution
- If the population follows a normal distribution then the Central Limit Theorem is not needed as $\overline{X}$ will be normal automatically
  - This is important as you might be asked whether the Central Limit Theorem was needed

### Worked Example

The integers 1 to 29 are written on counters and placed in a bag. The expected value when one is picked at random is 15 and the variance is 70. Susie randomly picks 40 integers, returning the counter after each selection.

Find the probability that the mean of Susie's 40 numbers is less than 13. Explain whether it was necessary to use the Central Limit Theorem in your calculation.

Let X be the number selected

$\overline{X}$ is the mean of the 40 numbers

X is not normally distributed so need to use the Central Limit Theorem as the sample size is large

$\overline{X} \sim N\left(15, \frac{70}{40}\right)$

Divide variance by sample size

$P(\overline{X} < 13) = P\left(Z < \frac{13 - 15}{\sqrt{\frac{70}{40}}}\right) = P(Z < -1.512)$

0.0653

## Examiner Tips and Tricks

- If asked to explain whether it was necessary to use the Central Limit Theorem, check whether the population follows a normal distribution, if it does not then check the size of the sample. If your answer is yes comment on both of these things.

Your notes

Your notes

# Confidence Interval for μ

## What is a confidence interval?

- It is **impossible** to find the **exact value** of a population parameter when taking a sample
- The best we can do is find an **interval** for which the exact value is likely to lie within
  - This is called a **confidence interval**
- The **confidence level** of a confidence interval is the **probability** that the **interval contains** the **population parameter**
  - Be careful with wording – the population parameter is a fixed value so it does not make sense to talk about the probability that it lies within an interval
    - Instead we talk about the probability an interval contains the parameter
  - Suppose samples were collected and a 95% confidence interval for a population parameter was constructed for each sample, then for every 100 intervals we would **expect on average 95** of them to contain the parameter
    - 95 out of 100 is **not guaranteed** – it is possible all of them could contain the parameter
    - It is possible (though **very unlikely**) that none of them contains the parameter

## What affects the width of a confidence interval?

- The **width** of a confidence interval is the **range of the values** in the interval
- The **confidence level** affects the width
  - Increasing the confidence level will increase the width
  - Decreasing the confidence level will decrease the width
- The **size of the sample** affects the width
  - Increasing the sample size will decrease the width
  - Decreasing the sample size will increase the width

## How do I calculate a confidence interval for the population mean (μ)?

- For this course we only construct **symmetrical confidence intervals** for the mean of a population when:
  - The **variance of the population** is **known**
  - The population follows a **normal distribution**

- The population does not follow a normal distribution but the **sample size is big** enough to use the **Central Limit Theorem**
- The confidence interval can be found using the **formula**

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

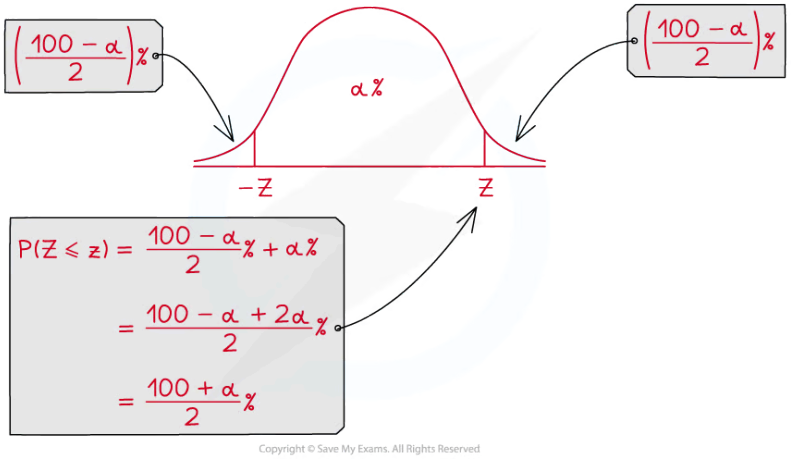
- $n$ is the size of the sample
- $\sigma$ is the standard deviation of the population
- $\bar{x}$ is the mean of the sample
- $z$ is the value such that $\mathrm{P}(-z < Z < z) = \alpha\%$ where $\alpha$ is the confidence level
- The width of the confidence interval is

$$2 \times z \times \frac{\sigma}{\sqrt{n}}$$

- Note that the sample mean is the **midpoint** of the confidence interval and does not affect the width

## How do I find the z-value for a confidence interval?

- You use the standard normal distribution $Z \sim N(0, 1^2)$
- If the confidence level is % we find such that $\mathrm{P}(-z < Z < z) = \alpha\%$
- To do this find $P(Z < z)$ - it can be shown that this is $\left(\frac{\alpha + 100}{2}\right)\%$





- The z-values for common confidence levels are:
  - For 90%: $\mathrm{P}(Z < z) = 0.95$ so $z = 1.645$
  - For 95%: $\mathrm{P}(Z < z) = 0.975$ so $z = 1.960$
  - For 99%: $\mathrm{P}(Z < z) = 0.995$ so $z = 2.576$

Your notes

- z-values for most confidence intervals you will need to work with will be given in the table of critical values

Your notes

## How can I interpret a confidence interval?

- After you have calculated a confidence interval for $\mu$ you might be expected to comment on the possibility of $\mu$ being a specific value
- If the value which is claimed to be $\mu$ is **within** the confidence interval then there is **evidence** to **support the claim**
- If the value is **outside** the interval then there is **not enough evidence** to support the claim

### Worked Example

The battery life of a certain brand of phone, $L$ hours, is known to follow a normal distribution with mean $\mu$ and variance 16. Jonny takes a random sample of 20 phones and calculates the mean battery life to be 20.3 hours.

Calculate a 95% confidence interval for μ.

Write down known values

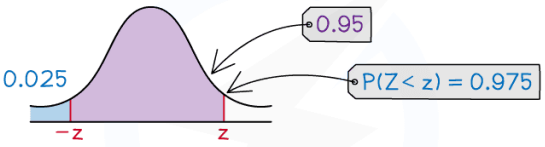
$n = 20$

$\bar{l} = 20.3$

$\sigma^2 = 16 \Rightarrow \sigma = 4$

Find the z−value

$P(-z < Z < z) = 0.95$t




$z = 1.960$

Use the formula "$\bar{x} \pm z \times \frac{\sigma}{\sqrt{n}}$"

Lower bound $= 20.3 - 1.96 \times \frac{4}{\sqrt{20}} = 18.546...$

Upper bound $= 20.3 + 1.96 \times \frac{4}{\sqrt{20}} = 22.053...$

$18.5 \leqslant \mu \leqslant 22.1$ hours (3sf)



### Examiner Tips and Tricks

- Always check whether the population follows a normal distribution, if it does not then you will have to state that the Central Limit Theorem is being used (as the

sample size should be big enough). Take care with the fiddly bits:

- You need to square-root the sample size
- You need to use the standard deviation so you might also need to square-root the variance

Your notes

# Confidence Interval for p

## How do I calculate a confidence interval for the proportion (p) of a population?

- If we want to find estimate the **proportion of a population** that fulfil a certain criteria we can construct a **confidence interval** based on the proportion of a sample who fulfil that criteria
  - The proportion is between 0 and 1
- For this course we only construct **symmetrical confidence intervals** for the proportion of a population provided that the **sample size is large** enough to use the **Central Limit Theorem**
- The confidence interval can be found using the formula

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

- $n$ is the size of the sample
- $\hat{p}$ (or $p_s$) is the proportion of the sample
- z is the value such that $\mathrm{P}(-z < Z < z) = \alpha\%$ where $\alpha$ is the confidence level
- In the **formula booklet** you are given the **distribution of the sample proportion** which might help you remember the formula for the confidence interval

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

- The **width** of the confidence interval is

$$2 \times z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Note that the sample proportion is the midpoint of the confidence interval but it **also affects** the width

## How can I interpret a confidence interval?

- Interpreting a confidence interval for p works in the same way as a confidence interval for $\mu$
- The only additional part is that you might get asked to see whether an experiment is fair
  - Find the probability of the outcome as though it were fair

- For example – a fair coin will have a 0.5 chance of landing on each side
- Use a sample to calculate a confidence interval
- See if the probability is in the confidence interval
  - If it is not then there is sufficient evidence to suggest that the experiment is not fair

Your notes

## Worked Example

Misty selects 50 fish at random from a lake. 17 of the 50 fish are trout.

Calculate a 90% confidence interval for the proportion of the fish in the lake that are trout.

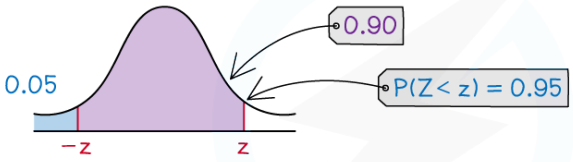
Write down known values

$n = 50$ ← Large so can use the Central Limit Theorem

$p_s = \frac{17}{50} = 0.34$

Find the z−value

$P(-z < Z < z) = 0.90$




$z = 1.645$

Use the formula "$P_s \pm \sqrt{\frac{P_s(1-P_s)}{n}}$"

Lower bound $= 0.34 - 1.645\sqrt{\frac{0.34 \times 0.66}{50}} = 0.2297...$

Upper bound $= 0.34 + 1.645\sqrt{\frac{0.34 \times 0.66}{50}} = 0.4502...$

$0.230 < p < 0.450$ (3sf)



## Examiner Tips and Tricks

- Remember that the confidence interval is not guaranteed to contain the true population proportion. When interpreting your answers use phrases like "there is evidence to support...".