Preface

This booklet contains all the **Deferred Material** from Strand 1 of the Leaving Certificate Higher Level Course. This material was introduced in September 2013 for examination in June 2015 and onwards.

This chapter is printed as a supplement for those students who bought *Text & Tests 5* at the beginning of 5th Year in September 2013. All future editions of *Text & Tests 5* will contain this new chapter.

Chapter 5 also contains a section on the Central Limit Theorem which was belatedly added to the originally published course.

O. D. Morris
Frances O’Regan

June 2014
Section 5.1  The sampling distribution of the mean – The Central Limit Theorem

In Section 4.4 it was stated that the purpose of sampling is to obtain information about a whole population by surveying a small part of the population. This small part is called a sample.

When we select a sample from a population, and study it, we hope that it is representative of the population as a whole. To ensure that it is representative, it must be a random sample. By a random sample we mean that

(i) every member of the population has an equal chance of being selected
(ii) the selections are made independently.

A very important part of the work of a statistician involves drawing conclusions about a population based on evidence gathered from the sample. This process is known as statistical inference.

Parameters – Statistics

It is known that the mean height of men in Ireland is 176 cm. The mean height of a sample of Munster rugby players is 186 cm. The value 176 cm is called a parameter as it is a numerical property of a population. The value 186 cm is called a statistic as it is a numerical property of a sample.

A parameter is a numerical property of a population. A statistic is a numerical property of a sample.

The sampling distribution of the mean

If we are interested in the weights, for example, of all sixteen-year-olds in Ireland, we generally require the mean and standard deviations of these weights. We use the symbols

(i) $\mu$ to denote the population mean
(ii) $\sigma$ to denote the population standard deviation.
In such a large population it would be impossible to obtain the weight of each person and so the values of \( \mu \) and \( \sigma \) will not be known. However, if we take a random sample of this population, we can get approximate values for \( \mu \) and \( \sigma \). Obviously, the larger the sample, the more accurate we would expect the approximations to be.

If we take a large number of different random samples of size \( n \), each sample will have its own mean, \( \bar{x} \), and standard deviation, \( \sigma_x \).

Some of these samples are illustrated on the right.

The different means of these samples are called the **sample means**.

If a large number of samples of the same size are taken, you get a correspondingly large number of means. These means form their own distribution giving us the **distribution of sample mean**. This distribution is also called the **sampling distribution of the mean**.

The following example illustrates the shape a distribution might take when different samples (of the same size) from a population are selected.

**Example 1**

A population consists of five digits 2, 4, 6, 8, 10.

(i) Write down all the possible samples of 2 different digits that can occur if random samples are taken.

(ii) Find the mean of each sample and plot the distribution of the sample means.

(iii) Compare the value of the mean of the sample means with the value of the population mean.

(i) The possible samples are:
   (2, 4), (2, 6), (2, 8), (2, 10), (4, 6), (4, 8), (4, 10), (6, 8), (6, 10), (8, 10)

(ii) Their means are: 3, 4, 5, 6, 5, 6, 7, 7, 8, 9

The distribution of the sample means is plotted below.

(iii) The mean of the population is 6.

The mean of the sample means is also 6.

Thus the mean of the sample means and the population mean are equal.
If you examine the distribution of the sample means plotted on the right in the worked example above, you will notice that it begins to approximate to a normal distribution. In this case the sample size was only 2.

However, as the sample size $n$ increases the closer the distribution will approximate to a normal distribution. Also the mean of the sampling distribution will be the same as the mean of the population.

The successive diagrams below illustrate the shape of the sampling distribution of means resulting from different-sized samples from a given population with a normal distribution.

Distribution when $n = 2, 5$ and 25.

From the diagrams, you can see that if samples are taken from a normal population, the sampling distribution of means is normal for any sample size.

As $n$ increases, the curve representing the sampling distribution of the mean gets taller and narrower. These diagrams also show how the standard deviation decreases as $n$ increases. The sample means will be packed tightly around the population mean. The larger the samples become, the tighter the means will be packed.

From the worked example and from the three diagrams shown above, we can see that when samples are taken from a population, the sampling distribution of the mean takes on the characteristics of a normal curve as the sample size increases. This observation leads us to one of the most important theorems in statistics that is widely used in sampling. It is called the Central Limit Theorem and is stated more formally below.

The Central Limit Theorem

If a random sample of size $n$ with mean $\bar{x}$ is taken from a population with mean $\mu$ and standard deviation $\sigma$, then

- If the sample size is large ($n \geq 30$), the distribution of the sample means will approximate to a normal distribution regardless of what the population distribution is.
- The mean of the distribution will be the same as the population mean $\mu$.
- The standard deviation of the sampling distribution (denoted by $\sigma_{\bar{x}}$) is given by $\frac{\sigma}{\sqrt{n}}$.

[$\frac{\sigma}{\sqrt{n}}$ is often referred to as the standard error of the mean.]

As $n$ increases, the standard error gets smaller.

- If the underlying population is normal, the sampling distribution of the mean will always have a normal distribution even if the sample size is small ($<30$).
The diagram on the right illustrates how the distribution of the sample mean approximates to a normal distribution even when the underlying population is skewed.

When dealing with the sampling distribution of the mean, we convert the given units to standard units using the formula given on the right.

\[ z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \]

**Example 2**

A random sample of 250 is selected from a population having mean 30 and standard deviation 5.
Find the probability that the sample mean is greater than 30.5.

Since \( n = 250 \), the sample mean is normally distributed since \( n \geq 30 \).
Changing to standard units we get:

\[ z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{30.5 - 30}{5/\sqrt{250}} = \frac{0.5}{0.3162} = 1.581 \]

\[ z = 1.581 \]

Now \( P(x > 30.5) = P(z > 1.581) \)
\[ = 1 - P(z < 1.581) \]
\[ = 1 - 0.9429 \]
\[ = 0.0571 \]

The probability that the mean is greater than 30.5 is 0.571.

**Example 3**

A normal distribution has a mean of 40 and a standard deviation of 4.
If 25 items are drawn at random, find the probability that their mean lies between 38 and 40.5.
Converting the given units to standard units we get:

\[ z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \]

For \( x = 38 \), \( z = \frac{38 - 40}{4} = \frac{-2}{0.8} = -2.5 \)

For \( x = 40.5 \), \( z = \frac{40.5 - 40}{4} = \frac{0.5}{0.8} = 0.625 \)

\[
P(38 < x < 40.5) = P(-2.5 < z < 0.625) = P(z < 0.625) - P(z < -2.5) = P(z < 0.625) - [1 - P(z < 2.5)] = 0.7324 - [1 - 0.99379] = 0.7324 - 0.00621 = 0.7262
\]

\( \Rightarrow P(\text{mean lies between 38 and 40.5}) = 0.7262. \)

### Example 4

A population is normally distributed with mean 12 and standard deviation 3. Find the sample size such that \( P(\bar{x} > 12.5) = 0.05 \), where \( \bar{x} \) is the sample mean.

\[
P(z > z_1) = 0.05 \Rightarrow P(z < z_1) = 0.95 \Rightarrow z_1 = 1.645
\]

\[
z_1 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow 1.645 = \frac{12.5 - 12}{3/\sqrt{n}} \Rightarrow 1.645 = \frac{\sqrt{n}(12.5 - 12)}{3} \Rightarrow \sqrt{n} = \frac{(1.645)3}{0.5} \Rightarrow n = 97.42 \text{ i.e. 98 ... round up}
\]

The required sample size is 98.
Example 5

A company installs new machines for packing peanuts. The company claims that the machines fill packets with a mean mass of 500 g and a standard deviation of 18 g.

To test the company’s claim several samples of size 40 packets are taken and their mean masses, \( \bar{x} \) grams, are recorded.

(i) Describe the sampling distribution of \( \bar{x} \) and explain your answer, referring to the theorem you have used.

(ii) Write down the mean and standard deviation of the distribution of \( \bar{x} \).

(iii) Draw a rough sketch of the sampling distribution of \( \bar{x} \).

(iv) Find the probability that the mean of the distribution of \( \bar{x} \) is less than 496.

(v) What sample size \( n \) is required so that \( P(\bar{x}) > 0.06? \)

(i) The sampling distribution of \( \bar{x} \) is approximately normal as the sample size of 40 is sufficiently large (i.e. \( \geq 30 \)) to apply The Central Limit Theorem.

(ii) The mean of the distribution of the sample means is 18 g, the same as the population mean.

The standard deviation (or standard error) is

\[
\frac{\sigma}{\sqrt{n}} = \frac{18}{\sqrt{40}} = 2.846
\]

\[= 2.85 \]

(iii) A sketch of the distribution of \( \bar{x} \) is shown below.

(iv) Converting the given units to z-scores, we use \( z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \).

For \( x = 496 \),

\[
z = \frac{496 - 500}{18/\sqrt{40}} = \frac{-4}{2.846} = -1.405
\]

\[P(\bar{x} < 496) = P(z < -1.405)
= 1 - P(z < 1.405)
= 1 - 0.9265
= 0.0735 \]

The probability that \( \bar{x} < 496 \) = 0.0735 or 7.35%. 

(v) \( P(z > z_1) = 0.06 \)
\( P(z < z_1) = 1 - 0.06 = 0.94 \)
\( \Rightarrow z_1 = 1.56 \)
\[
z_1 = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \Rightarrow 1.56 = \frac{503 - 500}{18 / \sqrt{n}} = \frac{3\sqrt{n}}{18} = \frac{\sqrt{n}}{6}
\]
\( \Rightarrow 1.56 = \frac{\sqrt{n}}{6} \)
\( \Rightarrow \sqrt{n} = 6(1.56) = 9.36 \)
\( \Rightarrow n = (9.36)^2 \)
\( \Rightarrow n = 87.6 \approx 88 \quad \text{... round up} \)

The sample size required is 88.

**Exercise 5.1**

1. Fill in the correct word or symbol to complete the following statements:
   (i) When a large number of samples of size \( n \) are taken from a population, then the distribution of \( \bar{x} \), the sample mean, is known as the _______ ________ of the mean.
   (ii) As the sample size increases, the standard deviation of the sampling distribution of the sample means will ________.
   (iii) If the mean of the underlying population is \( \mu \), the mean of the sampling distribution of the means is ________.
   (iv) If the standard deviation of a population is \( \sigma \) and samples of size \( n \) are taken from it, then the standard deviation of the distribution of the sample means is ________.

2. The diagram on the right shows two curves. One of these curves represents a distribution and the other represents the distribution of the sample means of size \( n \) taken from this distribution. Which curve represents the distribution of the sample means?

3. Samples of size 36 are taken from a population with mean 12 and standard deviation 2. The sampling distribution of the means are plotted in a curve.
   (i) Describe the shape of this curve naming the theorem you have used to support your description.
   (ii) Explain why the theorem you have mentioned can be applied when the shape of the underlying population is unknown.
   (iii) Write down the mean and standard deviation of the sampling distribution of the mean.
4. A population consists of the elements \{4, 6, 8, 10\}.
   (i) Write down all possible samples of size 2 (chosen with replacement) from this population.
   (ii) Give the sample mean, \( \bar{x} \), for each pair.
   (iii) Are each of the values you have found a statistic or a parameter?
   (iv) Show that the mean of all possible samples of size 2 equals the mean of the population.

5. Explain the difference between a parameter and a statistic.

6. The diagram on the right shows two curves \( A \) and \( B \). Diagram \( A \) represents the distribution of a population and diagram \( B \) represents the distribution of the means from a large number of samples of size 40.
   (i) Is distribution \( A \) skewed positively or negatively?
   (ii) Describe distribution \( B \).
   (iii) Explain why the Central Limit Theorem can be used to describe distribution \( B \) even though the underlying population is not normally distributed.

7. A random sample of size 36 is chosen from a population with a mean of 12 and a standard deviation of 3.
   Find the probability that the sample mean is greater than 13.

8. A random sample of size 15 is taken from a normal distribution with mean 60 and standard deviation 4.
   Find the probability that the mean of the sample is less than 58.

9. Men have a mean height of 176 cm with standard deviation 11 cm.
   Find the probability that the mean of a random sample of 80 men
   (i) exceeds 177 cm
   (ii) is less than 174.8 cm.

10. At a certain college, students spend on average 4.2 hours per week at a computer terminal, with a standard deviation of 1.8 hours.
    (i) Find the standard error for a random sample of 36 students.
    (ii) Find the probability that the average time spent using a computer terminal is
         (a) at least 4.8 hours
         (b) between 4.1 and 4.5 hours.

11. The sugar content per litre bottle of a soft drink is known to be distributed with mean 5.8 and standard deviation 1.2. A sample of 900 bottles is taken at random and the sugar content of each bottle is measured.
    Estimate to 3 decimal places the probability that the mean sugar content of the 900 bottles will be less than 5.85.
12. A firm produces alternators for cars. The alternators are known to have a mean lifetime of 8 years with standard deviation 6 months. Forty samples of 144 alternators produced by the firm are tested. Estimate the number of samples which would be expected to have a mean lifetime of more than 8 years and 1 month.

13. A random sample of size 10 is taken from a normal distribution with mean 200 and standard deviation 10. Find the probability that the sample mean lies outside the range 198 to 205.

14. In the given diagram, curve $\text{A}$ represents a normal distribution. Curve $\text{B}$ represents the sampling distribution of means taken from samples of size 36. The distribution represented by $\text{A}$ has mean $\mu = 80$ and standard deviation $\sigma = 8$. The point $C$ represents the mean of both distributions. The point $D$ represents the value of the variable that is two standard deviations from $C$ in distribution $\text{A}$. The point $E$ represents the value of the variable that is one standard error from $C$ in distribution $\text{B}$. Write down the values of $C$, $D$ and $E$.

15. A normal distribution has mean 75 and standard deviation 9. A sample of size $n$ is selected at random and the mean of this sample is $\bar{x}$. Find $n$ if $P(\bar{x} > 73) = 0.8708$.

16. A normal distribution has a mean of 30 and a standard deviation of $\sqrt{5}$.
   (i) Find the probability that the mean of a random sample of 40 exceeds 30.5.
   (ii) Find the value of $n$ such that the probability that the mean of a sample of size $n$ exceeds 30.4 is less than 0.01.

17. Free-range eggs supplied by a health food cooperative have a mean weight of 52 g with a standard deviation of 4 g. Assuming the weights are normally distributed find the probability that:
   (i) a randomly selected egg will weigh more than 60 g
   (ii) the mean weight of five randomly selected eggs will be between 50 g and 55 g
   (iii) the mean weight of 90 randomly selected eggs will be between 52.1 g and 52.2 g.
Which of your answers would be unchanged if the weights are not normally distributed?
Section 5.2 Confidence interval for a mean

In Section 5.1, the Central Limit Theorem was used to show that the sampling distribution of the mean approximates to a normal distribution for large $n$ ($n \geq 30$). In this section we introduce a different way of presenting information provided by a sample mean to estimate the mean of the population from which the sample came.

If samples of size $n$ are taken from a population, the means of the samples will vary. To accommodate this variety, we introduce the concept of a confidence interval. This interval will produce a range of values in which we are 'quite confident' the population mean $\mu$ lies. The endpoints of this interval are called confidence limits.

But how do we measure this confidence? The degree of confidence is generally given as a percentage. These percentages are generally 90%, 95% and 99%.

The most commonly used measure of confidence is a 95% confidence level. This means that there is a 95% probability that the population mean lies in the given interval.

In the standard normal distribution, we require the values of $z$ such that 95% of the population lies in the interval $-z_1 \leq z \leq z_1$. The work involved in finding the value of $z$, is shown below.

We use the standard normal tables on pages 36 and 37 of Formulae and Tables.

From the given diagram,

$$P(z \leq z_1) = 0.95 + 0.025 = 0.975$$

From the tables

$$z_1 = 1.96$$

$$\Rightarrow -z_1 = -1.96$$

Thus in the normal distribution, 95% of the population lies within 1.96 standard deviations of the mean. Since the sample mean is normally distributed, 95% of the population will lie in the interval

$$\bar{x} \pm 1.96 \sigma_x,$$

where $\sigma_x$ is the standard error of the mean.

If $\mu$ is the population mean, then 95% of the sample means lie in the interval

$$\bar{x} - 1.96 \sigma_x < \mu < \bar{x} + 1.96 \sigma_x$$

where $\sigma_x = \sigma / \sqrt{n}$, $\sigma$ being the standard deviation of the population.

This can be written as $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$, which are the end-points (or confidence limits) of the mean.
Confidence Interval for Mean

If $\bar{x}$ is the mean of a random sample of size $n$ taken from a population with a normal distribution with known standard deviation $\sigma$, then the end-points of the 95% confidence interval for $\mu$, the population mean, are given by

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

Note: If $\sigma$, the standard deviation of the population, is not given, use the standard deviation of the sample as an approximation.

Example 1

A random sample of 400 oranges was taken from a large consignment with unknown mean $\mu$ and standard deviation 15 grams. The mean weight of the random sample was 81.4 grams.

Find a 95% confidence interval for the mean weight of the oranges in the consignment.

The 95% confidence interval for $\mu$ is

$$
\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 81.4 \pm 1.96 \frac{15}{\sqrt{400}} \ldots \sigma = 15 \text{ and } n = 400
$$

$$
= 81.4 \pm 1.96(0.75)
$$

$$
= 81.4 \pm 1.47
$$

$$
= 79.93, 82.87
$$

$\Rightarrow 79.93 < \mu < 82.87$

The mean of the consignment lies between 79.93 g and 82.87 g.

Example 2

A certain type of tennis ball is known to have a height of bounce which is normally distributed with standard deviation 2 cm. A sample of 60 such tennis balls is tested and the mean height of the bounce of the sample is 140 cm.

(i) Find a 95% confidence interval for the mean height of the bounce of this type of tennis ball.

(ii) Explain what is meant by a “95% confidence interval”.

(iii) If a tennis ball is selected at random, what is the probability that its bounce is outside the confidence interval found in (i) above?
(i) The 95% confidence interval is given by
\[ \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \]
\[ = 140 \pm 1.96\left(\frac{2}{\sqrt{60}}\right) \]
\[ = 140 \pm 1.96(0.258) \]
\[ = 140 \pm 0.506 \]
\[ = 140.506, 139.494 \]
The 95% confidence interval is 140.506 < \mu < 139.494.

(ii) A “95% confidence interval”, means that on 95 occasions out of 100 the interval will contain the true population mean.

(iii) \( P(\text{ball bounce lies outside 95% confidence interval}) = \frac{5}{100} = \frac{1}{20} \).

**Example 3**

The heights of people have a standard deviation of 11.5 cm. It is required to estimate the mean height of people, with 95% confidence, to within ±0.4 cm. What sample size should be taken in order to achieve this estimate?

Let \( \mu \) be the mean height of people.

The 95% confidence limits for \( \mu \) are \( \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \).

\[ \Rightarrow \pm 1.96 \frac{\sigma}{\sqrt{n}} = \pm 0.4 \ldots \text{standard error is} \pm 0.4 \text{ cm} \]

\[ \Rightarrow 1.96\left(\frac{11.5}{\sqrt{n}}\right) = 0.4 \]

\[ \Rightarrow \sqrt{n} = \frac{11.5(1.96)}{0.4} \]

\[ \Rightarrow \sqrt{n} = 56.35 \]

\[ \Rightarrow n = (56.35)^2 = 3175.3 \]

Therefore, a sample of at least 3176 should be taken.
Example 4

On the basis of the results obtained from a random sample of 100 men from a particular district, the 95% confidence interval for the mean height of the men in the district is found to be (177.22 cm, 179.18 cm). Find the value of \( \bar{x} \), the mean of the sample, and \( \sigma \), the standard deviation of the normal population from which the sample is drawn.

The 95% confidence interval is given by
\[
\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = (177.22,179.18)
\]
\[
\Rightarrow \bar{x} + 1.96 \frac{\sigma}{10} = 179.18 \quad \cdots \quad 1
\]
and \( \bar{x} - 1.96 \frac{\sigma}{10} = 177.22 \quad \cdots \quad 2 \)

Adding 1 and 2:
\[
2\bar{x} = 356.4
\]
\[
\bar{x} = 178.2
\]

Subtracting 1 and 2:
\[
2(1.96) \frac{\sigma}{10} = 1.96
\]
\[
\frac{2\sigma}{10} = 1 \Rightarrow \sigma = 5
\]

The sample mean \( \bar{x} = 178.2 \) cm.
The population standard deviation is 5 cm.

Exercise 5.2

1. A population has mean \( \mu \) and standard deviation 12.
A random sample of 800 from this population has mean 63.
Find a 95% confidence interval for \( \mu \).

2. The weights of dairy cows are known to have a standard deviation of 42 kg.
A random sample of 280 dairy cows has a mean weight of 284 kg.
Find a 95% confidence limit for the mean weight of all the cows.

3. Seventy packs of butter, selected at random from a large batch delivered to a supermarket, are weighed. The mean weight is found to be 227 g and the standard deviation is found to be 7.5 g.
   (i) Calculate a 95% confidence interval for the mean weight of all packs in the batch.
   (ii) If one pack is selected at random from the seventy packs, find the probability that its weight is not in the given interval.

4. In a random sample of 100 students taking a state examination, it was found that the mean mark was 62.7 with a standard deviation of 9.2 marks.
Find the 95% confidence limits for the mean score of all the students who took the examination.
5. The weight of vitamin E in a capsule manufactured by a drug company is normally distributed with standard deviation 0.04 mg. A random sample of 12 capsules was analysed and the mean weight of vitamin E was found to be 5.12 mg.
   (i) Calculate a 95% confidence interval for the population mean weight of vitamin E per capsule.
   (ii) Give the values of the end-points of the interval, correct to three significant figures.
   (iii) Explain what is meant by “95% confidence”?

6. A bank selected a random sample of 400 customers and found that they had a mean credit of €280 with a standard deviation of €105 in their accounts. Calculate a 95% confidence interval for the mean credit of all the bank’s customers.

7. Shoe shop staff routinely measure the lengths of their customers’ feet. Measurements of the length of one foot (without shoes) from each of 180 adult male customers yielded a mean length of 29.2 cm and a standard deviation of 1.47 cm.
   (i) Calculate a 95% confidence interval for the mean length of male feet.
   (ii) Why was it not necessary to assume that the lengths of feet are normally distributed in order to calculate the confidence interval in (i) above?

8. A random sample of 64 sweets is selected from a large batch. The sweets are found to have a mean weight of 0.932 grams and a standard deviation of 0.1 grams.
   (i) Calculate the standard error of the mean.
   (ii) What is the best estimate for μ, the mean of the large batch?
   (iii) Construct a 95% confidence interval for μ.
   (iv) What would happen if a sample size of 100 was selected rather than a sample of 64?
   (v) What conclusion can you draw from your result in part (iv)?

9. A random sample of 240 cars had a mean age of 4.6 years with a standard deviation of 2.5 years.
   (i) Give a 95% confidence interval for the mean age of all cars.
   (ii) What size of sample would be needed to estimate the mean age, with 95% confidence, to within ±0.2 years?

10. 150 boxes of cereal of a certain brand are weighed and the mean weight is 748 grams with standard deviation 3.6 grams.
    (i) Find a 95% confidence interval for the mean weight of all boxes of cereal of that brand.
    (ii) What size of sample would be needed to estimate the mean weight, with 95% confidence, to within ±1.5 grams?

11. Eighty people were asked to measure their pulse rates when they woke up in the morning. The mean was 69 beats per minute and the standard deviation 4 beats.
    (i) Find a 95% confidence interval for the population mean.
    (ii) What size of sample would be needed to estimate the mean number of beats, with 95% confidence, to within ±1.5 beats?
12. The weights of pebbles on a beach are distributed with mean 48.6 g and standard deviation 8.5 g.
   A random sample of 50 pebbles is chosen.
   (i) Find the probability that the mean weight will be less than 49 g.
   (ii) Find the limits within which the central 95% of such sample means would lie.
   (iii) How large a sample would be needed in order that the central 95% of sample means would lie in an interval of width at most 4 g?

13. The 95% confidence interval for the mean mark of a group of students is (54.09, 60.71).
   This interval is based on the results from a random sample of 80 students.
   (i) Find $\bar{x}$, the mean of the sample.
   (ii) Find $\sigma$, the standard deviation of the normal population from which the sample is taken.

**Section 5.3 Confidence interval for a proportion**

In Section 4.4 of Chapter 4 it was shown how to find the confidence interval for a population proportion using the margin of error $\frac{1}{\sqrt{n}}$, where $n$ is the sample size.

This confidence interval is shown again on the right.

The 95% confidence interval for a proportion $p$ is

$$\hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

Here is a reminder of what a proportion is!

If 150 television viewers are interviewed in a sample survey and 63 say they like a new situation comedy, then $\frac{63}{150} = 0.42$ is the proportion of the sample who like the new show. This sample proportion, $\hat{p}$, is used as an estimate of the true population proportion $p$ of television viewers who like the new show.

The notation $\hat{p}$ is used to denote **sample proportion**.
The notation $p$ is used to denote **population proportion**.

Since $p$ is generally not known, $\hat{p}$ is used as an estimator for the true population proportion, $p$.

If many samples of the same size are taken from a population, each sample will produce a different (but similar) proportion. All these proportions form their own distribution called the **sampling distribution of the proportion**.

The **standard error**, $\sigma_p$, of this distribution is given on page 34 of *Formulae and Tables* and is shown on the right.

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$
In this section we will use the Standard Normal Tables (rather than the margin of error, \(\frac{1}{\sqrt{n}}\)) to get a more accurate confidence interval for a population proportion.

Since the 95% level of confidence will be used, the diagram on the right will remind us that 95% of a normal distribution lies within 1.96 standard deviations of the mean.

If \(\hat{p}\) is the sample proportion and \(p\) is the population proportion, then the 95% confidence interval for \(p\) is given by

\[
\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}
\]

This can be written more concisely as

\[
\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}.
\]

The 95% confidence interval for a population proportion

\[
\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}.
\]

Note: An increase in confidence levels results in an increase in the interval width.

**Example 1**

In a survey carried out in a large city, 170 households out of a random sample of 250 owned at least one pet.

(i) Find the standard error of the sampling distribution of the proportion at the 95% confidence level.

(ii) Find the 95% confidence interval for the proportion of households in the city who own at least one pet.

(i) The sample proportion \(\hat{p} = \frac{170}{250} = 0.68\)

Standard error \(\sigma_p = \sqrt{\frac{p(1-p)}{n}}\)

\[
= \sqrt{\frac{0.68(1-0.68)}{250}}
\]

\[
\sigma_p = 0.029
\]
(ii) The 95% confidence interval is given by
\[
\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}
\]
\[
= 0.68 \pm 1.96(0.029) \ldots \text{from (i) above}
\]
\[
= 0.68 \pm 0.0568
\]
\[
= (0.6232, 0.7368) \text{ or about (62%, 74%)}
\]

**Example 2**

A random sample of 250 cars were surveyed passing a certain junction and 36 were found to have K registrations.

(i) Determine a 95% confidence interval for the proportion of cars in that area that have a K registration.

(ii) What sample size would have to be taken in order to estimate the percentage to within ±2%?

(i) The sample proportion \( \hat{p} = \frac{36}{250} = 0.144 \).

The 95% confidence interval is given by
\[
\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}
\]
\[
= 0.144 \pm 1.96\sqrt{\frac{0.144(1-0.144)}{250}}
\]
\[
= 0.144 \pm 1.96(0.0222)
\]
\[
= 0.144 \pm 0.0435
\]
\[
= 0.100, 0.1875
\]

The 95% confidence interval is (0.100, 0.1875).

(ii) Let \( n \) be the sample size.

We require \( n \) such that \( \hat{p} \pm \sqrt{\frac{p(1-p)}{n}} = \hat{p} \pm 0.02 \ldots 2\% = 0.02 \)

\[
\therefore \sqrt{\frac{0.144(1-0.144)}{n}} = 0.02
\]
\[
\sqrt{\frac{(0.144)(0.856)}{n}} = 0.02
\]
\[
\frac{0.123264}{n} = (0.02)^2
\]
\[
n = \frac{0.123264}{(0.02)^2} = 308.16
\]

So a sample size of 309 would have to be taken.
Exercise 5.3

1. A manufacturer wants to assess the proportion of defective items in a large batch produced by a particular machine. He tests a random sample of 300 items and finds that 45 are defective. Calculate a 95% confidence interval for the proportion of defective items in the complete batch.

2. In order to assess the probability of a successful outcome, an experiment is performed 200 times and the number of successful outcomes is found to be 72. Find a 95% confidence interval for \( p \), the probability of a successful outcome.

3. A market researcher carries out a survey in order to determine the popularity of SUDZ washing powder in the Cork area. He visits every house in a large housing estate in Cork and asks the question: “Do you use SUDZ washing powder?” Of 235 people questioned, 75 answered “YES”. Treating the sample as being random, calculate a 95% confidence interval for the proportion of households in the Cork area which use SUDZ.

4. An importer has ordered a large consignment of tomatoes. When it arrives, he examines a randomly chosen sample of 50 boxes and finds that 12 contain at least one bad tomato. Assuming that these boxes may be regarded as being a random sample from the boxes in the consignment, obtain an approximate 95% confidence interval for the proportion of boxes containing at least one bad tomato, giving your confidence limits correct to three decimal places.

5. If 400 persons, constituting a random sample, are given a flu vaccine and 136 of them experienced some discomfort, construct a 95% large-sample confidence interval for the corresponding true proportion.

6. A random sample of 120 library books is taken as they are borrowed. They are classified as fiction or non-fiction, and hardback or paperback. 88 books are found to be fiction, and of these, 74 are paperback. Find a 95% confidence limit for:
   (i) the proportion of books borrowed that are fiction
   (ii) the proportion of fiction books borrowed that are paperback.

7. In a sample of 400 shops taken in 2012, it was discovered that 136 of them sold carpets at below the list prices which had been recommended by manufacturers.
   (i) Estimate the percentage of all carpet-selling shops selling below list price.
   (ii) Calculate the 95% confidence limits for this estimate, and explain briefly what these mean.
   (iii) What size sample would have to be taken in order to estimate the percentage to within ±2%?
8. In a random sample of 1,200 voters interviewed nationwide, only 324 felt that the salaries of certain government officials should be raised. Construct a 95% confidence interval of the corresponding true proportion.

9. In a market research survey, 15 people out of a random sample of 100 from a certain area said that they used a particular brand of soap.
   (i) Calculate a 95% confidence interval for the proportion of people who use this brand of soap.
   (ii) What size sample would need to be taken in order to estimate the percentage to within $\pm 1\frac{1}{2}\%$? Give your answer correct to the nearest 10.

Section 5.4 Hypothesis testing for a population mean

Hypothesis testing

In Sections 5.2 and 5.3 we dealt with confidence intervals, one of the two most common types of statistical inference.
The second type of statistical inference has a different objective. It is called hypothesis testing.
Its purpose is to test the truth or otherwise of a claim, statement or hypothesis made about a population parameter.

An hypothesis is a statement or conjecture made about some characteristic or parameter of a population.

Here is an example of an hypothesis:

‘The mean age of men on their wedding day is 32 years.’

An hypothesis test is a statistical method of proving the truth or otherwise of this statement.

It has already been shown that in any normal distribution 95% of the population lies within 1.96 standard deviations of the mean, that is, 95% of the population will be in the interval $\mu \pm 1.96\sigma$.

If we are dealing with a normal distribution and an experiment produces a result which is outside the interval $\mu \pm 1.96\sigma$, we would be inclined to suspect that factors other than chance are involved in the result. For example, some form of bias may be present. If we toss a coin 100 times we would ‘expect’ heads to occur 50 times.

Do we conclude that the coin is biased if heads occur 60 times?

Is the unexpected result more than mere chance?

To answer this question we start with the assumption, or hypothesis, that the coin is not biased.

This assumption is called the null hypothesis, denoted by $H_0$.

Usually the null hypothesis is a statement of “no difference”, “no effect” or “no change”.

An hypothesis test is then carried out to accept or reject the null hypothesis.

In this test, we speak of rejecting the null hypothesis ‘at a certain level’.

This ‘certain level’ is called the level of significance.

The 5% level of significance is by far the most commonly-used one.

It is the only one that we deal with in our course.
The 5% level of significance means that the result obtained is likely to occur on only 5 occasions out of 100.

At the 5% level of significance, the set of values, \( z > 1.96 \) or \( z < -1.96 \), is known as the critical region and the boundaries of the critical region are called the critical values.

If the values of \( z \) are in the critical region (i.e. \( z > 1.96 \) or \( z < -1.96 \)), we reject the null hypothesis and conclude that factors other than chance are involved.

The critical regions at the 5% level of significance are shown below.

Hypothesis Testing
At the 5% level of significance, the null hypothesis is rejected if \( z < -1.96 \) or \( z > 1.96 \)

Hypothesis testing for a population mean
When a population is very large, it is generally not practical to find the true mean and standard deviation of the total population. However, assumptions are often made about these values and their validity is tested based on observations made from random samples taken from the population.

Take, for example, machines designed to produce batteries which last for 120 hours with a standard deviation 4 hours. What conclusions can we come to about one of these machines if a random sample of 50 batteries produced by it had a mean life of 121 hours?

We now begin the process of investigating whether these machines are producing the type of battery they were designed to produce. This process is called hypothesis testing.

Here are the basic steps of a hypothesis test:

1. Write down \( H_0 \), the null hypothesis, and \( H_1 \), the alternative hypothesis.
   - \( H_0 \): The mean life of a battery is 120 hours.
   - \( H_1 \): The mean life of a battery is not 120 hours.

2. State the significance level, \( \alpha \).
   The significance level on our course is 5% (\( \alpha = 0.05 \)).
   This means that if \( z < -1.96 \) or \( z > 1.96 \), we reject the null hypothesis and accept the alternative hypothesis.

3. Calculate the value of the test statistic.
   This involves converting the given units to \( z \)-units.
To convert the given units to standard units we use

\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \]

where

- \( \bar{x} \) = the sample mean
- \( \mu \) = population mean
- \( \sigma \) = population standard deviation
- \( n \) = size of sample

For the machine mentioned above,

\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{121 - 120}{\frac{4}{\sqrt{50}}} = \frac{1}{0.566} = 1.767 \]

The test statistic is \( z = 1.767 \).

4. Come to a conclusion.

Since \( z = 1.767 \) does not lie outside the range \(-1.96 < z < 1.96\) it is not in the critical region. So we accept the null hypothesis which states that the mean life of a battery is 120 hours.

Note: If \( \sigma \), the standard deviation of the population is not given, use \( \sigma / \sqrt{n} \) the standard deviation of the sample instead.

---

**Example 1**

Over the years, a market gardener found that the mean yield from his tomato plants was 1.83 kg per plant with a standard deviation of 0.35 kg per plant.

One year he planted 600 of a new variety and these yielded 1.87 kg per plant.

At the 5% level of significance, test whether the mean yield from the new plants is different from his normal variety.

1. **H\(_0\):** The mean \( \mu \) is 1.83.
   **H\(_1\):** The mean \( \mu \) is not 1.83.

2. The level of significance is 5%.
   The critical region is \( z < -1.96 \) or \( z > 1.96 \).

3. Calculate the test statistic by converting to standard units.

\[ z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \]

\[ \bar{x} = 1.87 \quad \mu = 1.83 \]

\[ n = 600 \quad \sigma = 0.35 \]

\[ z = \frac{1.87 - 1.83}{\frac{0.35}{\sqrt{600}}} = \frac{0.04}{0.0143} = 2.797 \]

Since \( z = 2.797 \) and 2.797 > 1.96, we reject the null hypothesis and conclude that the new variety is different from the normal variety.
**Using p-values**

Suppose we carry out an hypothesis test and find the test statistic to be \( z = 2.16 \).
Since 2.16 is greater than 1.96, we reject the null hypothesis at the 5% level of significance \( (\alpha = 0.05) \).

Instead of comparing \( z = 2.16 \) with \( z = 1.96 \) (and \( z = -1.96 \)), we compare the total area of the two coloured regions below with the specific level of significance, \( \alpha = 0.05 \).

We use pages 36 and 37 of *Formulae and Tables* to find the probability that \( z \leq -2.16 \) or \( z \geq 2.16 \).

\[
P(z \leq -2.16) + P(z \geq 2.16) = 2P(z \geq 2.16) = 2[1 - P(z \leq 2.16)] = 2[1 - 0.9846] = 2[0.0154] = 0.0308
\]

The shaded areas above are referred to as the **p-value**, or probability-value corresponding to the observed value of the test statistic.

The value 0.0308 found above is the **p-value** that corresponds to the test statistic \( z = |2.16| \).

The **p-value** 0.0308 is interpreted as the **lowest level of significance** at which the null hypothesis could have been rejected.

With a test statistic of \( z = 2.16 \), we would certainly have rejected the null hypothesis at the specified level of significance \( (\alpha = 0.5) \).

The **p-value** of 0.0308 gives us a **specific** or more precise level of significance.

The **smaller** the **p-value** is, the **stronger** is the evidence against \( H_0 \) provided by the data.

---

**The p-value of a Test Statistic**

The p-value is the sum of the two shaded areas.

\[
p-value = 2 \times P(z > |z_1|),
\]

where \( z_1 \) is the test statistic.
Example 2

Calculate the p-value for the sample statistic \( z = -2.08 \).

Sample statistic is \( z = -2.08 \).
The sum of the probabilities that

\[
    z > 2.08 \quad \text{and} \quad z < -2.08
\]

is the p-value.

\[
p-value = 2 \times P(z > |2.08|) \\
= 2 \times [1 - P(z < 2.08)] \\
= 2 \times [1 - 0.9812] \\
= 2(0.0188) \\
p-value = 0.0376
\]

Steps involved in a Test of Significance using a p-value
1. Write down the null hypothesis \( H_0 \) and the alternative hypothesis \( H_1 \).
2. State the significance level \( \alpha \). (On our course \( \alpha = 0.5 \).)
3. Calculate the test statistic.
4. Find the p-value that corresponds to the test statistic.
5. If the p-value > 0.05, the result is not significant and we do not reject the null hypothesis \( H_0 \).
   
   If the p-value \( \leq 0.05 \), we reject the null hypothesis \( H_0 \) in favour of the alternative hypothesis \( H_1 \).

Example 3

A random sample of 36 observations is to be taken from a distribution with standard deviation 10. In the past, the distribution has had a mean of 83, but it is believed that the mean may have changed.

When the sample was taken it was found to have a mean of 86.2.

(i) State \( H_0 \) and \( H_1 \).
(ii) Calculate the value of the test statistic.
(iii) Calculate the p-value for the test statistic.
(iv) Use the p-value to state if the result is significant at the 5% level of significance.
    Explain your conclusion.
(i) $H_0$: Mean $\mu = 83$
$H_1$: Mean $\mu \neq 83$
(ii) Test statistic $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$

$$z = \frac{86.2 - 83}{10 / \sqrt{36}} = \frac{3.2 \times \sqrt{36}}{10} = 1.92$$

The test statistic is $z = 1.92$
(iii) The $p$-value $= 2 \times P(z > 1.92)$

$$= 2 \times [1 - P(z < 1.92)]$$

$$= 2 \times [1 - 0.9726]$$

$$= 2(0.0274) = 0.0548$$

(iv) As the $p$-value is not less than or equal to 0.05, the result is not significant; we do not reject the null hypothesis.

**Exercise 5.4**

1. A normal distribution is thought to have a mean of 50.
   A random sample of 100 gave a mean of 52.4 and a standard deviation of 14.3.
   Is there evidence to suggest that the true mean is different from the assumed mean at the 5% level of significance?

2. Over a long period the scores obtained in a particular intelligence test were normally distributed with mean score 70 and standard deviation 6.
   When a test was taken by a random sample of 64 students, the mean score was 68.
   Is there sufficient evidence, at the 5% level of significance, that these students differ from the normal students?

3. The management of a large hospital states that the mean age of its patients is 45 years.
   The HSE statistics department decides to test this claim about the mean age of the patients.
   It took a random sample of 100 patients and found that the mean age was 48.4 years with a standard deviation of 18 years.
   (i) What is the null hypothesis?
   (ii) State the alternative hypothesis.
   (iii) Work out the test statistic for the sample mean.
   (iv) At the 5% level of significance, is there evidence to show that the mean age of the patients is not 45 years?
   Give a reason for your conclusion.
4. A particular machine produces metal rods which are normally distributed with a mean length of 210 cm and with a standard deviation of 6 cm. The machine is serviced and a sample is taken to investigate if the mean length has changed. The sample of 100 rods gave a mean length of 211.5 cm.
   (i) What is the null hypothesis?
   (ii) What is the alternative hypothesis?
   (iii) Work out the test statistic for the sample mean.
   (iv) At the 5% level of significance, is there evidence of a change in the mean length of rods produced by the machine? Explain your conclusion.

5. Mice kept under laboratory conditions have a mean lifespan of 258 days and a standard deviation of 45 days. 64 of these mice, selected at random, were each given a measured dose of a certain drug each day, and the mean lifespan for this group was 269 days. At the 5% level of significance, is there evidence to suggest that the drug has altered the mean lifespan of the mice?

6. In 1970 the average number of children per family in a certain town was 3.8 with a standard deviation of 0.6. In 1980 a random sample of 40 families had a total of 144 children. At the 5% level of significance, is there evidence to conclude that the mean number of children per family had changed?

7. The mean mark for all students taking a certain Leaving Certificate subject was 48.7. In a particular town, 120 students took this examination. The mean mark of these students was 46.5 with a standard deviation of 9.5. At the 5% level of significance, is there evidence to suggest that the mean mark of the students of this town differs from the rest of the population?

8. In each of the following, the z-score for a sample mean is given. Work out the corresponding p-value for each test statistic.
   (i) $z = 1.73$  (ii) $z = -1.91$  (iii) $z = -1.65$  (iv) $z = -2.06$

9. ‘Standard’ batteries have a mean lifetime of 85 hours with a standard deviation 12 hours. A sample of 200 ‘long-life’ batteries had a mean lifetime of 86.5 hours.
   (i) Calculate the sample statistic for this sample.
   (ii) Work out the corresponding p-value for this sample statistic.
   (iii) Is the result significant at the 5% level of significance?

10. Experience has shown that the scores obtained in a particular test are normally distributed with mean score 70 and standard deviation 6. When the test is taken by a random sample of 36 students, the mean is 68.5.
(i) Calculate the sample statistic for this sample.
(ii) Calculate the $p$-value for this sample statistic.
(iii) Use the $p$-value you have found to investigate if the mean score of the sample differs from the mean score of the population at the 5% level of significance.

11. The security department of a warehouse wants to know whether the average time required by the night watchman to walk his round is 12.0 minutes. In a random sample of 36 rounds, the night watchman averaged 12.3 minutes with a standard deviation of 1.2 minutes.
   (i) Calculate the test statistic for this sample.
   (ii) Can we reject the null hypothesis that $\mu = 12.0$ minutes at the 5% level of significance?
   (iii) Work out the $p$-value that corresponds to the test statistic found in (i) above.
   (iv) If this $p$-value is used, do you reach the same conclusion with regard to significance at the 5% level?

12. The lengths of metal bars produced by a particular machine are normally distributed with mean length 420 cm and standard deviation 12 cm. The machine is serviced, after which a sample of 100 bars gives a mean length of 423 cm.
   (i) Calculate the sample statistic for this sample.
   (ii) Work out the $p$-value for this sample statistic.
   (iii) Use this $p$-value to determine if there is evidence, at the 5% level, of a change in the mean length of the bars produced by the machine, assuming that the standard deviation remains the same.

13. A machine is designed to produce screws with a stated mean length of 5 mm. A random sample of 400 screws produced by the machine is found to have a mean length of 5.008 mm and a standard deviation of 0.072 mm. Estimate the standard error of the mean, and obtain an approximate 95% confidence interval for the mean of the whole output of this machine. Investigate if the mean of the sample differs significantly from the stated mean at the 5% level of significance.
Test yourself 5
A – questions

1. The weights of a large collection of bags of potatoes have a mean of 25 kg and a standard deviation of $\sqrt{5}$ kg.
   Estimate, to 2 decimal places, the probability that a random sample of 50 bags will have a mean weight of between 24.5 kg and 25.5 kg.

2. A random sample of size 20 is taken from a population of size 80 (with replacement). Find the mean and standard error of the sample if the population is normally distributed with mean 2.85 and standard deviation 0.07.

3. The pulse-rate of a sample of 32 people was measured. The mean was found to be 26.2 with standard deviation $\sigma = 5.15$. Calculate the 95% confidence interval for the population mean.

4. A machine is regulated to dispense liquid into cartons in such a way that the amount of liquid dispensed on each occasion is normally distributed with a standard deviation of 20 ml. Find the confidence limits for the mean amount of liquid dispensed if a random sample of 40 cartons had an average content of 266 ml.

5. Among the first 150 customers at a new snack bar, 90 order coffee. Assuming that this is a random sample from the population of future customers, estimate a 95% confidence interval for the proportion of future customers who will order coffee.

6. A sample poll of 100 voters chosen at random from all voters in a given constituency indicated that 55% of them were in favour of candidate A. Find the 95% confidence interval for the proportion of all the voters in the district in favour of this candidate.

7. Irish third-level students are known to have a mean height of 176 cm with a standard deviation 11 cm. A random sample of 60 equivalent German students had a mean height of 179 cm. Does this suggest that the mean height of German students differs from that of Irish students at the 95% confidence level?

8. Jars of honey are filled by a machine. It has been found that the quantity of honey in a jar has a mean of 460.3 g with a standard deviation of 3.2 g. It is believed that the machine controls have been altered in such a way that, although the standard deviation is unaltered, the mean quantity may have changed. A random sample of 60 jars is taken and the mean quantity of honey per jar is found to be 461.2 g.
(i) State the null and alternative hypotheses.
(ii) Calculate the sample statistic for the mean.
(iii) Is there evidence, at the 5% level of significance, that the sample mean is different from the population mean?

9. A firm produces batteries which are known to have a mean lifetime of 96 hours. Forty samples of 36 batteries each are tested.
   (i) Describe the sampling distribution of the means of these samples, mentioning the theorem you have used to justify your answer.
   (ii) Explain why the theorem you have mentioned can be applied when the shape of the underlying population is not known.
   (iii) Estimate the number of samples in which the average lifetime of the 36 batteries is greater than 98 hours if the standard deviation of the batteries is 6 hours.

10. Draw a rough sketch of the normal curve showing the critical regions, at the 5% level of significance, of a hypothesis test.
    (i) Clearly indicate the rejection regions.
    (ii) What are the critical z-values for the limits of these rejection regions?
    (iii) For a z-value of 1.6, estimate the corresponding p-value for this statistic.

B – questions

1. A large number of random samples of size $n$ are taken from a normal distribution with a mean of 74 and a standard deviation of 6. The means, $\bar{x}$, of these samples are calculated. Find the sample size $n$ required to ensure that the probability of $\bar{x} > 72$ is 0.854.

2. The weights of bags of fertiliser may be modelled by a normal distribution with mean 12.1 kg and standard deviation 0.4 kg. Find the probability that:
   (i) a randomly selected bag will weigh less than 12.0 kg,
   (ii) the mean weight of four bags selected at random will weigh more than 12.0 kg,
   (iii) the mean weight of 100 bags will be between 12.0 and 12.1 kg.
   How would your answer to (iii) be affected if the normal distribution was not a good model for the weights of the bags? Explain your answer.

3. A plant produces steel sheets whose weights are known to be normally distributed with a standard deviation of 2.4 kg. A random sample of 36 sheets had a mean weight of 31.4 kg. Find a 95% confidence interval for the mean weight of sheets produced by the plant.
4. The residents of a rural area are being asked for their views on a plan to build a wind farm in their area. Environmental campaigners claim that 20% of the residents are against the plan.

(i) State one reason why surveying a random sample of 30 residents will allow reliable conclusions to be drawn.

(ii) Using a 5% significance level, calculate a 95% confidence interval for the population proportion against the plan.

5. (i) Explain briefly what is meant by the term “95% confidence interval”.

(ii) A car manufacturing company tested a random sample of 150 cars of the same model to estimate the mean number of kilometres travelled per litre of petrol consumption for all cars of that model. The sample mean of kilometres travelled per litre consumed was 13.52 and the standard deviation was 2.23.

Form a 95% confidence interval for the mean number of kilometres travelled per litre of petrol consumed for all cars of that make.

Give all calculations correct to two places of decimal.

6. A neurologist wants to test the effect a new drug has on response times. 100 rats are injected with a unit dose of this drug and the response times are recorded. The neurologist knows that the mean response time for rats not injected with the drug is 1.2 seconds.

The mean response time of the 100 rats injected with the drug is 1.05 seconds with a sample standard deviation of 0.5 seconds.

(i) State the null and alternative hypotheses for this test.

(ii) Determine the critical region at the 5% level of significance and illustrate your answer with a sketch.

(iii) Calculate the test statistic and answer the question “Do you think that the drug has an effect on the response time at the 5% level of significance?”

(iv) Calculate the $p$-value for the test statistic and interpret this value.

7. A school of motoring claims that 80% of its clients are successful in their first driving test. A person who did not believe this claim took a random sample of 72 clients and found that 50 of these had been successful in their first driving test.

(i) Using $\frac{1}{\sqrt{n}}$, write down the margin of error.

(ii) Calculate the sample proportion as a decimal correct to two decimal places.

(iii) Write down the confidence interval, at the 95% level of confidence, in terms of $\hat{p}$ and $n$.

(iv) Can the school’s claim be upheld at the 95% level of confidence?

[Note: See Section 4.4, page 166.]
8. A market gardener sells carrots in 25 kg sacks. The wholesaler suspects that the true mean weight is not 25 kg. He weighs a random sample of 50 sacks and finds that the mean weight is 24.5 kg with a standard deviation of 1.5 kg.
   (i) State the null and alternative hypotheses.
   (ii) Calculate the sample statistic for the sample.
   (iii) Calculate the \( p \)-value for this statistic.
   (iv) Is the wholesaler’s suspicion justified at the 5% level of significance?
   (v) Complete the following sentence:
       “The \( p \)-value is the \__________ level of significance at which the null hypothesis could have been \__________.”

9. The weights of male students at a large university are normally distributed with a mean of 68 kg and a standard deviation of 3 kg. Eighty samples of 25 students are picked at random (with replacement).
   (i) Find the mean and standard error of the resulting sampling distribution.
   (ii) In how many of the samples would you expect the sample mean to be less than 67.5 kg?

C – questions

1. A company instals a new machine in a factory. The company claims that the machine will fill bags with wholemeal flour having a mean weight of 500 g and a standard deviation of 18 g. 36 bags are checked in a random sample to test this claim. Their mean weight is 505 g.
   (i) State the null and the alternative hypotheses.
   (ii) Calculate the test statistic for the sample mean.
   (iii) Find the \( p \)-value that corresponds to the test statistic.
   (iv) Is the result significant at the 5% level of significance? Explain your answer.

2. Among 80 fish caught in a certain lake, 28 were inedible as a result of the chemical pollution of their environment.
   (i) Work out the standard error for this proportion.
   (ii) Construct a 95% confidence interval for the true proportion of fish in this lake which are inedible as a result of chemical pollution.

3. The 95% confidence interval for the mean weight, in grams, of a consignment of oranges is (79.93, 82.87). This result is based on a random sample of 400 oranges. Using this confidence interval, find
   (i) \( \bar{x} \), the mean of the sample
   (ii) \( \sigma \), the standard deviation of the normal population from which the sample is taken.
4. The masses of loaves from a certain bakery are normally distributed with mean 500 grams and standard deviation 20 grams.
   (i) Determine what percentage of the output would fall below 475 grams and what percentage would be above 530 grams.
   (ii) A sample of 40 loaves yielded a mean mass of 495 grams. Calculate the sample statistic for the mean.
   (iii) Calculate the p-value for this sample statistic.
   (iv) Does the p-value found above provide evidence that the mean weight of loaves from this sample is different from the mean of 500 g at the 5% level of significance?

5. (a) Write down the mean and standard deviation of the distribution of the means of all possible samples of size \( n \) taken from an infinite population having mean \( \mu \) and standard deviation \( \sigma \).
   Describe the shape of this distribution of sample means when
   (i) \( n \) is large
   (ii) the distribution of the population is normal.
   Explain briefly how the Central Limit Theorem can be applied to (i) and (ii) above.
   (b) The standard deviation of all till receipts at a supermarket during 2013 was €8.50 and the mean of the receipts was €37.
      (i) Find the probability that the mean of a random sample of 100 till receipts is greater than €37.50.
      (ii) Find the value of \( n \) such that the probability that the mean of the sample of size \( n \) exceeds €37.50 is less than 0.06.

6. The distribution of the hourly earnings of all employees in Ireland in October 2009 is shown in the diagram. It can be seen that the distribution is positively skewed.
   The mean is €22.05.
   The median is €17.82.
   The standard deviation is €10.64.
   The lower quartile is €12.80.
   The upper quartile is €26.05.
   (i) If an employee is selected at random from this population, what is the probability that the employee earns more than €12.80?
   (ii) If six employees are selected at random from this population, what is the probability that exactly four of them had hourly earnings of more than €12.80?
   In a computer simulation, random samples of size 200 are repeatedly selected from this population and the mean of each sample is recorded. A thousand such sample means are recorded.
   (iii) Describe the expected distribution of these sample means. Your description should refer to the shape of the distribution and to its mean and standard deviation.
   (iv) How many of the sample means would you expect to be greater than €23?
7. The contents of a bag of oats are normally distributed with mean 3.05 kg and standard deviation 0.08 kg.
   (i) What proportion of bags contains less than 3.11 kg?
   (ii) What proportion of bags contains between 3.00 kg and 3.15 kg?
   (iii) Without using tables, write down the weight that is exceeded by 97.5% of the bags.
        (Use the Empirical Rule.)
   (iv) If 6 bags are selected at random, what is the probability that the mean weight of
        the contents will be between 3.00 kg and 3.15 kg?

8. A gas supplier maintains a team of engineers who are available to deal with leaks reported by customers. Most reported leaks can be dealt with quickly, but some require a long time. The time (excluding travelling time) taken to deal with reported leaks is found to have a mean of 65 minutes and a standard deviation of 60 minutes.
   (a) Assuming that the times may be modelled by a normal distribution, estimate the probability that:
       (i) it will take more than 185 minutes to deal with a reported leak,
       (ii) it will take between 50 minutes and 125 minutes to deal with a reported leak,
       (iii) the mean time to deal with a random sample of 90 reported leaks is less than 70 minutes.
   (b) A statistician, consulted by the gas supplier, stated that, as the times had a mean of 65 minutes and a standard deviation of 60 minutes, the normal distribution would not provide an adequate model.
       (i) Explain the reason for the statistician's statement.
       (ii) Give a reason why, despite the statistician's statement, your answer to (a) (iii) is still valid.

9. Hens’ eggs have masses which may be said to have a normal distribution about a mean mass of 60 g and a standard deviation of 15 g.
   Eggs of mass less than 45 g are classified as small. The remainder are classified into two further divisions called standard and large.
   (i) If an egg is picked at random from the batch, find the probability that it is small.
   (ii) A sample of 50 eggs are selected at random.
        Find the probability that the mean of the sample is less than 58 g.
   (iii) It is desired that the standard and large classes should have about the same number of eggs in each. Estimate the mass at which the division should be made. Give your answer to the nearest gram.
Chapter 5: Inferential statistics

Exercise 5.1

1. (i) Sampling distribution
(ii) decrease
(iii) \( \mu \)
(iv) \( \frac{\sigma}{\sqrt{n}} \)
2. Represents the distribution of the sample means
3. (i) Normal distribution; The Central Limit Theorem
(ii) Sample size is sufficiently large (i.e. >30)
(iii) Mean = 12;
   Standard deviation = \( \frac{2}{\sqrt{36}} = \frac{1}{3} \)
4. (i) (4, 6), (4, 8), (4, 10), (6, 8), (6, 10), (8, 10)
(ii) Means are: 5, 6, 7, 7, 8, 9
(iii) Statistic
(iv) Both = 7
5. A parameter is a numerical property of a population.
   A statistic is a numerical property of a sample.
6. (i) Positively
(ii) Normal distribution
(iii) Sample size is sufficiently large (\( n > 30 \))
   to apply the theorem.
7. 0.0228
8. 0.0262
9. (i) 0.209 (ii) 0.1635
10. (i) 0.3 (ii) (a) 0.0228 (b) 0.4706
11. 0.894
12. \( P(\bar{x} > 8 \text{ years } 1 \text{ month}) = 0.0228; \)
   Number of samples = 1
13. 0.3206
14. C = 80, D = 96, E = 78\( \frac{2}{3} \)
15. \( n = 26 \)
16. (i) 0.0793 (ii) \( n > 170 \)
17. (i) 0.0228 (ii) 0.822 (iii) 0.086; Answer (iii)

Exercise 5.2

1. \( 62.17 < \mu < 63.83 \)
2. \( 279.1 < \mu < 288.9 \)
3. (i) \( 225.2 < \mu < 228.8 \)
   (ii) 5%
4. \( 60.9 < \mu < 64.5 \)
5. (i) \( 5.097 < \mu < 5.143 \)
   (ii) 5.10 and 5.14
   (iii) The “95% confidence” means that the
   mean lies in the range 5.10 to 5.14
   95 times out of 100.
6. €269.71 < \( \mu < €290.29 \)
7. (i) 28.99 cm < \( \mu < 29.41 \) cm
   (ii) No, as the sample size (180) is sufficiently
   large to apply The Central Limit Theorem.
8. (i) 0.0125
(ii) 0.932 g
(iii) 0.9075
(iv) The confidence interval would be
   \[ 0.9124, 0.9516 \]
(v) A larger sample results in a smaller
   confidence interval.
9. (i) 4.28
(ii) 601 cars
10. (i) 747.42
(ii) \( n = 23 \)
11. (i) 68.12
(ii) \( n = 28 \)
12. (i) 0.629
   (ii) 46.2 g < \( \mu < 51.0 \) g
   (iii) At least 70
13. (i) \( \bar{x} = 57.4 \)
   (ii) \( \sigma = 15.1 \)

Exercise 5.3

1. 0.1096 < \( p < 0.1904 \)
2. 0.293 < \( p < 0.427 \)
3. 0.260 < \( p < 0.379 \)
4. 0.122 < \( p < 0.358 \)
5. 0.294 < \( p < 0.386 \)
6. (i) 0.654 < \( p < 0.812 \)
   (ii) 0.765 < \( p < 0.917 \)
7. (i) 34%
   (ii) 29.4% < \( p < 38.6\% \); On 95% of samples
   the true proportion will lie in this interval.
   (iii) 2156
8. 0.245 < \( p < 0.295 \)
9. (i) 0.08 < \( p < 0.22 \)
   (ii) 2180
**Exercise 5.4**

1. No; $z = 1.68$; so not greater than 1.96

2. Yes; $z = -2.66$

3. (i) $H_0$: The mean age of the patients is 45 years.
   $H_1$: The mean age of the patients is not 45 years.
   (iii) $z = 1.89$
   (iv) No; $z = 1.89$; so not greater than 1.96

4. (i) $N_0$: The mean length is 210 cm
   $N_1$: The mean length is not 210 cm
   (iii) Yes: $z = 2.5$
   As $z = 2.5$ is in the critical region, we reject the null hypothesis.

5. No: $z = 1.955$; so $z$ not greater than 1.96

6. Yes: $z = -2.11$, which is $<-1.96$

7. Yes: $z = -2.54$, which is less than $-1.96$

8. (i) 0.0836
   (ii) 0.0562
   (iii) 0.099
   (iv) 0.0394

9. (i) $z = 1.77$
   (ii) $p$-value = 0.0768
   (iii) No, as $p$-value $< 0.05$

10. (i) $z = -1.5$
    (ii) $p$-value = 0.1336
    (iii) No, as $p$-value $> 0.05$

11. (i) $z = 1.5$
    (ii) No, as $z > 1.96$
    (iii) $p$-value = 0.1336
    (iv) Yes, as $p$-value is not less than 0.05; so we accept $H_0$ that the mean time required is 12 mins.

12. (i) $z = 2.5$
    (ii) 0.0124
    (iii) Yes, as $p$-value $< 0.05$

13. Standard error = 0.0036; $4.993 < \mu < 5.007$; Yes as $z = 2.22$ and so $z > 1.96$.

**Test yourself 5**

**A – questions**

1. 0.89

2. Mean $\mu = 2.85$;
   Standard error $\frac{0.07}{\sqrt{20}} = 0.016$

3. $24.42 < \mu < 27.98$

4. $260.0 \text{ ml} < \mu < 272.2 \text{ ml}$

5. $0.522 < p < 0.678$

6. $0.45 < p < 0.65$

7. Yes; $z = 2.113$ and $z > 1.96$

8. (i) The mean 460.3 g has not changed.
   (ii) $z = 2.18$
   (iii) Since $z = 2.18$ and $1.96 \Rightarrow$ the new mean is different from the known mean.

9. (i) Normal distribution; Central Limit Theorem
   (ii) Sample size is large (i.e. $> 30$)
   (iii) 39 or 40 samples

10. (i) 0.0026
    (ii) $z = -1.96$ and $z > 1.96$
    (iv) $p$-value $= 0.1096$

**B – questions**

1. $n = 10$

2. (i) 0.401
    (ii) 0.691
    (iii) 0.494; unaffected as $n > 30$

3. $30.6 \text{ kg} < \mu < 32.2 \text{ kg}$

4. (i) Since $n \geq 30$, the Central Limit Theorem can be applied.
    (ii) 0.057 < $p < 0.343$

5. (i) If 100 samples of the same size are taken, then the true population mean or proportion will lie in the given interval on 95 occasions out of 100.
    (ii) $3.16 \text{ km} < \mu < 13.88$

6. (i) $N_0$: $\mu = 1.2 \text{ sec}; N_1$: $\mu \neq 1.2 \text{ sec}$
    (ii) Critical regions are: $z < -1.96$ and $z > 1.96$
    (iii) $z = -3$. Yes as $z < -1.96$ and so is in the critical region.
    (iv) $p$-value = 0.0026
    Since $p < 0.05$, we reject the null hypothesis that $\mu = 1.2 \text{ sec}$. 
5. (a) Mean = \( \mu \); Standard deviation = \( \frac{\sigma}{\sqrt{n}} \)
   (i) Normal distribution.
   (ii) Normal distribution.
   The Central Limit Theorem can be applied to any distribution if \( n \) is large (\( n > 30 \)).
   The distribution of the sample means will always be normal when the underlying population is normal.
   (b) (i) 0.278 (ii) 694

6. (i) \( \frac{3}{4} \)
   (ii) 0.2966
   (iii) Normal distribution; \( \mu = 22.05 \); \( \sigma = 0.7525 \)
   (iv) 104

7. (i) 0.773
   (ii) 0.631
   (iii) 2.89 kg
   (iv) 0.936

8. (a) (i) 0.0228
   (ii) 0.440
   (iii) 0.785
   (b) (i) Mean only 1.08 standard deviations above zero.
   For a normal distribution this gives a probability of about 0.16 of negative times, which are impossible.
   (ii) Large sample \( \Rightarrow \) mean approximately normally distributed.

9. (i) 68 kg; 0.6 kg
   (ii) 17 samples
Leaving Certificate Higher Level Maths
Contains all the Deferred Material and Central Limit Theorem for examination in 2015 and onwards

Statistics: Chapter 5 – Inferential Statistics