REVISION SHEET – STATISTICS 2 (MEI)

BIVARIATE DATA

The main ideas are:

- Scatter Diagrams and Lines of Best Fit
- Pearson's Product Moment Correlation
- Spearman's Ranking
- The Least Squares Regression Line

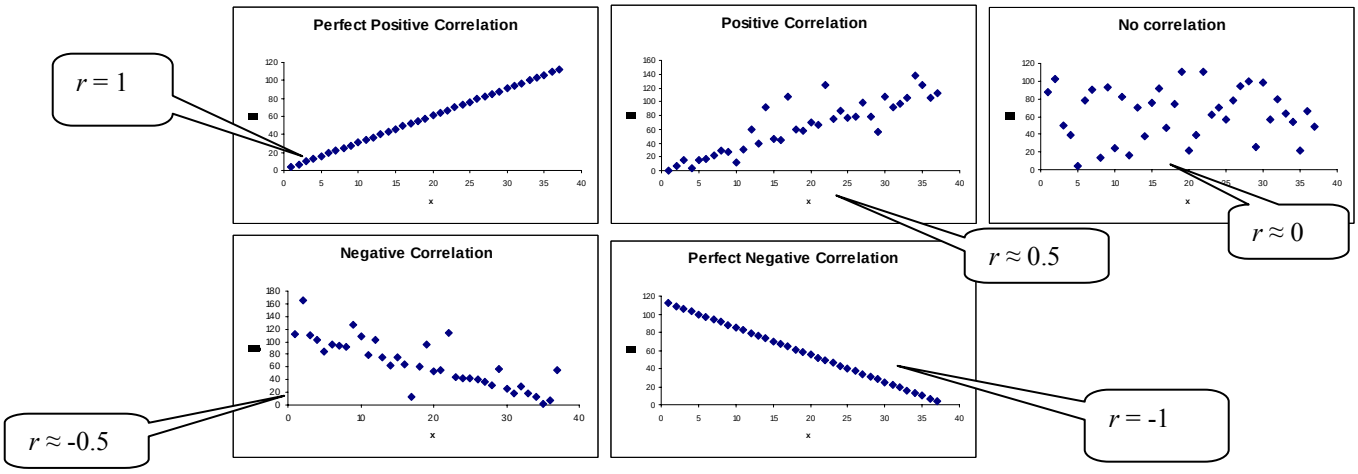
Before the exam you should know:

- Know when to use Pearson's product moment correlation coefficient.
- How to use summary statistics such as $\sum x, \sum x^2, \sum y, \sum y^2, \sum xy$ to calculate S_{xx}, S_{yy}, S_{xy}.
- Know how to recognise when a 1 or 2-tail test is required and apply it to the PPMCC and Spearman's Ranking.
- What is meant by a residue and the "least squares" regression line.

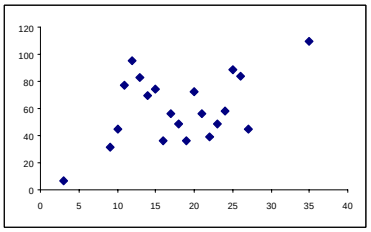
Scatter Diagrams

With Bivariate Data we are usually trying to investigate whether there is a correlation between the two underlying variable, usually called x and y.

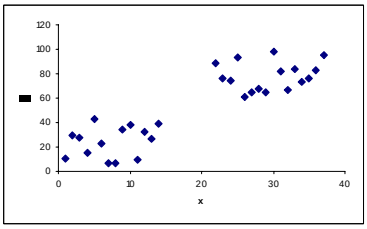
Pearson's product moment correlation coefficient, r, is a number between -1 and +1 which can be calculated as a measure of the correlation in a population of bivariate data.



Beware of diagrams which appear to indicate a linear correlation but in fact to not:



Here two outliers give the impression that there is a linear relationship where in fact there is no correlation.



Here there are 2 distinct groups, neither of which have a correlation.

Product Moment Correlation

Pearson's product Moment Correlation Coefficient:

$$S_{xy} = \sum (x-\bar{x})(y-\bar{y}) = \sum xy - n\bar{x}\bar{y}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}} \qquad \text{where : } S_{xx} = \sum (x-\bar{x})^2 = \sum x^2 - n\bar{x}^2$$

$$S_{yy} = \sum (y-\bar{y})^2 = \sum_{i=1}^{n} y^2 - n\bar{y}^2$$

A value of +1 means perfect positive correlation, a value close to 0 means no correlation and a value of -1 means perfect negative correlation. The closer the value of *r* is to +1 or -1, the stronger the correlation.

Example

A 'games' commentator wants to see if there is any correlation between ability at chess and at bridge.

A random sample of eight people, who play both chess and bridge, were chosen and their grades in chess and bridge were as follows:

Player	A	B	C	D	E	F	G	H
Chess grade *x*	160	187	129	162	149	151	189	158
Bridge grade *y*	75	100	75	85	80	70	95	80

Using a calculator:

$n = 8,\quad \Sigma x = 1285,\quad \Sigma y = 660,\quad \Sigma x^2 = 209141,\quad \Sigma y^2 = 55200,\quad \Sigma xy = 107230\quad \bar{x} = 160.625,\quad \bar{y} = 82.5$

$$r = \frac{107230 - 8\times 160.625\times 82.5}{\sqrt{(209141 - 8\times 160.625^2)(55200 - 8\times 82.5^2)}} = 0.850 \text{ (3 s.f.)}$$

Rank Correlation

Spearman's Rank Correlation

This is given by: $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$, where *d* represents the difference in ranks for each of the *n* pairs of rankings. Spearman's coefficient of rank correlation can be used to investigate whether there is a general increase or decrease (i.e. non-linear correlation), which is not possible with Pearson's product moment correlation coefficient.

The Least Squares Regression Line

This is a line of best fit which produces the least possible value of the sum of the squares of the residuals (the vertical distance between the point and the line of best fit).

It is given by: $y - \bar{y} = \frac{S_{xy}}{S_{xx}}(x-\bar{x})$ Alternatively, $y = a + bx$ where, $b = \frac{S_{xy}}{S_{xx}}, a = \bar{y} - b\bar{x}$

Predicted values

For any pair of values (*x*, *y*), the *predicted value* of *y* is given by $\hat{y}$ **= *a* + *bx*.**

If the regression line is a good fit to the data, the equation may be used to predict *y* values for *x* values within the given domain, i.e. *interpolation*.

It is unwise to use the equation for predictions if the regression line is *not* a good fit for any part of the domain (set of *x* values) or the *x* value is outside the given domain, i.e. the equation is used for *extrapolation*.

The corresponding residual = $\varepsilon = y - \hat{y} = y - (a + bx)$. The sum of the residuals = $\Sigma\varepsilon = 0$

The least squares regression line minimises the sum of the squares of the residuals, $\Sigma\varepsilon^2$.

Acknowledgement: Some material on these pages was originally created by Bob Francis and we acknowledge his permission to reproduce such material in this revision sheet.

REVISION SHEET – STATISTICS 2 (MEI)

HYPOTHESIS TESTING & CONTINGENCY TABLES

The main ideas are:

- Hypothesis Testing Normal Distribution
- χ^2 and Contingency Tables

Before the exam you should know:

- About hypothesis testing for the mean using the Normal Distribution.
- About using a known and an estimated standard deviation.
- The χ^2 test for independence in a contingency table.

Hypothesis Testing

A ***null hypothesis*** (H_0) is tested against an ***alternative hypothesis*** (H_1) at a particular ***significance level***.
According to given criteria, the null hypothesis is either rejected is not rejected.
The hypothesis test can be either 1-tailed or 2-tailed.

Sample data, drawn from the parent population, may be used to carry out a hypothesis test on the null hypothesis that the population mean has some particular value, μ_0.

Hypothesis testing procedure

(1) Establish null and alternative hypotheses:
H_0: $\mu = \mu_0$
H_1: $\mu < \mu_0$ *or* $\mu > \mu_0$ (1-tail test);
or $\mu \neq \mu_0$ (2-tail test)

(2) Decide on the significance level: *s%*

(3) Collect data (independent and at random): obtain sample of size *n* from the parent population and calculate mean $\bar{x}$.

(4) Conduct test:
Calculate test statistic: $z = \dfrac{\bar{x} - \mu_0}{\dfrac{\sigma}{\sqrt{n}}}$
[if σ unknown, use *s* – provided *n* is large]
1-tail: H_1: $\mu < \mu_0$ – compare *z* with $\Phi^{-1}(s\%)$
1-tail: H_1: $\mu > \mu_0$
– compare *z* with $\Phi^{-1}((100 - s)\%)$
2-tail: H_1: $\mu \neq \mu_0$
if $\bar{x} < \mu_0$ compare *z* with $\Phi^{-1}(\frac{1}{2}s\%)$
if $\bar{x} > \mu_0$ compare *z* with $\Phi^{-1}((100 - \frac{1}{2}s)\%)$

(5) Interpret result in terms of the original claim:
1-tail: if $z < \Phi^{-1}(s\%)$ reject H_0
1-tail: if $z > \Phi^{-1}((100 - s)\%)$ reject H_0
2-tail: if $z < \Phi^{-1}(\frac{1}{2}s\%)$ *or* $z > \Phi^{-1}((100 - \frac{1}{2}s)\%)$ reject H_0
Finally present conclusion in context of problem.

Distribution of Sample Means

For samples of size *n* drawn from a Normal distribution with mean μ and finite variance σ^2, [$X \sim N(\mu, \sigma^2)$] the distribution of sample means, $\bar{X}$, is Normal with mean μ and variance $\dfrac{\sigma^2}{n}$, i.e. $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$.

The standard error of the mean *(i.e. the standard deviation of the sample means)* is given by $\dfrac{\sigma}{\sqrt{n}}$.

Example 1

The packaging on a type of electric light bulb states that the average lifetime of bulbs is 1000 hours. A consumer association thinks that this is an overestimate and tests a sample of 100 bulbs, recording the life-time, *x* hours, of each bulb.

Assuming the distribution of lifetimes is Normal, test the consumer association's claim at the 5% level.

Hypothesis testing procedure

(1) H_0: $\mu = 1000$; H_1: $\mu < 1000$ (1-tail test)

(2) Significance level: 5%

(3) Summary statistics for the lifetimes, *x*, of a random sample of 100 bulbs:
$n = 100$, $\Sigma x = 99860$, $\Sigma x^2 = 99725047$
$\bar{x} = 99860 \div 100 = 998.6$;
$s = \sqrt{\dfrac{99725047 - 100 \times 998.6^2}{99}} = 7$

(4) Test statistic: $z = \dfrac{998.6 - 1000}{\dfrac{7}{\sqrt{100}}} = -2$
Critical value in lower tail: $\Phi^{-1}(0.05) = -1.645$, which is greater than –2.

(5) Since $-2 < -1.645$, there is sufficient evidence to reject H_0, i.e. the consumer association's claim that the average life-time is less than 1000 hours is upheld at the 5% significance level.

Contingency Tables

An $m \times n$ ***contingency table*** results when two variables are measured on a sample, with the first variable having m possible categories of results and the second variable having n possible categories.

Each cell contains an *observed frequency* (f_o), with which that pair of categories of values of the two variables occurs in the sample.

Marginal row and column totals are used to calculate *expected frequencies* (f_e).

Hypothesis Testing

A ***null hypothesis*** (H_0) is tested against an ***alternative hypothesis*** (H_1) at a particular ***significance level*** with a number of ***degrees of freedom***.

According to given criteria, the null hypothesis is either rejected is not rejected.
The hypothesis test is always 1-tailed.

Sample data, drawn from the parent population, may be used to carry out a hypothesis test on the null hypothesis that there is *no association* between the two variables, i.e. are *independen*t.

Hypothesis testing procedure

(1) Establish null and alternative hypotheses:
H_0: no association between the variables,
H_1: the variables are *not* independent**.**

(2) Decide on the significance level: *s%*

(3) Collect data in the form of an m by n contingency table of observed frequencies (f_o)

(4) Conduct test:
Calculate marginal row and column totals.
Calculate expected frequencies (f_e):

$$f_e = \frac{\text{row total} \times \text{column total}}{\text{total sample size}}$$

Calculate test statistic $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$

Find degrees of freedom $v = (m-1)(n-1)$
Compare with critical value from tables, dependant on significance level *s*% and d.o.f. v.

(5) Interpret result in terms of the original claim:

If test statistic χ^2 > critical value, then reject H_0 (i.e. accept H_1)

If test statistic χ^2 < critical value, then do not reject H_0 (i.e. accept H_0)

Present conclusion in context of problem.

(6) Discuss conclusions in terms of which cells make the greatest contribution to the total value of the test statistic.

Example 2

A personnel manager is investigating whether there is any association between the length of service of the employees and the type of training they receive.

Carry out a hypothesis at the (a) 5% and (b) 1% significance level, to determine if there is any association between length of service and type of training.

Hypothesis testing procedure

(1) H_0: the variables are independent,
H_1: the variables are *not* independent**.**

(2) Significance level = 5% (and 1%)

(3) Records of a random sample of 200 employees are shown in the following ***contingency table*** of observed frequencies (f_o):

Type of training	*Length of service*			
	Short	Medium	Long	*Totals*
Induction course	14	23	13	50
Initial on-the-job	12	7	13	32
Continuous	28	32	58	118
Totals	54	62	84	200

(4) Marginal row and column totals are shown above.

Expected frequencies (f_e):

Type of training	*Length of service*			
	Short	Medium	Long	*Totals*
Induction course	13.5	15.5	21	50
Initial on-the-job	8.64	9.92	13.44	32
Continuous	31.86	36.58	49.56	118
Totals	54	62	84	200

Contributions to χ^2 : $\frac{(f_o - f_e)^2}{f_e}$

0.019	3.629	3.048
1.307	0.86	0.014
0.468	0.573	1.437

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

$= 0.019 + 3.629 + \ldots + 0.573 + 1.437$
$= 11.354$

Degrees of freedom $v = (3-1)(3-1) = 4$
Critical values: (a) 5%: 9.488, (b) 1%: 13.28

(5) At 5% level of significance:

(a) Since 11.354 > 9.488, reject H_0 (i.e. accept H_1), there is an association between the length of service and the type of training.

(b) Since 11.354 < 13.28, reject H_1 (i.e. accept H_0), there is no association between the length of service and the type of training.

(6) The cells with the largest values are medium/ induction and long/ induction, so medium and long service may seem to be associated, respectively, with more than and fewer than expected employees with induction-only training.

Acknowledgement: Material in this revision sheet was originally created by Bob Francis and we acknowledge his permission to reproduce it here.

REVISION SHEET – STATISTICS 2 (MEI)

NORMAL DISTRIBUTION

The main ideas are:

- Properties of the Normal Distribution
- Mean, SD and Var
- Approximating from other distributions

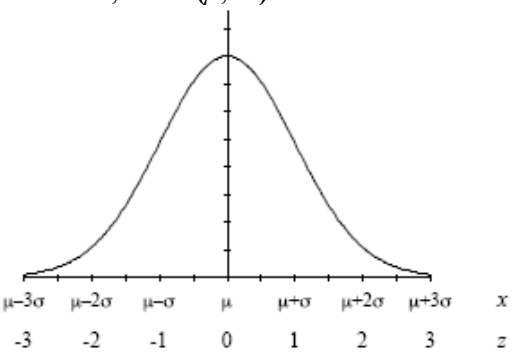
Before the exam you should know:

- All of the properties of the Normal Distribution.
- How to use the relevant tables.
- How to calculate mean, standard deviation and variance.
- How to approximate to the normal distribution from others.

Definition

A continuous random variable X which is bellshaped and has mean (expectation) μ and standard deviation σ is said to follow a **Normal Distribution** with **parameters** μ and σ.

In shorthand, $\boldsymbol{X \sim N(\mu, \sigma^2)}$



This may be given in *standardised* form by using the transformation

$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \sigma z + \mu \text{, where } Z \sim \mathbf{N}(0, 1)$$

Calculating Probabilities

The area to the left of the value z, representing $P(Z \le z)$, is denoted by $\Phi(z)$ and is read from tables for $z \ge 0$.

Useful techniques for $z \ge 0$:

- $P(Z > z) = 1 - P(Z \le z)$
- $P(Z > -z) = P(Z \le z)$
- $P(Z < -z) = 1 - P(Z \le z)$

The *inverse normal tables* may be used to find $z = \Phi^{-1}(p)$ for $p \ge 0.5$. For $p < 0.5$, use symmetry properties of the Normal distribution.

99.73% of values lie within 3 s.d. of the mean

Estimating μ and/or σ

Use (simultaneous) equations of the form: $x = \sigma z + \mu$ for matching (x, z) pairs – where z is given or may be deduced from $\Phi^{-1}(p)$ for given value(s) of x.

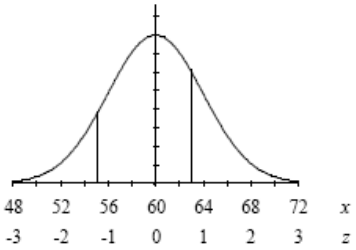
Example 1

$X \sim N(60, 16) \Rightarrow z = \frac{x-60}{4}$;

find (a) $P(X < 66)$, (b) $P(X \ge 66)$, (c) $P(55 \le X \le 63)$, (d) x_0 s.t. $P(X > x_0) = 99\%$

(a) $P(X < 66) = P(Z < 1.5) =$ **0.9332**

(b) $P(X \ge 66) = 1 - P(X < 66) = 1 - 0.9332 =$ **0.0668**



(c) $P(55 \le X \le 63) = P(-1.25 \le Z \le 0.75)$
$= P(Z \le 0.75) - P(Z < -1.25)$
$= P(Z \le 0.75) - P(Z > 1.25)$
$= P(Z \le 0.75) - [1 - P(Z \le 1.25)]$
$= 0.7734 - [1 - 0.8944] =$ **0.6678**

(d) $P(Z > -2.326) = 0.99$ from tables

Since $z = \frac{x-60}{4}$, $x = 4z + 60$

$\Rightarrow x_0 = 60 + 4\times(-2.326) =$ **50.7** (to 3 s.f.)

Example 2

For a certain type of apple, 20% have a mass greater than 130g and 30% have a mass less than 110g.

(a) Estimate μ and σ.

(b) When 5 apples are chosen at random, find the probability that all five have a mass exceeding 115g

(a) $P(Z > 0.8416) = 0.2$ $(X = 130)$
$P(Z < -0.5244) = 0.3$ $(X = 110)$
$\Rightarrow 130 = 0.8416\,\sigma + \mu$
$110 = -0.5244\,\sigma + \mu$
Solving equations simultaneously gives: $\boldsymbol{\mu = 117.68}$, $\boldsymbol{\sigma = 14.64}$

(b) $X \sim N(117.68, 14.64^2) \Rightarrow z = \frac{x-117.68}{14.64}$;
$P(X > 115)^5 = P(Z > -0.183)^5 = 0.5726^5 =$ **0.0616** (to 3 s.f.)

Acknowledgement: Material on this page was originally created by Bob Francis and we acknowledge his permission to reproduce it here.

Further examples

Example 1
Suppose that $X \sim N(12,4)$. Calculate $P(X<13)$.

Solution

$$P(X<13)=\Phi\left(\frac{13-12}{2}\right)=\Phi(0.5)=$$

This is the number of standard deviations between 13 and the mean, 12.

Example 2
Suppose that the time taken for a journey to work is normally distributed with a mean of 20 and a standard deviation of 3. Calculate the probability that a journey to work takes between 17 and 21 minutes.

Solution
Let X = time taken for the journey to work. Then $X \sim N(20,9)$.

$$P(17<X<21)=\text{Area A}\ -\ \text{Area B}$$

This is the number of standard deviations between 21 and the mean, 20.

This is the number of standard deviations between 17 and the mean, 20.

$$=\Phi\left(\frac{21-20}{3}\right)-\left(1-\Phi\left(\frac{20-17}{3}\right)\right)$$
$$=\Phi\left(\frac{1}{3}\right)-\left(1-\Phi(1)\right)$$
$$=0.6301-(1-0.8413)$$
$$=0.4714$$

Example 3
Suppose that $X \sim N(\mu,16)$. If $P(X>10)=0.3$, find μ.

Solution
Since $P(X>10)=0.3$ μ must be less than 10.

So, $$0.3=P(X>10)=1-\Phi\left(\frac{10-\mu}{4}\right) \text{ and so } \Phi\left(\frac{10-\mu}{4}\right)=0.7.$$

This gives that:

$$\frac{10-\mu}{4}=\Phi^{-1}(0.7)=0.5244$$
$$\Rightarrow \mu=10-(4\times 0.5244)=7.9024$$

Approximating the Binomial Distribution

Suppose $X \sim \text{Binomial}(n,p)$.
If n is large and p is not too close to 0 or 1 (i.e. the distribution is reasonably symmetrical), then using the mean (np) and variance (npq) of a binomial distribution we can approximate using the normal distribution.

$$X \sim N(np,npq)$$

Approximating the Poisson Distribution

Suppose $X \sim Poisson(\lambda)$
If λ is large, then the Poisson distribution is reasonably symmetrical.
Then using the mean (λ) and variance (λ) of a Poisson distribution we can approximate using the normal distribution.

$$X \sim N(\lambda,\lambda)$$

Important: In both cases above we are using a continuous distribution to approximate a discrete one and as such we must use continuity correcting when calculating a probability. Make sure you understand how to do this.

REVISION SHEET – STATISTICS 2 (MEI)

POISSON DISTRIBUTION

The main ideas are:

- Calculations using the Poisson Distribution
- Modelling the Binomial distribution with the Poisson distribution

Before the exam you should know:

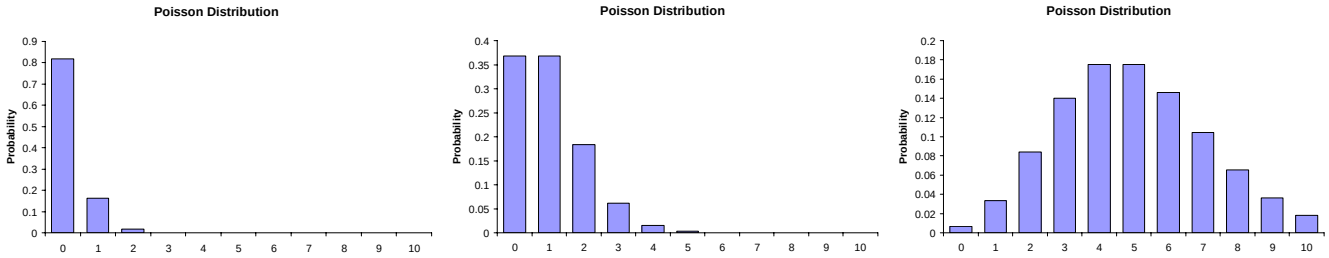
- When the Poisson distribution is an appropriate model for a given situation.
- The relationship $e^y = x \Leftrightarrow y = \ln x$, this is sometimes useful in questions.
- How to use the formula $P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}$ (without getting confused between λ and r).
- How to look up $P(X \le r)$ in the tables given.

Poisson Distribution

This models events which are random, independent, which occur singly and with a uniform likelihood.

If $X \sim$ Poisson(λ) then: $P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$, where E($X$) = $\mu = \lambda$ and Var(X) = $\sigma^2 = \lambda$.

The Poisson Distribution for various values of λ is shown below.



$\lambda = 0.2$ $\lambda = 1$ $\lambda = 5$

Calculations using the Poisson Distribution

You should be able to use the formula $P(X = r) = e^{-\lambda}\frac{\lambda^r}{r!}$ and the cumulative Poisson tables (which give $P(X \le r)$ for various values of λ) to find simple probabilities.

Example 1

The number of goals, X, scored by a team playing at home in the Premier League is modelled by a Poisson distribution with a mean of 1.6. What is the probability that the team scores

a) 3 goals in a game

b) More than 4 goals in a game

Solution

a) The probability of the team scoring 3 goals in a game is: $P(X = 3) = e^{-1.6}\frac{1.6^3}{3!} = 0.138$ (to 3 d.p).

b) The probability of the team scoring more than 4 goals in a game is:

$$P(X > 4) = 1 - P(X \le 4) = 1 - 0.9763 = 0.0237$$

More Complicated Questions

In other questions you will need to use the following properties of the Poisson Distribution:

If $X \sim \text{Poisson}(\lambda_1)$ and $Y \sim \text{Poisson}(\lambda_2)$ then: $nX \sim \text{Poisson}(n\lambda_1)$ and $X+Y \sim \text{Poisson}(\lambda_1+\lambda_2)$

Example 2
The mean number of burgers sold per minute at a snack bar is 0.7. The mean number of hotdogs sold per minute is 0.5. Using a Poisson distribution model calculate the probability that the burger bar sells:

a) 5 burgers in a 5 minute period.
b) No hot dogs or burgers in a 1 minute period.

Solution

a) The mean number of burgers sold in one minute is 0.7. Therefore the mean number of burgers sold in five minutes is $0.7 \times 5 = 3.5$. So, $5X$ is the number of burgers sold in 5 minutes and we have that $5X \sim \text{Poisson}(3.5)$. Therefore,

$$P(5X = 5) = e^{-3.5}\frac{3.5^5}{5!} = 0.132 \text{ (to 3 d.p)}$$

b) If X is the number of burgers sold in a minute then $X \sim \text{Poisson}(0.7)$. If Y is the number of burgers sold in a minute then $Y \sim \text{Poisson}(0.5)$. So the total number of hotdogs and burgers sold in a minute is $X + Y$ and $X+Y \sim \text{Poisson}(0.7 + 0.5 = 1.2)$. Therefore,

$$P(X + Y = 0) = e^{-1.2}\frac{1.2^0}{0!} = 0.301$$

Approximating the Binomial Distribution with the Poisson Distribution

If $X \sim \text{Binomial}(n, p)$ a Poisson approximation of $X \sim \text{Poisson}(np)$ can be used when

- n is large
- p is small (i.e. it is a rare event)

but it is only useful if np is not too large.

For example if $n = 1000$, $p = 0.002$, then $np = 2$. Under the binomial distribution $X \sim \text{Binomial}(1000, 0.02)$

$$P(X = 10) = {}^{1000}C_{10} \times 0.002^{10} \times 0.998^{990} = 0.000037 \text{ to (6 d.p)}$$

With the Poisson Distribution $X \sim \text{Poisson}(2)$

$$P(X = 10) = e^{-2}\frac{2^{10}}{10!} = 0.000038 \text{ (to 6 d.p)}$$

The difference between these two values is only 0.000001