MEI

Mathematics in Education and Industry

MEI Structured Mathematics

Module Summary Sheets

Statistics 2
(Version B: reference to new book)

Topic 1: The Poisson Distribution

Topic 2: The Normal Distribution

Topic 3: Samples and Hypothesis Testing

1. Test for population mean of a Normal distribution
2. Contingency Tables and the Chi-squared Test

Topic 4: Bivariate Data

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 S2 Topic 1 – ***The Poisson Distribution***

References: Chapter 1 Pages 1-4

Example 1.1 Page 4

Exercise 1A Q. 7

The Poisson distribution is a discrete random variable X where

$$P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}$$

The parameter, λ, is the mean of the distribution.
We write $X \sim$ Poisson(λ).

The distribution may be used to model the number of occurences of an event in a given interval provided the occurrences are:
(i) random, (ii) independent,
(iii) occurring at a fixed average rate.

Mean, $E(X) = \lambda$, Variance, $Var(X) = \lambda$
Mean ≈ Variance is a quick way of seeing if a Poisson model might be appropriate for some data.

It is possible to ca.culate terms of the Poisson distribution by a recurrence relationship.

E.g. $P(X = r) = \frac{\lambda^r}{r!} \times e^{-\lambda}$;

$$P(X = r+1) = \frac{\lambda^{r+1}}{(r+1)!} \times e^{-\lambda} = \frac{\lambda}{(r+1)} \times P(X = r)$$

Care needs to be taken over the cumulation of errors.

E.g. The mean number of telephone calls to an office is 2 every 10 minutes.
The probability distribution (Poisson(2)) is as follows:

$P(X = 0) = e^{-2} = 0.1353$

$P(X = 1) = 2e^{-2} = 0.2707$

$P(X = 2) = \frac{2^2}{2!}e^{-2} = 2e^{-2} = 0.2707$

$P(X = 3) = \frac{2^3}{3!}e^{-2} = 0.1804$

$P(X = 4) = \frac{2^4}{4!}e^{-2} = 0.0902$

$P(X > 4) = 1 - \text{sum of above} = 0.0527$

E.g. For $\lambda = 2$, find P(3) from tables.

Note that $P(3) = e^{-2}\frac{2^3}{3!} = 0.1353 \times \frac{8}{6}$

$= 0.1804$

From tables, $P(3) = P(X \le 3) - P(X \le 2)$

$= 0.8571 - 0.6767 = 0.1804$

References: Chapter 1 Pages 5-6

Example 1.2 Page 6

Exercise 1A Q. 4, 5, 11

Use of cumulative probability tables

Cumulative Poisson probability tables are on pages 40-42 of the Students' Handbook and are available in the examinations.

They give cumulative probabilities, i.e. $P(X \le r)$.

So $P(X = r) = P(X \le r) - P(X \le r-1)$

For $\lambda = 1.8$ (Page 40), the second and third entry of the tables give

$P(X \le 1) = 0.4628$, $P(X \le 2) = 0.7306$

i.e. $P(X = 2) = P(X \le 2) - P(X \le 1) = 0.7306 - 0.4628 = 0.2678$

E.g. Vehicles passing along a road were counted and categorised as either private or commercial; on average there were 3 private and 2 commercial vehicles passing a point every minute. It is assumed that the distributions are independent of each other. Find the probability that in a given minute there will be (i) no private vehicles, (ii) no vehicles, (iii) exactly one vehicle.
For the private vehicles, $\lambda_p = 3$ and for commercial vehicles $\lambda_c = 2$.
The distribution may be modelled by the Poisson distribution so that for private vehicles, $X \sim$ Poisson(3) and for commercial vehicles, $Y \sim$ Poisson(2).
For all vehicles, $Z = X + Y \sim$ Poisson(3+2) = Poisson(5)

(i) $P(X = 0) = e^{-3} = 0.0498$
(ii) $P(Z = 0) = e^{-5} = 0.0067$
(iii) $P(Z = 1) = 5e^{-5} = 0.0337$
Note that P(1 veh) = P(1 priv). P(0 comm) + P(0 priv) . P (1 comm)
$= 3e^{-3}.e^{-2} + e^{-3}.2e^{-2}$
$= (2 + 3)e^{-5} = 5e^{-5} = 0.0337$

References: Chapter 1 Pages 12-15

Exercise 1B Q. 4, 8

Sum of Poisson distributions
Two or more Poisson distributions can be combined by addition providing they are independent of each other.

If $X \sim$ Poisson(λ) and $Y \sim$ Poisson(μ), then $X + Y \sim$ Poisson($\lambda + \mu$).

N.B. *You may only add two Poisson distributions in this way if they are independent of each other. There is no corresponding result for subtraction.*

References: Chapter 1 Pages 18-22

Exercise 1C Q. 1(i), 3, 7

Approximation to the binomial distribution
The Poisson distribution may be used as an approximation to the binomial distribution B(n, p) when

(i) n is large, (ii) p is small (so the event is rare).

Then $\lambda = np$.
Note that the Normal distribution is likely to be a good approximation if np is large.

E.g. Some equivalent values:

For $X \sim$B(40, 0.03), $np = 1.2$;
$P(0) = (0.97)^{40} = 0.296 \approx 0.3$;
For $X \sim$Poisson(1.2), $P(0) = e^{-1.2}$
$= 0.301 \approx 0.3$

For $X \sim$B(80, 0.02), $np = 1.6$;
$P(0) = (0.98)^{80} = 0.199 \approx 0.2$;
For $X \sim$Poisson(1.6), $P(0) = e^{-1.6}$
$= 0.202 \approx 0.2$

Exercise 1D Q. 4, 5

Summary S2 Topic 2 – ***The Normal Distribution***

References: Chapter 2 Pages 32-44

The Normal distribution, N(μ,σ^2), is a continuous, symmetric distribution with mean μ and standard deviation σ. The standard Normal distribution N(0,1) has mean 0 and standard deviation 1.

Example 2.1 Page 35

P($X<x_1$) is represented by the area under the curve below x_1.
(It is a special case of a continuous probability density function which is a topic in Statistics 3.)
The area under the standard Normal distribution curve can be found from tables.

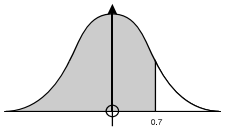
Exercise 2A Q. 4

To find the area under any other Normal distribution curve, the values need to be standardised by the formula

$$z = \frac{x - \mu}{\sigma}$$

For N(0,1), P($Z < z_1$) can be found from tables (Students Handbook, page 44)

E.g. P($Z < 0.7$) = 0.7580
P($Z > 0.7$) = 1 – 0.7580 = 0.2420



For N(2,9) [$\mu = 2$, $\sigma = 3$]
P($X < 5$) = P($Z < z_1$) where $z_1 = {}^{(5-2)}/_{3} = 1$
= 0.8413

References: Chapter 2 Pages 49, 50

Modelling
Many distributions in the real world, such as adult heights or intelligence quotients, can be modelled well by a Normal distribution with appropriate mean and variance.
Given the mean μ and standard deviation σ, the Normal distribution N(μ,σ^2) may often be used.

Example 2.3 Page 49

When the underlying distribution is discrete then the Normal distribution may often be used, but in this case a continuity correction must be applied. This requires us to take the mid-point between successive possible values when working with continuous distribution tables.

Exercise 2B Q. 1

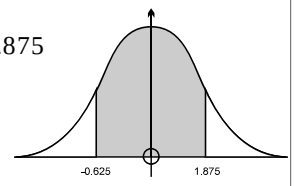
E.g. P(X) > 30 means P(X > 30.5) if X can take only integer values.

E.g. The distribution of masses of adult males may be modelled by a Normal distribution with mean 75 kg and standard deviation 8 kg. Find the probability that a man chosen at random will have mass between 70 kg and 90 kg.

We require P($70<X<90$) = P($z_1 < Z < z_2$)

where $z_1 = \frac{70-75}{8} = -0.625$

and $z_2 = \frac{90-75}{8} = 1.875$



$\Rightarrow$ P($70 < X < 90$)
= P($Z < 1.875$) – P$\left(Z < -0.625\right)$
= P($Z < 1.875$) – $\left(1 - \text{P}(Z < 0.625)\right)$
= 0.9696 – (1 – 0.7340) = 0.7036

References: Chapter 2 Pages 50-52

The Normal approximation to the binomial distribution.
This is a valid process provided
(i) n is large,
(ii) p is not too close to 0 or 1.

Exercise 2B Q. 4

Mean = np, Variance = npq.
The approximation will be N(np, npq).

A continuity correction must be applied because we are approximating a discrete distribution by a continuous distribution.

E.g. Find the probability that when a die is thrown 30 times there are at least 10 sixes.
Using the binomial distribution requires P(30 sixes) + P(29 sixes) + ...+ P(10 sixes).
However, using N(np, npq) where n = 30 and $p = {}^{1}/_{6}$, gives N(5, 4.167).
P($X > 9.5$) = P($Z > z_1$)

where $z_1 = {}^{(9.5-5)}/_{2.04} = 2.205$

= 1 – 0.9863 = 0.0137

N.B. a continuity correction is applied because the original distribution (binomial) is being approximated by a continuous distribution (Normal).

References: Chapter 2 Pages 52-54

The Normal approximation to the Poisson distribution.
This is a valid process provided λ is sufficiently large for the distribution to be reasonably symmetric. A good guideline is if λ is at least 10.

Exercise 2B Q. 11

For a Poisson distribution, mean = variance = λ
The approximation will be N(λ,λ).

Exercise 2C Q. 1, 6, 7

As with the Binomial Distribution, a continuity correction must be applied because we are approximating a discrete distribution by a continuous distribution.

E.g. A large firm has 50 telephone lines. On average, 40 lines are in use at once and the distribution may be modelled by Poisson(40). Find the probability of there not being enough lines.

The distribution is Poisson(40).

Approximate by N(40,40).

Then we require P($X > 50$) = P($Z > z_1$)

where $z_1 = \frac{50.5-40}{\sqrt{40}} = 1.66$

$\Rightarrow$ P($X > 50$) = 1 – 0.9515 = 0.0485

Summary S2 Topic 3 – ***Samples and Hypothesis Testing***

1: Estimating the population mean of a Normal distribution

References: Chapter 3 Pages 68-71

Example 3.1 Page 70

The distribution of sample means

If a population may be modelled by a Normal distribution and samples of size n are taken from the population, then the distribution of means of these samples is also Normal.

If the parent population is $N(\mu, \sigma^2)$ then the sampling distribution of means is $N\left(\mu, \frac{\sigma^2}{n}\right)$.

E.g. If the parent population is N(10, 16) and a sample of size 25 has mean 8.6, then this value comes from the sampling distribution of means which is N(10, 0.64).

E.g. It is thought that the parent population is Normally distributed with mean 20.
A random sample of 50 data items has a sample mean of 24.2 and s.d. 8.3.
Is there any evidence at the 0.1% significance level that the mean of the population is not 20?

$H_0 : \mu = 20$

$H_1 : \mu \neq 20$

(Note that although the mean of the sample is greater than the proposed mean, we do not have $\mu > 20$ because of the wording of the question.)

$$\Rightarrow z = \frac{x - \bar{\mu}}{\sigma/\sqrt{n}} = \frac{24.2 - 20}{8.3/\sqrt{50}} = 3.578$$

Critical value from tables for two-tailed, 0.1% significance level is 3.27

Since 3.578>3.27 we reject H_0 in favour of H_1.

There is evidence that the mean of the population is not 20.

References: Chapter 3 Pages 71-73

Exercise 3A Q. 1(i),(iii)

Hypothesis test for the mean using the Normal distribution

Tests on the mean using a single sample.

H_0 is $\mu = \mu_0$ where μ_0 is some specified value.

H_1 may be one tailed: $\mu < \mu_0$ or $\mu > \mu_0$

or two tailed: $\mu \neq \mu_0$.

In other words, given the mean of the sample taken we ask the question, "Could the mean of the parent population be what we think it is?"

Suppose the parent population is $N(\mu, \sigma^2)$, then the sampling distribution of means is $N\left(\mu, \frac{\sigma^2}{n}\right)$.The critical values are therefore $\mu \pm k\frac{\sigma}{\sqrt{n}}$ where the value of k depends on the level of significance and whether it is a one or two-tailed test.

We therefore calculate the value $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ and compare it to the value found in tables.

Alternatively, if the value of the mean of the sample lies inside the acceptance region then we would accept H_0, but if it lay in the critical region then we would reject H_0 in favour of H_1.

Alternatively, calculate the probability that the value is greater than the value found and see if it less than the significance level be used.

E.g. A population has variance 16. It is required to test at the 0.5% level of significance whether the mean of the population could be 10 or whether it is less than this. A random sample size 25 has a mean of 8.6.

$H_0 : \mu = 10$

$H_1 : \mu < 10$

$k = 2.58$ (for 0.5% level, 1-tailed test)

Critical value is $10 - 2.58 \times \frac{\sigma}{\sqrt{n}} = 10 - 2.58 \times \frac{4}{5} = 7.936$

Since $8.6 > 7.936$ we accept H_0; there is no evidence at the 0.5% level of significance that the mean is less than 10.

Alternatively, if the mean is 10 then the sampling distribution of means is N(10, 0.64)

Then $P(\bar{X} \leq 8.6) = 1 - \Phi\left(\frac{10 - 8.6}{0.8}\right) = 1 - \Phi(1.75)$

$= 1 - 0.9599 = 0.0401.$

Since $0.0401 > 0.005$ we accept H_0

References: Chapter 3 Pages 73-74

Example 3.3 Page 74

Exercise 3A Q. 6

Known and estimated standard deviation

The hypothesis test described above requires the value of the standard deviation of the parent population.

In reality the standard deviation of the parent population will usually not be known and will have to be estimated from the sample data.

If the sample size is sufficiently large, the s.d. of the sample may be used as the s.d. of the parent population.

A good guideline is to require $n \geq 50$.

Summary S2 Topic 3 – ***Samples and Hypothesis Testing***
2: Contingency Tables and the Chi-squared Test

References: Chapter 3 Pages 81-85

Contingency Tables

Suppose the elements of a population have 2 sets of distinct characteristics $\{X, Y\}$, each set containing a finite number of discrete characteristics $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots\ldots, y_n\}$ then each element of the population will have a pair of characteristics (x_i, y_j).

The frequency of these $m \times n$ pairs (x_i, y_j) can be tabulated into an ***m* × *n* contingency table**.

	y_1	y_2		y_n
x_1	$f_{1,1}$	$f_{1,2}$		$f_{1,n}$
x_2	$f_{2,1}$	$f_{2,2}$		$f_{2,n}$
x_m	$f_{m,1}$	$f_{m,2}$		$f_{m,n}$

The **marginal totals** are the sum of the rows and the sum of the columns and it is usual to add a row and a column for these.

The requirement is to determine the extent to which the variables are related.

If they are not related but independent, then theoretical probabilities can be estimated from the sample data.

You now have two tables, one containing the actual (observed) frequencies and the other containing the estimated expected frequencies based on the assumption that the variables are independent.

The hypothesis test

H_0: The variables are not associated.

H_1: The variables are associated.

References: Chapter Pages 87-92

The χ^2 Statistic (Chi-squared statistic)

This statistic measures how far apart are the set of observed and expected frequencies.

$$X^2 = \sum \frac{(\text{observed frequency - expected frequency})^2}{\text{expected frequency}}$$

$$= \sum \frac{(f_o - f_e)^2}{f_e}$$

References: Chapter 3 Page 85

Degrees of freedom

The distribution depends on the number of free variables there are, called the degrees of freedom, ν.

This is the number of cells less the number of restrictions placed on the data.

For a 2×2 table such as the example given the number of cells to be filled is 4, but the overall total is 50 which is a restriction and the proportions for each variable were also estimated from the data, giving two further restrictions. So the number of degrees of freedom in the example is 1.

Exercise 3B Q. 4, 5

In general the number of degrees of freedom for an $m \times n$ table is $(m - 1)(n - 1)$.

Exercise 3C Q. 1, 8

E.g. a group of 50 students was selected at random from the whole population of students at a College. Each was asked whether they drove to College or not and whether they lived more than or less than 10 km from the College. The results are shown in this table.

	Nearer than 10 km	Further than 10 km	
Drives	11	17	**28**
Does not drive	15	7	**22**
	26	**24**	**50**

E.g. If driving to College and the distance lived are not associated events then if one student is chosen at random the estimated probabilities are

$$\text{P(drives)} = \frac{28}{50}, \text{P(lives further than 10 km)} = \frac{24}{50}$$

and P(drives and lives further than 10 km) = $\frac{28}{50} \times \frac{24}{50}$

$= 0.2688$

So out of 50 people we would expect 50×0.2688

$= 13.44$

In a similar way the entries in the other three boxes are calculated to give the following:

	Nearer than 10 km	Further than 10 km	
Drives	14.56	13.44	**28**
Does not drive	11.44	10.56	**22**
	26	**24**	**50**

We test the hypotheses:

H_0: the two events are not associated

H_1: The two events are associated.

$$X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(11-14.56)^2}{14.56} + \frac{(17-13.44)^2}{13.44} + \frac{(15-11.44)^2}{11.44} + \frac{(7-10.56)^2}{10.56}$$

$$= 0.8704 + 0.9430 + 1.1078 + 1.2002$$

$$= 4.1214$$

If the test is at the 5% level then the tables on page 45 of the MEI Students' Handbook gives the critical value of 3.841 ($\nu = 1$).

Since $4.1214 > 3.841$ we reject the null hypothesis, H_0, and conclude that there is evidence that the two events are associated.

If the test were at the 1% significance level then we would conclude that there was not enough evidence to reject the null hypothesis.

Summary S2 Topic 4 – ***Bivariate Data – 1***

References:
Chapter 4
Pages 104-109

Bivariate Data are pairs of values (x, y) associated with a single item.
e.g. lengths and widths of leaves.
The individual variables x and y may be discrete or continuous.

A scatter diagram is obtained by plotting the points (x_1, y_1), (x_2, y_2) etc.
Correlation is a measure of the linear association between the variables.

A line of best fit is a line drawn to fit the set of data points as closely as possible.

This line will pass through the mean point $(\bar{x}, \bar{y})$ where

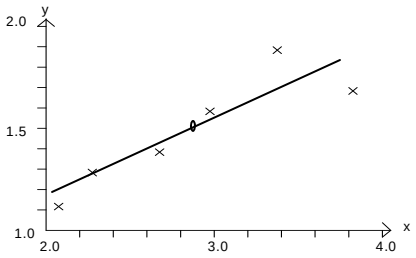
$\bar{x}$ is the mean of the x values and $\bar{y}$ is the mean of the y values.

There is said to be **perfect correlation** if all the points lie on a line.

Example 1
The length (x cm) and width (y cm) of leaves of a tree were measured and recorded as follows:

x	2.1	2.3	2.7	3.0	3.4	3.9
y	1.1	1.3	1.4	1.6	1.9	1.7

The scatter graph is drawn as shown.



The mean point is $(\bar{x}, \bar{y})$ which is (2.9, 1.5)
The line of best fit is drawn through the point (2.9,1.5)

Correlation and Regression
If the x and y values are both regarded as values of random variables, then the analysis is correlation. Choose a sample from a population and measure two attributes.
If the x value is non-random (e.g. time at fixed intervals) then the analysis is regression. Choose the value of one variable and measure the corresponding value of another.

E.g. 50 students are selected at random and their heights and weights are measured. This will require correlation analysis.
A ball is bounced 5 times from each of a number of different heights and the height is recorded. This will require regression analysis.

References:
Chapter 4
Pages 110-111

Notation for n pairs of observations (x, y).

$$S_{xy} = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right),\ S_{xx} = \sum\left(x_i - \bar{x}\right)\left(x_i - \bar{x}\right)$$

$$S_{yy} = \sum\left(y_i - \bar{y}\right)\left(y_i - \bar{y}\right)$$

The alternative form for S_{xy} is

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = \sum x_i y_i - n\bar{x}\bar{y}$$

For the data above:

$$\sum x = 17.4; \quad \sum y = 9.0; \quad \sum xy = 26.97$$

$$S_{xy} = 26.97 - \frac{17.4 \times 9.0}{6} \Rightarrow S_{xy} = 0.87$$

Pearson's Product Moment Correlation Coefficient provides a standardised measure of covariance.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}.S_{yy}}} = \frac{\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum\left(x_i - \bar{x}\right)^2 \sum\left(y_i - \bar{y}\right)^2}}$$

$$= \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum x_i^2 - n\bar{x}^2\right)\left(\sum y_i^2 - n\bar{y}^2\right)}}$$

The pmcc lies between -1 and +1.

For the data above:

$$\sum x^2 = 52.76 \Rightarrow S_{xx} = 2.3$$

$$\sum y^2 = 13.92 \Rightarrow S_{yy} = 0.42$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{0.87}{\sqrt{2.3 \times 0.42}} = 0.885$$

r can be found directly with an appropriate calculator.

References:
Chapter 4
Pages 111-114

Example 4.1
Page 112

Exercise 4A
Q. 2

Statistics 2
Version B: page 6
Competence statements b1, b2, b3

References: Chapter 4 Pages 118-124

Exercise 4C Q. 3

Testing a parent population correlation by means of a sample where *r* has been found

The value of *r* found for a sample can be used to test hypotheses about the value of ρ, the correlation in the parent population.

Conditions:

(i) the values of *x* and *y* must be taken from a bivariate Normal distribution,

(ii) the data must be a random sample.

An indication that a bivariate Normal distribution is a valid model is shown by a scatter plot which is roughly elliptical with the points denser near the middle.

$H_0: \rho = 0$ There is no correlation between the two variables.

$H_1: \rho \neq 0$ There is correlation between the two variables (2 - tailed test.)

Or :

$H_1: \rho > 0$ There is positive correlation between the two variables (1 - tailed test.)

Or :

$H_1: \rho < 0$ There is negative correlation between the two variables (1 - tailed test.)

For the data of Example 1:

$r = 0.885$

We wish to test the hypothesis that there is positive correlation between lengths and widths of the leaves of the tree.

$H_o: \rho = 0$ There is no correlation between the two variables.

$H_1: \rho > 0$ There is positive correlation between the two variables (1-tailed test).

From the Students' Handbook, the critical value for $n = 8$ at 5% level (one tailed test) is 0.6215.

Since 0.885>0.6215 there is evidence that H_0 can be rejected and that there is positive correlation between the two variables.

References: Chapter 4 Pages 132-134

Example 4.3 Page 134

Exercise 4C Q. 5

Spearman's coefficient of rank correlation

If the data do not look linear when plotted on a scatter graph (but appear to be reasonably monotomic), or if the rank order instead of the values is given, then the Pearson correlation coefficient is not appropriate.

Instead, Spearman's rank correlation coefficient should be used. It is usually calculated using the formula

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where *d* is the difference in ranks for each data pair.

This coefficient is used:

(i) when only ranked data are available,

(ii) the data cannot be assumed to be linear.

In the latter case, the data should be ranked.

Where r_s has been found the hypothesis test is set up in the same way. The condition here is that the sample is random.

Make sure that you use the right tables!

Tied Ranks

If two ranks are tied in, say, the 3rd place then each should be given the rank $3^1/_2$.

Example 2

2 judges ranked 5 competitors as follows:

Competitor	A	B	C	D	E
Judge 1	1	3	4	2	5
Judge 2	2	3	1	4	5
d	1	0	-3	2	0
d^2	1	0	9	4	0

$$\sum_1^5 d^2 = 14 \Rightarrow r_s = 1 - \frac{6x14}{5x24} = 0.3$$

For the data of example 2:

$r_s = 0.3$

$H_o: \rho = 0$ There is no correlation between the two variables.

$H_1: \rho > 0$ There is positive correlation between the two variables (1-tailed test).

For $n = 5$ at the 5% level (1 tailed test), the critical value is 0.9.

Since $0.3 < 0.9$ we are unable to reject H_o and conclude that there is no evidence to suggest correlation.

Statistics 2 Version B: page 7 Competence statements b4, b5, b6, b7, b8

References: Chapter 4 Pages 142-144

Exercise 4D Q. 4

The least squares regression line

For each value of *x* the value of *y* given and the value on the line may be different by an amount called the **residual**.

If the data pair is (x_i,y_i) where the line of best fit is $y = a + bx$ then $y_i - (a + bx_i) = e_i$ giving the residual e_i.

The least squares regression line is the line that minimises the sum of the squares of the residuals.

The equation of the line is $y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x - \bar{x})$

E.g. For the data

x	1	2	3	4	5
y	1.1	2.4	3.6	4.7	6.1

$$\sum x = 15,\ \sum y = 17.9,\ \sum x^2 = 55,\ \sum xy = 65$$

$$\Rightarrow \bar{x} = \frac{15}{5} = 3,\ \bar{y} = \frac{17.9}{5} = 3.58$$

$$\Rightarrow S_{xx} = \sum x^2 - n\bar{x}^2 = 55 - 6\times 3^2 = 10$$

$$\Rightarrow S_{xy} = \sum xy - n\bar{x}\bar{y} = 65 - 5\times 3\times 3.58 = 11.3$$

$$\Rightarrow y - 3.58 = \frac{11.3}{10}(x-3) \Rightarrow y = 1.13x + 0.19$$