MEI
Mathematics in Education and Industry

MEI Structured Mathematics

Module Summary Sheets

Statistics 3

(Version B: reference to new book)

Topic 1: Continuous random variables

Topic 2: Expectation algebra

Topic 3: Sampling

Topic 4: Interpreting sample data using the Normal distribution

Topic 5: Interpreting sample data using the *t* distribution

Topic 6: Non-parametric tests of location

Topic 7: The chi squared distribution

MEI, Oak House, 9 Epsom Centre, White Horse Business Park, Trowbridge, Wiltshire. BA14 0XG.
Company No. 3265490 England and Wales Registered with the Charity Commission, number 1058911
Tel: 01225 776776. Fax: 01225 775755

Summary S3 Topic 1: Continuous random variables - 1

References:
Chapter 1
Pages 1-8

Continuous random variables are used to model continuous data.

Example 1.2
Page 6

Probability density function

The p.d.f., f(x), of the continuous random variable, X, has the following properties:

(i) f(x) is always ≥ 0

(ii) the total area under f(x) is 1; i.e. $\int_{\text{all } x} \mathrm{f}(x)\,\mathrm{d}x = 1$

(iii) $\mathrm{P}(a \leq x \leq b) = \int_a^b \mathrm{f}(x)\,\mathrm{d}x$

The probability that X takes any specific value is zero. i.e. $\mathrm{P}(X = x) = 0$

Exercise 1A
Q. 4, 8

References:
Chapter 1
Pages 13-15

Example 1.5
Page 15

Expectation and variance

The **mean**, or **expected value**, or **expectation** , of X is given by

$$\mu = \mathrm{E}(X) = \int_{\text{all } x} x\mathrm{f}(x)\,\mathrm{d}x$$

The variance is given by

$$\mathrm{Var}(X) = \int_{\text{all } x} (x-\mu)^2\,\mathrm{f}(x)\,\mathrm{d}x = \int_{\text{all } x} x^2\mathrm{f}(x)\,\mathrm{d}x - \mu^2$$

These expressions are similar to those for discrete random variables (See Statistics 1).

Exercise 1B
Q. 7

References:
Chapter 1
Pages 16-18

Example 1.6
Page 17

Median and Mode

For continuous random variables, the **median** is the value that separates the area under the graph of the p.d.f. f(x) into two equal parts, each of area 0.5. The **mode** is the value at which the maximum value of f(x) occurs.

The **median** is the solution of $\int_{-\infty}^{m} \mathrm{f}(x)\,\mathrm{d}x = \frac{1}{2}$

or the solution of $\int_{m}^{\infty} \mathrm{f}(x)\,\mathrm{d}x = \frac{1}{2}$.

Exercise 1B
Q. 2, 5

The **mode** is the value of x which maximises f(x). Depending on the shape of the graph of the function f(x), this may or may not be a point at which f '(x) = 0.

E.g. 1. A continuous random variable has the p.d.f. given by $\mathrm{f}(x) = kx(5-x)$ for $0 \leq x \leq 5$. (N.B. $\mathrm{f}(x) \geq 0$)
Find (i) the value of the constant k,
(ii) $\mathrm{P}(1 \leq X \leq 2)$.

(i) $\int_0^5 kx(5-x)\,\mathrm{d}x = k\int_0^5 (5x-x^2)\,\mathrm{d}x = k\left[\frac{5x^2}{2} - \frac{x^3}{3}\right]_0^5$

$$= k\left(\frac{5^3}{2} - \frac{5^3}{3}\right) = \frac{125k}{6} = 1$$

$$\Rightarrow k = \frac{6}{125}$$

(ii) $\mathrm{P}(1 \leq X \leq 2) = k\int_1^2 (5x-x^2)\,\mathrm{d}x = k\left[\frac{5x^2}{2} - \frac{x^3}{3}\right]_1^2$

$$= \frac{6}{125}\left(\left(10 - \frac{8}{3}\right) - \left(\frac{5}{2} - \frac{1}{3}\right)\right) = \frac{6}{125}\left(\frac{22}{3} - \frac{13}{6}\right)$$

$$= \frac{6}{125}.\frac{31}{6} = \frac{31}{125}$$

E.g. For the example above, find the mean and variance.

$$\mathrm{E}(X) = \int_0^5 kx^2(5-x)\,\mathrm{d}x = k\int_0^5 (5x^2 - x^3)\,\mathrm{d}x = k\left[\frac{5x^3}{3} - \frac{x^4}{4}\right]_0^5$$

$$= k\left(\frac{5^4}{3} - \frac{5^4}{4}\right) = \frac{625k}{12} = \frac{5}{2}$$

$$\mathrm{Var}(X) = \int_0^5 kx^3(5-x)\,\mathrm{d}x - \mu^2 = k\int_0^5 (5x^3 - x^4)\,\mathrm{d}x - \mu^2$$

$$= k\left[\frac{5x^4}{4} - \frac{x^5}{5}\right]_0^5 - \left(\frac{5}{2}\right)^2 = k\left(\frac{5^5}{4} - \frac{5^5}{5}\right) - \frac{25}{4}$$

$$= \frac{5^5k}{20} - \frac{25}{4} = \frac{25\times 6}{20} - \frac{125}{20} = \frac{25}{20} = \frac{5}{4}$$

E.g. For the example above, the p.d.f. is a parabola.
Therefore the median is the midpoint, because of symmetry, and the mode is the maximum value which is also the mid-point.
So in this case median = mode = 2.5

E.g. Show that $\mathrm{f}(x) = \frac{1}{8}x$ for $0 \leq x \leq 4$, f(x) = 0 elsewhere, satisfies the conditions for a p.d.f. and find the median and mode. (N.B. $\mathrm{f}(x) \geq 0$)

$$\int_0^4 \frac{1}{8}x\,\mathrm{d}x = \left[\frac{x^2}{16}\right]_0^4 = 1$$

Median is m where $\int_0^m \frac{1}{8}x\,\mathrm{d}x = 0.5 \Rightarrow \left[\frac{x^2}{16}\right]_0^m = 0.5$

$$\Rightarrow \frac{m^2}{16} = 0.5 \Rightarrow m^2 = 8 \Rightarrow m \approx 2.83 (3 \text{ s.f.})$$

There is no local maximum for this p.d.f. and so the mode is the largest value of x, at the upper limit, x = 4.

Summary S3 Topic 1: **Continuous random variables - 2**

References:
Chapter 1
Pages 18-21

Example 1.7
Page 19

Exercise 1B
Q. 8

References:
Chapter 1
Pages 27-30

Example 1.8
Page 27

Exercise 1C
Q. 2, 4

References:
Chapter 1
Pages 34-42

Example 1.10
Page 39

Exercise 1D
Q. 2, 11

The Rectangular distribution is also called the continuous uniform distribution.

If $X \sim U(a,b)$ then $\mathrm{f}(x) = \dfrac{1}{b-a}$ where $a < x < b$

and $\mathrm{f}(x) = 0$ for all other values of x.

$\mathrm{E}(X) = \dfrac{a+b}{2}$ and $\mathrm{Var}(X) = \dfrac{(b-a)^2}{12}$

$\mathrm{P}(c < X < d) = \dfrac{d-c}{(b-a)}$ provided $a < c < d < b$

Functions of *X*

If the random variable X has p.d.f. $\mathrm{f}(x)$ and $\mathrm{g}[X]$ is a function of X then

$$\mathrm{E}(\mathrm{g}[X]) = \int_{\text{all } x} \mathrm{g}[X]\,\mathrm{f}(x)\,\mathrm{d}x$$

$$\mathrm{Var}(\mathrm{g}[X]) = \int_{\text{all } x} (\mathrm{g}[x])^2\,\mathrm{f}(x)\,\mathrm{d}x - \{\mathrm{E}(\mathrm{g}[X])\}^2$$

General results

$\mathrm{E}(c) = c, \qquad \mathrm{Var}[c] = 0$

$\mathrm{E}(aX) = a\mathrm{E}(X) \qquad \mathrm{Var}(aX) = a^2\mathrm{Var}(X)$

$\mathrm{E}(aX+b) = a\mathrm{E}(X)+b \qquad \mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$

$\mathrm{E}[\mathrm{g}(X) + \mathrm{h}(X)] = \mathrm{E}[\mathrm{g}(X)] + \mathrm{E}[\mathrm{h}(X)]$

The cumulative distribution function (c.d.f.) of the random variable X is defined by

$$\mathrm{F}(x) = \mathrm{P}(X \le x) = \int_{-\infty}^{x} \mathrm{f}(t)\,\mathrm{d}t$$

$\mathrm{F}(x)$ increases from 0 to 1 as x ranges from $-\infty$ to $+\infty$.

Conversely, $\mathrm{f}(x) = \dfrac{\mathrm{d}}{\mathrm{d}x}\mathrm{F}(x) = \mathrm{F}'(x)$.

The median of X, m, is a root of $\mathrm{F}(m) = \dfrac{1}{2}$.

Statistics 3
Version B: page 3
Competence statements R1, 2, 3, 4, 5
© MEI

E.g. 1. A Uniform distribution over $[0,3]$ has

$$\mathrm{f}(x) = \frac{1}{3}$$

$\mathrm{E}(X) = \dfrac{0+3}{2} = 1.5, \quad \mathrm{E}(X) = \dfrac{(3-0)^2}{12} = 0.75$

$\mathrm{P}(1 < X < 2) = \dfrac{2-1}{3-0} = \dfrac{1}{3}$

E.g. For the example above, find
(i) E(2*X* + 5), (ii) Var(2*X* + 5)
(iii) E(2*X* – 5), (iv) Var (2*X* – 5)

(i) $\mathrm{E}(2X + 5) = 2\mathrm{E}(X) + 5 = 2 \times 1.5 + 5 = 8$
(ii) $\mathrm{Var}(2X + 5) = 4\mathrm{Var}(X) = 4 \times 0.75 = 3$
(iii) $\mathrm{E}(2X - 5) = 2\mathrm{E}(X) - 5 = 2 \times 1.5 - 5 = -2$
(iv) $\mathrm{Var}(2X - 5) = 4\mathrm{Var}(X) = 4 \times 0.75 = 3$

E.g. The p.d.f. of a functon is given by

$\mathrm{f}(x) = kx$ for $0 \le x \le 2$, $\mathrm{f}(x) = 0$ elsewhere.

Find
(i) the value of k,
(ii) $\mathrm{E}(X)$,
(iii) $\mathrm{Var}(X)$,
(iv) the c.d.f., $\mathrm{F}(x)$,
(v) the median.

(i) $\displaystyle\int_0^2 kx\,\mathrm{d}x = \left[\frac{kx^2}{2}\right]_0^2 = 2k = 1 \Rightarrow k = \frac{1}{2}$

(ii) $\displaystyle\mathrm{E}(X) = \int_0^2 x.kx\,\mathrm{d}x = \left[\frac{kx^3}{3}\right]_0^2 = \frac{8k}{3} = \frac{4}{3}$

(iii) $\displaystyle\mathrm{Var}(X) = \int_0^2 x^2.kx\,\mathrm{d}x - \mathrm{E}(X)^2 = \left[\frac{kx^4}{4}\right]_0^2 - \frac{16}{9}$

$$= 2 - \frac{16}{9} = \frac{2}{9}$$

(iv) the c.d.f., $\displaystyle\mathrm{F}(x) = \int_0^x kt\,\mathrm{d}t = \left[\frac{kt^2}{2}\right]_0^x = \frac{1}{4}x^2$

(v) the median, m, is a root of the equation $\mathrm{F}(m) = \dfrac{1}{2}$.

$\Rightarrow \dfrac{1}{4}m^2 = \dfrac{1}{2} \Rightarrow m^2 = 2 \Rightarrow m = \sqrt{2} \approx 1.41$

(given that $m \ge 0$)

Alternatively, m satisfies $\displaystyle\int_0^m kx\,\mathrm{d}x = \frac{1}{2}$

$$\Rightarrow \left[\frac{x^2}{4}\right]_0^m = \frac{1}{2} \Rightarrow m^2 = 2$$

Summary S3 Topic 2: **Expectation algebra**

References:
Chapter 2
Pages 53-54

Example 2.1
Page 54

Expectation of a function of *X*, E(g[*X*])
g[*X*] is a function of the random variable *X* .
If *X* is discrete then E(g[*X*]) is given by

$$\mathrm{E}(\mathrm{g}[X]) = \sum \mathrm{g}[x_i] \times \mathrm{P}(X = x_i)$$

If *X* is continuous then E[g(*X*)] is given by

$$\mathrm{E}(\mathrm{g}[X]) = \int_{-\infty}^{\infty} \mathrm{g}[x]\mathrm{f}(x)\ \mathrm{d}x$$

References:
Chapter 2
Pages 55-61

Exercise 2A
Q. 1, 7

Exercise 2B
Q. 1

Algebraic results

$\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$

$\mathrm{E}(\mathrm{f}[X] + \mathrm{g}[X]) = \mathrm{E}(\mathrm{f}[X]) + \mathrm{E}(\mathrm{g}[X])$

If, in addition, the random variables *X* and *Y* are independent then
$\mathrm{Var}(aX \pm bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$

In particular, if the random variables X_1, X_2 are independent,

$$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$$
$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

$$\mathrm{E}(X_1 - X_2) = \mathrm{E}(X_1) - \mathrm{E}(X_2)$$
$$\mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

References:
Chapter 2
Pages 61-63
Pages 65-68

Example 2.6
Page 62

In general, if the random variables X_1, X_2, … are independent,
$\mathrm{E}(a_1X_1 \pm a_2X_2 \pm \ldots\ldots \pm a_nX_n)$
$= a_1\mathrm{E}(X_1) \pm a_2\mathrm{E}(X_2) \pm \ldots\ldots \pm a_nE(X_n)$
$\mathrm{Var}(a_1X_1 \pm a_2X_2 \pm \ldots\ldots \pm a_nX_n)$
$= a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + \ldots\ldots + a_n^2Var(X_n)$

Exercise 2B
Q. 2, 8

Exercise 2C
Q. 3, 5

Sums and differences of Normal distributions
If the independent random variables X_1 and X_2 are Normally distributed, then the distributions of $aX_1 + bX_2$ and $aX_1 - bX_2$ are also Normally distributed.

References:
Chapter 2
Pages 75, 76

Example 2.11
Page 76

Distribution of sample mean
If the population mean is μ and the variance is σ^2 and a random sample of *n* items, $x_1, x_2, \ldots, x_n$, from the population is taken, then

$$\bar{x} = \frac{x_1 + x_2 + \ldots.. + x_n}{n}$$

$\bar{x}$ is a value of the random variable $\bar{X} = \dfrac{X_1 + X_2 + \ldots.. + X_n}{n}$

$$\Rightarrow \mathrm{E}(\bar{X}) = \frac{1}{n}\mathrm{E}(X_1) + \frac{1}{n}\mathrm{E}(X_2) + \ldots\ldots + \frac{1}{n}\mathrm{E}(X_n)$$
$$= \frac{1}{n}.\mu + \frac{1}{n}.\mu + \ldots. + \frac{1}{n}.\mu = \mu$$

Given also that $X_1, X_2, \ldots. X_n$ are independent,

$$\mathrm{Var}(\bar{X}) = \left(\frac{1}{n}\right)^2 \mathrm{Var}\left(X_1 + X_2 + \ldots. + X_n\right)$$
$$= \frac{1}{n^2}\left(\mathrm{Var}\left(X_1\right) + \mathrm{Var}\left(X_2\right) + \ldots + \mathrm{Var}\left(X_n\right)\right)$$
$$= \frac{1}{n^2}.\left(\sigma^2 + \sigma^2 + \ldots + \sigma^2\right) = \frac{1}{n^2} \times n\sigma^2 = \frac{1}{n}\sigma^2$$

E.g. The random variable *X* has a probability distribution as shown.

x	1	3	5
P($X = x$)	0.5	0.3	0.2

Find E(2*X* + 3) by two methods.

Method 1.
By the work of Statistics 1;
E(2*X* + 3) = 2E(*X*) + 3
And E(*X*) = 1 × 0.5 + 3 × 0.3 + 5 × 0.2 = 2.4
⇒ E(2*X* + 3) = 2× 2.4 + 3 = 7.8

Method 2.
g[*X*] = 2*X* + 3 gives the table as shown.

x	1	3	5
g[x]	5	9	13
P($X = x$)	0.5	0.3	0.2

(When x = 1, g(x) = 2 ×1 + 3 = 5, etc.)
⇒ E(2*X* + 3) = 5 × 0.5 + 9 × 0.3 + 13 × 0.2
= 7.8

E.g The random variable *X* is Normally distributed.
Then E(2*X*) = 2E(*X*), Var(2*X*) = 4Var(*X*)

The independent random variables X_1 and X_2 are drawn from the same random variable *X*.
Then $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 2\mathrm{E}(X)$
and
$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\mathrm{Var}(X)$

E.g. The distribution of masses in kilograms of male students at a college is N(75, 5) and of females is N(70,8).
A male and a female enter a lift together. Assuming the two students to be a random choice from the populations, find the probability that
(i) the mass of the two students is less than 155 kg,
(ii) the female is heavier than the male.

(i) $X_m \sim \mathrm{N}(75{,}5),\ X_f \sim \mathrm{N}(70{,}8)$
$\Rightarrow X_m + X_f \sim \mathrm{N}(145{,}13)$
$\Rightarrow \mathrm{P}(X_m + X_f < 155) = \Phi\left(\dfrac{155 - 145}{\sqrt{13}}\right)$
$= \Phi\left(2.7735\right) = 0.9972$
(ii) $X_m - X_f \sim \mathrm{N}(5{,}13)$
$\Rightarrow \mathrm{P}(X_m - X_f < 0) = \Phi\left(\dfrac{-5}{\sqrt{13}}\right)$
$= 1 - \Phi\left(1.3868\right) = 1 - 0.9172 = 0.0828$

Statistics 3
Version B: page 4
Competence statements a1, 2, 3, 4

Summary S3 Topic 3: Sampling

References:
Chapter 3
Pages 80,81

Terms
The *population* or *parent population* is the set of all data in which you are interested.
A *sample* is a set of data drawn from the parent population in order to investigate the population.
Usually the population is too large (or even infinite) or too difficult (or costly) to collect or to investigate and so we investigate using a smaller set which needs to be as representative as possible.
A representation of the items available to be sampled is called the *sampling frame*.
The proportion of available items that are actually sampled is called the *sampling fraction.*

References:
Chapter 3
Page 81

Notation
A statistical value that describes the set (for instance mean or variance) is called a **(sample) statistic** if it is found from the sample and a **(population) parameter** if it is found from the population.
For a population, greek letters are used.

e.g. mean = μ, variance = σ^2

For a sample, roman letters are used.

e.g. mean = $\bar{x}$, variance = s^2

When the parameters of a sample are used for the population then they are called *estimates*. Upper case letters, *X*, *Y*, etc are used for random variables and lower case letters, *x*, *y*, etc for particular values.

References:
Chapter 3
Pages 81-84

Sampling
The reason for taking a sample is:
(i) In order to estimate the values of the parameters of the population when to find the actual values is too costly, time consuming or impossible.
(ii) To conduct a hypothesis test.

Questions to ask.
Are the data relevant?
Are the data biased?
Does the method of collecting the data create bias or distortion?
Is the sample large enough?
Is the sampling procedure appropriate?

References:
Chapter 3
Pages 84-86

Exercise 3A
Q. 2, 7

Sampling techniques
Simple random sampling. Every possible sample of given size is equally likely to be selected.
Stratified sampling. When the population is split into distinct strata, then separate samples (often simple random, and sometimes in proportion) are chosen from each strata.
Cluster sampling. The population is split into groups or clusters. A simple random sample (often only of size one) of clusters is chosen. The chosen clusters are then surveyed - often by studying all the elements in them, or by taking random samples within them.
Systematic sampling. This is a method of choosing a sample from a sampling frame by a systematic method - e.g. taking every 3rd member of a list.
Quota sampling. A sample is taken that fulfils some criteria of quotas of certain sub-groups within the population.

E.g. If a sample of size 10 is chosen from a population of size 200 then the sampling fraction is

$$\frac{10}{200} = \frac{1}{20}$$

To have the same sampling fraction for a population of size 5000 requires a sample of size

$$5000 \times \frac{1}{20} = 250$$

Simple random sampling
E.g. The choice of a simple random sample may be made using random numbers from the Students Handbook (page 52).
For numbers between 1 and 100 take numbers in pairs (taking 00 to be 100) or add 1 to each number generated.
For numbers between 1 and 800 take digits in threes and reject any number obtained that is greater than 800; also reject 000.

Random numbers can also be generated from most calculators. These are generally a three digit decimal number between 0 and 1. If you want numbers between 1 and 800, multiply by 800, ignore the decimal part and add 1 to each.

Stratified Sampling
E.g. Split a student population into male and female.
E.g. Split a work force into the “managers” and “machine operators”.

Cluster Sampling
E.g. To investigate all students studying Mathematics in Colleges of Further Education, take as a sample all students (or a simple random sample of students) in a random selection of colleges. The colleges are the clusters.

Systematic Sampling
E.g. Select a sample of size 10 from 200.
$^{200}/_{10}$ = 20. Take a number randomly in the range 1 – 20 then add 20 to each number.
E.g. 13, so take the 13th, 33rd, 53rd, etc.

Quota Sampling
E.g. A poll is to be carried out on the population of the UK on drinking habits.
A number of people are asked to conduct the poll. Each one has to interview 20 females aged under 20, 20 who are in the age bracket 20 – 40 and 20 aged over 40. They have to make the same selection for men.

Statistics 3
Version B: page 5
Competence statements I1, 2
© MEI

Summary S3 Topic 4: ***Interpreting sample data using the Normal distribution***

References:
Chapter 4
Pages 91-92

The Central Limit Theorem
If $X_1, X_2, \ldots X_n$ are independent and identically distributed and come from a (usually unknown) distribution with mean μ and variance σ^2 then approximately:

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ and } T = X_1 + X_2 + \ldots + X_n \sim \mathrm{N}(n\mu, n\sigma^2)$$

The approximation to a Normal distribution improves as the sample size, n, increases.
Usually a sample size of at least 25 is adequate.
Such samples are called "large".

$\frac{\sigma}{\sqrt{n}}$ is the standard deviation of $\bar{X}$. It is sometimes called the **standard error of the mean.**

E.g. Assume that the marks of all candidates in Statistics 3 in a particular college are Normally distributed with variance 100.
If samples of size 20 are taken then the Standard Error of the mean is

$$\frac{10}{\sqrt{20}} = 2.236$$

If samples of size 200 are taken then the Standard Error of the mean is

$$\frac{10}{\sqrt{200}} = 0.7071$$

References:
Chapter 4
Pages 93-98

Example 4.1
Page 98

Exercise 4A
Q. 1, 3

Confidence intervals
A **95% confidence interval** for the mean of the population is an interval calculated from a random sample. A different sample would give a different interval. If lots of samples were taken from the population, 95% of the intervals calculated would
contain the fixed, but unknown, mean of the population.
If we have a sample of size n from a Normal distribution whose variance is known, the symmetrical 100(1-α)% confidence limits for the mean of the population are given by

$$\bar{x} \pm k\frac{\sigma}{\sqrt{n}} \text{ where } \mathrm{P}(-k < Z < k) = 1 - \alpha$$

Values of k will depend on the confidence level

Confidence level	k
90%	1.645
95%	1.96
99%	2.576

The confidence interval is from $\bar{x} - k\frac{\sigma}{\sqrt{n}}$ to $\bar{x} + k\frac{\sigma}{\sqrt{n}}$

E.g. In the college above, the mean mark of one sample of 200 candidates is 58, then the 95% confidence limits are

$$58 \pm \frac{1.96 \times 10}{\sqrt{200}}$$

i.e. 56.6 and 59.4

E.g. If $\bar{X} \sim \mathrm{N}(\mu, 25)$ and a random sample of size 16 from it has mean 40, then the 99% confidence interval for μ is

$$40 \pm 2.576 \times \frac{\sqrt{25}}{\sqrt{16}} = 40 \pm 3.22$$

i.e. 36.88 to 43.22.

E.g. If $\bar{X} \sim \mathrm{N}(\mu, 25)$ and a random sample of size 100 from it has mean 40, then the 99% confidence interval for μ is

$$40 \pm 2.576 \times \frac{\sqrt{25}}{\sqrt{100}} = 40 \pm 1.29$$

i.e. 38.71 to 41.29.

References:
Chapter 4
Page 95

Known and estimated standard deviation
In the work above the value of the standard deviation of the population is needed. If the standard deviation is not known then it has to be estimated from the sample. In this case it is acceptable to use the s.d. from the sample as an estimate and assume that the sampling distribution of means is approximately Normal provided that the sample size is at sufficiently large (a sample size of about 50 is often adequate.)

References:
Chapter 4
Page 98

Exercise 4A
Q. 4

Size of sample.
In all investigations where a sample is required, the larger the sample the more confidence you have in the results, but this needs to be balanced against the cost and the time of collecting .
The width of the confidence interval is inversely proportional to the square root of n, so we need to quadruple the size of the sample in order to halve the width.

E.g. In the college above, the investigator would like to be able to estimate the mean mark within 2 marks with 99% confidence. What size sample should she take?

The 99% confidence interval for the mean is given by $\bar{x} - 2.58\frac{\sigma}{\sqrt{n}}$ to $\bar{x} + 2.58\frac{\sigma}{\sqrt{n}}$,

where $\sigma = 10$.

$$\Rightarrow \frac{25.8}{\sqrt{n}} \le 2$$

$$\Rightarrow n \ge \left(\frac{25.8}{2}\right)^2 \Rightarrow n \ge 166.4$$

$\Rightarrow$ at least 167

Statistics 3
Version B: page 6
Competence statements I3, 4, 5, 6, 7

Summary S3 Topic 5: ***Interpreting sample data using the t distribution - 1***

References:
Chapter 5
Page 106-107

Interpreting the sample data
For situations where the population mean, μ, and variance, σ^2 are both unknown, sample data may be interpreted using the *t* distribution provided the distribution from which they are drawn is Normal.
The *t* distribution is only needed if the sample size is small; for large samples the Normal distribution may be used because of the Central Limit Theorem.

References:
Chapter 5
Pages 107-108

Degrees of freedom
The number of free variables in the situation is called the degrees of freedom.
As the sample size increases, the estimate s will tend to approach the true value of σ and the relevant *t* distribution approaches the standard Normal distribution. The *t* distribution which we use is specified by its "degrees of freedom", v. In this situation, $v = n - 1$.

E.g. A machine packs bags of sugar of nominal mass 1000g. A sample of size 50 has mean mass 999.3g and s.d. 2.3g. Is there evidence at the 5% level that bags are being underfilled?
H_0: $\mu = 1000$
H_1: $\mu < 1000$
The value of the test statistic is

$$z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{999.3 - 1000}{2.3/\sqrt{50}} = -2.152$$

Because the sample size is large enough, we may use Normal Tables which gives the 5% critical value to be -1.645. We therefore reject H_0 - there is evidence to suggest that the bags are being underfilled.

References:
Chapter 5
Pages 108-110

Confidence intervals
Just as in the case above, if we have a random sample of size n from a Normal population where we do not know the value of σ, the symmetric $100(1-\alpha)$ % confidence limits for the mean of the population are given by

$$\bar{x} \pm \frac{ks}{\sqrt{n}}$$

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$ is an "unbiased" estimate of the (unknown) population variance from the sample data.

If you were to take lots of samples and calculate s^2 for each sample the values would centre on the correct value, σ^2.

k is found from t tables with $n - 1$ degrees of freedom.
Then $P(-k < t_{n-1} < k) = 1 - \alpha$
If the sample size is large the central limit theorem means that the confidence limits are

$$\bar{x} \pm \frac{ks}{\sqrt{n}}$$

where k is found from the Normal table and s^2 is used in place of the population variance if the latter is not known.

Exercise 5A
Q. 2, 4

E.g. 6 students are asked to estimate a time interval of 1 minute. Their estimates, in seconds, are:
55.2, 67.3, 53.9, 59.2, 65.0, 63.7
Assuming that these form a random sample from a Normal population, the 90% confidence limits for the mean are:

$$\bar{x} \pm \frac{ks}{\sqrt{n}} = 60.72 \pm 2.015 \times \frac{5.47}{\sqrt{6}} = 56.2 \text{ to } 65.2$$

where $\bar{x}$ and s (the sample s.d. with divisor n - 1) from a calculator are 60.72 and 5.47 and k from t_5 tables is 2.015.

References:
Chapter 5
Pages 110-112

Hypothesis test using the *t* distribution
As before, H_0 is $\mu = \mu_0$ and H_1 may be one tailed, $\mu < \mu_0$ or $\mu > \mu_0$, or two tailed: $\mu \neq \mu_0$.

The test statistic is $$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

with s as above and the t tables with $n - 1$ degrees of freedom are used. This t test may only be used for random samples from a Normal population.

Example 5.1
Page 111

Exercise 5A
Q. 3, 6

E.g. The mean mass of adult females of a certain species of bird is 16.2g.
An ornithologist finds four similar looking birds of masses 15.9, 17.8, 18.2, 19.7g. Is there evidence at the 10% level that they may be from a different species?
H_0: $\mu = 16.2$
H_1: $\mu \neq 16.2$
Assuming a random sample from a Normal population the value of the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{17.9 - 16.2}{1.564/\sqrt{4}} = 2.174$$

From t_3 tables, the 10% critical value is 2.353 so we accept H_0 at the 10% level.

Statistics 3
Version B: page 7
Competence statements I7, 8, 9, 10

Summary S3 Topic 5: ***Interpreting sample data using the t distribution - 2***

References:
Chapter 5
Page 121-124

The paired sample *t* test

The test in this case is on a hypothesis related to the difference in 2 characteristics of each member of a population. Each element therefore produces a pair of values.

If the random variable, D, is the difference (assumed to be Normal) and S^2 is the unbiased estimator of its variance based on a sample of size n then the test statistic, T, is given by

$$T = \frac{\bar{D}}{S/\sqrt{n}}$$

which has a t distribution with $n - 1$ degrees of freedom.

References:
Chapter 5
Page 126

Important assumptions

1. The sample is random.
2. The difference is Normally distributed.

References:
Chapter 5
Page 127

Example 5.2
Page 127

Testing for a non-zero value, *k*, of the difference of two means with paired samples

The test statistic in this case is

$$T = \frac{\bar{D} - k}{S/\sqrt{n}}$$

where H_0: $\mu_D = k$ H_1: $\mu_D \neq k$ (or $\mu_D < k$ or $\mu_D > k$)

Once again, we assume D to be Normally distributed - T then has a t distribution with $n - 1$ degrees of freedom.

References:
Chapter 5
Page 128

Exercise 5B
Q. 2, 5

Confidence intervals for the difference of two means from a paired sample

The symmetrical confidence interval for the mean difference between the characteristics is

$$\bar{d} \pm k \times \frac{s}{\sqrt{n}}$$

where the value of k is found from the t distribution tables (Students handbook, page 45) for the appropriate level of significance and the number of degrees of freedom.

E.g. 6 cars are used to assess the effect of a new fuel additive designed to reduce fuel consumption.

Under the same driving conditions, the total consumption figures, in litres per 100 km, are as follows.

(x = without additive, y = with additive)

Car	A	B	C	D	E	F
x	9.2	8.3	5.8	5.9	4.6	7.7
y	8.3	7.5	5.1	5.4	4.8	7.8
d	-0.9	-0.8	-0.7	-0.5	0.2	0.1

Is there evidence at the 5% level that the additive is effective?

$H_0: \mu_D = 0$ (no effect)

$H_1: \mu_D < 0$ (reduced fuel consumption)

For these data $\sum d = -2.6 \Rightarrow \bar{d} = -0.433$

$\sum d^2 = 2.24 \Rightarrow s = 0.472$

$$\Rightarrow \text{Test statistic}, t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{-0.433}{0.472/\sqrt{6}} = -2.247$$

For t_5, 5%, 1-tail test the critical value = -2.015. Since $-2.247 < -2.015$, we reject H_0 in favour of H_1; there is evidence to suggest that consumption is less with the additive.

E.g. For the same data, is there evidence at the 5% level of significance that the additive reduces the fuel consumption by at least 0.25 litres per 100 km?

$H_0: \mu_D = -0.25$

$H_1: \mu_D < -0.25$

$$\Rightarrow \text{Test statistic}, t = \frac{\bar{d} - (-0.25)}{s/\sqrt{n}} = -0.950$$

As before, the critical value is -2.015 and since $-0.950 > -2.015$ we accept H_0.

E.g. For the same data, the 90% confidence limits for the mean change in fuel consumption with the additive are

$$\bar{d} \pm 2.015 \times \frac{s}{\sqrt{n}} = -0.433 \pm 2.015 \times \frac{0.472}{\sqrt{6}}$$

i.e. -0.821 and -0.044

Summary S3 Topic 6: **Non-parametric tests of location**

References:
Chapter 6
Pages 135-136

The sign test
This is the simplest of tests. It seeks to discover a trend of those in a sample who agree or disagree or who say yes or no.
The hypothesis is usually that equal numbers fall into the two categories.
On that basis, the distribution will be B(n, 0.5) and a test can be carried out using the cumulative binomial tables (as in the module Statistics 1).

References:
Chapter 6
Pages 137-143

The Wilcoxon single sample test
The survey may be more complicated than simply agree or disagree. There may be categories such as “strongly agree” or a value may be given, say 1 - 10.
This Wilcoxon test takes the “difference” of the value of the response from the median value and then ranks their absolute values.
If the median is the “usual” response, then the distribution will be symmetric about this value. (So one above and one below will receive equal rankings.) Also the sum of rankings above the median will be equal to the sum of rankings below.
The assumptions are as follows:
The random variable is symmetrically distributed about the median, M.
The data form a random sample of size n. If any value is equal to the median, then remove it and reduce the value of n.
H_0: The population median equals M.
H_1: can be one tailed (the population median $> M$ or $< M$) or two tailed (the population median $\neq M$).
Calculation is as follows:
Calculate the absolute differences between each sample value and the hypothesised median, deleting those that are equal to this median.
Rank these values from 1 to n giving the lowest rank to the smallest absolute difference. If any are equal then give each the average of the ranking positions that they occupy together.
Calculate the sum W_+ of the ranks of the values above m, and the sum, W_- of the ranks of the sample values below M.
Check that $W_+ + W_-$ equals the sum of integers 1,…, n which is

$$\frac{1}{2}n(n+1)$$

For the two tailed hypothesis, the test statistic, W is the smaller of W_+ and W_- . For the one tailed test W is one or the other as appropriate.
Critical values are given in the table in the Students’ Handbook (Page 52) for the smallest value that one of the sums could be before rejecting the null hypothesis that there is no strong opinion either way.

Exercise 6A
Q. 1, 4

Example 6.1
Page 142

References:
Chapter 6
Pages 149-150

The Wilcoxon paired sample test.
This concerns the testing of two conditions, giving a pair of values for each member of a population.
Find the differences between the values in the two conditions.
Use the Wilcoxon single sample test with the hypothesis that these differences have the suggested median (often zero).
The assumption on the distribution is that the differences between the values of the random variables in the two conditions are symmetrically distributed about the differences in medians.

Example 6.3
Page 151

Exercise 6B
Q. 2, 4

E.g. 10 people taste butter and a butter substitute. They are asked to identify which is which. If 8 are correct, is there evidence at the 5% level that people can tell butter from the substitute?
H_0: There is no evidence that people can tell the difference
H_1: There is evidence that people can tell the difference.
From B(10, 0.5) tables the 1-tail 5% critical region is {9, 10}; since 8 is not in this region we can accept $H_{0;}$ there is no evidence at the 5% level that people can tell butter from the substitute.

E.g. A newspaper article claimed that teachers work an average of 38 hours a week. 10 teachers were asked how many hours they worked in one week and their responses were as follows.
34 36 38 39 42 45 47 52 53 55.
Test, at the 5% level of significance, the newspaper article’s claim, using the median as “average” against the alternative hypothesis that teachers work longer hours than 38 hours a week.
H_0: $M = 38$ H_1: $M > 38$
Values of M – 38:
–4 –2 1 4 7 9 14 15 17
Ranked absolute differences are:
3.5 2 1 3.5 5 6 7 8 9
$W_- = 5.5$ $W_+ = 39.5$

Check: $5.5 + 39.5 = \frac{1}{2}\times 9\times 10$

For $n = 9$, the 5% 1-tail critical value is 8; since 5.5 < 8 we accept H_1 - there is evidence that teachers work longer hours.

E.g. 7 people are asked to have 10 goes at writing the word “Scissors” with their non-dominant hand.

Person	A	B	C	D	E	F	G
Time for first try	3.2	3.7	4.2	5.2	3.8	3.4	4.0
Time for 10th try	2.6	3.4	4.3	4.7	3.2	3.6	3.2
Difference	0.6	0.3	-0.1	0.5	0.6	-0.2	0.8
Ranked absolute difference	5.5	3	1	4	5.5	2	7

Is there evidence, at the 5% level of significance, that they have improved with practice?
H_0: $M = 0$, H_1: $M > 0$ (i.e. 10th try is quicker than 1st.)
$W_- = 1 + 2 = 3$
For $n = 7$, the 5% 1-tail critical value is 3; we thus accept H_1 - there is evidence that they improve with practice.

Summary S3 Topic 7: The Chi squared distribution - 1

References:
Chapter 7
Pages 157-158

Example 7.2
Page 163

Exercise 7A
Q. 4

The Chi squared distribution

The Chi squared *(χ^2)* goodness of fit test may be used to see if a particular model (e.g. binomial, Poisson, Normal) is appropriate for the given data set. If any of the parameters of the model are not specified in the null hypothesis (e.g. the mean of the Poisson model) then they are estimated from the data.
The expected frequencies for each of the possible values are calculated using the model. In the case of models such as the Poisson model, where there are infinitely many possible values, the "tail" values will be put together in one group.

The test is approximate. This approximation is considered "good" if all of the expected frequencies are at least 5. If this is not the case, classes are pooled/ combined until this condition is satisfied.

The test statistic is $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

where f_0 is the observed frequency for a class and f_e is the expected frequency, as calculated from the model.

This has a chi squared distribution if the model is correct.

References:
Chapter 7
Page 158

Significance levels

The significance level of the test is determined by your original hypothesis and not by what you see in the data. Therefore the significance level of a test must be set at the beginning.

References:
Chapter 7
Page 159

Degrees of freedom

The degree of freedom is the number of cells minus one (for the totals to agree) minus one for each parameter of the model which was estimated from the data.

E.g. Poisson(λ): v = No. of cells – 1 if λ is given.
v = No. of cells – 2 if λ is estimated from the data.

References:
Chapter 7
Page 160

Alternative notation

Rather than f_0 and f_e we sometimes use E_i and O_i to represent the expected and observed frequency in the ith class respectively.

Then $X^2 = \sum \frac{(E_i - O_i)^2}{E_i}$

Example 1: Five identical dice are thrown 150 times and the number of sixes is recorded.

No. of sixes	0	1	2	3	4	5
Observed frequency	49	62	24	12	3	0

Test if the dice are unbiased.

H_0: The outcome can be modelled by the binomial distribution B(5,$^1/_6$)
H_1: The outcome cannot be modelled by the binomial distribution. B(5,$^1/_6$)
Expected frequencies are calculated from the binomial distribution:

E.g. for 2 sixes, expected number $= 150 \times \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 24.1$

Expected values = 60.3, 60.3, 24.1, 4.8, 0.5, 0.0

The last 3 values are pooled to give a value greater than 5.

Sixes	0	1	2	3+
f_o	49	62	24	15
f_e	60.3	60.3	24.1	5.3
$f_o - f_o$	-11.3	1.7	-0.1	9.7
$(f_o - f_e)^2$	127.69	2.89	0.01	94.09
$\frac{(f_o - f_e)^2}{f_e}$	2.11	0.05	0.00	17.75

⇒ Test statistic, $X^2 = 19.91$ $v = 4 - 1$ (same total) = 3

⇒ Reject H_0 at 5% level since critical value = 19.91 > 7.81

E.g. The number of cars passing an observer in 64 one-minute intervals is given in the following table.

No. of cars	0	1	2	3	4
Frequency	20	19	13	10	2

Test whether these data could have come from a Poisson Distribution.
Sample mean = 1.3
H_0: The distribution is Poisson with mean 1.3
H_1: The distribution is not Poisson with mean 1.3
The table (with the last two groups combined) is as follows:

No of cars	0	1	2	3+
f_o	20	19	13	12
f_e	17.44	22.67	14.74	9.16
$f_o - f_e$	2.56	-3.67	-1.74	1.85
$(f_o - f_e)^2$	6.55	13.47	3.03	8.12
$\frac{(f_o - f_e)^2}{f_e}$	0.375	0.594	0.205	0.888

Expected frequencies are calculated by $64 \times e^{-1.3}\frac{1.3^r}{r!}$

⇒ Test statistic, $X^2 = 2.06$, $v = 4 - 2$ (as λ is estimated from the sample)

⇒ Accept H_0 at 5% level since critical value = 2.06 < 5.99

Statistics 3
Version B: page 10
Competence statements I13

Summary S3 Topic 7: ***The Chi squared distribution - 2***

References:
Chapter 7
Pages 161-167

Example 7.1
Page 161

Example 7.2
Page 163

Using the test

1. Decide on the model.
2. Set up the hypotheses and decide on the level of significance.
3. Collect the data.
4. Calculate the expected frequencies.
5. Check the size of the classes and pool if necessary.
6. Calculate X^2. .
7. Work out the degrees of freedom.
8. Find the critical value from tables and compare.
9. State your conclusion.

E.g. Test the following data at the 5% level of significance to see if a N(18, 36) is a good model.

Classes	f_o
$x < 5$	3
$5 \le x < 10$	15
$10 \le x < 15$	40
$15 \le x < 20$	36
$20 \le x < 25$	18
$x \ge 25$	8
Total	120

H_0: the data can be modelled by a N(18, 36) distribution
H_1: the data cannot be modelled by a N(18, 36) distribution.

Expected frequencies are calculated by finding the Normal probability of being in the particular class and multiplying by 120.
The first two classes need to be pooled as the frequency of the first is less than 5

E.g. For the 20 - 25 class $f_e = \left(\Phi\left(\frac{25-18}{6}\right) - \Phi\left(\frac{20-18}{6}\right)\right) \times 120$
$= 29.73$

Classes	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$(f_o - f_e)^2/f_e$
$x < 10$	18	10.95	7.05	49.70	4.539
$10 \le x < 15$	40	26.08	13.92	193.79	7.431
$15 \le x < 20$	36	38.64	-2.64	6.98	0.181
$20 \le x < 25$	18	29.73	-11.73	137.65	4.630
$x \ge 25$	8	14.60	-6.60	43.57	2.984
Total	120	120			19.765

Test statistic, $X^2 = 19.765$ with $v = 5 - 1$ (same total) = 4
Critical value from tables = 9.488
Since 19.765 > 9.488 we reject H_0 in favour of H_1: there is evidence at the 5% level of significance that the data comes from a distribution other than N(18, 36).

References:
Chapter 7
Pages 177-178

Example 7.4
Page 177

Exercise 7B
Q. 2

Testing for a given Normal distribution
To test for N(μ, σ^2) carry out the steps as above where the expected frequencies are derived from the standardised Normal distribution tables. To do this you need to decide the number of classes to be used and the width of each. You need to choose these so that each class has a minimum of 5 members.
The degrees of freedom are the number of classes minus 1, because the sum of the observed and expected frequencies are the same.

References:
Chapter 7
Pages 179-181

Example 7.5
Page 179

Exercise 7B
Q. 4

Testing for Normality
First calculate an estimate of the mean and variance of the population given the data in the sample.
Then carry out the steps as above.
The degrees of freedom are two less than above as the mean and variance have been determined by the sample.

References:
Chapter 7
Page 182

Example 7.6
Page 182

Other distributions
You may be asked to use the Chi squared test to test for the following distributions:

- Uniform
- Binomial
- Poisson
- Normal (with mean and variance known and unknown)

The text also gives examples of exponential distributions.

E.g. The days of birth of 100 footballers are:

	Sat	Sun	Mon	Tues	Wed	Thurs	Frid
f_o	20	17	11	15	10	13	14
f_e	14.3	14.3	14.3	14.3	14.3	14.3	14.3

Test at the 5% level of significance that these data can be considered to be a uniform distribution.
H_0: Birthdays are uniformly distributed through the week
H_1: Birthdays are not uniformly distributed through the week.

For these data: $\sum \frac{(f_o - f_e)^2}{f_e} = 4.995$

$v = 7 - 1$ (same total) = 6; critical value = 12.59
Since 4.995 < 12.59 we accept H_0: we conclude that there is evidence to support the hypothesis that the data are uniformly distributed.

Exercise 7B
Q. 5

References:
Chapter 7
Pages 168 - 169

The left hand tail
Although the χ^2 test is a 1-tailed test (upper tail), very low values for χ^2 which would occur in the corresponding lower tail are ***sometimes*** considered as indicating a suspicious data set when the data fit is "too good to be true."

E.g. in Example 1 on the previous page, had the observed frequencies been 60, 60, 24, 6; for left hand tail, using the calculated frequencies of the table gives $X^2 = 0.096$.
For $v = 3$ and the same level of significance, critical value = 0.115.
Since 0.096 < 0.115 this gives rise to suspicion that the data are “too good to be true”.