# DP IB Maths: AI HL

Your notes

## 4.9 Further Normal Distribution (inc Central Limit Theorem)

### Contents

# 4.9.1 Sample Mean Distribution

Your notes

## Combinations of Normal Variables

### What is a linear combination of normal random variables?

- Suppose you have *n* **independent** normal random variables $X_i \sim \mathrm{N}(\mu_i,\ \sigma_i^2)$ for $i = 1, 2, 3, \ldots, n$
- A linear combination is of the form $X = a_1X_1 + a_2X_2 + \ \ldots + a_nX_n + b$ where $a_i$ and *b* are constants
- The mean and variance can be calculated using results from random variables
  - $\mathrm{E}(X) = a_1\mu_1 + a_2\mu_2 + \ \ldots + a_n\mu_n + b$
  - $\mathrm{Var}(X) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ \ldots + a_n^2\sigma_n^2$
    - The variables **need to be independent** for this result to be true
- A **linear combination of *n* independent normal random variables** is also a n**ormal random variable** itself
  - $X \sim \mathrm{N}(a_1\mu_1 + a_2\mu_2 + \ \ldots + a_n\mu_n + b,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ \ldots + a_n^2\sigma_n^2)$
  - This can be used to find probabilities when combining normal random variables

### What is meant by the sample mean distribution?

- Suppose you have a population with distribution *X* and you take a random sample with *n* observations $X_1, X_2, \ldots, X_n$
- The sample mean distribution is the distribution of the values of the sample mean
  - $\bar{X} = \dfrac{X_1 + X_2 + \ \ldots + X_n}{n}$
- For an individual sample the sample mean $\bar{x}$ can be calculated
  - This is also called a point estimate
  - $\bar{X}$ is the distribution of the point estimates

### What does the sample mean distribution look like when *X* is normally distributed?

- If the population is normally distributed then the sample mean distribution is also normally distributed
- $\mathrm{E}(\bar{X}) = \mathrm{E}\left(\dfrac{X_1 + X_2 + \ \ldots + X_n}{n}\right) = \dfrac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ \ldots + \mathrm{E}(X_n)}{n} = \dfrac{\mu + \mu + \ \ldots + \mu}{n} = \dfrac{n\mu}{n} = \mu$
- $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\dfrac{X_1 + X_2 + \ \ldots + X_n}{n}\right) = \dfrac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ \ldots + \mathrm{Var}(X_n)}{n^2} = \dfrac{\sigma^2 + \sigma^2 + \ \ldots + \sigma^2}{n^2} = \dfrac{n\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$
- Therefore you divide the variance of the population by the size of the sample to get the variance of the sample mean distribution

- $X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow \overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Your notes

Your notes

## Worked example

Amber makes a cup of tea using a hot drink vending machine. When the hot water button is pressed the machine dispenses $W$ ml of hot water and when the milk button is pressed the machine dispenses $M$ ml of milk. It is known that $W \sim \text{N}(100,\ 15^2)$ and $M \sim \text{N}(10,\ 2^2)$.

To make a cup of tea Amber presses the hot water button three times and the milk button twice. The total amount of liquid in Amber's cup is modelled by $C$ ml.

a) Write down the distribution of $C$.

$C = W_1 + W_2 + W_3 + M_1 + M_2$

$E(C) = E(W_1) + E(W_2) + E(W_3) + E(M_1) + E(M_2)$

$\mu = 100 + 100 + 100 + 10 + 10 = 320$

$\text{Var}(C) = \text{Var}(W_1) + \text{Var}(W_2) + \text{Var}(W_3) + \text{Var}(M_1) + \text{Var}(M_2)$

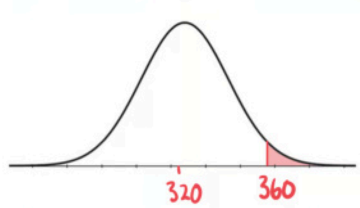
$\sigma^2 = 15^2 + 15^2 + 15^2 + 2^2 + 2^2 = 683$

A linear combination of normal variables is also a normal variable

$C \sim \text{N}(320, 683)$

b) Find the probability that the total amount of liquid in Amber's cup exceeds 360 ml.

Use normal distribution on GDC




$\mu = 320 \quad \sigma = \sqrt{683}$

Lower $= 360$ Upper $= 9999\ldots$

$P(C > 360) = 0.062939\ldots$

$P(C > 360) = 0.0629$ (3 s.f.)

c) Amber makes 15 cups of tea and calculates the mean $\overline{C}$. Write down the distribution of $\overline{C}$.

Your notes

$$\bar{C} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\boxed{\bar{C} \sim N\left(320, \frac{683}{15}\right)}$$

# Central Limit Theorem

Your notes

## What is the Central Limit Theorem?

- The Central Limit Theorem says that if a sufficiently large random sample is taken from any distribution $X$ then the sample mean distribution $\overline{X}$ can be approximated by a normal distribution
  - In your exam $n$ > 30 will be considered sufficiently large for the sample size
- Therefore $\overline{X}$ can be modelled by $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - $\mu$ is the mean of $X$
  - $\sigma^2$ is the variance of $X$
  - $n$ is the size of the sample

## Do I always need to use the Central Limit Theorem when working with the sample mean distribution?

- No – the Central Limit Theorem is not needed when the population is normally distributed
- If $X$ is **normally distributed** then $\overline{X}$ is normally distributed
  - This is true regardless of the size of the sample
  - The **Central Limit Theorem is not needed**
- If $X$ is **not normally distributed** then $\overline{X}$ is approximately normally distributed
  - Provided the sample size is large enough
  - The **Central Limit Theorem is needed**

Your notes

## Worked example

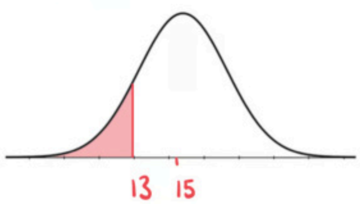
The integers 1 to 29 are written on counters and placed in a bag. The expected value when one is picked at random is 15 and the variance is 70. Susie randomly picks 40 integers, returning the counter after each selection.

a) Find the probability that the mean of Susie's 40 numbers is less than 13.

Let $\bar{X}$ be the mean of 40 numbers

n large $\Rightarrow$ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ $\quad \bar{X} \sim N(15, \frac{70}{40})$

Use normal distribution on GDC




$\mu = 15 \quad \sigma = \sqrt{\frac{70}{40}}$

Lower $= -999\ldots$ Upper $= 13$

$P(\bar{X} < 13) = 0.065285\ldots$

$P(\bar{X} < 13) = 0.0653$ (3sf)

b) Explain whether it was necessary to use the Central Limit Theorem in your calculation.

The Central Limit Theorem was necessary as the random variable for the number picked out of the bag is not normally distributed.

# 4.9.2 Confidence Interval for the Mean

Your notes

## Confidence Interval for μ

### What is a confidence interval?

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

### How do I find a confidence interval for the population mean ($\mu$)?

- You will be given data using a **sample from a population**
  - The population will be **normally distributed**
    - If not then the sample size should be large enough so you can use the **Central Limit Theorem**
- You will use the **interval functions** on your calculator
- Use a ***z*-interval** if the **population variance** is **known** $\sigma^2$
  - On your GDC enter:
    - the standard deviation $\sigma$ and the confidence level $\alpha$%
    - EITHER the raw data
    - OR the sample mean $\overline{X}$ and the sample size $n$
- Use a ***t*-interval** if the **population variance** is **unknown**
  - In this case the test uses the **unbiased estimate** for the variance $s_{n-1}^2$
  - On your GDC enter:
    - the confidence level $\alpha$%
    - EITHER the raw data
    - OR the sample mean $\overline{X}$, the value of $s_{n-1}$ and the sample size $n$
- Your GDC will give you the lower and upper bounds of the interval
  - It can be written as $a < \mu < b$

### 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 with **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 can I interpret a confidence interval?

- After you have found a confidence interval for $\mu$ you might be expected to **comment on the claim** for a value of $\mu$
- If the claimed value is **within** the confidence interval then there is **not enough evidence to reject the claim**
  - Therefore the **claim is supported**
- If the claimed value is **outside** the interval then there is **sufficient evidence to reject the claim**
  - The value is **unlikely to be correct**

Your notes

### Worked example

Cara wants to check the mean weight of burgers sold by a butcher. The weights of the burgers are assumed to be normally distributed. Cara takes a random sample of 12 burgers and finds that the mean weight is 293 grams and the standard deviation of the sample is 5.5 grams.

a) Find a 95% confidence interval for the population mean, giving your answer to 4 significant figures.

The population variance is unknown so use a t-interval

Formula booklet

| | |
|---|---|
| Unbiased estimate of population variance $s_{n-1}^2$ | $s_{n-1}^2 = \frac{n}{n-1} s_n^2$ |

$s_{n-1}^2 = \frac{12}{11} \times 5.5^2 = 33$

Enter data into GDC

Confidence level = 0.95 $\quad \bar{x} = 293 \quad s_{n-1} = \sqrt{33} \quad n = 12$

Lower: 289.35...

Upper: 296.64...

**$289.4 < \mu < 296.6$ (4 sf)**

b) The butcher claims the burgers weigh 300 grams. Comment on this claim with reference to the confidence interval.

**300 is above the confidence interval which suggests that the butcher's claim is not true.**