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

Purchasers have the licence to make multiple copies for use within a single establishment

© MEI November, 2005
The Poisson distribution is a discrete random variable $X$ where $P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$.

The parameter, $\lambda$, is the mean of the distribution. We write $X \sim \text{Poisson}(\lambda)$.

The distribution may be used to model the number of occurrences 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, $\text{Var}(X) = \lambda$.

Mean $\approx$ Variance is a quick way of seeing if a Poisson model might be appropriate for some data.

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

E.g. $P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}$;
$P(X = r + 1) = \frac{\lambda^{r+1} e^{-\lambda}}{(r+1)!} = \frac{\lambda}{r+1} P(X = r)$

Care needs to be taken over the cumulation of errors.

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. $P(X \leq r)$.

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

For $\lambda = 1.8$ (Page 40), the second and third entry of the tables give $P(X \leq 2) = 0.7306$;
and $P(X \leq 1) = 0.6767$.

i.e. $P(X = 2) = P(X \leq 2) - P(X \leq 1)$.

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

If $X \sim \text{Poisson}(\lambda)$ and $Y \sim \text{Poisson}(\mu)$, then $X + Y \sim \text{Poisson}(\lambda + \mu)$.

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.

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.

Example 1.1

Exercise 1A
Q. 7

Exercise 1B
Q. 4, 8

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

Exercise 1D
Q. 4, 5

References:
Chapter 1
Pages 1-4

References:
Chapter 1
Pages 5-6

References:
Chapter 1
Pages 12-15

References:
Chapter 1
Pages 18-22

Statistics 2
Version B: page 2

Competence statements P1, P2, P3, P4, P5

© MEI
The Normal distribution, \(N(\mu, \sigma^2)\), is a continuous, symmetric distribution with mean \(\mu\) and standard deviation \(\sigma\). The standard Normal distribution \(N(0,1)\) has mean 0 and standard deviation 1.

\[ P(X < x_1) \] is represented by the area under the curve below \(x_1\).

(\text{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.

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

\[ z = \frac{x - \mu}{\sigma} \]

### 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 \(\mu\) and standard deviation \(\sigma\), the Normal distribution \(N(\mu, \sigma^2)\) may often be used.

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.

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

### 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.

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. 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 \(z_2 = \frac{90 - 75}{8} = 1.875\)

\[ \Rightarrow P(70 < X < 90) = P(Z < -0.625) - P(Z < 1.875) \]

\[ = 0.7580 - (1 - P(Z < 0.625)) \]

\[ = 0.7580 - 0.2420 = 0.5160 \]

**E.g.** Find the probability that when a die is thrown 30 times there are at least 10 sixes. Using the binomial distribution requires \(P(30\text{ sixes}) + P(29\text{ sixes}) + \ldots + P(10\text{ sixes})\). However, using \(N(np, npq)\) where \(n = 30\) and \(p = \frac{1}{6}\), gives \(N(5, 4.167)\).

\(P(X > 9.5) = P(Z > z_1)\)

where \(z_1 = \frac{9.5 - 5}{\sqrt{4}} = 2.205\)

\[ = 1 - 0.9863 = 0.0137 \]

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

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 \(z_1 = \frac{50.5 - 40}{\sqrt{40}} = 1.66\)

\[ \Rightarrow P(X > 50) = 1 - 0.9515 = 0.0485 \]
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 \( N\left(\mu, \frac{\sigma^2}{n}\right) \).

Hypothesis test for the mean using the Normal distribution
Tests on the mean using a single sample. 

\( H_0 \) is \( \mu = \mu_0 \) where \( \mu_0 \) 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.

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 \( n \geq 50 \).

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.)

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

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.

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 \( 10 - \frac{2.58 \times 4}{5} = 10 - \frac{10.92}{5} = 7.936 \)

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 \( \text{P}(\bar{X} \leq 8.6) = 1 - \Phi\left(10 - \frac{8.6}{0.8}\right) = 1 - \Phi(1.75) = 1 - 0.9599 = 0.0401 \)

Since 0.0401 > 0.005 we accept \( H_0 \)
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 \( Y = \{y_1, y_2, \ldots, y_n\} \) 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 \( m \times n \) contingency table.

<table>
<thead>
<tr>
<th></th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( f_{1,1} )</td>
<td>( f_{1,2} )</td>
<td>( f_{1,n} )</td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( f_{2,1} )</td>
<td>( f_{2,2} )</td>
<td>( f_{2,n} )</td>
</tr>
<tr>
<td>( x_m )</td>
<td>( f_{m,1} )</td>
<td>( f_{m,2} )</td>
<td>( f_{m,n} )</td>
</tr>
</tbody>
</table>

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.

The \( \chi^2 \) Statistic (Chi-squared statistic)
This statistic measures how far apart are the set of observed and expected frequencies.
\[
\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}
\]
\[
= \sum \frac{(f_x - f_e)^2}{f_e}
\]

Degrees of freedom
The distribution depends on the number of free variables there are, called the degrees of freedom, \( \nu \).
This is the number of cells less the number of restrictions placed on the data.

For a \( 2 \times 2 \) 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.
In general the number of degrees of freedom for an \( m \times n \) table is \((m - 1)(n - 1)\).

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.

<table>
<thead>
<tr>
<th></th>
<th>Nearer than 10 km</th>
<th>Further than 10 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drives</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Does not drive</td>
<td>15</td>
<td>7</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(f_x - f_e)^2}{f_e} = \frac{(11 - 14.56)^2}{14.56} + \frac{(17 - 13.44)^2}{13.44} + \frac{(15 - 11.44)^2}{11.44} + \frac{(26 - 10.56)^2}{10.56}
\]
\[
= 0.8704 + 0.9430 + 1.1078 + 1.2002
\]
\[
= 4.1214
\]

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.

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
\[
P(\text{drives}) = \frac{28}{50}, P(\text{lives further than 10 km}) = \frac{24}{50}
\]
and
\[
P(\text{drives and lives further than 10 km}) = \frac{28}{50} \times \frac{24}{50} = 0.2688
\]

So out of 50 people we would expect \( 50 \times 0.2688 = 13.44 \)

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

<table>
<thead>
<tr>
<th></th>
<th>Nearer than 10 km</th>
<th>Further than 10 km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drives</td>
<td>14.56</td>
<td>13.44</td>
</tr>
<tr>
<td>Does not drive</td>
<td>11.44</td>
<td>10.56</td>
</tr>
</tbody>
</table>

\[
\chi^2 = \sum \frac{(f_x - f_e)^2}{f_e} = \frac{(11 - 14.56)^2}{14.56} + \frac{(17 - 13.44)^2}{13.44} + \frac{(15 - 11.44)^2}{11.44} + \frac{(26 - 10.56)^2}{10.56}
\]
\[
= 0.8704 + 0.9430 + 1.1078 + 1.2002
\]
\[
= 4.1214
\]

We test the hypotheses:
H$_0$: the two events are not associated
H$_1$: The two events are associated.

Exercise 3B
Q. 4, 5

Exercise 3C
Q. 1, 8

Statistics 2
Version B: page 5
Competence statements H1, H2
© MEI
**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.

**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.

**Notation for \(n\) pairs of observations \((x_i, y_i)\).**

\[
S_{xx} = \sum (x_i - \bar{x})^2, \quad S_{yy} = \sum (y_i - \bar{y})^2
\]

The alternative form for \(S_{xy}\) is

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

**Pearson’s Product Moment Correlation Coefficient**
provides a standardised measure of covariance:

\[
r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)(\sum y_i^2 - n \bar{y}^2)}}
\]

The pmcc lies between -1 and +1.

**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 \((2.9, 1.5)\)

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.

For the data above:

\[
\sum x = 17.4; \quad \sum y = 9.0; \quad \sum xy = 26.97 \\
S_{xy} = 26.97 - \frac{17.4 \times 9.0}{6} \Rightarrow S_{xy} = 0.87
\]

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 104-109

Example 4.1
Page 112

Exercise 4A
Q. 2

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

\[ H_0: \rho = 0 \quad \text{There is no correlation between the two variables.} \]

\[ H_1: \rho > 0 \quad \text{There is positive correlation between the two variables (1-tailed test).} \]

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.

**Example 2**

2 judges ranked 5 competitors as follows:

<table>
<thead>
<tr>
<th>Competitor</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Judge 2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ d_1 = 0 \quad d_2 = 1 \]

\[ \sum d^2 = 14 \quad r_s = 1 - \frac{6 \times 14}{5 \times 24} = 0.3 \]

For the data of example 2:

\[ r_s = 0.3 \]

\[ H_0: \rho = 0 \quad \text{There is no correlation between the two variables.} \]

\[ H_1: \rho > 0 \quad \text{There is positive correlation between the two variables (1-tailed test).} \]

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 \( H_0 \) and conclude that there is no evidence to suggest correlation.

**Exercise 4C**

Q. 3

Example 4.3

Page 134

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

\[ r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)} \]

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 \( 3 \frac{1}{2} \).

**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 \( y - \bar{y} = \frac{S_y}{S_x}(x - \bar{x}) \)