# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 1**

**Question:**

Given the random variables $X \sim \mathrm{N}(80, 3^2)$ and $Y \sim \mathrm{N}(50, 2^2)$ where $X$ and $Y$ are independent find the distribution of $W$ where:

**a** $W = X + Y$,

**b** $W = X - Y$.

**Solution:**

**a**
$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(X) + \mathrm{E}(Y) \\ &= 80 + 50 \\ &= 130 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(W) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ &= 9 + 4 \\ &= 13 \end{aligned}$$

$$W \sim \mathrm{N}(130, 13)$$

**b**
$$\begin{aligned} \mathrm{E}(W) &= \mathrm{E}(X) - \mathrm{E}(Y) \\ &= 80 - 50 \\ &= 30 \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(W) &= \mathrm{Var}(X) + \mathrm{Var}(Y) \\ &= 9 + 4 \\ &= 13 \end{aligned}$$

$$W \sim \mathrm{N}(30, 13)$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 2**

**Question:**

Given the random variables $X \sim \mathrm{N}(45,6), Y \sim \mathrm{N}(54,4)$ and $W \sim \mathrm{N}(49,8)$ where $X$, $Y$ and $W$ are independent, find the distribution of $R$ where $R = X + Y + W$.

**Solution:**

$$\begin{aligned}
\mathrm{E}(R) &= \mathrm{E}(X) + \mathrm{E}(Y) + \mathrm{E}(W) \\
&= 45 + 54 + 49 \\
&= 148 \\
\mathrm{Var}(R) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(W) \\
&= 6 + 4 + 8 \\
&= 18 \\
R &\sim \mathrm{N}(148, 18)
\end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 3**

**Question:**

$X_1$ and $X_2$ are independent normal random variables. $X_1 \sim \mathrm{N}(60, 25)$ and $X_2 \sim \mathrm{N}(50,16)$. Find the distribution of $T$ where:

**a** $T = 3X_1$,

**b** $T = 7X_2$,

**c** $T = 3X_1 + 7X_2$,

**d** $T = X_1 - 2X_2$.

**Solution:**

**a**
$$\begin{aligned}\mathrm{E}(T) &= 3\mathrm{E}(X_1)\\ &= 3\times 60\\ &= 180\\ \mathrm{Var}\,(T) &= 9\ \mathrm{Var}(X_1)\\ &= 9\times 25\\ &= 225\end{aligned}$$
$$T \sim \mathrm{N}(180,\ 225) \text{ or } \mathrm{N}(180,\ 15^2)$$

**b**
$$\begin{aligned}\mathrm{E}(T) &= 7\mathrm{E}(X_2)\\ &= 7\times 50\\ &= 350\\ \mathrm{Var}\,(T) &= 49\ \mathrm{Var}(X_2)\\ &= 49\times 16\\ &= 784\end{aligned}$$
$$T \sim \mathrm{N}(350,\ 784) \text{ or } \mathrm{N}(350,\ 28^2)$$

**c**
$$\begin{aligned}\mathrm{E}(T) &= \mathrm{E}(3X_1) + \mathrm{E}(7X_2)\\ &= 180 + 350\\ &= 530\\ \mathrm{Var}\,(T) &= \mathrm{Var}(3X_1) + \mathrm{Var}(7X_2)\\ &= 225 + 784\\ &= 1009\end{aligned}$$
$$T \sim \mathrm{N}(530,\ 1009)$$

**d**
$$\begin{aligned}\mathrm{E}(T) &= \mathrm{E}(X_1) - 2\mathrm{E}(X_2)\\ &= 60 - 2\times 50\\ &= -40\\ \mathrm{Var}\,(T) &= \mathrm{Var}(X_1) + 4\mathrm{Var}(X_2)\\ &= 25 + 4\times 16\\ &= 89\end{aligned}$$
$$T \sim \mathrm{N}(-40,\ 89)$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 4**

**Question:**

$Y_1, Y_2$ and $Y_3$ are independent normal random variables.

$Y_1 \sim \text{N}(8,2), Y_2 \sim \text{N}(12,3)$ and $Y_3 \sim \text{N}(15,4)$. Find the distribution of $A$ where:

**a** $A = Y_1 + Y_2 + Y_3$,

**b** $A = Y_3 - Y_1$,

**c** $A = Y_1 - Y_2 + 3Y_3$,

**d** $A = 3Y_1 + 4Y_3$,

**e** $A = 2Y_1 - Y_2 + Y_3$.

**Solution:**

**a**
$$\begin{aligned}\mathrm{E}(A) &= \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \mathrm{E}(Y_3)\\ &= 8 + 12 + 15\\ &= 35\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(A) &= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \mathrm{Var}(Y_3)\\ &= 2 + 3 + 4\\ &= 9\end{aligned}$$
$$A \sim \mathrm{N}(35, 9) \text{ or } \mathrm{N}(35, 3^2)$$

**b**
$$\begin{aligned}\mathrm{E}(A) &= \mathrm{E}(Y_3) - \mathrm{E}(Y_1)\\ &= 15 - 8\\ &= 7\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(A) &= \mathrm{Var}(Y_3) + \mathrm{Var}(Y_1)\\ &= 4 + 2\\ &= 6\end{aligned}$$
$$A \sim \mathrm{N}(7, 6)$$

**c**
$$\begin{aligned}\mathrm{E}(A) &= \mathrm{E}(Y_1) - \mathrm{E}(Y_2) + 3\mathrm{E}(Y_3)\\ &= 8 - 12 + 3\times 15\\ &= 41\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(A) &= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + 9\mathrm{Var}(Y_3)\\ &= 2 + 3 + 9\times 4\\ &= 41\end{aligned}$$
$$A \sim \mathrm{N}(41, 41)$$

**d**
$$\begin{aligned}\mathrm{E}(A) &= 3\mathrm{E}(Y_1) + 4\mathrm{E}(Y_3)\\ &= 3\times 8 + 4\times 15\\ &= 84\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(A) &= 9\mathrm{Var}(Y_1) + 16\mathrm{Var}(Y_3)\\ &= 9\times 2 + 16\times 4\\ &= 82\end{aligned}$$
$$A \sim \mathrm{N}(84, 82)$$

**e**
$$\begin{aligned}\mathrm{E}(A) &= 2\mathrm{E}(Y_1) - \mathrm{E}(Y_2) + 3\mathrm{E}(Y_3)\\ &= 2\times 8 - 12 + 15\\ &= 19\end{aligned}$$
$$\begin{aligned}\mathrm{Var}(A) &= 4\mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \mathrm{Var}(Y_3)\\ &= 4\times 2 + 3 + 4\\ &= 15\end{aligned}$$
$$A \sim \mathrm{N}(19, 15)$$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 5**

**Question:**

$A$, $B$ and $C$ are independent normal random variables.
$A \sim \mathrm{N}(50, 6)$, $B \sim \mathrm{N}(60, 8)$ and $C \sim \mathrm{N}(80, 10)$. Find:

**a** $\mathrm{P}(A + B < 115)$,

**b** $\mathrm{P}(A + B + C > 198)$,

**c** $\mathrm{P}(B + C < 138)$,

**d** $\mathrm{P}(2A + B - C < 70)$,

**e** $\mathrm{P}(A + 3B - C > 140)$,

**f** $\mathrm{P}(105 < (A + B) < 116)$.

**Solution:**

**a** $A+B \sim \mathrm{N}(50+60,\ 6+8) = \mathrm{N}(110,\ 14)$

$$\mathrm{P}(A+B < 115) = \mathrm{P}\left(z < \frac{115-110}{\sqrt{14}}\right)$$
$$= \mathrm{P}\ (z < 1.34)$$
$$= 0.9099\ (\ 0.9093)$$

Answers which round to (awrt) 0.91

**b** $A+B+C \sim \mathrm{N}(50+60+80,\ 6+8+10) = \mathrm{N}(190,\ 24)$

$$\mathrm{P}(A+B+C > 198) = \mathrm{P}\left(z > \frac{198-190}{\sqrt{24}}\right)$$
$$= \mathrm{P}\ (z < 1.63)$$
$$= \ 1 - 0.9484$$
$$= 0.0516\ \ (0.0512)$$

**c** $B+C \sim \mathrm{N}(60+80,\ 8+10) = \mathrm{N}(140,\ 18)$

$$\mathrm{P}(B+C < 138) = \mathrm{P}\left(z < \frac{138-140}{\sqrt{18}}\right)$$
$$= \mathrm{P}\ (z < -0.47)$$
$$= 1 - 0.6808$$
$$= 0.3192\ (0.3186)$$

Awrt 0.319

**d** $2A+B-C \sim \mathrm{N}(2\times 50+60-80,\ 4\times 6+8+10) = \mathrm{N}(80,\ 42)$

$$\mathrm{P}(2A+B-C < 70) = \mathrm{P}\left(z < \frac{70-80}{\sqrt{42}}\right)$$
$$= \mathrm{P}\ (z < -1.54)$$
$$= \ 1 - 0.9382$$
$$= 0.0618\ \ (0.0614)$$

**e** $A+3B-C \sim \mathrm{N}(50+3\times 60-80,\ 6+9\ \times\ 8+10) = \mathrm{N}(150,\ 88)$

$$\mathrm{P}(A+3B-C > 140) = \mathrm{P}\left(z > \frac{140-150}{\sqrt{88}}\right)$$
$$= \mathrm{P}\ (z > -1.07)$$
$$= \ 0.8577\ (0.8568)$$

Awrt 0.858 (0.857)

**f** $A+B \sim \mathrm{N}(50+60,\ 6+8) = \mathrm{N}(110,\ 14)$

$$\mathrm{P}(105 < A+B < 116) = \mathrm{P}\left(\frac{105-110}{\sqrt{14}} < z < \frac{116-110}{\sqrt{14}}\right)$$
$$= \mathrm{P}\ (-1.34 < z < 1.60)$$
$$= 0.9452 - (1 - 0.9099)$$
$$= 0.8551\ (\ 0.8549)$$

Awrt 0.855

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 6**

**Question:**

Given the random variables $X \sim N(20, 5)$ and $Y \sim N(10, 4)$ where $X$ and $Y$ are independent, find

**a** $E(X - Y)$,

**b** $Var(X - Y)$,

**c** $P(13 < X - Y < 16)$. ***E***

**Solution:**

**a** $E(X - Y) = 20 - 10 = 10$

**b** $Var(X - Y) = 5 + 4 = 9$

**c** $X - Y \sim N(10, 9)$

$$\begin{aligned} P(13 < X - Y < 16) &= P(X - Y < 16) - P(X - Y < 13) \\ &= P(z < 2) - P(z < 1) \\ &= 0.9772 - 0.8413 \\ &= 0.1359 \\ &\text{Awrt } 0.136 \end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 7**

**Question:**

The random variable $R$ is defined as $R = X + 4Y$ where $X \sim N(8, 2^2)$, $Y \sim N(14, 3^2)$ and $X$ and $Y$ are independent.
Find

**a** $E(R)$,

**b** $Var(R)$,

**c** $P(R < 41)$.

The random variables $Y_1$, $Y_2$ and $Y_3$ are independent and each has the same distribution as $Y$. The random variable $S$ is defined as

$$S = \sum_{i=1}^{3} Y_i - \frac{1}{2}X.$$

**d** Find Var ($S$). ***E***

**Solution:**

**a** $E(R) = E(X) + 4E(Y) = 8 + (4 \times 14) = 64$

**b** $\text{Var}(R) = \text{Var}(X) + 16\ \text{Var}(Y) = 2^2 + (16 \times 3^2)$
$= 148$

**c** $P(R < 41) = P\left(Z < \frac{41 - 64}{\sqrt{148}}\right) = P(Z < -1.89)$
$= 0.0294\ (0.0293)$

**d** $S = Y_1 + Y_2 + Y_3 - 0.5X$
$\text{Var}(S) = 3\ \text{Var}(Y) + (\frac{1}{2})^2\ \text{Var}(X)$
$= 27 + 1$
$= 28$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 8**

**Question:**

A factory makes steel rods and steel tubes. The diameter of a steel rod is normally distributed with mean 3.55 cm and standard deviation 0.02 cm. The internal diameter of a steel tube is normally distributed with mean 3.60 cm and standard deviation 0.02 cm.
A rod and a tube are selected at random. Find the probability that the rod cannot pass through the tube. ***E***

**Solution:**

$T \sim \mathrm{N}(3.60, 0.02^2)$ $R \sim \mathrm{N}(3.55, 0.02^2)$
$\mathrm{P}(T < R) = \mathrm{P}(T - R < 0)$
$\mathrm{E}(T - R) = 3.60 - 3.55 = 0.05$
$\mathrm{Var}(T - R) = 0.02^2 + 0.02^2 = 0.0008$

$$\begin{aligned}\mathrm{P}(T - R < 0) &= \mathrm{P}\left(z < \frac{0 - 0.05}{\sqrt{0.0008}}\right)\\ &= \mathrm{P}(z < -1.77)\\ &= 1 - 0.9616\\ &= 0.0384\ (0.0385)\end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 9**

**Question:**

The weight of a randomly selected tin of jam is normally distributed with a mean weight of 1 kg and a standard deviation of 12 g. The tins are packed in boxes of 6 and the weight of the box is normally distributed with mean weight 250 g and standard deviation 10 g. Find the probability that a randomly chosen box of 6 tins will weigh less than 6.2 kg. ***E***

**Solution:**

$T \sim \mathrm{N}(1000, 12^2) \quad B \sim \mathrm{N}(250, 10^2)$

$Y = T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + B_1$

$\mathrm{E}(Y) = 6 \times 1000 + 250 = 6250$

$\mathrm{Var}(Y) = 6 \times 12^2 + 10^2 = 964$

$$\begin{aligned} \mathrm{P}(Y < 6200) &= \mathrm{P}\left(z < \frac{6200 - 6250}{\sqrt{964}}\right) \\ &= \mathrm{P}(z < -1.61) \\ &= 1 - 0.9463 \\ &= 0.0537 \end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 10**

**Question:**

The thickness of paperback books can be modelled as a normal random variable with mean 2.1 cm and variance $0.39\ \text{cm}^2$. The thickness of hardback books can be modelled as a normal random variable with mean 4.0 cm and variance $1.56\ \text{cm}^2$. A small bookshelf is 30 cm long.

**a** Find the probability that a random sample of
 **i** 15 paperback books can be placed side-by-side on the bookshelf,
 **ii** 5 hardback and 5 paperback books can be placed side-by-side on the bookshelf.

**b** Find the shortest length of bookshelf needed so that there is at least a 99% chance that it will hold a random sample of 15 paperback books. ***E***

**Solution:**

$P \sim N(2.1, 0.39)$ $H \sim N(4.0, 1.56)$

**a i** $Y = P_1 + P_2 + P_3 + \dots + P_{15}$

$$E(Y) = 15 \times 2.1 = 31.5$$
$$Var(Y) = 15 \times 0.39 = 5.85$$
$$P(Y < 30) = P\left(z < \frac{30 - 31.5}{\sqrt{5.85}}\right)$$
$$= P(z < -0.62)$$
$$= 1 - 0.7324$$
$$= 0.2676$$

Awrt 0.268

**ii** $Y = P_1 + P_2 + \dots + P_5 + H_1 + H_2 + \dots + H_5$

$$E(Y) = 5 \times 2.1 + 5 \times 4.0 = 30.5$$
$$Var(Y) = 5 \times 0.39 + 5 \times 1.56 = 9.75$$
$$P(Y < 30) = P\left(z < \frac{30 - 30.5}{\sqrt{9.75}}\right)$$
$$= P(z < -0.16)$$
$$= 1 - 0.5636$$
$$= 0.4364$$

Awrt 0.436

**b** $Y = P_1 + P_2 + P_3 + \dots + P_{15}$

$$E(Y) = 15 \times 2.1 = 31.5$$
$$Var(Y) = 15 \times 0.39 = 5.85$$
$$P(Y < x) > 0.99$$
$$P\left(z < \frac{x - 31.5}{\sqrt{5.85}}\right) > 0.99$$
$$\frac{x - 31.5}{\sqrt{5.85}} > 2.3263$$
$$x = 37.1$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 11**

**Question:**

A sweet manufacturer produces two varieties of fruit sweet, Xtras and Yummies. The weights, $X$ and $Y$ in grams, of randomly selected Xtras and Yummies are such that

$X \sim \mathrm{N}(30,25)$ and $Y \sim \mathrm{N}(32,16)$.

**a** Find the probability that the weight of two randomly selected Yummies will differ by more than 5 g.

One sweet of each variety is selected at random.

**b** Find the probability that the Yummy sweet weighs more than the Xtra.

A packet contains 6 Xtras and 4 Yummies.

**c** Find the probability that the average weight of the sweets in the packet lies between 28 g and 33 g. ***E***

**Solution:**

**a** $W = Y_1 - Y_2$

$\mathrm{E}(W) = \mathrm{E}(Y_1) - \mathrm{E}(Y_2) = 0$

$\mathrm{Var}(W) = \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) = 32$

$$\mathrm{P}(|W| > 5) = \mathrm{P}(W < -5) + \mathrm{P}(W > 5)$$
$$= \mathrm{P}\left(z < \frac{-5-0}{\sqrt{32}}\right) + \mathrm{P}\left(z > \frac{5-0}{\sqrt{32}}\right)$$
$$= \mathrm{P}(z < -0.88) + \mathrm{P}(z > 0.88)$$
$$= (1 - 0.8106) + (1 - 0.8106)$$
$$= 0.3788\ (0.3768)$$

Awrt 0.379/0.377

**b** $W = Y_1 - X_1$

$\mathrm{E}(W) = \mathrm{E}(Y_1) - \mathrm{E}(X_1) = 2$

$\mathrm{Var}(W) = \mathrm{Var}(Y_1) + \mathrm{Var}(X_1) = 41$

$$\mathrm{P}(W > 0) = \mathrm{P}\left(z > \frac{0-2}{\sqrt{41}}\right)$$
$$= \mathrm{P}(z > -0.31)$$
$$= 0.6217\ (0.6226)$$

Awrt 0.622/0.623

**c** $W = X_1 + X_2 + \ldots + X_6 + Y_1 + Y_2 + \ldots + Y_4$

$\mathrm{E}(W) = 6\mathrm{E}(X) + 4\mathrm{E}(Y) = 308$

$\mathrm{Var}(W) = 6\mathrm{Var}(X) + 4\mathrm{Var}(Y) = 214$

Since the total weight of the 10 sweets must be between 280 g and 330 g

$$\mathrm{P}(28 < \overline{W} < 33) = \mathrm{P}\left(\frac{280-308}{\sqrt{214}} < z < \frac{330-308}{\sqrt{214}}\right)$$
$$= \mathrm{P}(-1.91 < z < 1.50)$$
$$= 0.9332 - (1 - 0.9719)$$
$$= 0.9051\ (0.9059)$$

Awrt 0.905/0.906

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Combinations of random variables**
**Exercise A, Question 12**

**Question:**

If $X_1, X_2, \ldots, X_n$, are independent random variables, each with mean $\mu$ and variance $\sigma^2$, and the random variable $Z$ is defined as $Z = X_1 + X_2 + \ldots + X_n$, show that $\mathrm{E}(Z) = n\mu$ and $\mathrm{Var}(Z) = n\sigma^2$.

A certain brand of biscuit is individually wrapped. The weight of a biscuit can be taken to be normally distributed with mean 75 g and standard deviation 5 g. The weight of an individual wrapping is normally distributed with mean 10 g and standard deviation 2 g. Six of these individually wrapped biscuits are then packed together. The weight of the packing material is a normal random variable with mean 40 g and standard deviation 3 g. Find, to 3 decimal places, the probability that the total weight of the packet lies between 535 g and 565 g. ***E***

**Solution:**

$$\begin{aligned}\mathrm{E}(Z) &= \mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_n)\\ &= \mu + \mu + \ldots + \mu\\ &= n\mu\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n)\\ &= \sigma^2 + \sigma^2 + \ldots + \sigma^2\\ &= n\sigma^2\end{aligned}$$

$B \sim \mathrm{N}(75, 5^2) \quad W \sim \mathrm{N}(10, 2^2) \quad X \sim \mathrm{N}(40, 3^2)$

$W = B_1 + B_2 + \ldots + B_6 + W_1 + W_2 + \ldots + W_6 + X$

$\mathrm{E}(W) = 6\mathrm{E}(B) + 6\mathrm{E}(W) + \mathrm{E}(X) = 550$

$\mathrm{Var}(W) = 6\mathrm{Var}(B) + 6\mathrm{Var}(W) + \mathrm{Var}(X) = 183$

$$\begin{aligned}\mathrm{P}(535 < W < 565) &= \mathrm{P}\left(\frac{535-550}{\sqrt{183}} < z < \frac{565-550}{\sqrt{183}}\right)\\ &= \mathrm{P}(-1.11 < z < 1.11)\\ &= 0.8665 - (1 - 0.8665)\\ &= 0.733\ (0.732)\end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 1**

**Question:**

Explain briefly what is meant by the term sampling and give three advantages of taking a sample as opposed to a census.

**Solution:**

The selection of individual elements of a population. Advantages: low cost, results obtained faster than a census, represents whole population, can be more reliable.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 2**

**Question:**

Define what is meant by a census. By referring to specific examples, suggest two reasons why a census might be used.

**Solution:**

Every item is observed/measured, e.g. National census for forecasting school places.
Census of a nursery school for numbers of carriers of a virus.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 3**

**Question:**

A factory makes safety harnesses for climbers and has an order to supply 3000 harnesses. The buyer wishes to know that the load at which the harness breaks exceeds a certain figure.
Suggest a reason why a census would not be used for this purpose.

**Solution:**

All the harnesses would be destroyed in testing them.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 4**

**Question:**

Explain:

**a** why a sample might be preferred to a census,
**b** what you understand by a sampling frame,
**c** what effect the size of the population has on the size of the sampling frame,
**d** what effect the variability of the population has on the size of the sampling frame.

**Solution:**

**a** Cheaper, quicker.

**b** The sampling frame is a list of all the members in the population (i.e. all the sampling units).

**c** The larger the population, the longer the list, so the larger the sampling frame.

**d** The variability of the population doesn't change how many members are in it, the list remains the same size so the size of the sampling frame isn't affected.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 5**

**Question:**

Using the random numbers 4 and 3 to give you the column and line respectively in the random number table (table $x$, p. 139), select a sample of size 6 from the numbers:

**a** 0–99
**b** 50–150
**c** 1–600

**Solution:**

**a** 01, 06, 64, 65, 94, 41 or 01, 18, 70, 97, 99, 56: others are possible.

**b** 079, 056, 110, 086, 143, 108 or 136, 097, 148, 069, 137, 123; others are possible.

**c** 010, 441, 172, 193, 249, 569 or 010, 184, 561, 547, 570, 278.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise A, Question 6**

**Question:**

A company wishes to do consumer marketing research using a certain town. Suggest a suitable sampling frame and describe in detail a way of selecting a sample of 400 people aged over 21.

**Solution:**

Electoral roll, could use systematic sampling (large number in population)
(see next section).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 1**

**Question:**

Explain briefly the difference between a census and a sample survey.
Write brief notes on:

**a** simple random sampling,
**b** stratified sampling,
**c** systematic sampling,
**d** quota sampling.

Your notes should include the definition, and any advantages and disadvantages associated with each method of sampling

**Solution:**

A census is where information is obtained from every member of the population.
A sample is a sub-set of the population's information.

**a** A simple random sample is a sample that is taken so that every member of the population has an equal chance of being included, all sub-groups of size $n$ are equally likely to be chosen and sampling is taken without replacement. It can be difficult in a large population.

**b** A stratified sample is where the population is divided into mutully exclusive groups and is where the proportion of the sample in each of these groups is the same as the proportion with which the group occurs in the population. It is good in easily stratified populations but you need to know the structure of the population. It is the most representative statistically. Within each strata members are selected by random sampling.

**c** Choosing at regular intervals from an ordered list; good for large populations and easy to use; bias in ordered list can be a problem.

**d** A quota sample has the proportions as for a stratified sample (see part **b**), but the 'quota' in each group allows the interviewer to choose who to interview to fill the quota required in the group. This is faster and cheaper than stratified sampling but can introduce 'interviewer bias' in who is interviewed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 2**

**Question:**

**a** Explain the purpose of stratification in carrying out a sample survey.
**b** The headteacher of an infant school wishes to take a stratified sample of 20% of the pupils at his school. The school has the following numbers of pupils.

| Year 1 | Year 2 | Year 3 |
|---|---|---|
| 40 | 60 | 80 |

Work out how many pupils in each age group there will be in the sample.

**Solution:**

**a** Divides population into mutually exclusive groups, where the proportion in each group in the sample is the same as that in the population. It is used so that the 'results' obtained will represent the 'results' of the whole population as closely as possible because this 'stratification' reduces bias.

**b** Year1: 8, Year 2: 12 and Year 3: 16

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 3**

**Question:**

A survey is to be done on the adult population of a certain city suburb, the population of which is 2000. An ordered list of the inhabitants is available.

**a** What sampling method would you use and why?

**b** What condition would have to be applied to your ordered list if the selection is to be truly random?

**Solution:**

**a** Systematic sampling would be easy to use.

**b** The ordered list would need to be truly random.

Note that if the structure of the sample of 2000 were known, it would be possible (and desirable) to take a stratified sample. With such a small sampling frame it would be possible to choose a simple random sample using random number tables, which would be truly random!

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 4**

**Question:**

In a marketing sample survey the sales of cigarettes in a variety of outlets is to be investigated. The outlets consist of small kiosks selling cigarettes and tobacco only, tobacconist's shops that sell cigarettes and related products and shops that sell cigarettes and other unrelated products.

**a** Suggest the most suitable form of taking a random sample.

**b** Explain how you would conduct the sample survey.

**c** What are the advantages and disadvantages of the method chosen?

**Solution:**

**a** Stratified – three strata.

**b** Questionnaire to 10% of each strata.

**c** Advantages: information on each strata, results should reflect those of the population.

Disadvantages: not suitable if sample size is large, strata may overlap if not already defined.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 5**

**Question:**

**a** Explain briefly:
  **i** why it is often desirable to take samples,
  **ii** what you understand by a sampling frame.

**b** State two circumstances when you would consider using
  **i** systematic sampling,
  **ii** stratification when sampling from a population,
  **iii** quota sampling.

**Solution:**

**a** **i** Cost, takes less time than a census.

  **ii** List of sampling units used to represent the population.

**b** Many possible answers, e.g.

  **i** shoppers at a supermarket when considering shopping
  **ii** students in a secondary school when looking at school meals
  **iii** voters asked which party they are going to vote for.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise B, Question 6**

**Question:**

A factory manager wants to get information about the ways his workers travel to work. There are 480 workers in the factory, and each has a clocking in number. The numbers go from 1 to 480. Explain how the manager could take a systematic sample of size 30 from these workers.

**Solution:**

$$k = \frac{480}{30} = 16$$

Randomly select a number between 1 and 16. Starting with the worker with this clock number select the workers that have every 16th clock number after this.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 1**

**Question:**

Using the random numbers on page 139, and starting at the top of the 11th column with the number 88 and working down, a simple random sample (without replacement) of size 10 was taken of numbers between 0 and 75 inclusive. The first two numbers were 17 and 52.

**a** Find the other eight numbers in the sample.

**b** Explain, with the aid of a practical situation, how this set of random numbers could be used to take a sample of size 10.

**Solution:**

**a** 12, 60, 73, 9, 41, 20, 04, 36

**b** Say the population was a school year of size 76, each member of this population is written down in alphabetical order and numbered. Students whose numbers correspond to the numbers from the table are selected. Repeats are ignored.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 2**

**Question:**

**a** Give one advantage and one disadvantage of using
  **i** a census,
  **ii** a sample survey.
**b** It is decided to take a sample of 100 from a population consisting of 500 elements. Explain how you would obtain a simple random sample without replacement from this population.

**Solution:**

**a** **i** Advantage: very accurate; disadvantage: expensive (time consuming).

  **ii** Advantage: easier data collection (quick, cheap); disadvantage: possible bias.

**b** Assign unique 3-digit identifiers 000, 001, ..., 499 to each member of the population. Work along rows of random number tables generating 3-digit numbers. If these correspond to an identifier then include the corresponding member in the sample; ignore repeats and numbers greater than 499. Repeat this process until the sample contains 100 members.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 3**

**Question:**

**a** Explain briefly what you understand by
  **i** a population,
  **ii** a sampling frame.

**b** A market research organisation wants to take a sample of
  **i** owners of diesel motor cars in the UK,
  **ii** persons living in Oxford who suffered from injuries to the back during July 1996.

Suggest a suitable sampling frame in each case.

**Solution:**

**a** **i** Collection of individual items.

**ii** List of sampling units.

**b** **i** List of registered owners from DVLA.

**ii** List of people visiting a doctor's clinic in Oxford in July 1996.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 4**

**Question:**

A gym keeps a numbered alphabetical list of their 200 clients.
Explain how you would choose a simple random sample of 40 clients.

**Solution:**

Assign unique 3-digit identifiers 000, 001, .., 199 to the clients. Work along random tables rows generating 3-digit numbers – include the members corresponding to these numbers in the sample, ignoring repeats and numbers larger than 199. Repeat this until the sample includes 40 members.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 5**

**Question:**

Write down one advantage and one disadvantage of using:

**a** stratified sampling,

**b** simple random sampling.

**Solution:**

**a** Advantage – the results are the most representative of the population since the structure of the sample reflects the structure of the population.
Disadvantage – you need to know the structure of the population before you can take a stratified sample.

**b** Advantage – quick and cheap.
Disadvantage – can introduce bias (e.g. if the sample, by chance, only includes very tall people in an investigation into heights of students).

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Sampling**

**Exercise C, Question 6**

**Question:**

The managing director of a factory wants to know what the workers think about the factory canteen facilities. One hundred people work in the offices and 200 work on the shop floor.

He decides to ask the people who work in the offices.

**a** Suggest reasons why this is likely to produce a biased sample.

**b** Explain briefly how the factory manager could select a sample of 30 workers using:

**i** systematic sampling,

**ii** stratified sampling,

**iii** quota sampling.

**Solution:**

**a** People not in office not represented.

**b** **i** Get a list of the 300 workers at the factory. $\frac{300}{30} = 10$ so choose one of the first ten workers on the list at random and every subsequent 10th worker on the list, e.g. if person 7 is chosen, then the sample includes workers 7, 17, 27, ..., 297.

**ii** The population contains 100 office workers ($\frac{1}{3}$ of population) and 200 shop floor workers ($\frac{2}{3}$ of population).

The sample should contain $\frac{1}{3} \times 30 = 10$ office workers and $\frac{2}{3} \times 30 = 20$ shop floor workers. The 10 office workers in the sample should be a simple random sample of the 100 office workers. The 20 shop floor workers should be a simple random sample of the 200 shop floor workers.

**iii** Decide the categories e.g. age, gender, office/non-office and set a quota for each in proportion to their numbers in the population. Interview workers until quotas are full.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 7**

**Question:**

A garden centre employs 150 workers. Sixty-five of the workers are women and 85 are men. Explain briefly how you would take a random sample of 30 workers using stratified sampling.

**Solution:**

List separately men and women. take a simple sample of $\frac{65 \times 30}{150} = 13$ women and 17 men.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 8**

**Question:**

The 240 members of a bowling club are listed alphabetically in the club's membership book. The committee wishes to select a sample of 30 members to fill in a questionnaire about the facilities the club has to offer.

**a** Explain how the committee could use a table of random numbers to take a systematic sample.

**b** Give one advantage of this method over taking a simple random sample.

**Solution:**

**a** Label members $1 \rightarrow 240$. Use random numbers to select first from 1–8. Select every 8th member (e.g. 6,14, 22, …)

**b** For example: more convenient, efficient, faster, simpler to carry out etc.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Sampling
**Exercise C, Question 9**

**Question:**

**a** Explain briefly what you understand by:

  **i** a population,

  **ii** a sample.

**b** Give one advantage and one disadvantage of taking a sample.

**Solution:**

**a** **i** The group of all the individuals or items of interest in the investigation.

  **ii** A selection of individual members of the population.

**b** Advantage: reduction in the amount of data to be analysed; disadvantage: loss of reliability and accuracy (relative to whole population).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 10**

**Question:**

A college of 3000 students has students registered in four departments: arts, science, education and crafts. The principal wishes to take a sample from the student population to gain information about the likely student response to a rearrangement of the college timetable in order to hold lectures on Wednesday, previously reserved for sports.
What sampling method would you advise the principal to use? Give reasons to justify your choice.

**Solution:**

Stratified – gives all groups an equal chance to give their views.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Sampling**
**Exercise C, Question 11**

**Question:**

As part of her statistics project, Deepa decided to estimate the amount of time A-level students at her school spent on private study each week. She took a random sample of students from those studying arts subjects, science subjects and a mixture of arts and science subjects. Each student kept a record of the time they spent on private study during the third week of term.

**a** Write down the name of the sampling method used by Deepa.

**b** Give a reason for using this method and give one advantage this method has over simple random sampling.

**Solution:**

**a** Stratified sampling

**b** Uses naturally occurring (strata) groupings. The results are more likely to represent the views of the population since the sample reflects its structure.

e.g. variance of estimator of population mean is usually reduced, either individual strata estimates available.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Sampling
**Exercise C, Question 12**

**Question:**

There are 64 girls and 56 boys in a school.
Explain briefly how you could take a random sample of 15 pupils using

**a** a simple random sample,

**b** a stratified sample.

**Solution:**

**a** Allocate a number between 1 and 120 to each pupil.
Use random number tables, computer or calculator to select 15 different numbers between 1 and 120 (or equivalent).

Pupils corresponding to these numbers become the sample.

**b** Allocate numbers 1–64 to girls and 65–120 to boys. Idea of different sets for boys and girls

Select $\frac{64}{120} \times 15 = 8$ different random numbers between 1–64 for girls

Select 7 different random numbers between 65–120 for boys.
Include the corresponding boys and girls in the sample.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise A, Question 1**

**Question:**

The random variable $H \sim \mathrm{N}(\mu, \sigma^2)$ represents the height of a variety of flower where $\mu, \sigma^2$ are unknown population parameters.

A random sample of 5 flowers of this variety are measured and their height, in cm, is given below.

$h_1 = 35.1, h_2 = 32.3, h_3 = 34.5, h_4 = 37.4, h_5 = 32.8$

Determine which of the following are statistics.

**a** $\sum_{i=1}^{5}\left(X_i - \mu\right)$

**b** $\sum_{i=1}^{5}\frac{\left(X_i - \bar{X}\right)^2}{4}$

**c** $\sum\left|\frac{X_i - \mu}{\sigma}\right|$

**d** $X_1 - X_5$

**Solution:**

**a** Not a statistic since $\mu$ not known

**b** Is a statistic – no unknown parameters

**c** Not a statistic since $\mu$, $\sigma$ not known

**d** Is a statistic – no unknown parameters

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise A, Question 2**

**Question:**

A random sample of 6 apples are weighed and their weights, $x_i$ g, are recorded

$x_1 = 168, x_2 = 185, x_3 = 161, x_4 = 172, x_5 = 187, x_6 = 176$

Calculate the values of the following statistics.

**a** $\dfrac{X_6 + X_1}{2}$

**b** $\displaystyle\sum_{i=1}^{6} \frac{\left(X_i - \bar{X}\right)^2}{6}$

**c** $\dfrac{\sum_{i=1}^{6} X_i^2}{\sum_{i=1}^{6} X_i}$

**Solution:**

**a** $\dfrac{X_6 + X_1}{2} = \dfrac{176 + 168}{2} = 172$

**b** $\bar{X} = \dfrac{168 + 185 + \cdots + 176}{6} = \dfrac{1049}{6} = 174.83$

$$\sum \frac{\left(X_i - \bar{X}\right)^2}{6} = 83.138\cdots \quad \left(\sigma x^2 \text{ on a calculator}\right)$$
$$= 83.1 \text{ (3 s.f.)}$$

**c**

$$\frac{\sum X_i^2}{\sum X_i} = \frac{183\,899}{1049}$$
$$= 175.308\cdots$$
$$= 175 \text{ (3 s.f.)}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise A, Question 3**

**Question:**

The lengths of nails produced by a certain machine are normally distributed with a mean $\mu$ and standard deviation $\sigma$. A random sample of 10 nails is taken and their lengths $\{X_1, X_2, X_3, \ldots, X_{10}\}$ are measured.

**i** Write down the distributions of the following:

**a** $\sum_1^{10} X_i$

**b** $\dfrac{2X_1 + 3X_{10}}{5}$

**c** $\sum_1^{10}(X_i - \mu)$

**d** $\bar{X}$

**e** $\sum_1^{5} X_i - \sum_6^{10} X_i$

**f** $\sum_1^{10}\left(\dfrac{X_i - \mu}{\sigma}\right)$

**ii** State which of the above are statistics.

**Solution:**

**i** **a** $\sum X_i \sim \text{N}(10\mu, 10\sigma^2)$

**b** $\dfrac{2X_1 + 3X_{10}}{5} \sim \text{N}\left(\mu, \dfrac{13}{25}\sigma^2\right)$, $\quad \dfrac{13}{25} = \dfrac{2^2 + 3^2}{5^2}$

**c** $\text{E}(X_i - \mu) = 0 \quad \text{Var}(X_i - \mu) = \text{Var}(X_i) = \sigma^2$

$\therefore \sum (X_i - \mu) \sim \text{N}(0, 10\sigma^2)$

**d** $\bar{X} = \dfrac{\sum X_i}{n} \sim \text{N}\left(\mu, \dfrac{\sigma^2}{10}\right) \quad (n = 10)$

**e** $\sum_{i=1}^{5} X_i - \sum_{i=6}^{10} X_i \sim \text{N}(5\mu, 5\sigma^2) - \text{N}(5\mu, 5\sigma^2)$

$\therefore$ combined distribution $\sim \text{N}(0, 10\sigma^2)$

[Remember $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$ ]

**f** $\dfrac{X_i - \mu}{\sigma} \sim \text{N}(0, 1^2) \quad \therefore \sum\left(\dfrac{X_i - \mu}{\sigma}\right) \sim \text{N}(0, 10)$

**ii** **a**, **b**, **d**, **e** are statistics since they do not contain $\mu$ or $\sigma$, the unknown population parameters.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise A, Question 4**

**Question:**

A large bag of coins contains 1p, 5p and 10p coins in the ratio 2:2:1.

**a** Find the mean $\mu$ and the variance $\sigma^2$ for the value of coins in this population.

A random sample of two coins is taken and their values $X_1$ and $X_2$ are recorded.

**b** List all possible samples.

**c** Find the sampling distribution for the mean $\bar{X} = \frac{X_1 + X_2}{2}$.

**d** Hence show that $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{2}/2$

**Solution:**

**a** $X =$ value of a coin.

| $x$ | 1 | 5 | 10 |
|---|---|---|---|
| $\mathrm{P}(X = x)$ | $\frac{2}{5}$ | $\frac{2}{5}$ | $\frac{1}{5}$ |

$$\therefore \mu = \mathrm{E}(X) = \frac{2}{5} + \frac{10}{5} + \frac{10}{5} = \frac{22}{5} \text{ or } 4.4$$

$$\mathrm{E}(X^2) = 1^2 \times \frac{2}{5} + 25 \times \frac{2}{5} + 100 \times \frac{1}{5} = \frac{152}{5}$$

$$\therefore \sigma^2 = \mathrm{E}(X^2) - \mu^2 = \frac{152}{5} - \frac{22^2}{25} = 11.04 \text{ or } \frac{276}{25}$$

**b** $\{1, 1\}$ $\{1, 5\}^{\times 2}$ $\{1, 10\}^{\times 2}$

$\{5, 5\}$ $\{5, 10\}^{\times 2}$

$\{10, 10\}$

**c**

| $\bar{x}$ | 1 | 3 | 5 | 5.5 | 7.5 | 10 |
|---|---|---|---|---|---|---|
| $\mathrm{P}(\bar{X} = \bar{x})$ | $\frac{4}{25}$ | $\frac{8}{25}$ | $\frac{4}{25}$ | $\frac{4}{25}$ | $\frac{4}{25}$ | $\frac{1}{25}$ |

$$\left[\begin{aligned} \text{e.g. } \mathrm{P}(\bar{X} = 5.5) &= \mathrm{P}(X_i = (nX_2 = 10) + \mathrm{P}(X_i = 10n\ X_2 = 1) \\ &= \frac{2}{5} \times \frac{1}{5} + \frac{1}{5} \times \frac{2}{5} = \frac{4}{25} \end{aligned}\right]$$

**d** $\mathrm{E}(\bar{X}) = 1 \times \frac{4}{25} + 3 \times \frac{8}{25} + \cdots + 10 \times \frac{1}{25} = 4.4 = \mu$

$$\mathrm{Var}(\bar{X}) = 1^2 \times \frac{4}{25} + 3^2 \times \frac{8}{25} + \cdots + 10^2 \times \frac{1}{25} - 4.4^2 = 5.52 = \frac{\sigma^2}{2}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 1**

**Question:**

Find unbiased estimates of the mean and variance of the populations from which the following random samples have been taken:

**a** 21.3; 19.6; 18.5; 22.3; 17.4; 16.3; 18.9; 17.6; 18.7; 16.5; 19.3; 21.8; 20.1; 22.0

**b** 1; 2; 5; 1; 6; 4; 1; 3; 2; 8; 5; 6; 2; 4; 3; 1

**c** 120.4; 230.6; 356.1; 129.8; 185.6; 147.6; 258.3; 329.7; 249.3

**d** 0.862; 0.754; 0.459; 0.473; 0.493; 0.681; 0.743; 0.469; 0.538; 0.361.

**Solution:**

Unbiased estimate of mean $= \bar{x} = \dfrac{\sum x}{n}$

Unbiased estimate of variance $= S^2 = \dfrac{\sum x^2 - n\bar{x}^2}{n-1}$

**a** $\sum x = 270.3, \sum x^2 = 5270.49, n = 14$

$\therefore\ \bar{x} = 19.3, S^2 = 3.98$

**b** $\sum x = 54, \sum x^2 = 252, n = 16$

$\therefore\ \bar{x} = 3.375, S^2 = 4.65$

**c** $\sum x = 2007.4, \sum x^2 = 505\,132.36, n = 9$

$\therefore\ \bar{x} = 223, S^2 = 7174$

**d** $\sum x = 5.833, \sum x^2 = 3.644555, n = 10$

$\therefore\ \bar{x} = 0.5833, S^2 = 0.0269$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 2**

**Question:**

Find unbiased estimates of the mean and the variance of the populations from which random samples with the following summaries have been taken.

**a** $n = 120$ $\Sigma x = 4368$ $\Sigma x^2 = 162\,466$

**b** $n = 30$ $\Sigma x = 270$ $\Sigma x^2 = 2546$

**c** $n = 1037$ $\Sigma x = 1140.7$ $\Sigma x^2 = 1278.08$

**d** $n = 15$ $\Sigma x = 168$ $\Sigma x^2 = 1913$

**Solution:**

**a** $\bar{x} = \frac{\sum x}{n} = \frac{4368}{120} = 36.4$

$$S^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{162\,466 - 120 \times 36.4^2}{119} = 29.166\ldots$$
$$= 29.2 \text{ (3 s.f.)}$$

**b** $\bar{x} = \frac{\sum x}{n} = \frac{270}{30} = 9$

$$S^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{2546 - 30 \times 9^2}{29} = 4$$

**c** $\bar{x} = \frac{\sum x}{n} = \frac{1140.7}{1037} = 1.1$

$$S^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{1278.08 - 1037 \times 1.1^2}{1036} = 0.0225$$

**d** $\bar{x} = \frac{\sum x}{n} = \frac{168}{15} = 11.2$

$$S^2 = \frac{\sum x^2 - n\bar{x}^2}{n-1} = \frac{1913 - 15 \times 11.2^2}{14} = 2.24285\ldots$$
$$= 2.24 \text{ (3 s.f.)}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 3**

**Question:**

The concentrations, in mg per litre, of a trace element in 7 randomly chosen samples of water from a spring were:
240.8 237.3 236.7 236.6 234.2 233.9 232.5.
Determine unbiased estimates of the mean and the variance of the concentration of the trace element per litre of water from the spring.

**Solution:**

$$\begin{aligned}\sum x &= 1652, \sum x^2 = 389\,917.48, n = 7\\ \therefore\ \bar{x} &= \frac{1652}{7} = 236\\ S^2 &= \frac{389\,917.48 - 7 \times 236^2}{6}\\ &= 7.58\end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 4**

**Question:**

Cartons of orange are filled by a machine. A sample of 10 cartons selected at random from the production contained the following quantities of orange (in ml).

| | | | | |
|---|---|---|---|---|
| 201.2 | 205.0 | 209.1 | 202.3 | 204.6 |
| 206.4 | 210.1 | 201.9 | 203.7 | 207.3 |

Calculate unbiased estimates of the mean and variance of the population from which this sample was taken.

**Solution:**

$$\begin{aligned}\sum x &= 2051.6,\ \sum x^2 = 420\,989.26,\ n = 10\\ \bar{x} &= 205.16 = 205\ (3\text{ s.f.})\\ S^2 &= \frac{420\,989.26 - 10\times\bar{x}^2}{9}\\ &= 9.22266\ldots = 9.22\ (3\text{ s.f.})\end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 5**

**Question:**

A manufacturer of self-assembly furniture required bolts of two lengths, 5 cm and 10 cm, in the ratio 2 : 1 respectively.

**a** Find the mean $\mu$ and the variance $\sigma^2$ for the lengths of bolts in this population.

A random sample of three bolts is selected from a large box containing bolts in the required ratio.

**b** List all possible samples.

**c** Find the sampling distribution for the mean $\bar{X}$.

**d** Hence find $\mathrm{E}(\bar{X})$ and $\mathrm{Var}(\bar{X})$.

**e** Find the sampling distribution for the mode $M$.

**f** Hence find $\mathrm{E}(M)$ and $\mathrm{Var}(M)$.

**g** Find the bias when $M$ is used as an estimator of the population mode.

**Solution:**

$X = \text{length of a bolt}$

| $x$ | 5 | 10 |
|---|---|---|
| $P(X = x)$ | $\frac{2}{3}$ | $\frac{1}{3}$ |

**a** $\mu = 5 \times \frac{2}{3} + 10 \times \frac{1}{3} = \frac{20}{3}$

$$\sigma^2 = 25 \times \frac{2}{3} + 100 \times \frac{1}{3} - \left(\frac{20}{3}\right)^2 = \frac{50}{9}$$

**b** $\{5, 5, 5\}$

$\{5, 5, 10\}^{\times 3}$ $\{5, 10, 10\}^{\times 3}$

$\{10, 10, 10\}$

**c**

| $\bar{x}$ | 5 | $\frac{20}{3}$ | $\frac{25}{3}$ | 10 |
|---|---|---|---|---|
| $P(\bar{X} = \bar{x})$ | $\frac{8}{27}$ | $\frac{12}{27}$ | $\frac{6}{27}$ | $\frac{1}{27}$ |

**d** $E(\bar{X}) = 5 \times \frac{8}{27} + \frac{20}{3} \times \frac{12}{27} + \cdots + 10 \times \frac{1}{27} = \frac{20}{3} = \mu$

$$\text{Var}(\bar{X}) = 5^2 \times \frac{8}{27} + \cdots + 10^2 \times \frac{1}{27} - \left(\frac{20}{3}\right)^2 = \frac{50}{27} = \frac{\sigma^2}{3}$$

**e**

| $m$ | 5 | 10 |
|---|---|---|
| $P(M = m)$ | $\frac{20}{27}$ | $\frac{7}{27}$ |

$P(M = 10)$ is cases $\{5, 10, 10\}$ and $\{10, 10, 10\}$

**f** $E(M) = 5 \times \frac{20}{27} + 10 \times \frac{7}{27} = \frac{170}{27} = 6.296\ldots$

$$\text{Var}(M) = 25 \times \frac{20}{27} + 100 \times \frac{7}{27} - \left(\frac{170}{27}\right)^2 = \frac{3500}{729} = 4.80\ldots$$

**g** $\text{Bias} = E(M) - 5 = 1.296\ldots = 1.30 \text{ (3 s.f.)}$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 6**

**Question:**

A biased six-sided die has probability $p$ of landing on a six.
Every day, for a period of 25 days, the die is rolled 10 times and the number of sixes $X$ is recorded giving rise to a sample $X_1, X_2, \ldots, X_{25}$.

**a** Write down $\mathrm{E}(X)$ in terms of $p$.

**b** Show that the sample mean $\bar{X}$ is a biased estimator of $p$ and find the bias.

**c** Suggest a suitable unbiased estimator of $p$.

**Solution:**

$X \sim B(10, p)$

**a** $\mathrm{E}(X) = np = 10p$

**b** $\bar{X} = \dfrac{X_1 + \ldots + X_{25}}{25}$

$$\mathrm{E}(\bar{X}) = \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ldots + \mathrm{E}(X_{25})}{25} = \frac{\mu + \mu + \ldots + \mu}{25} = \frac{25\mu}{25} = \mu$$

$\therefore \bar{X}$ is an unbiased estimator of $\mu$

But $\mathrm{E}(\bar{X}) = 10p$ $\therefore \bar{X}$ is a *biased* estimator of $p$.

so $\text{bias} = 10p - p = 9p$

**c** $\mathrm{E}\left(\dfrac{\bar{X}}{10}\right) = \dfrac{1}{10}\mathrm{E}(\bar{X}) = p$

$\therefore \dfrac{\bar{X}}{10}$ is an unbiased estimator of $p$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise B, Question 7**

**Question:**

The random variable $X \sim \mathrm{U}[-\alpha, \alpha]$

**a** Find $\mathrm{E}(X)$ and $\mathrm{E}(X^2)$.

A random sample $X_1, X_2, X_3$ is taken and the statistic $Y = X_1^2 + X_2^2 + X_3^2$ is calculated.

**b** Show that $Y$ is an unbiased estimator of $\alpha^2$.

**Solution:**

$X \sim \mathrm{U}[-\alpha, \alpha]$

**a** $\mathrm{E}(X) = \dfrac{-\alpha + \alpha}{2} = 0$

$$\mathrm{Var}(X) = \frac{(\alpha - (-\alpha))^2}{12} = \frac{4\alpha^2}{12} = \frac{\alpha^2}{3}$$

$\mu$ and $\sigma^2$ for $\mathrm{U}[a, b]$ are given in formula booklet under S2.

$$\therefore \mathrm{E}(X^2) = \mathrm{Var}(X) + [\mathrm{E}(X)]^2$$

$$\therefore \mathrm{E}(X^2) = \frac{\alpha^2}{3} + 0 = \frac{\alpha^2}{3}$$

**b** $Y = X_1^2 + X_2^2 + X_3^2$

$$\mathrm{E}(Y) = \mathrm{E}(X_1^2) + \mathrm{E}(X_2^2) + \mathrm{E}(X_3^2)$$

$$= \frac{\alpha^2}{3} \times 3 = \alpha^2$$

$\therefore$ $Y$ is an unbiased estimator of $\alpha^2$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 1**

**Question:**

John and Mary each independently took a random sample of sixth-formers in their college and asked them how much money, in pounds, they earned last week. John used his sample of size 20 to obtain unbiased estimates of the mean and variance of the amount earned by a sixth-former at their college last week. He obtained values of $\bar{x} = 15.5$ and $S_x^2 = 8.0$.

Mary's sample of size 30 can be summarised as $\Sigma y = 486$ and $\Sigma y^2 = 8222$.

**a** Use Mary's sample to find unbiased estimates of $\mu$ and $\sigma^2$.

**b** Combine the samples and use all 50 observations to obtain further unbiased estimates of $\mu$ and $\sigma^2$.

**c** Find the standard error of the mean for each of these estimates of $\mu$.

**d** Comment on which estimate of $\mu$ you would prefer to use.

**Solution:**

**a** $\bar{y} = \frac{486}{30} = 16.2$

$$S_y^2 = \frac{8222 - 30 \times 16.2^2}{29} = 12.0275\ldots = 12.0 \text{ (3 s.f.)}$$

**b** Let $\sum w = \sum x + \sum y$

$\bar{x} = 15.5 \Rightarrow \sum x = 15.5 \times 20 = 310$

$\therefore \sum w = 796$

$S_x^2 = 8.0 \Rightarrow \sum x^2 = 8 \times 19 + 20 \times 15.5^2 = 4957$

$\therefore \sum w^2 = 13179$

$\therefore \bar{w} = \frac{796}{50} = 15.92$

$$S_w^2 = \frac{13179 - 50 \times 15.92^2}{49} = 10.340\ldots = 10.34$$

**c** Standard error of the mean is $\frac{S}{\sqrt{n}}$

$$\frac{S_x}{\sqrt{20}} = 0.632 \text{ (3 s.f.)}, \frac{S_y}{\sqrt{30}} = 0.633 \text{ (3 s.f.)}, \frac{S_w}{\sqrt{50}} = 0.455 \text{ (3 s.f.)}$$

**d** Prefer to use $\bar{w}$ since it is based on a larger sample size and has smallest standard error.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 2**

**Question:**

A machine operator checks a random sample of 20 bottles from a production line in order to estimate the mean volume of bottles (in $\text{cm}^3$) from this production run. The 20 values can be summarised as $\Sigma x = 1300$ and $\Sigma x^2 = 84\,685$.

**a** Use this sample to find unbiased estimates of $\mu$ and $\sigma^2$

A supervisor knows from experience that the standard deviation of volumes on this process, $\sigma$, should be $3\,\text{cm}^3$ and he wishes to have an estimate of $\mu$ that has a standard error of less than $0.5\,\text{cm}^3$.

**b** What size sample will he need to achieve this?

The supervisor takes a further sample of size 16 and finds $\Sigma x = 1060$.

**c** Combine the two samples to obtain a revised estimate of $\mu$.

**Solution:**

**a** $\bar{x} = \dfrac{1300}{20} = 65$

$$S_x^2 = \frac{84\,685 - 20 \times 65^2}{19} = 9.7368\cdots = 9.74 \text{ (3 s.f.)}$$

**b** $\dfrac{\sigma}{\sqrt{n}} < 0.5 \Rightarrow \dfrac{3}{\sqrt{n}} < 0.5$

$6 < \sqrt{n}$

$n > 36$

$\therefore$ need a sample of 37 or more

**c** Let $y$ = combined sample

$\sum y = 1300 + 1060$

$\sum y = 2360$ $\quad n = 36$

$\therefore \bar{y} = \dfrac{2360}{36} = 65.555\ldots$ or $65.6$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 3**

**Question:**

The heights of certain seedlings after growing for 10 weeks in a greenhouse have a standard deviation of 2.6 cm. Find the smallest sample that must be taken for the standard error of the mean to be less than 0.5 cm.

**Solution:**

$$\frac{\sigma}{\sqrt{n}} < 0.5$$

$\sigma = 2.6 \Rightarrow \quad \sqrt{n} > 2.6 \times 2 = 5.2$

i.e. $n > 27.04$

So need a sample of 28 (or more)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 4**

**Question:**

The hardness of a plastic compound was determined by measuring the indentation produced by a heavy pointed device.
The following observations in tenths of a millimetre were obtained:
4.7, 5.2, 5.4, 4.8, 4.5, 4.9, 4.5, 5.1, 5.0, 4.8.

**a** Estimate the mean indentation for this compound.

**b** Estimate the standard error of the mean.

**c** Estimate the size of sample required in order that in future the standard error of the mean should be just less than 0.05.

**Solution:**

Let $x =$ indentation

$\sum x = 48.9, \quad \sum x^2 = 239.89, n = 10$

**a** $\hat{\mu} = \bar{x} = \dfrac{48.9}{10} = 4.89$

**b** $\hat{\sigma}^2 = S^2 = \dfrac{239.89 - 10 \times 4.89^2}{9} = 0.08544...$

$\dfrac{S}{\sqrt{n}} = 0.092436... = 0.0924$ (3 s.f.)

**c** Require $\dfrac{S}{\sqrt{n}} = \dfrac{0.2923...}{\sqrt{n}} < 0.05$

$\Rightarrow \quad \sqrt{n} > 5.846...$

$n > 34.17...$

$\therefore$ need $n = 35$ (or more)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise C, Question 5**

**Question:**

Prospective army recruits receive a medical test. The probability of each recruit passing the test is $p$, independent of any other recruit. The medicals are carried out over two days and on the first day $n$ recruits are seen and on the next day $2n$ are seen. Let $X_1$ be the number of recruits who pass the test on the first day and let $X_2$ be the number passing on the second day.

**a** Write down $\mathrm{E}(X_1), \mathrm{E}(X_2), \mathrm{Var}(X_1)$ and $\mathrm{Var}(X_2)$.

**b** Show that $\frac{X_1}{n}$ and $\frac{X_2}{2n}$ are both unbiased estimates of $p$ and state, giving a reason, which you would prefer to use.

**c** Show that $X = \frac{1}{2}\left(\frac{X_1}{n} + \frac{X_2}{2n}\right)$ is an unbiased estimator of $p$.

**d** Show that $Y = \left(\frac{X_1 + X_2}{3n}\right)$ is an unbiased estimator of $p$.

**e** Which of the statistics $\frac{X_1}{n}$, $\frac{X_2}{2n}$, $X$ or $Y$ is the best estimator of $p$?

The statistic $T = \left(\frac{2X_1 + X_2}{3n}\right)$ is proposed as an estimator of $p$.

**f** Find the bias.

**Solution:**

$X_1 \sim \mathrm{B}(n, p) \qquad X_2 \sim \mathrm{B}(2n, p)$

**a** $\mathrm{E}(X_1) = np, \mathrm{E}(X_2) = 2np, \mathrm{Var}(X_1) = np(1-p), \mathrm{Var}(X_2) = 2np(1-p)$

**b** $\mathrm{E}\left(\frac{X_1}{n}\right) = \frac{\mathrm{E}(X_1)}{n} = \frac{np}{n} = p \therefore \frac{X_1}{n}$ is unbiased estimator of $p$

$\mathrm{E}\left(\frac{X_2}{2n}\right) = \frac{\mathrm{E}(X_2)}{2n} = \frac{2np}{2n} = p \therefore \frac{X_2}{2n}$ is unbiased estimator of $p$

Prefer $\frac{X_2}{2n}$ since based on a larger sample (and therefore will have smaller variance)

**c** $X = \frac{1}{2}\left(\frac{X_1}{n} + \frac{X_2}{2n}\right) \Rightarrow \mathrm{E}(X) = \frac{1}{2}\left[\frac{\mathrm{E}(X_1)}{n} + \frac{\mathrm{E}(X_2)}{2n}\right]$

$$= \frac{1}{2}\left[\frac{np}{n} + \frac{2np}{2n}\right]$$

$$= \frac{1}{2}[p + p] = p$$

$\therefore X$ is an unbiased estimator of $p$

**d** $Y = \left(\frac{X_1 + X_2}{3n}\right) \Rightarrow \mathrm{E}(Y) = \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2)}{3n} = \frac{np + 2np}{3n} = p$

$\therefore Y$ is an unbiased estimator of $p$

**e** $\mathrm{Var}\left(\frac{X_1}{n}\right) = \frac{1}{n^2} \quad \mathrm{Var}(X_1) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

$\mathrm{Var}\left(\frac{X_2}{2n}\right) = \frac{1}{4n^2} \quad \mathrm{Var}(X_2) = \frac{2np(1-p)}{4n^2} = \frac{p(1-p)}{2n}$

$\mathrm{Var}(X) = \frac{1}{4}\left[\mathrm{Var}\left(\frac{X_1}{n}\right) + \mathrm{Var}\left(\frac{X_2}{2n}\right)\right] = \frac{1}{4}\left[\frac{p(1-p)}{n} + \frac{p(1-p)}{2n}\right] = \frac{3p(1-p)}{8n}$

$\mathrm{Var}(Y) = \frac{1}{9n^2}\left[\mathrm{Var}(X_1) + \mathrm{Var}(X_2)\right] = \frac{1}{9n^2}[np(1-p) + 2np(1-p)]$

$\mathrm{Var}(Y) = \frac{3np(1-p)}{9n^2} = \frac{p(1-p)}{3n}$

$\therefore$ Var $(Y)$ is smallest so $Y$ is the best estimator.

**f** $T = \left(\frac{2X_1 + X_2}{3n}\right)$

$\mathrm{E}(T) = \frac{2\mathrm{E}(X_1) + \mathrm{E}(X_2)}{3n} = \frac{2np + 2np}{3n} = \frac{4p}{3}$

bias $= \mathrm{E}(T) - p$

$= \frac{p}{3}$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 6**

**Question:**

Two independent random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are taken from a population with mean $\mu$ and variance $\sigma^2$. The unbiased estimators $\bar{X}$ and $\bar{Y}$ of $\mu$ are calculated. A new unbiased estimator $T$ of $\mu$ is sought of the form $T = r\bar{X} + s\bar{Y}$.

**a** Show that, since $T$ is unbiased, $r + s = 1$.

**b** By writing $T = r\bar{X} + (1-r)\bar{Y}$, show that

$$\text{Var}(T) = \sigma^2\left[\frac{r^2}{n} + \frac{(1-r)^2}{m}\right]$$

**c** Show that the minimum variance of $T$ is when $r = \frac{n}{n+m}$.

**d** Find the best (in the sense of minimum variance) estimator of $\mu$ of the form $r\bar{X} + s\bar{Y}$.

**Solution:**

$\text{E}(\bar{X}) = \mu \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$

$\text{E}(\bar{Y}) = \mu \quad \text{Var}(\bar{Y}) = \frac{\sigma^2}{m}$

**a** $T = r\bar{X} + s\bar{Y}$
$\text{E}(T) = r\mu + s\mu = (r+s)\mu$
So if $T$ is unbiased $r + s = 1$

**b** $r + s = 1 \Rightarrow s = 1 - r$
$\therefore T = r\bar{X} + (1-r)\bar{Y}$

$$\text{Var}(T) = r^2\text{Var}(\bar{X}) + (1-r)^2\text{Var}(\bar{Y}) = r^2\frac{\sigma^2}{n} + (1-r)^2\frac{\sigma^2}{m}$$
$$= \sigma^2\left[\frac{r^2}{n} + \frac{(1-r)^2}{m}\right]$$

**c** $\frac{\text{d}}{\text{d}r}\text{Var}(T) = \sigma^2\left[\frac{2r}{n} + \frac{2(1-r)\times(-1)}{m}\right]$

$\therefore$ Var $(T)$ is a quadratic function of $r$ with positive $r^2$ term $\therefore$ min

$\frac{\text{d}}{\text{d}r}\text{Var}(T) = 0 \Rightarrow rm = (1-r)n$ or $r(m+n) = n$ or $r = \frac{n}{m+n}$

**d** Best estimator of form $T$ is

$$T = \frac{n}{m+n}\bar{X} + \frac{m}{m+n}\bar{Y} \quad \text{or} \quad \frac{n\bar{X} + m\bar{Y}}{m+n}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise C, Question 7**

**Question:**

A large bag of counters has 40% with the number 0 on, 40% with the number 2 on and 20% with the number 1.

**a** Find the mean $\mu$, and the variance $\sigma^2$, for this population of counters.

A random sample of size 3 is taken from the bag.

**b** List all possible samples.

**c** Find the sampling distribution for the mean $\bar{X}$.

**d** Find $\mathrm{E}(\bar{X})$ and $\mathrm{Var}(\bar{X})$.

**e** Find the sampling distribution for the median $N$.

**f** Hence find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$.

**g** Show that $N$ is an unbiased estimator of $\mu$.

**h** Explain which estimator, $\bar{X}$ or $N$, you would choose as an estimator of $\mu$.

**Solution:**

Let $X$ = number on a counter.

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(X=x)$ | 0.4 | 0.2 | 0.4 |

**a** $\mu = 1$ (by symmetry)

$\sigma^2 = 0 + 1\times 0.2 + 2^2 \times 0.4 - 1 = 0.8$ or $\frac{4}{5}$

**b** $\{0,0,0\}$ $\{0,0,1\}^{\times 3}$ $\{0,0,2\}^{\times 3}$

$\{1,1,1\}$ $\{1,1,0\}^{\times 3}$ $\{1,1,2\}^{\times 3}$

$\{2,2,2\}$ $\{2,2,0\}^{\times 3}$ $\{2,2,1\}^{\times 3}$ $\{0,1,2\}^{\times 3!=6}$

**c**

| $\bar{x}$ | 0 | $\frac{1}{3}$ | $\frac{2}{3}$ | 1 | $\frac{4}{3}$ | $\frac{5}{3}$ | 2 |
|---|---|---|---|---|---|---|---|
| $P(\bar{X}=\bar{x})$ | $\frac{8}{125}$ | $\frac{12}{125}$ | $\frac{30}{125}$ | $\frac{25}{125}$ | $\frac{30}{125}$ | $\frac{12}{125}$ | $\frac{8}{125}$ |

**d** $E(\bar{X}) = 1$ (by symmetry) $(= \mu)$

$$\text{Var}(\bar{X}) = 0 + \frac{1}{9}\times\frac{12}{125} + \frac{4}{9}\times\frac{30}{125} + \cdots + 4\times\frac{8}{125} - 1^2$$

$$= \frac{4}{15} \qquad \left(= \frac{\sigma^2}{3}\right)$$

**e**

| $n$ | 0 | 1 | 2 |
|---|---|---|---|
| $P(N=n)$ | $\frac{44}{125}$ | $\frac{37}{125}$ | $\frac{44}{125}$ |

e.g. $P(N=2)$ is cases
$\{2,2,2\}; \{2,2,0\}; \{2,2,1\}$

**f** $E(N) = 1$ (by symmetry)

$$\text{Var}(N) = 0 + 1^2 \times \frac{37}{125} + 2^2 \times \frac{44}{125} - 1^2 = \frac{88}{125} \qquad (= \sigma^2)$$

**g** $\because E(N) = 1 = \mu \therefore$ $N$ is an unbiased estimator of $\mu$.

**h** $\because \text{Var}(\bar{X}) < \text{Var}(N)$ choose $\bar{X}$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 1**

**Question:**

A sample of size 6 is taken from a normal distribution $N(10, 2^2)$. What is the probability that the sample mean exceeds 12?

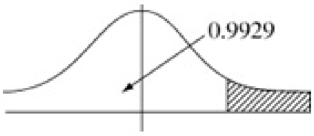
**Solution:**

$X \sim N(10, 2^2)$

$\overline{X} \sim N(10, \frac{4}{6})$

$$P(\overline{X} > 12) = P\left(Z > \frac{12-10}{\sqrt{\frac{4}{6}}}\right)$$

$$= P(Z > 2.449...)$$




Use 2.44

$= 1 - 0.9927$

$= 0.0073$ (tables)

or $= 0.00715...$ (from calculator)

So accept awrt 0.007

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 2**

**Question:**

A machine fills cartons in such a way that the amount of drink in each carton is distributed normally with a mean of 40 $\text{cm}^3$ and a standard deviation of 1.5 $\text{cm}^3$.

**a** A sample of four cartons is examined. Find the probability that the mean amount of drink is more than 40.5 $\text{cm}^3$

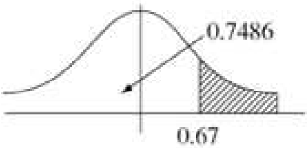
**b** A sample of 49 cartons is examined. Find the probability that the mean amount of drink is more than 40.5 $\text{cm}^3$ on this occasion.

**Solution:**

$D \sim \text{N}(40, 1.5^2)$

$\overline{D} \sim \text{N}(40, \frac{1.5^2}{4})$

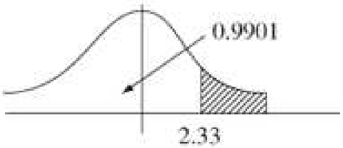
**a**




$$\begin{aligned} \text{P}(\overline{D} > 40.5) &= \text{P}\left(Z > \frac{40.5-40}{\frac{1.5}{2}}\right) \\ &= \text{P}(Z > \tfrac{2}{3}) && \text{Use } 0.67 \\ &= 1 - 0.7486 \\ &= 0.2514 && \text{(tables)} \\ \text{or} &= 0.25249\ldots && \text{(calculator)} \end{aligned}$$

So accept awrt 0.251 or 0.252

**b** $\overline{X} \sim \text{N}(40, \frac{1.5^2}{49})$




$$\begin{aligned} \text{P}(\overline{X} > 40.5) &= \text{P}\left(Z > \frac{40.5-40}{\frac{1.5}{7}}\right) \\ &= \text{P}(Z > 2.33\ldots) && \text{Use } 2.34 \\ &= 1 - 0.9904 \\ &= 0.0096 && \text{(tables)} \\ \text{or} &= 0.009815\ldots \end{aligned}$$

So accept awrt $0.0098 \sim 0.0096$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 3**

**Question:**

The lengths of bolts produced by a machine have an unknown distribution with mean 3.03 cm and standard deviation 0.20 cm. A sample of 100 bolts is taken.

**a** Estimate the probability that the mean length of this sample is less than 3 cm.

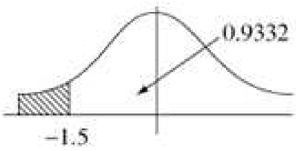
**b** What size sample is required if the probability that the mean is less than 3 cm is to be less than 1%?

**Solution:**

$$L \sim N(3.03, 0.20^2)$$

$$\bar{L} \simeq\sim N(3.03, \frac{0.20^2}{100})$$

**a**




$$\begin{aligned} P(\bar{L} < 3) &= P\left(Z < \frac{3-3.03}{0.02}\right) \\ &= P(Z < -1.5) \\ &= 1-0.9332 \\ &= 0.0668 \quad \text{(tables)} \\ \text{or} &= 0.066807\ldots \quad \text{(calculator)} \end{aligned}$$

So accept awrt 0.0668

**b** $P(\bar{L} < 3) = P\left(Z < \frac{3-3.03}{\frac{0.2}{\sqrt{n}}}\right) < 0.01$

$$\begin{aligned} \Rightarrow \frac{3-3.03}{\frac{0.2}{\sqrt{n}}} &< -2.3263 \\ \frac{0.03}{2.3263} &> \frac{0.2}{\sqrt{n}} \\ \sqrt{n} &> 15.50\ldots \\ n &> 240.51\ldots \end{aligned}$$

$\therefore n = 241$ (or more)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 4**

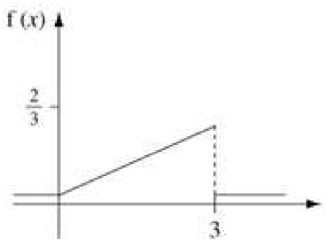
**Question:**

Forty observations are taken from a population with distribution given by the probability density function

$$f(x) = \begin{cases} \frac{2}{9}x, & 0 \le x \le 3, \\ 0, & \text{otherwise.} \end{cases}$$

**a** Find the mean and variance of this population.
**b** Find an estimate of the probability that the mean of the 40 observations is more than 2.10.

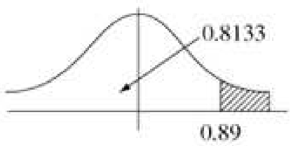
**Solution:**




**a** $\mu = 2$ (Centroid of $\Delta$) or from $\int_0^3 x \times \frac{2}{9} x \, dx$

$$\begin{aligned} \sigma^2 &= \int_0^3 \frac{2}{9} x^3 \, dx - 2^2 \\ &= \left[ \frac{2}{9} \times \frac{x^4}{4} \right]_0^3 - 4 \\ &= \frac{9}{2} - 0 - 4 = 0.5 \end{aligned}$$

**b** $\bar{X} \simeq\sim N\left(2, \frac{0.5}{40}\right)$




$$\begin{aligned} P(\bar{X} > 2.10) &= P\left( Z > \frac{2.10 - 2}{\sqrt{\frac{0.5}{40}}} \right) \\ &= P(Z > 0.8944...) \\ &= 1 - 0.8133 \quad \text{Use 0.89} \\ &= 0.1867 \quad \text{(tables)} \\ \text{or } &= 0.1855466... \quad \text{(calculator)} \end{aligned}$$

so accept awrt $0.186 \sim 0.187$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 5**

**Question:**

A fair die is rolled 35 times.

**a** Find the approximate probability that the mean of the 35 scores is more than 4.

**b** Find the approximate probability that the total of the 35 scores is less than 100.

**Solution:**

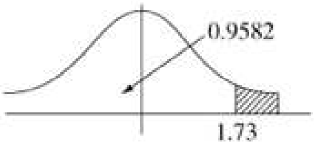
$S =$ score on die

$\mathrm{E}(S) = 3.5 \quad \mathrm{Var}(S) = \frac{35}{12}$

$\boxed{\text{See S1}}$

$$\bar{S} \simeq\sim \mathrm{N}\left(3.5, \frac{\left(\frac{35}{12}\right)}{35}\right)$$

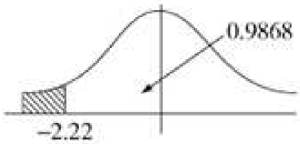
**a**




$$\begin{aligned} \mathrm{P}(\bar{S} > 4) &= \mathrm{P}\left(Z > \frac{4-3.5}{\sqrt{\frac{1}{12}}}\right) \quad \text{and } \frac{\frac{35}{12}}{35} = \frac{1}{12} \\ &= \mathrm{P}(Z > 1.732\ldots) \quad \text{Use } 1.73 \\ &= 1 - 0.9582 \\ &= 0.0418 \quad \text{(tables)} \\ \text{or} &= 0.041632\ldots \quad \text{(calculator)} \end{aligned}$$

so accept awrt 0.042

**b** $T =$ total score, $T = 35\bar{S}$




$$\begin{aligned} \mathrm{P}(T < 100) &= \mathrm{P}\left(\bar{S} < \frac{100}{35}\right) \\ &= \mathrm{P}\left(Z < \frac{\frac{100}{35} - 3.5}{\sqrt{\frac{1}{12}}}\right) \\ &= \mathrm{P}(Z < -2.2269\ldots) \quad \text{Use } -2.22 \\ &= 1 - 0.9868 \\ &= 0.0132 \quad \text{(tables)} \\ \text{or} &= 0.01297618\ldots \end{aligned}$$

so accept awrt 0.013

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 6**

**Question:**

The 25 children in a class each roll a fair die 30 times and record the number of sixes they obtain. Find an estimate of the probability that the mean number of sixes recorded for the class is less than 4.5.

**Solution:**

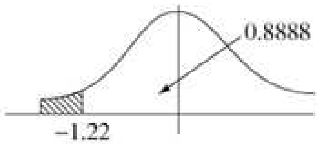
Let $S =$ number of sixes in 30 rolls of a die

$S \sim \mathrm{B}(30, \frac{1}{6})$

$\mu = 5 \quad \sigma^2 = \frac{25}{6}$

$\bar{S} \simeq\sim \mathrm{N}\left(5, \frac{\left(\frac{25}{6}\right)}{25}\right)$

Using formula in S2. mean $= np$
variance $= np(1-p)$




$$
\begin{aligned}
\mathrm{P}(\bar{S} < 4.5) &= \mathrm{P}\left(Z < \frac{4.5-5}{\sqrt{\frac{1}{6}}}\right) \\
&= \mathrm{P}(Z < -1.2247...) \qquad \text{Use } -1.22. \\
&= 1 - 0.8888 \\
&= 0.1112 \qquad \text{(tables)} \\
\text{or } &= 0.110335... \qquad \text{(calculator)}
\end{aligned}
$$

$\therefore$ accept awrt 0.11

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 7**

**Question:**

The error in mm made in measuring the length of a table has a uniform distribution over the range $[-5,5]$. The table is measured 20 times. Find an estimate of the probability that the mean error is less than $-1$ mm .

**Solution:**

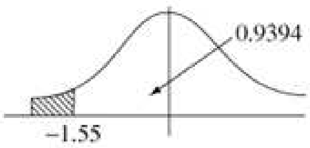
$E = \text{error}$

$E \sim \text{U}[-5,5]$

$\mu = 0 \quad \sigma^2 = \frac{(5-(-5))^2}{12} = \frac{100}{12} = \frac{25}{3}$

See formula in S2.

$$\bar{E} \simeq \sim \text{N}\left(0, \frac{\left(\frac{25}{3}\right)}{20}\right)$$




$$\begin{aligned} \text{P}(\bar{E} < -1) &= \text{P}\left(Z < \frac{-1-0}{\sqrt{\frac{5}{12}}}\right) \text{ since } \frac{\left(\frac{25}{3}\right)}{20} = \frac{5}{12} \\ &= \text{P}(Z < -1.5491\ldots) \quad \text{Use } -1.55 \\ &= 1 - 0.9394 \quad \text{(tables)} \\ &= 0.0606 \quad \text{(calculator)} \\ \text{or} &= 0.060667\ldots \end{aligned}$$

So accept awrt 0.061

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 8**

**Question:**

Telephone calls arrive at an exchange at an average, rate of two per minute. Over a period of 30 days a telephonist records the number of calls that arrive in the five-minute period before her break.

**a** Find an approximation for the probability that the total number of calls recorded is more than 350.

**b** Estimate the probability that the mean number of calls in the five-minute interval is less than 9.0.

**Solution:**

$C =$ number of calls that arrive in 5-minute period

$C \sim \text{Po}(10)$

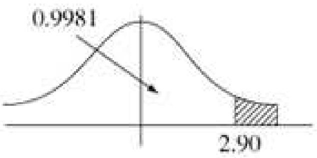
$n = 30$

$\bar{C} \simeq\sim \text{N}(10, \frac{10}{30})$

$T =$ total number of calls in 30 days

$T = 30\bar{C}$

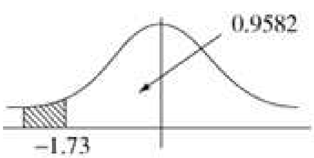
**a**




$$\begin{aligned} \therefore \text{P}(T > 350) &= \text{P}\left(\bar{C} > \frac{350}{30}\right) \\ &= \text{P}\left(Z > \frac{\frac{35}{3} - 10}{\sqrt{\frac{1}{3}}}\right) \\ &= \text{P}(Z > 2.8867...) && \text{Use } 2.90 \\ &= 1 - 0.9981 \\ &= 0.0019 && \text{(tables)} \\ \text{or} &= 0.001946... && \text{(calculator)} \end{aligned}$$

So accept awrt 0.0019

**b**




$$\begin{aligned} \text{P}(\bar{C} < 9) &= \text{P}\left(Z < \frac{9 - 10}{\sqrt{\frac{1}{3}}}\right) \\ &= \text{P}(Z < -1.732...) && \text{Use } -1.73 \\ &= 1 - 0.9582 \\ &= 0.0418 && \text{(tables)} \\ \text{or} &= 0.041632... && \text{(calculator)} \end{aligned}$$

So accept awrt 0.042

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 9**

**Question:**

How many times must a fair die be rolled in order for there to be a less than 1% chance that the mean of all the scores differs from 3.5 by more than 0.1?

**Solution:**

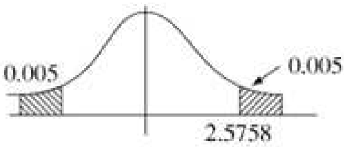
$S =$ score on a die

$\mu = 3.5 \quad \sigma^2 = \frac{35}{12}$

$\bar{S} \simeq \sim \mathrm{N}\left(3.5, \frac{35}{12n}\right)$

$\mathrm{P}(|\bar{S} - 3.5| > 0.1) < 0.01$

$= \mathrm{P}\left(|Z| > \frac{0.1}{\sqrt{\frac{35}{12n}}}\right) < 0.01$




$\Rightarrow \frac{0.1}{\sqrt{\frac{35}{12n}}} > 2.5758$

$\left(\frac{0.1}{2.5758}\right)^2 > \frac{35}{12n}$

$\therefore n > \frac{35}{12\left(\frac{0.1}{2.5758}\right)^2} = 1935.12\ldots$

$\therefore$ need $n$ at least 1936

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 10**

**Question:**

The heights of women in a certain area have a mean of 175 cm and a standard deviation of 2.5 cm. The heights of men in the same area have a mean of 177 cm and a standard deviation of 2.0 cm. Samples of 40 women and 50 men are taken and their heights are recorded. Find the probability that the mean height of the men is more than 3 cm greater than the mean height of the women.

**Solution:**

$W =$ height of a woman

$W \sim \mathrm{N}(175, 2.5^2)$

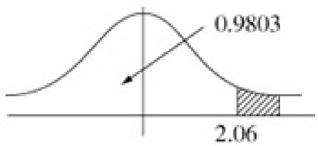
$M =$ height of a man

$M \sim \mathrm{N}(177, 2^2)$

$\overline{W} \sim \mathrm{N}\left(175, \frac{2.5^2}{40}\right) \quad \overline{M} \sim \mathrm{N}\left(177, \frac{2^2}{50}\right)$

$\mathrm{P}(\overline{M} - \overline{W} > 3)$ requires $X = \overline{M} - \overline{W}$

$$X \sim \mathrm{N}\left(2, \frac{2^2}{50} + \frac{2.5^2}{40}\right) \quad \frac{2^2}{50} + \frac{2.5^2}{40} = 0.23625$$




$$\begin{aligned}
\therefore \mathrm{P}(X > 3) &= \mathrm{P}\left(Z > \frac{3-2}{\sqrt{0.23625}}\right) \\
&= \mathrm{P}(Z > 2.0573\cdots) \quad \text{Use } 2.06 \\
&= 1 - 0.9803 \\
&= 0.0197 \quad \text{(tables)} \\
\text{or } &= 0.0198248\ldots \quad \text{(calculator)}
\end{aligned}$$

So accept awrt $0.0197 \sim 0.198$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 11**

**Question:**

A computer, in adding numbers, rounds each number off to the nearest integer. All the rounding errors are independent and come from a uniform distribution over the range $[-0.5, 0.5]$.

**a** Given that 1000 numbers are added, find the probability that the total error is greater than $+10$.

**b** Find how many numbers can be added together so that the probability that the magnitude of the total error is less than 10 is at least 0.95.

**Solution:**

$E = \text{rounding error}$

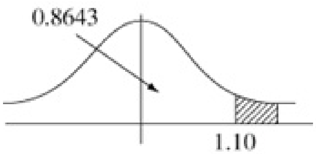
$E \sim \text{U}[-0.5, 0.5]$

$\mu = 0, \sigma^2 = \frac{1}{12}$

$n = 1000 \quad \bar{E} \sim \text{N}\left(0, \frac{1}{12\,000}\right)$

$T = \text{ total error} = 1000\bar{E}$

**a**




$$\text{P}(T > 10) = \text{P}\left(\bar{E} > \frac{10}{1000}\right)$$

$$= \text{P}\left(Z > \frac{0.01 - 0}{\sqrt{\frac{1}{12\,000}}}\right)$$

$$= \text{P}(Z > 1.09544...) \qquad \text{Use } 1.10$$

$$= 1 - 0.8643$$

$$= 0.1357 \qquad \text{(tables)}$$

$$\text{or} \quad = 0.136660... \qquad \text{(calculator)}$$

So accept awrt $0.136 \sim 0.137$

**b** $n = N, T' = N\bar{E}' \quad \text{where } \bar{E}' \sim \text{N}\left(0, \frac{1}{12N}\right)$

$$\text{P}(|T'| < 10) = \text{P}\left(|\bar{E}| < \frac{10}{N}\right) \geq 0.95$$

$$= \text{P}\left(|Z| < \frac{\frac{10}{N}}{\sqrt{\frac{1}{12N}}}\right) \geq 0.95$$

$$\Rightarrow \frac{10}{N\sqrt{\frac{1}{12N}}} \geq 1.96 \Rightarrow 10\sqrt{12} > 1.96\sqrt{N}$$

$$\sqrt{N} < \frac{10\sqrt{12}}{1.96}$$

$$\sqrt{N} < 17.673$$

$$N < 312.3$$

so need 312

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise D, Question 12**

**Question:**

An electrical company repairs very large numbers of television sets and wishes to estimate the mean time taken to repair a particular fault. It is known from previous research that the standard deviation of the time taken to repair this particular fault is 2.5 minutes.
The manager wishes to ensure that the probability that the estimate differs from the true mean by less than 30 seconds is 0.95.
Find how large a sample is required.

**Solution:**

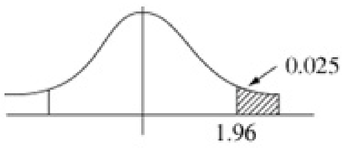
$T$ = time to repair fault

$T \sim \mathrm{N}(\mu, 2.5^2)$

$\overline{T} \simeq\sim \mathrm{N}(\mu, \frac{2.5^2}{n})$

$$\mathrm{P}(|\overline{T} - \mu| < 0.5) = 0.95$$

$$\Rightarrow \mathrm{P}\left(|Z| < \frac{0.5}{\left(\frac{2.5}{\sqrt{n}}\right)}\right) = 0.95$$




$$\therefore \frac{0.5}{\left(\frac{2.5}{\sqrt{n}}\right)} = 1.96$$

$$\sqrt{n} = \frac{1.96 \times 2.5}{0.5}$$

$$n = 9.8^2 = 96.04$$

So need 97 (accept 96)

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests

**Exercise E, Question 1**

**Question:**

A random sample of size 9 is taken from a normal distribution with variance 36. The sample mean is 128.

**a** Find a 95% confidence interval for the mean $\mu$ of the distribution.

**b** Find a 99% confidence interval for the mean $\mu$ of the distribution.

**Solution:**

$n = 9$, $\sigma^2 = 36$, $\bar{x} = 128$

**a** 95% C.I. for $\mu$ is $128 \pm 1.96 \times \frac{6}{\sqrt{9}} = (124.08, 131.92...)$

$= (124, 132)$ (3 s.f.)

**b** 99% C.I. for $\mu$ is $128 \pm 2.5758 \times \frac{6}{\sqrt{9}} = (122.84..., 133.15...)$

$= (123, 133)$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 2**

**Question:**

A random sample of size 25 is taken from a normal distribution with standard deviation 4. The sample mean is 85.

**a** Find a 90% confidence interval for the mean $\mu$ of the distribution.

**b** Find a 95% confidence interval for the mean $\mu$ of the distribution.

**Solution:**

$n = 25,\ \sigma = 4,\ \bar{x} = 85$

**a** 90% C.I. for $\mu$ is $85 \pm 1.6449 \times \frac{4}{\sqrt{25}} = (83.684\ldots, 86.315\ldots)$
$= (83.7, 86.3)$

**b** 95% C.I. for $\mu$ is $85 \pm 1.96 \times \frac{4}{\sqrt{25}} = (83.432, 86.568)$
$= (83.4, 86.6)$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 3**

**Question:**

A normal distribution has mean $\mu$ and variance 4.41. A random sample has the following values:
23.1, 21.8, 24.6, 22.5
Use this sample to find 98% confidence limits for the mean $\mu$.

**Solution:**

$n = 4$, $\sigma^2 = 4.41$, $\bar{x} = 23$

98% C.I. for $\mu$ is $23 \pm 2.3263 \times \sqrt{\dfrac{4.41}{4}}$

$= (20.557, 25.443)$

$= (20.6, 25.4)$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 4**

**Question:**

A normal distribution has standard deviation 15. Estimate the sample size required if the following confidence intervals for the mean should have width of less than 2.

**a** 90%
**b** 95%
**c** 99%

**Solution:**

$\sigma = 15$

C.I. is $\bar{x} \pm z \times \dfrac{\sigma}{\sqrt{n}}$

width $= \dfrac{2z\sigma}{\sqrt{n}}$

**a** $90\% \Rightarrow z = 1.6449 \quad \therefore \dfrac{2 \times 1.6449 \times 15}{\sqrt{n}} < 2$

$\Rightarrow \sqrt{n} > 24.67\ldots \quad \therefore n > 608.78\ldots$

So $n = 609$

**b** $95\% \Rightarrow z = 1.96 \therefore \dfrac{2 \times 1.96 \times 15}{\sqrt{n}} < 2$

$\Rightarrow \sqrt{n} > 1.96 \times 15 \therefore n > 864.36\ldots$

So $n = 865$

**c** $99\% \Rightarrow z = 2.5758 \quad \therefore \dfrac{2 \times 2.5758 \times 15}{\sqrt{n}} < 2$

$\Rightarrow \sqrt{n} > 2.5758 \times 15 \quad \therefore n > 1492.817\ldots$

So $n = 1493$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise E, Question 5**

**Question:**

Repeat Question 4 for a normal distribution with standard deviation 2.4 and a desired width of less than 0.8.

**Solution:**

$\sigma = 2.4$

$\text{width} = \dfrac{2z\sigma}{\sqrt{n}} < 0.8$

$\therefore \sqrt{n} > \dfrac{4.8z}{0.8} = 6z$

**a** $z = 1.6449 \quad \therefore n > 97.405\ldots \quad$ So $n = 98$

**b** $z = 1.96 \quad \therefore n > 138.29\ldots \quad$ So $n = 139$

**c** $z = 2.5758 \quad \therefore n > 238.85\ldots \quad$ So $n = 239$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise E, Question 6**

**Question:**

An experienced poultry farmer knows that the mean weight $\mu$ kg for a large population of chickens will vary from season to season but the standard deviation of the weights should remain at 0.70 kg. A random sample of 100 chickens is taken from the population and the weight $x$ kg of each chicken in the sample is recorded, giving $\Sigma x = 190.2$. Find a 95% confidence interval for $\mu$.

**Solution:**

$\sigma = 0.70,\ n = 100,\ \bar{x} = \frac{190.2}{100} = 1.902$

95% C.I. for $\mu$ is $\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{100}}$

$= \ 1.902 \pm 1.96 \times \frac{0.7}{10}$

$= \ (1.7648,\ 2.0392)$

$= \ (1.76,\ 2.04) \quad (3 \text{ s.f.})$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 7**

**Question:**

A railway watchdog is studying the number of seconds that express trains are late in arriving. Previous surveys have shown that the standard deviation is 50. A random sample of 200 trains was selected and gave rise to a mean of 310 seconds late. Find a 90% confidence interval for the mean number of seconds that express trains are late.

**Solution:**

$\sigma = 50 \quad n = 200 \quad \bar{x} = 310$

90% C.I. is $\bar{x} \pm 1.6449 \times \dfrac{\sigma}{\sqrt{200}}$

$= \left(310 \pm 1.6449 \times \dfrac{50}{\sqrt{200}}\right)$

$= (304.184\ldots, 315.815\ldots)$

$= (304, 316) \quad (3\text{ s.f.})$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 8**

**Question:**

An investigation was carried out into the total distance travelled by lorries in current use. The standard deviation can be assumed to be 15 000 km. A random sample of 80 lorries were stopped and their mean distance travelled was found to be 75 872 km. Find a 90% confidence interval for the mean distance travelled by lorries in current use.

**Solution:**

$\sigma = 15\,000 \quad n = 80 \quad \bar{x} = 75\,872$

90% C.I. is $\bar{x} \pm 1.6449 \times \dfrac{\sigma}{\sqrt{80}}$

$= 75\,872 \pm 1.6449 \times \dfrac{15\,000}{\sqrt{80}}$

$= (73\,113.41\ldots,\ 78\,630.58\ldots)$

$= (73\,113,\ 78\,631)$ (nearest integer)

or $(73\,100,\ 78\,600)$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 9**

**Question:**

It is known that each year the standard deviation of the marks in a certain examination is 13.5 but the mean mark $\mu$ will fluctuate. An examiner wishes to estimate the mean mark of all the candidates on the examination but he only has the marks of a sample of 250 candidates which give a sample mean of 68.4.

**a** What assumption about these candidates must the examiner make in order to use this sample mean to calculate a confidence interval for $\mu$?

**b** Assuming that the above assumption is justified, calculate a 95% confidence interval for $\mu$.

Later the examiner discovers that the actual value of $\mu$ was 65.3.

**c** What conclusions might the examiner draw about his sample?

**Solution:**

$\sigma = 13.5 \quad n = 250 \quad \bar{x} = 68.4$

**a** Must assume that these students form a random sample or that they are representative of the population.

**b** 95% C.I. is $68.4 \pm 1.96 \times \dfrac{13.5}{\sqrt{250}}$

$= (66.726\ldots, 70.073\ldots)$

$= (66.7, 70.1)$ (3 s.f.)

**c** If $\mu = 65.3$ that is outside the C.I. so the examiner's sample was not representative. The examiner marked more better than average candidates.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 10**

**Question:**

The number of hours for which an electronic device can retain information has a uniform distribution over the range $[\mu - 10, \mu + 10]$ but the value of $\mu$ is not known.

**a** Show that the variance of the number of hours the device can retain the information for is $\frac{100}{3}$.

A random sample of 120 devices were tested and the mean number of hours they were retained information for was 78.7.

**b** Find a 95% confidence interval for $\mu$.

**Solution:**

$\mathrm{H} \sim \mathrm{U}[\mu - 10, \mu + 10]$

**a** $\mathrm{E}(H) = \mu \quad \mathrm{Var}(H) = \frac{(20)^2}{12} = \frac{400}{12} = \frac{100}{3}$

Use formula in S2.

**b** $n = 120 \quad \bar{h} = 78.7$

$$95\% \text{ C.I. is } \bar{h} \pm 1.96 \times \frac{\sqrt{\frac{100}{3}}}{\sqrt{120}}$$

$$= \left(78.7 \pm 1.96 \times \sqrt{\frac{100}{360}}\right)$$

$$= (77.666\ldots, 79.733\ldots)$$

$$= (77.7, 79.7) \quad (3 \text{ s.f.})$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 11**

**Question:**

A statistics student calculated a 95% and a 99% confidence interval for the mean $\mu$ of a certain population but failed to label them. The two intervals were (22.7, 27.3) and (23.2, 26.8).

**a** State, with a reason, which interval is the 95% one.
**b** Estimate the standard error of the mean in this case.
**c** What was the student's unbiased estimate of the mean $\mu$ in this case?

**Solution:**

**a** (23.2, 26.8) is 95% C.I. since it is the narrower interval.

**b** $\bar{x} = \frac{1}{2}(23.2 + 26.8) = 25$

$\therefore 1.96 \frac{\sigma}{\sqrt{n}} = 25 - 23.2 = 1.8$

$\therefore \frac{\sigma}{\sqrt{n}} = 0.9183\ldots = 0.918 \quad (3\text{ s.f.})$

**c** $\hat{\mu} = \bar{x} = 25$ (mid-point of the intervals)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 12**

**Question:**

A 95% confidence interval for a mean $\mu$ is $85.3 \pm 2.35$. Find the following confidence intervals for $\mu$.

**a** 90%
**b** 98%
**c** 99%

**Solution:**

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} = 85.3 \pm 2.35$$

$$\therefore \bar{x} = 85.3 \quad \text{and} \quad \frac{\sigma}{\sqrt{n}} = \frac{2.35}{1.96} = 1.1989.$$

**a** 90% C.I. is $85.3 \pm 1.6449 \times \frac{2.35}{1.96}$

$= (83.327\ldots, 87.272\ldots)$

$= (83.3, 87.3) \quad (3\text{ s.f.})$

**b** 98% C.I. is $85.3 \pm 2.3263 \times \frac{2.35}{1.96}$

$= (82.510\ldots, 88.089\ldots)$

$= (82.5, 88.1) \quad (3\text{ s.f.})$

**c** 99% C.I. is $85.3 \pm 2.5758 \times \frac{2.35}{1.96}$

$= (82.211\ldots, 88.388\ldots)$

$= (82.2, 88.4) \quad (3\text{ s.f.})$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 13**

**Question:**

The managing director of a certain firm has commissioned a survey to estimate the mean expenditure of customers on electrical appliances. A random sample of 100 people were questioned and the research team presented the managing director with a 95% confidence interval of (£128.14, £141.86). The director says that this interval is too wide and wants a confidence interval of total width £10.

**a** Using the same value of $\bar{x}$, find the confidence limits in this case.

**b** Find the level of confidence for the interval in part **a**.

The managing director is still not happy and now wishes to know how large a sample would be required to obtain a 95% confidence interval of total width no more than £10.

**c** Find the smallest size of sample that will satisfy this request.

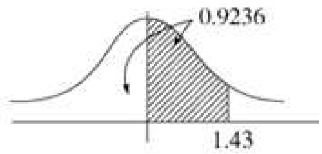
**Solution:**

**a** $\bar{x} = \frac{1}{2}(128.14 + 141.86) = \frac{270}{2} = 135$

$\therefore$ C.I. will be (130, 140)

**b** $z \times \frac{\sigma}{\sqrt{n}} = 5$ but $1.96 \frac{\sigma}{\sqrt{n}} = 6.86$

$\therefore z = \frac{5}{\left(\frac{6.86}{1.96}\right)} = 1.4285\ldots$ Use 1.43




$\therefore$ C.I. is $2 \times (0.9236 - 0.5)$
$= 0.8472$ (tables)
or $= 0.846872\ldots$ (calculator)

$\therefore$ C.I. is 85%

**c** Now we know $1.96 \frac{\sigma}{\sqrt{100}} = 6.86$

$\therefore \sigma = \frac{6.86 \times 10}{1.96} = 35$

and require $z \times \frac{\sigma}{\sqrt{n}} = 5$ where $z = 1.96$

$\therefore \frac{1.96 \times 35}{5} = \sqrt{n}$

$\Rightarrow n = 188.23\ldots$

$\therefore$ Need $n = 189$ or more

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise E, Question 14**

**Question:**

A plant produces steel sheets whose weights are known to be normally distributed with a standard deviation of 2.4 kg. A random sample of 36 sheets had a mean weight of 31.4 kg. Find 99% confidence limits for the population mean.

**Solution:**

$W \sim \mathrm{N}(\mu, 2.4^2) \quad n = 36 \quad \bar{w} = 31.4$

99% C.I. is $31.4 \pm 2.5758 \times \frac{2.4}{\sqrt{36}}$

$= (30.369\ldots, 32.430\ldots)$

$= (30.4, 32.4) \quad (3 \text{ s.f.})$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 15**

**Question:**

A machine is regulated to dispense liquid into cartons in such a way that the amount of liquid dispensed on each occasion is normally distributed with a standard deviation of 20 ml. Find 99% confidence limits for the mean amount of liquid dispensed if a random sample of 40 cartons had an average content of 266 ml.

**Solution:**

$\sigma = 20,\ n = 40,\ \bar{x} = 266$

99% C.I. is $266 \pm 2.5758 \times \dfrac{20}{\sqrt{40}}$

$= (257.854\ldots, 274.145\ldots)$

$= (258, 274)$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise E, Question 16**

**Question:**

**a** The error made when a certain instrument is used to measure the body length of a butterfly of a particular species is known to be normally distributed with mean 0 and standard deviation 1 mm. Calculate, to 3 decimal places, the probability that the error made when the instrument is used once is numerically less than 0.4mm.

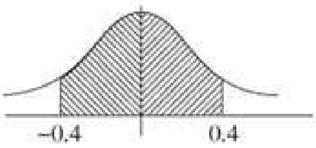
**b** Given that the body length of a butterfly is measured 9 times with the instrument, calculate, to 3 decimal places, the probability that the mean of the 9 readings will be within 0.5 mm of the true length.

**c** Given that the mean of the 9 readings was 22.53 mm, determine a 98% confidence interval for the true body length of the butterfly.

**Solution:**

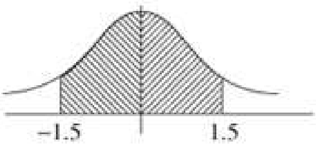
$E \sim \mathrm{N}(0, 1^2)$

**a** $\mathrm{P}(|E| < 0.4) = (0.6554 - 0.5) \times 2$
$= 0.3108$




**b** $\bar{E} \sim \mathrm{N}(0, \frac{1}{9})$

$\mathrm{P}(|\bar{E}| < 0.5) = \mathrm{P}(|Z| < \frac{0.5}{\sqrt{\frac{1}{9}}})$
$= (0.9332 - 0.5) \times 2$
$= 0.8664$




**c** 98% C.I. is $22.53 \pm 2.3263 \times \frac{1}{\sqrt{9}}$
$= (21.754\ldots, 23.305\ldots)$
$= (21.8, 23.3)$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 1**

**Question:**

A random sample of size $n$ is taken from a population having a normal distribution with mean $\mu$ and variance $\sigma^2$. Test the hypotheses at the stated levels of significance.

$H_0; \mu = 21$, $H_1 : \mu \neq 21$, $n = 20$, $\bar{x} = 21.2$, $\sigma = 1.5$, at the 5% level

**Solution:**

$H_0 : \mu = 21 \quad H_1 : \mu \neq 21$ 5% c.v. is $z = \pm 1.96$

$$\text{t.s.} = z = \frac{(21.2 - 21)}{\left(\frac{1.5}{\sqrt{20}}\right)} = 0.596\ldots < 1.96$$

Not significant

Accept $H_0$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 2**

**Question:**

A random sample of size $n$ is taken from a population having a normal distribution with mean $\mu$ and variance $\sigma^2$. Test the hypotheses at the stated levels of significance.

$\mathrm{H}_0: \mu = 100$, $\mathrm{H}_1: \mu < 100$, $n = 36$, $\bar{x} = 98.5$, $\sigma = 5.0$, at the 5% level

**Solution:**

$\mathrm{H}_0: \mu = 100 \quad \mathrm{H}_1: \mu < 100$ 5% c.v. is $z = -1.6449$

$$\text{t.s.} = z = \frac{(98.5 - 100)}{\left(\frac{5}{\sqrt{36}}\right)} = -1.8 < -1.6449$$ Significant

Reject $\mathrm{H}_0$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 3**

**Question:**

A random sample of size $n$ is taken from a population having a normal distribution with mean $\mu$ and variance $\sigma^2$. Test the hypotheses at the stated levels of significance.

$\mathrm{H}_0; \mu = 5, \mathrm{H}_1: \mu \neq 5, n = 25, \bar{x} = 6.1, \ \sigma = 3.0$, at the 5% level

**Solution:**

$\mathrm{H}_0 : \mu = 5 \quad \mathrm{H}_1 : \mu \neq 5$ 5% c.v. is $z = \pm 1.96$

$$\text{t.s.} = z = \frac{(6.1 - 5)}{\left(\frac{3}{\sqrt{25}}\right)} = 1.83\ldots < 1.96$$

Not significant

Accept $\mathrm{H}_0$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests

**Exercise F, Question 4**

**Question:**

A random sample of size $n$ is taken from a population having a normal distribution with mean $\mu$ and variance $\sigma^2$. Test the hypotheses at the stated levels of significance.

$H_0; \mu = 15$, $H_1 : \mu > 15$, $n = 40$, $\bar{x} = 16.5$, $\sigma = 3.5$, at the 1% level

**Solution:**

$H_0 : \mu = 15 \quad H_1 : \mu > 15$ 1% c.v. is $z = 2.3263$

$$\text{t.s.} = z = \frac{(16.5 - 15)}{\left(\frac{3.5}{\sqrt{40}}\right)} = 2.710\ldots > 2.3263$$

Significant

Reject $H_0$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise F, Question 5**

**Question:**

A random sample of size $n$ is taken from a population having a normal distribution with mean $\mu$ and variance $\sigma^2$. Test the hypotheses at the stated levels of significance.

$H_0; \mu = 50, H_1: \mu \neq 50, n = 60, \bar{x} = 48.9, \ \sigma = 4.0$, at the 1% level

**Solution:**

$H_0: \mu = 50 \quad H_1: \mu \neq 50$ — 1% c.v. is $z = \pm 2.5758$

$$\text{t.s.} = z = \frac{(48.9 - 50)}{\left(\frac{4}{\sqrt{60}}\right)} = -2.130\ldots > -2.5758$$

Not significant

Accept $H_0$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 6**

**Question:**

A sample of size $n$ is taken from a population having a $N(\mu, \sigma^2)$ distribution. Find the critical regions for the test statistic $\bar{X}$ in the following test.

$H_0 : \mu = 120, H_1 : \mu < 120, n = 30, \ \sigma = 2.0,$ at the 5% level

**Solution:**

$H_0 : \mu = 120 \quad H_1 : \mu < 120 \quad$ c.v. is $Z = -1.6449$

$$Z = \frac{(\bar{X} - 120)}{\left(\frac{2}{\sqrt{30}}\right)}$$

Reject $H_0$ for $Z < -1.6449$

$$\Rightarrow \quad \bar{X} < 120 - 1.6449 \times \frac{2}{\sqrt{30}}$$

i.e. $\bar{X} < 119.39\ldots$ or $119 \quad (3\text{ s.f.})$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 7**

**Question:**

A sample of size $n$ is taken from a population having a $N(\mu, \sigma^2)$ distribution. Find the critical regions for the test statistic $\bar{X}$ in the following test.

$H_0 : \mu = 12.5, H_1 : \mu > 12.5, n = 25, \sigma = 1.5$, at the 1% level

**Solution:**

$H_0 : \mu = 12.5 \quad H_1 : \mu > 12.5 \quad$ c.v. is $Z = 2.3263$

$$Z = \frac{(\bar{X} - 12.5)}{\left(\frac{1.5}{\sqrt{25}}\right)}$$

Reject $H_0$ for $Z > 2.3263$

$$\Rightarrow \quad \bar{X} > 12.5 + 2.3263 \times \frac{1.5}{\sqrt{25}}$$

i.e. $\bar{X} > 13.19789$

$\bar{X} > 13.2$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 8**

**Question:**

A sample of size $n$ is taken from a population having a $N(\mu, \sigma^2)$ distribution. Find the critical regions for the test statistic $\bar{X}$ in the following test.

$H_0: \mu = 85, H_1: \mu < 85, n = 50, \sigma = 4.0,$, at the 10% level

**Solution:**

$H_0: \mu = 85 \quad H_1: \mu < 85 \quad$ c.v. is $Z = -1.2816$

$$Z = \frac{(\bar{X} - 85)}{\left(\frac{4}{\sqrt{50}}\right)}$$

Reject $H_0$ for $Z < -1.2816$

$$\Rightarrow \quad \bar{X} < 85 - 1.2816 \times \frac{4}{\sqrt{50}}$$

i.e. $\bar{X} < 84.2750\ldots$ or $\bar{X} < 84.3$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 9**

**Question:**

A sample of size $n$ is taken from a population having a $N(\mu, \sigma^2)$ distribution. Find the critical regions for the test statistic $\bar{X}$ in the following test.

$H_0 : \mu = 0, H_1 : \mu \neq 0, n = 45, \sigma = 3.0,$ at the 5% level

**Solution:**

$H_0 : \mu = 0 \quad H_1 : \mu \neq 0 \quad$ c.v. is $Z = \pm 1.96$

$$Z = \frac{(\bar{X} - 0)}{\left(\frac{3}{\sqrt{45}}\right)}$$

Reject $H_0$ for $|Z| > 1.96$

$$\Rightarrow \quad \bar{X} > 0 + 1.96 \times \frac{3}{\sqrt{45}} \text{ or } \bar{X} < 0 - 1.96 \times \frac{3}{\sqrt{45}}$$

i.e. $\bar{X} > 0.87653\ldots$ or $\bar{X} < -0.87653\ldots$

i.e. $\bar{X} > 0.877$ or $\bar{X} < -0.877 \quad$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 10**

**Question:**

A sample of size $n$ is taken from a population having a $N(\mu, \sigma^2)$ distribution. Find the critical regions for the test statistic $\bar{X}$ in the following test.

$H_0 : \mu = -8$, $H_1 : \mu \neq -8$, $n = 20$, $\sigma = 1.2$, at the 1% level

**Solution:**

$H_0 : \mu = -8 \quad H_1 : \mu \neq -8 \quad$ c.v. is $Z = \pm 2.5758$

$$Z = \frac{\left(\bar{X} - (-8)\right)}{\left(\frac{1.2}{\sqrt{20}}\right)} \qquad \text{Reject } H_0 \text{ for } |Z| > 2.5758$$

$$\Rightarrow \bar{X} > -8 + 2.5758 \times \frac{1.2}{\sqrt{20}} \text{ or } \bar{X} < -8 - 2.5758 \times \frac{1.2}{\sqrt{20}}$$

$$\bar{X} > -7.3088\ldots \text{ or } \bar{X} < -8.6911\ldots$$

$$\text{i.e. } \bar{X} > -7.31 \text{ or } \bar{X} < -8.69 \quad (3 \text{ s.f.})$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 11**

**Question:**

The times taken for a capful of stain remover to remove a standard chocolate stain from a baby's bib are normally distributed with a mean of 185s and a standard deviation of 15s. The manufacturers of the stain remover claim to have developed a new formula which will shorten the time taken for a stain to be removed. A random sample of 25 capfuls of the new formula are tested and the mean time for the sample is 179s.
Test, at the 5% level, whether or not there is evidence that the new formula is an improvement.

**Solution:**

$\sigma = 15 \quad n = 25 \quad \bar{x} = 179$

$H_0 : \mu = 185$ (no improvement) $\qquad H_1 : \mu < 185$ (shorter time)

$$\text{t.s. is } z = \frac{(179 - 185)}{\left(\frac{15}{\sqrt{25}}\right)} = -2 < -1.6449, \quad 5\% \text{ c.v. is } z = -1.6449$$

Result is significant, so reject $H_0$
There is evidence that the new formula is an improvement.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 12**

**Question:**

The IQ scores of a population are normally distributed with a mean of 100 and standard deviation of 15. A psychologist wishes to test the theory that eating chocolate before sitting an IQ test improves your score. A random sample of 100 people are selected and they are each given a 100 g bar of chocolate to eat before taking a standard IQ test. Their mean score on the test was 102.5. Test the psychologist's theory at the 5% level.

**Solution:**

$\sigma = 15, n = 100, \bar{x} = 102.5$

$H_0 : \mu = 100 \quad H_1 : \mu > 100$ (improved score)

t.s. is $z = \dfrac{(102.5 - 100)}{\left(\dfrac{15}{\sqrt{100}}\right)} = 1.66 > 1.6449$, 5% c.v. is $z = 1.6449$

Result is significant so reject $H_0$.

There is evidence to support the psychologist's theory.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 13**

**Question:**

The diameters of circular cardboard drinks mats produced by a certain machine are normally distributed with a mean of 9 cm and a standard deviation of 0.15 cm. After the machine is serviced a random sample of 30 mats is selected and their diameters are measured to see if the mean diameter has altered. The mean of the sample was 8.95 cm. Test, at the 5% level, whether there is significant evidence of a change in the mean diameter of mats produced by the machine.

**Solution:**

$\sigma = 0.15, n = 30, \bar{x} = 8.95$

$\mathrm{H}_0 : \mu = 9$ (no change) $\mathrm{H}_1 : \mu \neq 9$ (change in area diameter)

t.s. is $z = \dfrac{(8.95 - 9)}{\left(\dfrac{0.15}{\sqrt{30}}\right)} = -1.8257$, 5% c.v. is $z = \pm 1.96$

$-1.8257\ldots > -1.96$ so result is not significant.

There is insufficient evidence of a change in mean diameter.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise F, Question 14**

**Question:**

**a** Research workers measured the body lengths, in mm, of 10 specimens of fish spawn of a certain species off the coast of Eastern Scotland and found these lengths to be

12.5 10.2 11.1 9.6 12.1 9.3 10.7 11.4 14.7 10.4

Obtain unbiased estimates for the mean and variance of the lengths of all such fish spawn off Eastern Scotland.

**b** Research shows that, for a very large number of specimens of spawn of this species off the coast of Wales, the mean body length is 10.2 mm. Assuming that the variance of the lengths of spawn off Eastern Scotland is 2.56, perform a significance test at the 5% level to decide whether the mean body length of fish spawn off the coast of Eastern Scotland is larger than that of fish spawn off the coast of Wales.

**Solution:**

$n = 10, \sum x = 112, \sum x^2 = 1277.26$

**a** $\bar{x} = \frac{112}{10} = 11.2$

$\sigma^2 = 2.56$

$s^2 = \frac{1277.26 - 10 \times 11.2^2}{9}$

$s^2 = 2.54$

**b** $H_0 : \mu = 10.2$

$H_1 : \mu > 10.2$ (Scottish fish are longer)

t.s. is $z = \frac{(11.2 - 10.2)}{\sqrt{\frac{2.56}{10}}} = 1.976\ldots$

5% c.v. is $z = 1.6449$

$1.976\ldots > 1.6449$ so the result is significant.

There is evidence that the fish off the coast of Eastern Scotland are longer than those off the coast of Wales.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise F, Question 15**

**Question:**

**a** Explain what you understand by the Central Limit Theorem.

**b** An electrical firm claims that the average lifetime of the bulbs it produces is 800 hours with a standard deviation of 42 hours. To test this claim a random sample of 120 bulbs was taken and these bulbs were found to have an average lifetime of 789 hours. Stating your hypotheses clearly and using a 5% level of significance, test the claim made by the electrical firm.

**Solution:**

**a** The Central Limit Theorem enables you to assume that $\bar{X}$ has a normal distribution no matter what the distribution of $X$ since it states that no matter what the distribution of the parent population, provided the size of randomly chosen samples is sufficiently large, then the distribution of the mean of such samples will be approximately normally distributed.

**b** $n = 120$, $\bar{x} = 789$, $\sigma = 42$

$\mathrm{H}_0 : \mu = 800$ (claim is true) $\mathrm{H}_1 : \mu \neq 800$ (claim is false)

t.s. is $z = \dfrac{(789 - 800)}{\left(\dfrac{42}{\sqrt{120}}\right)} = -2.869\ldots$, c.v. is $z = \pm 1.96$

$-2.869 < -1.96$ so the result is significant

There is evidence that the mean length of lifetime of the bulbs is not 800 hours.

Suspect the claim made by the firm.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise G, Question 1**

**Question:**

Carry out a test on the given hypotheses at the given level of significance. The population from which the random sample is drawn is normally distributed.
$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 > \mu_2$; $n_1 = 15$, $\sigma_1 = 5.0$, $n_2 = 20$, $\sigma_2 = 4.8$, $\bar{x}_1 = 23.8$ and $\bar{x}_2 = 21.5$
using a 5% level

**Solution:**

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 > \mu_2$ 5% c.v. is $z = 1.6449$

$$\text{t.s. is } z = \frac{(23.8 - 21.5) - 0}{\sqrt{\frac{5^2}{15} + \frac{4.8^2}{20}}} = 1.3699\ldots$$

$1.3699 < 1.6449$ so result is not significant, accept $H_0$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise G, Question 2**

**Question:**

Carry out a test on the given hypotheses at the given level of significance.
The population from which the random sample is drawn is normally distributed.
$H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$; $n_1 = 30$, $\sigma_1 = 4.2$, $n_2 = 25$, $\sigma_2 = 3.6$, $\bar{x}_1 = 49.6$ and $\bar{x}_2 = 51.7$
using a 5% level

**Solution:**

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$ 5% c.v. is $z = \pm 1.96$

t.s. is $z = \dfrac{(51.7 - 49.6) - 0}{\sqrt{\dfrac{4.2^2}{30} + \dfrac{3.6^2}{25}}}$ Choose $\bar{x}_2 - \bar{x}_1$ to get $z > 0$

t.s. is $z = 1.996\ldots \quad > 1.96$ so result is significant, reject $H_0$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise G, Question 3**

**Question:**

Carry out a test on the given hypotheses at the given level of significance. The population from which the random sample is drawn is normally distributed.
$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 < \mu_2$; $n_1 = 25$, $\sigma_1 = 0.81$, $n_2 = 36$, $\sigma_2 = 0.75$, $\bar{x}_1 = 3.62$ and $\bar{x}_2 = 4.11$ using a 1% level

**Solution:**

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 < \mu_2$ 1% c.v. is $z = -2.3263$

t.s. is $z = \dfrac{(3.62 - 4.11) - 0}{\sqrt{\dfrac{0.81^2}{25} + \dfrac{0.75^2}{36}}} = -2.3946\ldots$

t.s. is $-2.3946\ldots < -2.3263$ so result is significant, reject $H_0$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise G, Question 4**

**Question:**

Carry out a test on the given hypothesis at the given level of significance. What is the significance of the Central Limit Theorem in these three questions?
$\mathrm{H}_0 : \mu_1 = \mu_2$; $\mathrm{H}_1 : \mu_1 \neq \mu_2$; $n_1 = 85$, $\sigma_1 = 8.2$, $n_2 = 100$, $\sigma_2 = 11.3$, $\bar{x}_1 = 112.0$ and $\bar{x}_2 = 108.1$ using a 1% level

**Solution:**

$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \mathrm{H}_1 : \mu_1 \neq \mu_2$ 1% c.v. is $z = \pm 2.5758$

$$\text{t.s. is } z = \frac{(112.0 - 108.1) - 0}{\sqrt{\frac{8.2^2}{85} + \frac{11.3^2}{100}}} = 2.712\ldots > 2.5758$$

Significant result so reject $\mathrm{H}_0$.

Central Limit Theorem applies since $n_1, n_2$ are large and enables you to assume $\bar{X}_1$ and $\bar{X}_2$ are both normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise G, Question 5**

**Question:**

Carry out a test on the given hypothesis at the given level of significance. What is the significance of the Central Limit Theorem in these three questions?
$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 > \mu_2$; $n_1 = 100$, $\sigma_1 = 18.3$, $n_2 = 150$, $\sigma_2 = 15.4$, $\bar{x}_1 = 72.6$ and $\bar{x}_2 = 69.5$ using a 5% level

**Solution:**

$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 > \mu_2$ 5% c.v. is $z = 1.96$

$$\text{t.s. is } z = \frac{(72.6 - 69.5) - 0}{\sqrt{\frac{18.3^2}{100} + \frac{15.4^2}{150}}} = 1.396\ldots < 1.96$$

Result is not significant so accept $H_0$.

Central Limit Theorem applies since $n_1, n_2$ are both large and enables you to assume $\bar{X}_1$ and $\bar{X}_2$ are normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise G, Question 6**

**Question:**

Carry out a test on the given hypothesis at the given level of significance. What is the significance of the Central Limit Theorem in these three questions?
$H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 < \mu_2$; $n_1 = 120$, $\sigma_1 = 0.013$, $n_2 = 90$, $\sigma_2 = 0.015$, $\bar{x}_1 = 0.863$ and $\bar{x}_2 = 0.868$ using a 1% level

**Solution:**

$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2$ 1% c.v. is $z = -2.3263$

$$\text{t.s. is } z = \frac{(0.863 - 0.868) - 0}{\sqrt{\frac{0.013^2}{120} + \frac{0.015^2}{90}}} = -2.5291\ldots < -2.3263$$

Result is significant so reject $H_0$.

Central Limit Theorem is used to assume $\bar{X}_1$ and $\bar{X}_2$ are normally distributed since both samples are large.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests

**Exercise G, Question 7**

**Question:**

A certain factory has two machines designed to cut piping. The first machine works to a standard deviation of 0.011 cm and the second machine has a standard deviation of 0.015 cm. A random sample of 10 pieces of piping from the first machine has a mean length of 6.531 cm and a random sample of 15 pieces from the second machine has a length of 6.524 cm. Assuming that the lengths of piping follow a normal distribution, test, at the 5% level, whether or not the machines are producing piping of the same mean length.

**Solution:**

$\sigma_1 = 0.011 \quad n_1 = 10 \quad \bar{x}_1 = 6.531$

$\sigma_2 = 0.015 \quad n_2 = 15 \quad \bar{x}_2 = 6.524$

$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \mathrm{H}_1 : \mu_1 \neq \mu_2$, 5% c.v. is $z = \pm 1.96$

$$\text{t.s. is } z = \frac{(6.531 - 6.524) - 0}{\sqrt{\frac{0.011^2}{10} + \frac{0.015^2}{15}}} = 1.34466\ldots < 1.96$$

Not significant.

Accept $\mathrm{H}_0$.

There is insufficient evidence to suggest that the machines are producing pipes of different lengths.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests

**Exercise G, Question 8**

**Question:**

A certain health authority set up an investigation to examine the ages of mothers when they give birth to their first children.

A random sample of 250 first-time mothers from a certain year had a mean age of 22.45 years with a standard deviation of 2.9 years. A further random sample of 280 first-time mothers taken 10 years later had a mean age of 22.96 years with a standard deviation of 2.8 years.

**a** Test whether or not these figures suggest that there is a difference in the mean age of first-time mothers between these two dates.

**b** State any assumptions you have made about the distribution of ages of first-time mothers.

**Solution:**

$n_1 = 250 \quad \bar{x}_1 = 22.45 \quad S_1 = 2.9$

$n_2 = 280 \quad \bar{x}_2 = 22.96 \quad S_2 = 2.8$

Assume $S_i = \sigma_i$ since samples are large.

$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \mathrm{H}_1 : \mu_1 \neq \mu_2$

Use 5% significance level

c.v. is $Z = \pm 1.96$

**a** $$\text{t.s.} = Z = \frac{(22.96 - 22.45) - 0}{\sqrt{\frac{2.9^2}{250} + \frac{2.8^2}{280}}}$$

Use $\bar{x}_2 - \bar{x}_1$ to make

t.s. $= Z = 2.054\ldots \quad > 1.96$

Result is significant.

There is evidence of a difference in mean age of first-time mothers between these two dates.

**b** There is no need to have to assume that both populations were normally distributed since both samples were large so the Central Limit Theorem allows you to assume both sample means are normally distributed.

We have assumed that $S_1 = \sigma_1$ and $S_2 = \sigma_2$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise H, Question 1**

**Question:**

An experiment was conducted to compare the drying properties of two paints, Quickdry and Speedicover. In the experiment, 200 similar pieces of metal were painted, 100 randomly allocated to Quickdry and the rest to Speedicover.
The table below summarises the times, in minutes, taken for these pieces of metal to become touch-dry.

| | Quickdry | Speedicover |
|---|---|---|
| Mean | 28.7 | 30.6 |
| Standard deviation | 7.32 | 3.51 |

Using a 5% significance level, test whether or not the mean time for Quickdry to become touch-dry is less than that for Speedicover. State your hypotheses clearly. ***E***

**Solution:**

$n_Q = 100 \quad \bar{x}_Q = 28.7 \quad s_Q = 7.32$

$n_S = 100 \quad \bar{x}_S = 30.6 \quad s_S = 3.51$

$\mathrm{H}_0 : \mu_Q = \mu_S \quad \mathrm{H}_1 : \mu_Q < \mu_S$ (i.e. Quickdry dries in a shorter time than Speedicover.)

$$\text{t.s. is } z = \frac{(30.6 - 28.7) - 0}{\sqrt{\frac{7.32^2}{100} + \frac{3.51^2}{100}}}$$

Test $\mu_S > \mu_Q$ to get $z > 0$.

$$= 2.34$$

5% c.v. is $z = 1.6449$

t.s. is $2.34 > 1.6449$ so the result is significant.
There is evidence that Quickdry dries faster than Speedicover.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise H, Question 2**

**Question:**

A supermarket examined a random sample of 80 weekend shoppers' purchases and an independent random sample of 120 weekday shoppers' purchases. The results are summarised in the table below.

| | $n$ | $\bar{x}$ | $s$ |
|---|---|---|---|
| Weekend | 80 | 38.64 | 6.59 |
| Weekday | 120 | 40.13 | 8.23 |

**a** Stating your hypotheses clearly test, at the 5% level of significance, whether there is evidence that the mean expenditure in the week is more than at weekends.

**b** State an assumption you have made in carrying out this test.

**Solution:**

**a** $n_1 = 80 \quad \bar{x}_1 = 38.64 \quad s_1 = 6.59$

$n_2 = 120 \quad \bar{x}_2 = 40.13 \quad s_2 = 8.23$

$\mathrm{H}_0 : \mu_1 = \mu_2 \quad \mathrm{H}_1 : \mu_2 > \mu_1$ 5% c.v. is $z = 1.6449$

$$\text{t.s. is } z = \frac{(40.13 - 38.64) - 0}{\sqrt{\frac{6.59^2}{80} + \frac{8.23^2}{120}}} = 1.4159 < 1.6449$$

Not significant

There is insufficient evidence to confirm that mean expenditure in the week is more than at week ends.

**b** We have assumed that $s_1 = \sigma_1$ and $s_2 = \sigma_2$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise H, Question 3**

**Question:**

It is claimed that the masses of components produced in a small factory have a mean mass of 10 g. A random sample of 250 of these components is tested and the sample mean, $\bar{x}$, is 9.88 g and the standard deviation, $s$, is 1.12 g.

**a** Test, at the 5% level, whether or not there has been change in the mean mass of a component.

**b** State any assumptions you would make to carry out this test.

**Solution:**

$s = \sigma = 1.12, n = 250, \bar{x} = 9.88$

**a** $H_0 : \mu = 10 \quad H_1 : \mu \neq 10$ 5% c.v. is $z = \pm 1.96$

t.s. is $z = \dfrac{9.88 - 10}{\sqrt{\dfrac{1.12^2}{250}}} = -1.694\ldots > -1.96$

Not significant

Insufficient evidence to support a change in mean mass.

**b** We have assumed that $s = \sigma$ since $n$ is large.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise H, Question 4**

**Question:**

Two independent samples are taken from population $A$ and population $B$. Carry out the following tests using the information given.

**a** $H_0: \mu_A = \mu_B \quad H_1: \mu_A > \mu_B$ using a 1% level of significance

$n_A = 90,\ \ n_B = 110,\ \ \bar{x}_A = 84.1,\ \ \bar{x}_B = 87.9,\ \ s_A = 12.5,\ \ s_B = 14.6$

**b** $H_0: \mu_A - \mu_B = 2 \quad H_1: \mu_A - \mu_B > 2$ using a 5% level of significance

$n_A = 150,\ \ n_B = 200,\ \ \bar{x}_A = 125.1,\ \ \bar{x}_B = 119.3,\ \ s_A = 23.2,\ \ s_B = 18.4$

**c** State an assumption that you have made in carrying out these tests.

**Solution:**

**a** $H_0: \mu_A = \mu_B \quad H_1: \mu_A < \mu_B \quad$ c.v. is $z = -2.3263$

$$\text{t.s. is } z = \frac{(84.1 - 87.9) - 0}{\sqrt{\frac{12.5^2}{90} + \frac{14.6^2}{110}}} = -1.9825\ldots > -2.3263$$

Not significant so accept $H_0$.

**b** $H_0: \mu_A - \mu_B = 2 \quad H_1: \mu_A - \mu_B > 2 \quad$ c.v. is $z = 1.6449$

$$\text{t.s. is } z = \frac{(125.1 - 119.3) - 2}{\sqrt{\frac{23.2^2}{150} + \frac{18.4^2}{200}}} = 1.6535\ldots > 1.6449$$

Significant so reject $H_0$.

**c** We have assumed $s_A = \sigma_A$ and $s_B = \sigma_B$ since the samples are both large.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise H, Question 5**

**Question:**

A shopkeeper complains that the average weight of chocolate bars of a certain type that he is buying from a wholesaler is less than the stated value of 85.0 g. The shopkeeper weighed 100 bars from a large delivery and found that their weights had a mean of 83.6 g and a standard deviation of 7.2 g. Using a 5% significance level, determine whether or not the shopkeeper is justified in his complaint. State clearly the null and alternative hypotheses that you are using, and express your conclusion in words. ***E***

**Solution:**

$n = 100,\ \bar{x} = 83.6,\ s = 7.2$

$\mathrm{H_0} : \mu = 85 \quad \mathrm{H_1} : \mu < 85 \quad$ c.v. is $z = -1.6449$

$$\text{t.s. is } z = \frac{(83.6 - 85)}{\left(\frac{7.2}{\sqrt{100}}\right)} = -1.944\ldots < -1.6449$$

Significant

There is evidence that the weights of chocolate bars are less than the stated value.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests

**Exercise I, Question 1**

**Question:**

The breaking stresses of rubber bands are normally distributed. A company uses bands with a mean breaking stress of 46.50 N. A new supplier claims that they can supply bands that are stronger and provides a sample of 100 bands for the company to test. The company checked the breaking stress, $x$, for each of these 100 bands and the results are summarised as follows:

$n = 100 \quad \Sigma x = 4715 \quad \Sigma x^2 = 222\,910$

**a** Test, at the 5% level, whether or not there is evidence that the new bands are better.

**b** Find an approximate 95% confidence interval for the mean breaking stress of these new rubber bands.

**Solution:**

**a** $\bar{x} = \dfrac{4715}{100} = 47.15$

$$s = \sqrt{\frac{222\,910 - 100 \times 47.15^2}{99}} = 2.4572\ldots$$

$H_0: \mu = 46.50$ (no better) $\qquad H_1: \mu > 46.50$

$$\text{t.s. is } z = \frac{(47.15 - 46.50)}{\left(\dfrac{2.4572\ldots}{\sqrt{100}}\right)} = 2.645\ldots$$

5% c.v. is $z = 1.6449$

t.s. is $2.645 > 1.6449$

Result is significant so reject $H_0$.

There is evidence that the new bands are better.

**b** 95% C.I. is $\bar{x} \pm 1.96 \dfrac{\sigma}{\sqrt{n}}$

$= 47.15 \pm 1.96 \times \dfrac{2.4572\ldots}{\sqrt{100}}$

$= (46.6683\ldots, 47.6316\ldots)$

$= (46.7, 47.6) \quad (3 \text{ s.f.})$

N.B. Since $n$ is large, we have assumed $s = \sigma$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 2**

**Question:**

On each of 100 days a conservationist took a sample of 1 litre of water from a particular place along a river, and measured the amount, $x$ mg, of chlorine in the sample. The results she obtained are shown in the table.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| ***Number of days*** | 4 | 8 | 20 | 22 | 16 | 13 | 10 | 6 | 1 |

**a** Calculate the mean amount of chlorine present per litre of water, and estimate, to 3 decimal places, the standard error of this mean.

**b** Obtain approximate 98% confidence limits for the mean amount of chlorine present per litre of water.

Given that measurements at the same point under the same conditions are taken for a further 100 days,

**c** estimate, to 3 decimal places, the probability that the mean of these measurements will be greater than 4.6 mg per litre of water. ***E***

**Solution:**

$n = 100, \quad \sum x = 453, \quad \sum x^2 = 2391$

**a** $\bar{x} = \frac{453}{100} = 4.53$

$s = \sqrt{\frac{2391 - 100 \times 4.53^2}{99}} = 1.85022\ldots$

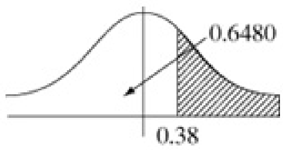
Standard error $= \frac{s}{\sqrt{n}} = 0.185$ (3 d.p.)

**b** 98% C.I. is $\bar{x} \pm 2.3263 \frac{\sigma}{\sqrt{n}}$ Use $\frac{s}{\sqrt{n}}$

$= (4.0995\ldots, 4.9604)$

$= (4.10, 4.96)$ (3 s.f.)

**c**




$\mathrm{P}(\bar{x} > 4.6) = \mathrm{P}\left(Z > \frac{4.6 - 4.53}{0.185\ldots}\right)$

$= \mathrm{P}(Z > 0.378\ldots)$ Use 0.38

$= 1 - 0.6480$

$= 0.3520$ (tables)

or $= 0.35259\ldots$ (calculator)

So accept awrt $0.352 \sim 0.353$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 3**

**Question:**

The amount, to the nearest mg, of a certain chemical in particles in the atmosphere at a meterological station was measured each day for 300 days. The results are shown in the table.

| **Amount of chemical (mg)** | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|
| **Number of days** | 5 | 42 | 210 | 31 | 12 |

Find the mean daily amount of chemical over the 300 days and estimate, to 2 decimal places, its standard error.

**Solution:**

$n = 300, \sum x = 4203, \sum x^2 = 59\,025$

**a** $\bar{x} = \dfrac{4203}{300} = 14.01$

$s = \sqrt{\dfrac{59\,025 - 300 \times 14.01^2}{299}} = 0.6866\ldots$

$\text{Standard error} = \dfrac{s}{\sqrt{n}} = 0.039643\ldots = 0.04 \text{ (2 d.p.)}$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**

**Exercise I, Question 4**

**Question:**

From time to time a firm manufacturing pre-packed furniture needs to check the mean distance between pairs of holes drilled by machine in pieces of chipboard to ensure that no change has occurred. It is known from experience that the standard deviation of the distance is 0.43 mm. The firm intends to take a random sample of size $n$, and to calculate a 99% confidence interval for the mean of the population. The width of this interval must be no more than 0.60 mm.

Calculate the minimum value of $n$. ***E***

**Solution:**

$\sigma = 0.43$

Width of 99% C.I. is $2 \times 2.5758 \frac{\sigma}{\sqrt{n}}$

Require $\frac{2 \times 2.5758 \times 0.43}{\sqrt{n}} < 0.60$

$\therefore \sqrt{n} > \frac{2 \times 2.5758 \times 0.43}{0.6} = 3.691\ldots$

$n > 13.63\ldots$

So the smallest value of $n$ is 14

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 5**

**Question:**

The times taken by five-year-old children to complete a certain task are normally distributed with a standard deviation of 8.0s. A random sample of 25 five-year-old children from school $A$ were given this task and their mean time was 44.2s.

**a** Find 95% confidence limits for the mean time taken by five-year-old children from school $A$ to complete this task.

The mean time for a random sample of 20 five-year-old children from school $B$ was 40.9 s. The headteacher of school $B$ concluded that the overall mean for school $B$ must be less than that of school $A$. Given that the two samples were independent,

**b** test the headteacher's conclusion using a 5% significance level. State your hypotheses clearly. ***E***

**Solution:**

$\sigma = 8.0$

$n_A = 25 \quad \bar{x}_A = 44.2$

**a** 95% C.I. is $44.2 \pm 1.96 \times \dfrac{8.0}{\sqrt{25}}$

$= (41.064, 47.336)$

$= (41.1, 47.3) \quad$ (3 s.f.)

**b** $n_B = 20 \quad \bar{x}_B = 40.9$

$H_0 : \mu_A = \mu_B \quad H_1 : \mu_B < \mu_A$ $\qquad$ 5% c.v. is $z = -1.6449$

$$\text{t.s. is } z = \frac{(40.9 - 44.2) - 0}{\sqrt{\dfrac{8^2}{20} + \dfrac{8^2}{25}}} = -1.375 > -1.6449$$

Not significant so accept $H_0$.

There is insufficient evidence to support the headteacher's claim.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise I, Question 6**

**Question:**

The random variable $X$ has a normal distribution with mean $\mu$ and standard deviation 2.

A random sample of 25 observations is taken and the sample mean $\bar{X}$ is calculated in order to test the null hypothesis $\mu = 7$ against the alternative hypothesis $\mu > 7$ using a 5% level of significance.

Find the critical region for $\bar{X}$. ***E***

**Solution:**

$X \sim \mathrm{N}(\mu, 2^2)$

$n = 25 \quad \bar{X} \sim \mathrm{N}\left(\mu, \left(\frac{2}{5}\right)^2\right)$

$\mathrm{H}_0 : \mu = 7 \quad \mathrm{H}_1 : \mu > 7$ c.v. is $z = 1.6449$

Reject $\mathrm{H}_0$ for $Z > 1.6449$

$$Z = \frac{(\bar{X} - 7)}{\left(\frac{2}{5}\right)} \Rightarrow \bar{X} > 7 + 1.6449 \times \frac{2}{5}$$

i.e. $\bar{X} > 7.65796$

i.e. $\bar{X} > 7.66$ (3 s.f.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 7**

**Question:**

A certain brand of mineral water comes in bottles. The amount of water in a bottle, in millilitres, follows a normal distribution of mean $\mu$ and standard deviation 2. The manufacturer claims that $\mu$ is 125. In order to maintain standards the manufacturer takes a sample of 15 bottles and calculates the mean amount of water per bottle to be 124.2 millilitres.
Test, at the 5% level, whether or not there is evidence that the value of $\mu$ is lower than the manufacturer's claim. State your hypotheses clearly. ***E***

**Solution:**

$B \sim \mathrm{N}(\mu, 2^2)$

$n = 15 \qquad \bar{B} \sim \mathrm{N}(\mu, \frac{4}{15}) \quad \bar{b} = 124.2$

$\mathrm{H}_0 : \mu = 125 \quad \mathrm{H}_1 : \mu < 125 \quad$ c.v. is $z = -1.6449$

$$\text{t.s. is } z = \frac{(124.2 - 125) - 0}{\sqrt{\frac{4}{15}}} = -1.5491\ldots > -1.6449$$

Not significant, so accept $\mathrm{H}_0$.

There is insufficient evidence to suggest that the mean contents of a bottle is lower than the manufacturer's claim.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 8**

**Question:**

The random variable $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$.

**a** Write down the distribution of the sample mean $\bar{X}$ of a random sample of size $n$.

An efficiency expert wishes to determine the mean time taken to drill a fixed number of holes in a metal sheet.

**b** Determine how large a random sample is needed so that the expert can be 95% certain that the sample mean time will differ from the true mean time by less than 15 seconds. Assume that it is known from previous studies that $\sigma = 40$ seconds. ***E***

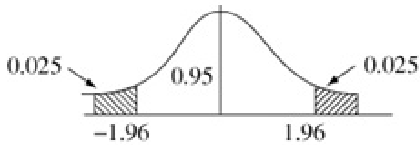
**Solution:**

$X \sim \mathrm{N}(\mu, \sigma^2)$

**a** $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

**b** $\mathrm{P}(|\bar{X} - \mu| < 15) = \mathrm{P}\left(|Z| < \frac{15}{\left(\frac{\sigma}{\sqrt{n}}\right)}\right)$

Require $\mathrm{P}\left(|Z| < \frac{15\sqrt{n}}{\sigma}\right) > 0.95$




i.e. $\frac{15\sqrt{n}}{\sigma} > 1.96$

$\sigma = 40 \Rightarrow \sqrt{n} > \frac{40 \times 1.96}{15} = 5.2266\ldots$

$\therefore n > 27.318\ldots$

So need $n = 28$ or more

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise I, Question 9**

**Question:**

A commuter regularly uses a train service which should arrive in London at 0931. He decided to test this stated arrival time. Each working day for a period of 4 weeks he recorded the number of minutes $x$ that the train was late on arrival in London. If the train arrived early then the value of $x$ was negative. His results are summarised as follows:

$n = 20$, $\Sigma x = 15.0$, $\Sigma x^2 = 103.21$.

**a** Calculate unbiased estimates of the mean and variance of the number of minutes late of his train service.

The random variable $X$ represents the number of minutes that the train is late on arriving in London. Records kept by the railway company show that over fairly short periods, the standard deviation of $X$ is 2.5 minutes. The commuter made 2 assumptions about the distribution of $X$ and the values obtained in the sample and went on to calculate a 95% confidence interval for the mean arrival time of this train service.

**b** State the two assumptions.

**c** Find the confidence interval.

**d** Given that the assumptions are reasonable, comment on the stated arrival time of the service. ***E***

**Solution:**

**a** $\bar{x} = \frac{\sum x}{n} = \frac{15.0}{20} = 0.75$

$s^2 = \frac{103.21 - 20 \times 0.75^2}{19} = 4.84$

**b** $\sigma = 2.5$

**i** assume that $X$ has a normal distribution

**ii** assume that the sample was random.

**c** 95% C.I. is $\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$

$= 0.75 \pm 1.96 \times \frac{2.5}{\sqrt{20}}$

$= (-0.34567\ldots, 1.8456\ldots)$

$= (-0.346, 1.85)$ (3 s.f.)

**d** Since 0 is in the interval it is reasonable to assume that trains do arrive on time.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Estimation, confidence intervals and tests
**Exercise I, Question 10**

**Question:**

The random variable $X$ is normally distributed with mean $\mu$ and variance $\sigma^2$.

**a** Write down the distribution of the sample mean $\bar{X}$ of a random sample of size $n$.

**b** Explain what you understand by a 95% confidence interval.

A garage sells both leaded and unleaded petrol. The distribution of the values of sales for each type is normal. During 1990 the standard deviation of individual sales of each type of petrol is £3.25. The mean of the individual sales of leaded petrol during this time is £8.72. A random sample of 100 individual sales of unleaded petrol gave a mean of £9.71.

Calculate

**c** an interval within which 90% of the sales of leaded petrol will lie,

**d** a 95% confidence interval for the mean sales of unleaded petrol.

The mean of the sales of unleaded petrol for 1989 was £9.10.

**e** Using a 5% significance level, investigate whether there is sufficient evidence to conclude that the mean of all the 1990 unleaded sales was *greater* than the mean of the 1989 sales.

**f** Find the size of the sample that should be taken so that the garage proprietor can be 95% certain that the sample mean of sales of unleaded petrol during 1990 will differ from the true mean by less than 50p. ***E***

**Solution:**

**a** $\bar{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$

**b** 95% C.I. is an interval within which we are 95% confident $\mu$ lies.

$L =$ sales of leaded petrol $\quad U =$ sales of unleaded petrol

$L \sim N(8.72, 3.25^2) \quad U \sim N(9.71, 3.25^2)$

**c** 90% of $L$ between $8.72 \pm 1.6449 \times 3.25$

$= (3.3740\ldots, 14.0659\ldots)$

$= (3.37, 14.1)$ (3 s.f.)

**d** $n = 100 \quad \bar{u} = 9.71$

95% C.I. for $\mu_u$ is: $9.71 \pm 1.96 \times \frac{3.25}{\sqrt{100}}$

$= (9.073, 10.347)$

$= (9.07, 10.35)$ (nearest penny)

**e** $\mathrm{H}_0 : \mu_u = 9.10$ (i.e. same as 1989) $\quad \mathrm{H}_1 : \mu_u > 9.10$ (1990 sales > 1989 sales)

5% c.v. is $z = 1.6449$

t.s. is $z = \frac{(9.71 - 9.10)}{\left(\frac{3.25}{\sqrt{100}}\right)} = 1.8769\ldots > 1.6449$ Significant so reject $\mathrm{H}_0$.

There is evidence that the mean sales of unleaded petrol in 1990 were greater than in 1989.

**f** $\mathrm{P}(|\bar{U} - \mu_u| < 0.50) > 0.95 \Rightarrow \frac{0.5\sqrt{n}}{3.25} > 1.96$

i.e. $\sqrt{n} > 12.74$ or $n > 162.30\ldots \therefore n = 163$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 11**

**Question:**

**a** Explain what is meant by a 98% confidence interval for a population mean.

The lengths, in cm, of the leaves of willow trees are known to be normally distributed with variance $1.33\ \text{cm}^2$.

A sample of 40 willow tree leaves is found to have a mean of 10.20 cm.

**b** Estimate, giving your answer to 3 decimal places, the standard error of the mean.

**c** Use this value to estimate symmetrical 95% confidence limits for the mean length of the population of willow tree leaves, giving your answer to 2 decimal places.

**d** Find the minimum size of the sample of leaves which must be taken if the width of the symmetrical 98% confidence interval for the population mean is at most 1.50 cm. ***E***

**Solution:**

**a** A 98% C.I. is an interval within which we are 98% sure the population mean will lie.

**b** $L$ = length of willow tree leaves $\qquad L \sim \text{N}(\mu, 1.33)$

$n = 40,\ \bar{L} = 10.20$

Standard error of the mean $= \dfrac{\sigma}{\sqrt{n}} = \dfrac{\sqrt{1.33}}{\sqrt{40}} = 0.18234\ldots$

$= 0.182$ (3 d.p.)

**c** 95% C.I. is $10.20 \pm 1.96 \times 0.182\cdots$

$= (9.8426\ldots, 10.5573\ldots)$

$= (9.84, 10.56)$ (2 d.p.)

**d** Width of 98% C.I. is $2 \times 2.3263 \times \dfrac{\sigma}{\sqrt{n}}$

$\therefore$ Require $\dfrac{2 \times 2.3263 \times \sqrt{1.33}}{\sqrt{n}} < 1.50$

or $\sqrt{n} > 3.57\ldots$

i.e. $n > 12.79\ldots$

$\therefore$ need $n = 13$

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 12**

**Question:**

The distance driven by a long distance lorry driver in a week is a normally distributed variable having mean 1130 km and standard deviation 106 km.

**a** Find, to 3 decimal places, the probability that in a given week he will drive less than 1000 km.

**b** Find, to 3 decimal places, the probability that in 20 weeks his average distance driven per week is more than 1200 km.

New driving regulations are introduced and, in the first 20 weeks after their introduction, he drives a total of 21 900 km.

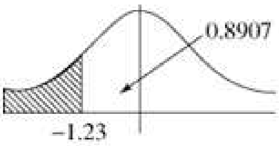
Assuming that the standard deviation of the weekly distances he drives is unchanged,

**c** test, at the 10% level of significance, if his mean weekly driving distance has been reduced. State clearly your null and alternative hypotheses. ***E***

**Solution:**

$D \sim \mathrm{N}(1130, 106^2)$

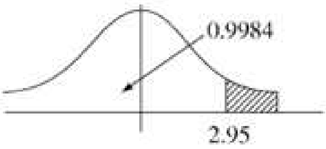
**a**




$$\begin{aligned} \mathrm{P}(D < 1000) &= \mathrm{P}\left(Z < \frac{1000-1130}{106}\right) \\ &= \mathrm{P}(Z < -1.2264\ldots) \quad \text{Use } -1.23 \\ &= 1 - 0.8907 \\ &= 0.1093 \quad \text{(tables)} \\ \text{or} &= 0.11002\ldots \quad \text{(calculator)} \end{aligned}$$

So accept 0.109 or 0.110 (3 d.p.)

**b** $n = 20 \quad \bar{D} \sim \mathrm{N}\left(1130, \frac{106^2}{20}\right)$




$$\begin{aligned} \mathrm{P}(\bar{D} > 1200) &= \mathrm{P}\left(Z > \frac{(1200-1130)}{\left(\frac{106}{\sqrt{20}}\right)}\right) \\ &= \mathrm{P}(Z > 2.953\ldots) \quad \text{Use } 2.95 \\ &= 1 - 0.9984 \\ &= 0.0016 \quad \text{(tables)} \\ \text{or} &= 0.001572\ldots \quad \text{(calculator)} \end{aligned}$$

So accept 0.002 (3 d.p.)

**c** $\mathrm{H}_0 : \mu = 1130$ (same) $\qquad \mathrm{H}_1 : \mu < 1130$ (reduced)

10% c.v. is $z = -1.2816$

$$\bar{d} = \frac{21\,900}{20} = 1095$$

$$\therefore \text{ t.s. is } z = \frac{(1095-1130)}{\left(\frac{106}{\sqrt{20}}\right)} = -1.4766\ldots < -1.2816$$

Significant so reject $\mathrm{H}_0$

There is evidence that his mean weekly driving distance has been reduced.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 13**

**Question:**

Climbing rope produced by a manufacturer is known to be such that one-metre lengths have breaking strengths that are normally distributed with mean 170.2 kg and standard deviation 10.5 kg. Find, to 3 decimal places, the probability that

**a** a one-metre length of rope chosen at random from those produced by the manufacturer will have a breaking strength of 175 kg to the nearest kg,

**b** a random sample of 50 one-metre lengths will have a mean breaking strength of more than 172.4 kg.

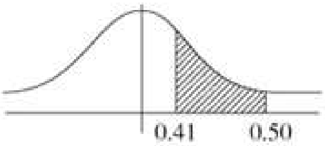
A new component material is added to the ropes being produced. The manufacturer believes that this will increase the mean breaking strength without changing the standard deviation. A random sample of 50 one-metre lengths of the new rope is found to have a mean breaking strength of 172.4 kg.

**c** Perform a significance test at the 5% level to decide whether this result provides sufficient evidence to confirm the manufacturer's belief that the mean breaking strength is increased. State clearly the null and alternative hypotheses that you are using. [E]

**Solution:**

$B =$ breaking strength $\sim N(170.2, 10.5^2)$

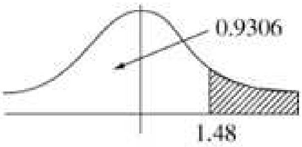
**a**




$$\begin{aligned} P(174.5 < B < 175.5) &= P(0.4095 < Z < 0.5047\cdots) \\ &= 0.6915 - 0.6591 \\ &= 0.0324 \quad \text{(tables)} \\ \text{or} &= 0.034214\ldots \quad \text{(calculator)} \end{aligned}$$

So accept $0.032 \sim 0.034$

**b** $n = 50$ $\quad \bar{B} \sim N\left(170.2, \dfrac{10.5^2}{50}\right)$




$$\begin{aligned} P(\bar{B} > 172.4) &= P\left(Z > \frac{172.4 - 170.2}{\frac{10.5}{\sqrt{50}}}\right) \\ &= P(Z > 1.4815\ldots) \quad \text{Use } 1.48 \\ &= 1 - 0.9306 \\ &= 0.0694 \quad \text{(tables)} \\ \text{or} &= 0.069229\ldots \quad \text{(calculator)} \end{aligned}$$

i.e. accept awrt 0.069

**c** $H_0: \mu = 170.2 \quad H_1: \mu > 170.2$ $\qquad$ 5% c.v. is $z = 1.6449$

t.s. is $z = \dfrac{(172.4 - 170.2)}{\left(\dfrac{10.5}{\sqrt{50}}\right)} = 1.4815\ldots \quad < 1.6449$ $\quad \therefore$ Not significant so accept $H_0$

Insufficient evidence of an increase in the mean breaking strength of climbing rope.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 14**

**Question:**

A machine fills 1 kg packets of sugar. The actual weight of sugar delivered to each packet can be assumed to be normally distributed. The manufacturer requires that,

**i** the mean weight of the contents of a packet is 1010 g, and

**ii** 95% of all packets filled by the machine contain between 1000 g and 1020 g of sugar.

**a** Show that this is equivalent to demanding that the variance of the sampling distribution, to 2 decimal places, is equal to $26.03\text{ g}^2$.

A sample of 8 packets was selected at random from those filled by the machine. The weights, in grams, of the contents of these packets were,

1012.6 1017.7 1015.2 1015.7 1020.9 1005.7 1009.9 1011.4.

Assuming that the variance of the actual weights is $26.03\text{ g}^2$,

**b** test at the 2% significance level, (stating clearly the null and alternative hypotheses that you are using) to decide whether this sample provides sufficient evidence to conclude that the machine is not fulfilling condition **i**. ***E***

**Solution:**

$W$ = weight of sugar

$W \sim \text{N}(1010, \sigma^2)$

**a** $\text{P}(1000 < W < 1020) = 0.95$

$\Rightarrow \quad 1.96\sigma = 10$

$\sigma = 5.102\ldots \quad \sigma^2 = 26.0308\ldots \therefore \sigma^2 = 26.03 \text{ (2 d.p.)}$

**b** $n = 8, \sum x = 8109.1, \left(\sum x^2 = 8\,219\,846.85\right)$

$\bar{x} = 1013.6375$

$\text{H}_0 : \mu = 1010 \quad \text{H}_1 : \mu \neq 1010$ 2% c.v. is $z = \pm 2.3263$

$$\text{t.s. is } z = \frac{(1013.6375 - 1010)}{\left(\frac{\sqrt{26.03}}{\sqrt{8}}\right)} = 2.0165\ldots < 2.3263$$

Not significant so accept $\text{H}_0$.

There is insufficient evidence of a deviation in mean from 1010, so we can assume that condition **i** is being met.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 15**

**Question:**

**a** Write down the mean and the variance of the distribution of the means of all possible samples of size $n$ taken from an infinite population having mean $\mu$ and variance $\sigma^2$.

**b** Describe the form of this distribution of sample means when

**i** $n$ is large,

**ii** the distribution of the population is normal.

The standard deviation of all the till receipts of a supermarket during 1984 was £4.25.

**c** Given that the mean of a random sample of 100 of the till receipts is £18.50, obtain an approximate 95% confidence interval for the mean of all the till receipts during 1984.

**d** Find the size of sample that should be taken so that the management can be 95% confident that the sample mean will not differ from the true mean by more than 50p.

**e** The mean of all the till receipts of the supermarket during 1983 was £19.40. Using a 5% significance level, investigate whether the sample in **a** above provides sufficient evidence to conclude that the mean of all the 1984 till receipts is different from that in 1983. ***E***

**Solution:**

**a** $E(\bar{X})=\mu \quad \text{Var}(\bar{X})=\dfrac{\sigma^2}{n}$

**b** **i** By Central Limit Theorem $\bar{X} \simeq \sim$ Normal i.e. $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

**ii** $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

$\sigma = 4.25$

**c** $n = 100 \quad \bar{x} = 18.50$

95% C.I. is $18.50 \pm 1.96 \times \dfrac{4.25}{\sqrt{100}}$

$= \quad (17.667, 19.333)$

$= \quad (17.7, 19.3) \quad (3\text{ s.f.})$

**d** $P(|\bar{X}-\mu| < 0.50) > 0.95$

i.e. $\dfrac{0.50 \times \sqrt{n}}{4.25} > 1.96$

i.e. $\sqrt{n} > 16.66\ldots$ i.e. $n > 277.55\ldots$

i.e. $n = 278$ or more

**e** $H_0: \mu = 19.4 \quad H_1: \mu \neq 19.4$ 5% c.v. is $z = \pm 1.96$

t.s. is $z = \dfrac{(18.50 - 19.4)}{\left(\dfrac{4.25}{\sqrt{100}}\right)} = -2.1176\ldots \quad < -1.96$ Significant so reject $H_0$.

There is evidence that the mean of till receipts in 1984 is different from the mean value in 1983.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 16**

**Question:**

The diameters of eggs of the little gull are approximately normally distributed with mean 4.11 cm and standard deviation 0.19 cm.

**a** Calculate the probability that an egg chosen at random has a diameter between 3.9 cm and 4.5 cm.

A sample of 8 little gull eggs was collected from a particular island and their diameters, in cm, were

4.4, 4.5, 4.1, 3.9, 4.4, 4.6, 4.5, 4.1

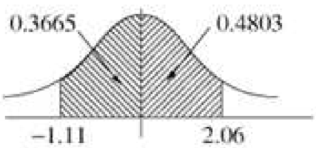
Assuming that the standard deviation of the diameters of eggs from the island is also 0.19 cm,

**b** test, at the 1% level, whether the results indicate that the mean diameter of little gull eggs on this island is different from elsewhere. ***E***

**Solution:**

$D$ = diameter $\sim N(4.11, 0.19^2)$

**a**




$P(3.9 < D < 4.5) = P(-1.105\ldots < Z < 2.052\ldots)$

$= 0.4803 + 0.3665$

$= 0.8468$ (tables)

or $= 0.845423\ldots$ (calculator)

so accept awrt $0.845 \sim 0.847$

**b** $\sigma = 0.19, n = 8, \sum x = 34.5, \bar{x} = 4.3125$

$H_0 : \mu = 4.11 \quad H_1 : \mu \neq 4.11$ 1% significance level c.v. is $z = \pm 2.5758$

t.s. is $z = \dfrac{(4.3125 - 4.11)}{\left(\dfrac{0.19}{\sqrt{8}}\right)} = 3.0145\ldots > 2.5758$ Significant so reject $H_0$.

There is evidence that the mean length of eggs from this island is different from elsewhere.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Estimation, confidence intervals and tests**
**Exercise I, Question 17**

**Question:**

Records of the diameters of spherical ball bearings produced on a certain machine indicate that the diameters are normally distributed with mean 0.824 cm and standard deviation 0.046 cm. Two hundred samples, each consisting of 100 ball bearings, are chosen.

**a** Calculate the expected number of the 200 samples having a mean diameter less than 0.823 cm.

On a certain day it was suspected that the machine was malfunctioning. It may be assumed that if the machine is malfunctioning it will change the mean of the diameters without changing their standard deviation. On that day a random sample of 100 ball bearings had mean diameter of 0.834 cm.

**b** Determine a 98% confidence interval for the mean diameter of the ball bearings being produced that day.

**c** Hence state whether or not you would conclude that the machine is malfunctioning on that day given that the significance level is 2%. ***E***

**Solution:**

$D =$ diameter $\sim \mathrm{N}(0.824, 0.046^2)$
$n = 100$

**a** $\mathrm{P}(\bar{D} < 0.823) = \mathrm{P}\left(Z < \dfrac{(0.823 - 0.824)}{\left(\dfrac{0.046}{\sqrt{100}}\right)}\right)$

$= \mathrm{P}(Z < -0.2173\ldots)$ Use $-0.22$
$= 1 - 0.5871$
$= 0.4129$ (tables)
or $= 0.41395\ldots$ (calculator)

i.e. accept awrt $0.413 \sim 0.414$
So out of $200 \simeq 83$ samples will have mean $< 0.823$

**b** $n = 100, \bar{d} = 0.834$

98% C.I. is $0.834 \pm 2.3263 \times \dfrac{0.046}{\sqrt{100}}$
$= (0.82329\ldots, 0.84470\ldots)$
$= (0.823, 0.845)$ (3 s.f.)

**c** Since 0.824 is *in* the C.I. we can conclude that there is insufficient evidence of a malfunction.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 1**

**Question:**

An octagonal die is thrown 500 times and the results are noted. It is assumed that the die is unbiased. A test is to be done to see whether the observed results differ from the expected ones. Write down a null hypothesis and an alternative hypothesis that can be used.

**Solution:**

$H_0$ : There is no difference between the observed and expected distributions.

$H_1$: There is a difference between the observed and expected distributions.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 2**

**Question:**

For 5 degrees of freedom find the critical value of $\chi^2$ which is exceeded with a probability of 5%.

**Solution:**

11.070

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 3**

**Question:**

Find the values of the following from the table on page 137.

**a** $\chi_5^2(5\%)$

**b** $\chi_8^2(1\%)$

**c** $\chi_{10}^2(10\%)$

**Solution:**

**i** 11.070
**ii** 20.090
**iii** 15.987

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Goodness of fit and contingency tables
**Exercise A, Question 4**

**Question:**

With $\nu = 10$ find the value of $\chi^2$ that is exceeded with 0.05 probability.

**Solution:**

18.307

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 5**

**Question:**

With $\nu = 8$ find the value of $\chi^2$ that is exceeded with 0.10 probability.

**Solution:**

13.362

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 6**

**Question:**

The random variable $Y$ has a $\chi^2$ distribution with 8 degrees of freedom. Find $y$ such that $\mathrm{P}(Y < y) = 0.99$.

**Solution:**

20.090

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 7**

**Question:**

The random variable $X$ has a $\chi^2$ distribution with 5 degrees of freedom. Find $x$ such that $P(X < x) = 0.95$.

**Solution:**

11.070

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise A, Question 8**

**Question:**

The random variable $Y$ has a $\chi^2$ distribution with 12 degrees of freedom.
Find:

**a** $y$ such that $P(Y < y) = 0.05$,

**b** $y$ such that $P(Y < y) = 0.95$.

**Solution:**

**a** 5.226

**b** 21.026

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 1**

**Question:**

The following table shows observed values for what is thought to be a discrete uniform distribution.

| $x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| **Frequency of $x$** | 12 | 24 | 18 | 20 | 25 | 17 | 21 | 23 |

**a** Calculate the expected frequencies and, using a 5% significance level, conduct a goodness of fit test.

**b** State your conclusions.

**Solution:**

**a** Total frequency $= 160$

Expected frequency for each value of $x = \frac{160}{8} = 20$

$H_0$: A discrete uniform distribution is a suitable model.

$H_1$: A discrete uniform distribution is *not* a suitable model.

Test statistic $\left(\chi^2\right) = \frac{(12-20)^2}{20} + \frac{(24-20)^2}{20} + \cdots + \frac{(23-20)^2}{20}$

t.s. $\left(\chi^2\right) \quad = 6.4$

Degrees of freedom $= 8 - 1 = 7$

Critical value (c.v.) $= \chi^2_7(5\%) = 14.067$

$\chi^2 <$ c.v. so accept $H_0$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

**b** There is no reason to doubt that the data could be from a discrete uniform distribution. A discrete uniform distribution is a good model.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**

**Exercise B, Question 2**

**Question:**

The following tables show observed values ($O$) and expected values ($E$) for a goodness of fit test of a binomial distribution model. The probability used in calculating the expected values has not been found from the observed values.

| | | | | | | |
|---|---|---|---|---|---|---|
| ***O*** | 17 | 28 | 32 | 15 | 5 | 3 |
| ***E*** | 19.69 | 34.74 | 27.59 | 12.98 | 4.01 | 0.99 |

**a** Conduct the test using a 5% significance level and state your conclusions.

**b** Suggest how the model might be improved.

**Solution:**

**a** $H_0$: A binomial distribution is a good model.

$H_1$: A binomial distribution is *not* a good model.

Test statistic $\chi^2 = \frac{(17-19.69)^2}{19.69} + \frac{(28-34.74)^2}{34.74} + \cdots + \frac{(8-5.00)^2}{5.00}$

The last 2 groups must be pooled to get $E$ as 5 or more

t.s. $\chi^2 = 4.49$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

There are 5 classes after pooling.

Degrees of freedom $= 5 - 1 = 4$

We only have 1 restriction, the total frequencies agree.

c.v. $= \chi^2_4(5\%) = 9.488$

$\chi^2 <$ c.v. so accept $H_0$

Conclude that there is no reason to doubt the data could be modelled by a binomial distribution.

**b** The mean of the original data could have been calculated and then used to find $p$ using mean $= np$.

This would increase the number of restrictions by 1 and hence reduce the degrees of freedom by 1.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 3**

**Question:**

The following table shows observed values for a distribution which it is thought may be modelled by a Poisson distribution.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | $> 5$ |
|---|---|---|---|---|---|---|---|
| **Frequency of $x$** | 12 | 23 | 24 | 24 | 12 | 5 | 0 |

A possible model is thought to be Po(2). From tables, the expected values are found to be as shown in the following table.

| $X$ | 0 | 1 | 2 | 3 | 4 | 5 | $> 5$ |
|---|---|---|---|---|---|---|---|
| **Expected frequency of $x$** | 13.53 | 27.07 | 27.07 | 18.04 | 9.02 | 3.61 | 1.66 |

**a** Conduct a goodness of fit test at the 5% significance level.

**b** It is suggested that the model could be improved by estimating the value of $\lambda$ from the observed results. What effect would this have on the number of constraints placed upon the degrees of freedom?

**Solution:**

**a** $H_0$: The data can be modelled by Po(2).

$H_1$: The data cannot be modelled by Po(2).

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 or more |
|---|---|---|---|---|---|---|
| $O$ | 12 | 23 | 24 | 24 | 12 | 5 |
| $E$ | 13.53 | 27.07 | 27.07 | 18.04 | 9.02 | 5.27 |

Pool the last 2 classes to get all the $E$ values to be 5 or more.

$$\text{Test statistic } (\chi^2) = \frac{(12-13.53)^2}{13.53} + \frac{(23-27.07)^2}{27.07} + \cdots + \frac{(5-5.27)^2}{5.27}$$

$$\text{t.s. } (\chi^2) = 4.10$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Classes after pooling.

Degrees of freedom $= 6 - 1 = 5$

One restriction, the total frequencies agree.

critical value $= \chi_5^2\,(5\%) = 11.070$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude that there is no reason to doubt the data can be modelled by Po(2).

There is no reason to reject $H_0$.

**b** If $\lambda$ is calculated then this becomes another restriction.

Degrees of freedom $= 6 - 2 = 4$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 4**

**Question:**

A mail order firm receives packets every day through the mail.
They think that their deliveries are uniformly distributed throughout the week. Tes this assertion, given that their deliveries over a 4-week period were as follows. Use a 0.05 significance level.

| **Day** | Mon | Tues | Wed | Thurs | Fri | Sat |
|---|---|---|---|---|---|---|
| **Frequency** | 15 | 23 | 19 | 20 | 14 | 11 |

**Solution:**

$H_0$: The data can be modelled by a discrete uniform distribution.
$H_1$: The data cannot be modelled by a discrete uniform distribution.

$$\text{Expected frequency} = \frac{102}{6} = 17$$

$$\text{Test statistic } \left(\chi^2\right) = \frac{(15-17)^2}{17} + \frac{(23-17)^2}{17} + \cdots + \frac{(11-17)^2}{17}$$

$$\text{t.s. } \left(\chi^2\right) = 5.765$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Degrees of freedom $= 6 - 1 = 5$

One restriction, total frequencies agree.

critical value $= \chi^2_5(5\%) = 11.070$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude that there is no reason to doubt the data can be modelled by a discrete uniform distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 5**

**Question:**

Over a period of 50 weeks the number of road accidents reported to a police station were as shown.

| Number of accidents | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Number of weeks | 15 | 13 | 9 | 13 | 0 |

**a** Find the mean number of accidents per week.
**b** Using this mean and a 0.10 significance level, test the assertion that these data are from a population with a Poisson distribution.

**Solution:**

**a** $\bar{X} = 1.4$

This is one restriction in part **b**.

**b** $H_0$: The data can be modelled by Po(1.4).

$H_1$: The data cannot be modelled by Po(1.4).

A Poisson distribution has no upper limit

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 or more |
|---|---|---|---|---|---|---|
| $O$ | 15 | 13 | 9 | 13 | 0 | 0 |
| $E$ | 12.330 | 17.262 | 12.083 | 5.639 | 1.974 | 0.712 |

50 – [total of other $E$'s in list]

Total frequency = 50

This is one restriction in part **b**.

$$E(0) = \left(e^{-1.4} \times \frac{1.4^0}{0!}\right) \times 50$$

The last 3 classes must be pooled so *all* the $E$'s are 5 or more.

| $x$ | 0 | 1 | 2 | 3 or more |
|---|---|---|---|---|
| $O$ | 15 | 13 | 9 | 13 |
| $E$ | 12.330 | 17.262 | 12.083 | 8.325 |

$$\text{Test statistic } (\chi^2) = \frac{(15-12.330)^2}{12.330} + \frac{(13-17.262)^2}{17.262} + \frac{(9-12.083)^2}{12.083} + \frac{(13-8.325)^2}{8.325}$$

$$\text{t.s. } (\chi^2) = 5.04$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Number of classes after pooling.

Number of restrictions.

$$\text{Degrees of freedom } = 4 - 2$$
$$= 2$$

$$\text{critical value } = \chi^2_2(10\%) = 4.605$$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude that there is evidence to suggest that the data cannot be modelled by Po(1.4).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 6**

**Question:**

A marksman fires 6 shots at a target and records the number $r$ of bull's-eyes hit. After a series of 100 such trials he analyses his scores, the frequencies being as follows.

| $r$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| **Frequency** | 0 | 26 | 36 | 20 | 10 | 6 | 2 |

**a** Estimate the probability of hitting a bull's-eye.
**b** Use a test at the 0.05 significance level to see if these results are consistent with the assumption of a binomial distribution.

**Solution:**

In part **b** this is 1 restriction since a parameter has been estimated by calculation in order to find $P$.

**a** $\bar{r} = 2.4$, $n = 6$

mean $= np$ $\quad 2.4 = 6 \times p$

$\therefore p = 0.4$

**b** Total frequency $= 100$

This is 1 restriction.

$H_0$: The data can be modelled by B(6, 0.4).

$H_1$: The data cannot be modelled by B(6, 0.4).

$$E(X=0) = 100 \times P(X=0) = 100 \times \binom{6}{0} \times 0.4^0 \times 0.6^6 = 4.666$$

$$E(X=1) = 100 \times P(X=1) = 100 \times \binom{6}{1} \times 0.4^1 \times 0.6^5 = 18.662$$

Pool to get $E \geq 5$.

$$E(X=2) = 100 \times P(X=2) = 100 \times \binom{6}{2} \times 0.4^2 \times 0.6^4 = 31.104$$

$$E(X=3) = 100 \times P(X=3) = 100 \times \binom{6}{3} \times 0.4^3 \times 0.6^3 = 27.648$$

$$E(X=4) = 100 \times P(X=4) = 100 \times \binom{6}{4} \times 0.4^4 \times 0.6^2 = 13.824$$

$$E(X=5) = 100 \times P(X=5) = 100 \times \binom{6}{5} \times 0.4^5 \times 0.6^1 = 3.6864$$

$$E(X=6) = 100 \times P(X=6) = 100 \times \binom{6}{6} \times 0.4^6 \times 0.6^0 = 0.4096$$

Pool to get $E \geq 5$.

| $x$ | 1 or less | 2 | 3 | 4 or more |
|---|---|---|---|---|
| $O$ | 26 | 36 | 20 | 18 |
| $E$ | 23.328 | 31.104 | 27.648 | 17.92 |

$$\text{Test statistic } (\chi^2) = \frac{(26-23.328)^2}{23.328} + \frac{(36-31.104)^2}{31.104} + \frac{(20-27.648)^2}{27.648} + \frac{(18-17.92)^2}{17.92}$$

t.s. $(\chi^2) = 3.19$

Degrees of freedom $= 4 - 2$

$= 2$

Number of classes after pooling.

Number of restrictions.

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_2^2(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data can be modelled by B(6, 0.4).

There is no reason to reject $H_0$.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Goodness of fit and contingency tables

**Exercise B, Question 7**

**Question:**

The table below shows the number of employees in thousands at five factories and the number of accidents in 3 years.

| Factory | A | B | C | D | E |
|---|---|---|---|---|---|
| Employees (thousands) | 4 | 3 | 5 | 1 | 2 |
| Accidents | 22 | 14 | 25 | 8 | 12 |

Using a 0.05 significance level, test the hypothesis that the number of accidents per 1000 employees is constant at each factory.

**Solution:**

$H_0$: The rate of accidents is constant at the factories.

$H_1$: The rate of accidents isn't constant at the factories.

Total number of accidents $= 81$

Total number of employees $= 15$(thousand)

$\therefore$ mean rate of accidents $= \frac{81}{15} = 5.4$ (per thousand)

This is 1 restriction, calculation of the mean rate.

| *Factory* | A | B | C | D | E |
|---|---|---|---|---|---|
| *O* | 22 | 14 | 25 | 8 | 12 |
| *E* | 21.6 | 16.2 | 27 | 5.4 | 10.8 |

$$\text{Test statistic } (\chi^2) = \frac{(22-21.6)^2}{21.6} + \frac{(14-16.2)^2}{16.2} + \cdots + \frac{(12-10.8)^2}{10.8}$$
$$= 1.84$$

Degrees of freedom $= 5 - 1 = 4$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_4^2(5\%) = 9.488$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt that accidents occur at a constant rate at the factories.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**

**Exercise B, Question 8**

**Question:**

In a test to determine the red blood cell count in a patient's blood sample, the number of cells in each of 80 squares is counted with the following results.

| **Number of cells per square, $x$** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| **Frequency, $f$** | 2 | 8 | 15 | 18 | 14 | 13 | 7 | 3 | 0 |

It is assumed that these will fit a Poisson distribution. Test this assertion at the 0.05 significance level.

**Solution:**

$H_0$: The data can be modelled by a Poisson distribution.

$H_1$: The data cannot be modelled by a Poisson distribution.

Total frequency $= 80$

This is 1 restriction.

Mean $= \lambda = 3.45$

The mean has been calculated so it is 1 restriction.

$$\mathrm{E}(X=0) = 80 \times \mathrm{P}(0) = 80 \times \left(e^{-3.45} \times \frac{3.45^0}{0!}\right) = 2.540$$

$$\mathrm{E}(X=1) = 80 \times \mathrm{P}(1) = 80 \times \left(e^{-3.45} \times \frac{3.45^1}{1!}\right) = 8.762$$

Pool to get $E$'s 5 or more.

Similarly $\mathrm{E}(2) = 15.114, \mathrm{E}(3) = 17.381, \mathrm{E}(4) = 14.991, \mathrm{E}(5) = 10.344$

$\mathrm{E}(6) = 5.948, \mathrm{E}(7) = 2.931$

$\mathrm{E}(8 \text{ or more}) = 80 - [\mathrm{E}(0) + \mathrm{E}(1) + \cdots + \mathrm{E}(7)] = 1.989$

The last 3 classes must be combined to get $E$ to be 5 or more.

| $X$ | 1 or less | 2 | 3 | 4 | 5 | 6 or more |
|---|---|---|---|---|---|---|
| $O$ | 10 | 15 | 18 | 14 | 13 | 10 |
| $E$ | 11.302 | 15.114 | 17.381 | 14.991 | 10.344 | 10.868 |

$$\text{Test statistic } (\chi^2) = \frac{(10-11.302)^2}{11.302} + \frac{(15-15.114)^2}{15.114} + \cdots + \frac{(10-10.868)^2}{10.868}$$

$$\text{t.s. } (\chi^2) = 0.990$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Degrees of freedom $= 6 - 2$
$= 4$

Number of classes after pooling.

critical value $= \chi^2_4(5\%) = 9.488$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt the data can be modelled by a Poisson distribution.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 9**

**Question:**

A factory has a machine. The number of times it broke down each week was recorded over 100 weeks with the following results.

| Number of times broken down | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 50 | 24 | 12 | 9 | 5 | 0 |

It is thought that the distribution is Poisson.

**a** Give reasons why this assumption might be made.

**b** Conduct a test at the 0.05 level of significance to see if the assumption is reasonable.

**Solution:**

**a** Breakdowns occur singly, independently and at random. They occur at a constant average rate.

**b** mean $= 0.95$ so let $\lambda = 0.95$

The mean has been calculated so it is 1 restriction.

$H_0$ : The data can be modelled by Po(0.95).

$H_1$: The data cannot be modelled by Po(0.95).

Total frequency $= 100$

This is 1 restriction. Total frequencies agree.

$$\mathrm{E}(X=0) = 100\times \mathrm{P}(X=0) = 100\times\left[e^{-0.95}\times\frac{0.95^0}{0!}\right] = 38.674$$

$$\mathrm{E}(X=1) = 100\times \mathrm{P}(X=1) = 100\times\left[e^{-0.95}\times\frac{0.95^1}{1!}\right] = 36.740$$

Similarly $\mathrm{E}(X=2) = 17.452, \mathrm{E}(X=3) = 5.526, \mathrm{E}(X=4) = 1.3125$.

There is no need to go further, as further terms are extremely small. Here we find E($X$ is 3 or more) to get all the $E$'s to be 5 or more.

| $X$ | 0 | 1 | 2 | 3 or more |
|---|---|---|---|---|
| $O$ | 50 | 24 | 12 | 14 |
| $E$ | 38.674 | 36.740 | 17.452 | 7.134 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(50-38.674)^2}{38.674}+\frac{(24-36.740)^2}{36.740}+\frac{(12-17.452)^2}{17.452}+\frac{(14-7.134)^2}{7.134}$$

$$\text{t.s. } \left(\chi^2\right) = 16.0$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Degrees of freedom $= 4 - 2$

$= 2$

Number of classes after pooling.

Number of restrictions.

critical value $= \chi^2_2(5\%) = 5.991$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude that breakdowns cannot modelled by Po(0.95).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 10**

**Question:**

In a lottery there are 505 prizes, and it is assumed that they will be uniformly distributed throughout the numbered tickets. An investigation gave the following:

| Ticket number | 1–1000 | 1001–2000 | 2001–3000 | 3001–4000 | 4001–5000 | 5001–6000 | 6001–7000 | 7001–8000 | 8001–9000 | 9001–10000 |
|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 56 | 49 | 35 | 47 | 63 | 58 | 44 | 52 | 51 | 50 |

Using a suitable test with a 0.05 significance level, and stating your null and alternative hypotheses, see if the assumption is reasonable.

**Solution:**

$H_0$: The prizes are uniformly distributed.

$H_1$: The prizes are not uniformly distributed.

$$\text{Total frequency} = 505$$

$$\therefore \text{Expected frequency for each class} = \frac{505}{10} = 50.5$$

This is 1 restriction, total frequencies agree.

$$\therefore \text{Test statistic } (\chi^2) = \frac{(56-50.5)^2}{50.5} + \frac{(49-50.5)^2}{50.5} + \cdots + \frac{(50-50.5)^2}{50.5}$$

$$\text{t.s. } (\chi^2) = 10.7$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$$\text{Degrees of freedom} = 10 - 1 = 9$$

$$\text{critical value} = \chi_9^2(5\%) = 16.919$$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt that prizes are distributed uniformly.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Goodness of fit and contingency tables
**Exercise B, Question 11**

**Question:**

Data were collected on the number of female puppies born in 200 litters of size 8. It was decided to test whether or not a binomial model with parameters $n = 8$ and $p = 0.5$ is a suitable model for these data. The following table shows the observed frequencies and the expected frequencies, to 2 decimal places, obtained in order to carry out this test.

| Number of females | Observed number of litters | Expected number of litters |
|---|---|---|
| 0 | 1 | 0.78 |
| 1 | 9 | 6.25 |
| 2 | 27 | 21.88 |
| 3 | 46 | $R$ |
| 4 | 49 | $S$ |
| 5 | 35 | $T$ |
| 6 | 26 | 21.88 |
| 7 | 5 | 6.25 |
| 8 | 2 | 0.78 |

**a** Find the values of $R$, $S$ and $T$.

**b** Carry out the test to determine whether or not this binomial model is a suitable one.
State your hypotheses clearly and use a 5% level of significance.

An alternative test might have involved estimating $p$ rather than assuming $p = 0.5$.

**c** Explain how this would have affected the test.

**Solution:**

**a** B(8, 0.5)

Total frequency = 200

This is 1 restriction, total frequencies agree.

$$R = 200 \times \mathrm{P}(X=3) = 200 \times \left[\binom{8}{3} \times 0.5^3 \times 0.5^5\right]$$

$R = 43.75$

$$S = 200 \times \mathrm{P}(X=4) = 200 \times \left[\binom{8}{4} \times 0.5^4 \times 0.5^4\right]$$

$S = 54.69$

$$T = 200 \times \mathrm{P}(X=5) = 200 \times \left[\binom{8}{5} \times 0.5^5 \times 0.5^3\right]$$

$T = 43.75$

**b** $\mathrm{H}_0$: B(8, 0.5) is a suitable model.

$\mathrm{H}_1$: B(8, 0.5) is not a suitable model.

| $X$ | 1 or fewer | 2 | 3 | 4 | 5 | 6 | 7 or more |
|---|---|---|---|---|---|---|---|
| $O$ | 10 | 27 | 46 | 49 | 35 | 26 | 7 |
| $E$ | 7.03 | 21.88 | 43.75 | 54.69 | 43.75 | 21.88 | 7.03 |

Pool the first 2 classes to get $E$ to be 5 or more.

Pool the last 2 classes to get $E$ to be 5 or more.

Test statistic $(\chi^2) = \dfrac{(10-7.03)^2}{7.03} + \dfrac{(27-21.88)^2}{21.88} + \cdots + \dfrac{(7-7.03)^2}{7.03}$

t.s. $(\chi^2) = 5.69$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Number of classes after pooling.

Degrees of freedom $= 7 - 1$

$= 6$

Only 1 restriction.

critical value $= \chi^2_6\,(5\%) = 10.645$

$\chi^2 <$ c.v. so accept $\mathrm{H}_0$.

Conclude there is no reason to doubt the data could be modelled by B(8, 0.5).

**c** We would have to calculate the mean which would give an extra 1 restriction. This would reduce the degrees of freedom by 1 so $\nu = 5$. The critical value would become $\chi^2_5\,(5\%) = 11.070$.

$\chi^2 <$ c.v. so no change in conclusion.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

### Goodness of fit and contingency tables
**Exercise B, Question 12**

**Question:**

A random sample of 300 football matches was taken and the number of goals scored in each match was recorded. The results are given in the table below.

| Number of goals | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 33 | 55 | 80 | 56 | 56 | 11 | 5 | 4 |

**a** Show that an unbiased estimate of the mean number of goals scored in a football match is 2.4 and find an unbiased estimate of the variance.

It is thought that a Poisson distribution might provide a good model for the number of goals per match.

**b** State briefly an implication of a Poisson model on goal scoring at football matches.

Using a Poisson distribution, with mean 2.4, expected frequencies were calculated as follows:

| Number of goals | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| Expected frequency | $s$ | 65.3 | $t$ | 62.7 | 37.6 | 18.1 | 7.2 | 2.5 |

**c** Find the values of $s$ and $t$.

**d** State clearly the hypotheses required to test whether or not a Poisson distribution provides a suitable model for these data.

In order to carry out this test the class for 7 goals is redefined as 7 or more goals.

**e** Find the expected frequency for this class.

The test statistic for the test in part **d** is 15.6 and the number of degrees of freedom used is 5.

**f** Explain fully why there are 5 degrees of freedom.

**g** Stating clearly the critical value used, carry out the test in part **d**, using a 5% level of significance.

**Solution:**

**a** $\text{mean} = \dfrac{[0\times33+1\times55+2\times80+3\times56+4\times56+5\times11+6\times5+7\times4]}{[33+55+80+56+56+11+5+4]}$

$\text{mean} = 2.4$

This is 1 restriction in part **d**.

Unbiased estimate of variance $= \dfrac{2426-300\times2.4^2}{299} = 2.4\ldots$ (a Poisson situation $\lambda = \sigma^2$)

**b** It assumes goals are scored independently and at random, at a constant average rate.

**c** Total frequency $= 300$

This is 1 restriction in part **d**.

$s = \mathrm{E}(X=0) = 300\times\mathrm{P}(X=0) = 300\times\left[e^{-2.4}\times\dfrac{2.4^0}{0!}\right] = 27.2$

$t = \mathrm{E}(X=2) = 300\times\mathrm{P}(X=2) = 300\times\left[e^{-2.4}\times\dfrac{2.4^2}{2!}\right] = 78.4$

**d** $\mathrm{H}_0$: Po(2.4) is a suitable model.

$\mathrm{H}_1$: Po(2.4) is not a suitable model.

**e** Expected frequency for '7 or more goals'

$= 300-[\mathrm{E}(0)+\mathrm{E}(1)+\mathrm{E}(2)+\mathrm{E}(3)+\mathrm{E}(4)+\mathrm{E}(5)+\mathrm{E}(6)]$

$= 300-[27.2+65.3+78.4+62.7+37.6+18.1+7.2]$

$= 3.5$

**f** This expected frequency is less than 5 so must be combined with $\mathrm{E}(X=6)$ to give the class '6 or more goals' which now has expected frequency $7.2+3.5=10.7$.
We now have 7 classes after pooling and 2 restrictions so degrees of freedom $= 7-2 = 5$.

**g** Test statistic $(\chi^2) = 15.6$

Given in the question.

critical value $= \chi^2_5\,(5\%) = 11.070$

$\chi^2 >$ c.v. so reject $\mathrm{H}_0$.

Conclude there is evidence the data cannot be modelled by Po(2.4).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise B, Question 13**

**Question:**

A student of botany believed that *multifolium uniflorum* plants grow in random positions in grassy meadowland. He recorded the number of plants in one square metre of grassy meadow, and repeated the procedure to obtain the 148 results in the table.

| Number of plants | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 or greater |
|---|---|---|---|---|---|---|---|---|
| Frequency | 9 | 24 | 43 | 34 | 21 | 15 | 2 | 0 |

**a** Show that, to two decimal places, the mean number of plants in one square metre is 2.59.

**b** Give a reason why the Poisson distribution might be an appropriate model for these data.

Using the Poisson model with mean 2.59, expected frequencies corresponding to the given frequencies were calculated, to two decimal places, and are shown in the table below.

| Number of plants | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 or greater |
|---|---|---|---|---|---|---|---|---|
| Expected frequencies | 11.10 | 28.76 | $s$ | 32.15 | 20.82 | 10.78 | 4.65 | $t$ |

**c** Find the values of $s$ and $t$ to two decimal places.

**d** Stating clearly your hypotheses, test at the 5% level of significance whether or not this Poisson model is supported by these data.

**Solution:**

**a** $\text{mean} = \dfrac{[0\times 9 + 1\times 24 + 2\times 43 + 3\times 34 + 4\times 21 + 5\times 15 + 6\times 2]}{[9+24+43+34+21+15+2]}$

mean = 2.5878…

mean = 2.59 (2 d.p.)

This is 1 restriction in part **d**.

**b** It is assumed plants occur at a constant average rate and occur independently and at random in the meadow.

Total frequency = 148

This is 1 restriction in part **c**.

**c** $s = \mathrm{E}(2) = 148\times \mathrm{P}(X=2) = 148\times\left[e^{-2.59}\times\dfrac{2.59^2}{2!}\right]$

$\therefore s = 37.24$ (2 d.p.)

$t = 148 - [11.10 + 28.76 + 37.24 + 32.15 + 20.82 + 10.78 + 4.65]$

$t = 2.50$

**d** $H_0$: Po(2.59) is a suitable model.

$H_1$: Po(2.59) is not a suitable model.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 or more |
|---|---|---|---|---|---|---|---|
| $O$ | 9 | 24 | 43 | 34 | 21 | 15 | 2 |
| $E$ | 11.10 | 28.76 | 37.24 | 32.15 | 20.82 | 10.78 | 7.15 |

Pool the last 2 classes to get $E$ to be 5 or more.

Test statistic $(\chi^2) = \dfrac{(9-11.10)^2}{11.10} + \dfrac{(24-28.76)^2}{28.76} + \cdots + \dfrac{(2-7.15)^2}{7.15}$

t.s. $(\chi^2) = 7.55$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Number of classes after pooling.

Degrees of freedom $= 7 - 2$

$= 5$

Number of restrictions.

critical value $= \chi^2_5(5\%) = 11.070$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt the data can be modelled by Po(2.59).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 1**

**Question:**

The diameters of a random sample of 30 mass-produced components were measured by checking their diameter with gauges. Of the 30, 18 passed through a 4.0 mm gauge and of these 6 failed to pass through a 3.5 mm gauge. Test the hypothesis that the diameters of the components were a sample of a normal population with mean 3.8 mm and standard deviation 0.5 mm. Use a 5% significance level for your test.

**Solution:**

$H_0$: The data can be modelled by $N(3.8, 0.5^2)$.
$H_1$: The data cannot be modelled by $N(3.8, 0.5^2)$.

| $x$ | $x < 3.5$ | $3.5 \le x < 4.0$ | $x \ge 4.0$ |
|---|---|---|---|
| 0 | 12 | 6 | 12 |

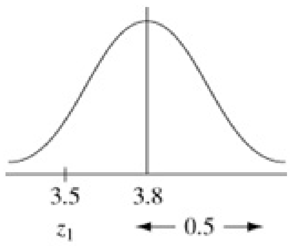
You need to put the given data into a table as shown first.

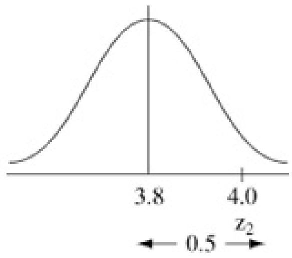
Total frequency = 30

This is 1 restriction.

$$z_1 = \frac{(3.5-3.8)}{0.5} = -0.6$$

$$z_2 = \frac{(4.0-3.8)}{0.5} = 0.4$$




$$\begin{aligned}
\therefore P(X < 3.5) &= \Phi(-0.6) = 0.2743 \\
\text{so } E(X < 3.5) &= 30 \times 0.2743 = 8.229 \\
P(3.5 \le x \le 4.0) &= \Phi(0.4) - \Phi(-0.6) \\
&= 0.6554 - 0.2743 \\
&= 0.3811 \\
\therefore E(3.5 \le x \le 4.0) &= 30 \times 0.3811 = 11.433 \\
\therefore E(X > 4.0) &= 30 - [8.229 + 11.433] = 10.338
\end{aligned}$$

The total frequency must total 30.

| $x$ | $x < 3.5$ | $3.5 \le x < 4.0$ | $x \ge 4.0$ |
|---|---|---|---|
| 0 | 12 | 6 | 12 |
| $E$ | 8.229 | 11.433 | 10.338 |

$$\text{Test statistic } (\chi^2) = \frac{(6-8.229)^2}{8.229} + \frac{(12-11.433)^2}{11.433} + \frac{(12-10.338)^2}{10.338}$$

$$\text{t.s. } (\chi^2) = 0.899$$

Number of classes.

Degrees of freedom $= 3 - 1$
$= 2$

Number of restrictions.

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_2^2(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by $N(3.8, 0.5^2)$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 2**

**Question:**

An egg producer takes a sample of 150 eggs from his flock of chickens and grades them into classes according to their weights as follows.

| Class | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Weight | 66–70 | 61–65 | 56–60 | 51–55 | 46–50 |
| Frequency | 10 | 32 | 67 | 29 | 12 |

Does this distribution fit a normal distribution of mean 58 g and standard deviation 4 g? Use a 5% significance level for your test.

**Solution:**

$H_0$: The data can be modelled by $N(58, 4^2)$.
$H_1$: The data cannot be modelled by $N(58, 4^2)$.

Total frequency $= 150$

This is 1 restriction.

| Class $a$ | Class $b$ | $z = \dfrac{b-\mu}{\sigma}$ (rounded values) | $P(a \le X \le b)$ | Expected frequency ($E$) |
|---|---|---|---|---|
| | $< 45.5$ | $-3.13$ | $1 - 0.9991 = 0.0009$ | 0.135 |
| 45.5 | 50.5 | $-1.88$ | $= 0.0291$ | 4.365 |
| 50.5 | 55.5 | $-0.63$ | $= 0.2343$ | 35.145 |
| 55.5 | 60.5 | 0.63 | $0.7357 - 0.2643 = 0.4714$ | 70.710 |
| 60.5 | 65.5 | 1.88 | $0.9700 - 0.7357 = 0.2343$ | 35.145 |
| 65.5 | 70.5 | 3.13 | $0.9991 - 0.9900 = 0.0291$ | 4.365 |
| $> 70.5$ | | | $1 - 0.9991 = 0.0009$ | 0.135 |

These 3 classes must be pooled to get $E$ to be 5 or more.

These 3 classes must be combined to get $E$ to be 5 or more.

The $Z$-scores are symmetrical for the classes so we can deduce the probabilities for the first 3 classes since they are the same as for the last 3 classes. This is also true for the expected frequencies.

| Class | 55.5 or less | 55.5 to 65.5 | 65.5 or more |
|---|---|---|---|
| $O$ | 42 | 67 | 41 |
| $E$ | 39.645 | 70.710 | 39.645 |

$$\text{Test statistic } (\chi^2) = \frac{(42-39.645)^2}{39.645} + \frac{(67-70.710)^2}{70.710} + \frac{(41-39.645)^2}{39.645}$$

$$\text{t.s. } (\chi^2) = 0.381$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

A $\chi^2$ value in the range 0.28–0.39 would be accepted.

Number of classes after pooling.

Degrees of freedom $= 3 - 1$
$= 2$

Number of restrictions.

Critical value $= \chi^2_2(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by $N(58, 4^2)$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 3**

**Question:**

A sample of 100 apples is taken from a load. The apples have the following distribution of sizes.

| Diameter (cm) | $\le 6$ | 7 | 8 | 9 | $\ge 10$ |
|---|---|---|---|---|---|
| Frequency | 8 | 29 | 38 | 16 | 9 |

It is thought that they come from a normal distribution with mean diameter of 8 cm and a standard deviation of 0.9 cm. Test this assertion using a 0.05 level of significance.

**Solution:**

$H_0$: The data can be modelled by $N(8, 0.9^2)$.

$H_1$: The data cannot be modelled by $N(8, 0.9^2)$.

Total frequency $= 100$

This is 1 restriction.

| Class $a$ | Class $b$ | $z = \frac{b-\mu}{\sigma}$ (rounded values) | $P(a \le X \le b)$ | $E$ |
|---|---|---|---|---|
| | $\le 6.5$ | $-1.67$ | 0.0475 | 4.75 |
| 6.5 | 7.5 | $-0.56$ | 0.2402 | 24.02 |
| 7.5 | 8.5 | 0.56 | $0.7123 - 0.2877 = 0.4246$ | 42.46 |
| 8.5 | 9.5 | 1.67 | $0.9525 - 0.7123 = 0.2402$ | 24.02 |
| $\ge 9.5$ | | | $1 - 0.9525 = 0.0475$ | 4.75 |

These classes must be pooled to get all the $E$'s to be 5 or more.

Same here.

Because the $Z$-scores are symmetrical we can fill in the probabilities for the first two classes with the values from the last 2 classes. We can do the same with the $E$'s.

| Class | 7 or less | 8 | 9 or more |
|---|---|---|---|
| $O$ | 37 | 38 | 25 |
| $E$ | 28.77 | 42.46 | 28.77 |

$$\text{Test statistic } (\chi^2) = \frac{(37-28.77)^2}{28.77} + \frac{(38-42.46)^2}{42.46} + \frac{(25-28.77)^2}{28.77}$$
$$= 3.32$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$\chi^2$ values accepted in a range 3.2–3.33.

Degrees of freedom $= 3 - 1$

Number of classes after pooling.

Restriction.

$= 2$

critical value $= \chi^2_2(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by $N(8, 0.9^2)$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 4**

**Question:**

A shop owner found that the number of cans of a particular drink sold per day during 100 days in summer was as follows.

| **Drinks per day, *d*** | 0– | 10– | 20– | 30– | 40– |
|---|---|---|---|---|---|
| **Frequency of *d*** | 10 | 24 | 45 | 14 | 7 |

It is thought that these data can be modelled by a normal distribution.

**a** Estimate values of $\mu$ and $\sigma$ and conduct a goodness of fit test using 1% significance level.

**b** Explain how the shopkeeper might use this model.

**Solution:**

**a**

| Mid-interval value | 5 | 15 | 25 | 35 | 45 |
|---|---|---|---|---|---|
| Frequency | 10 | 24 | 45 | 14 | 7 |

mean $= \bar{x} = 23.4$
standard deviation $= S = 10.222$
Total frequency $= 100$

Get these directly from the calculator.

This is 1 restriction.

Because the mean *and* standard deviation have been estimated by *calculation* they give a further 2 restrictions. Here we have 3 restrictions in total.

$H_0$: The data can may be modelled by $N(23.4, 10.222^2)$.

$H_1$: The data cannot may be modelled by $N(23.4, 10.222^2)$.

| Class $a$ | Class $b$ | $z = \frac{b-\mu}{\sigma}$ (rounded values) | $P(a \le X \le b)$ | $E$ |
|---|---|---|---|---|
| | $\le 10$ | $-1.31$ | $0.0951$ | 9.51 |
| 10 | 20 | $-0.33$ | $0.3707 - 0.0951 = 0.2756$ | 27.56 |
| 20 | 30 | 0.65 | $0.7422 - 0.3707 = 0.3715$ | 37.15 |
| 30 | 40 | 1.62 | $0.9474 - 0.7422 = 0.2052$ | 20.52 |
| $\ge 40$ | | | $1 - 0.9474 = 0.0526$ | 5.26 |

No pooling needed as *all* the $E$'s are 5 or more.

| Class | $\le 10$ | 10– | 20– | 30– | 40 or more |
|---|---|---|---|---|---|
| $O$ | 10 | 24 | 45 | 14 | 7 |
| $E$ | 9.51 | 27.56 | 37.15 | 20.52 | 5.26 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(10-9.51)^2}{9.51} + \frac{(24-27.56)^2}{27.56} + \cdots + \frac{(7-5.26)^2}{5.26}$$

$$\text{t.s. } \left(\chi^2\right) = 4.79$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$\chi^2$ values between 4.79 and 4.94 accepted.

Number of classes.

Degrees of freedom $= 5 - 3$
$= 2$

Number of restrictions.

critical value $= \chi_2^2\,(1\%) = 9.210$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by $N(23.4, 10.222^2)$.

**b** The shopkeeper could use it to work out the maximum number of cans of the drink they need to keep in stock in order to cope with demand.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 5**

**Question:**

An outfitter sells boys' raincoats and these are stocked in four sizes.
Size 1 fits boys up to 1.25 m in height. Size 2 fits boy from 1.26 to 1.31 m. Size 3 fits boys from 1.32 to 1.37 m. Size 4 fits those over 1.38 m. To assist the outfitter in deciding the stock levels he should order each year, the heights of 120 boys in the right age range were measured with the following results.

| Height, $h$(m) | 1.20–1.22 | 1.23–1.25 | 1.26–1.28 | 1.29–1.31 | 1.32–1.34 |
|---|---|---|---|---|---|
| Frequency of $h$ | 9 | 9 | 18 | 23 | 20 |
| Height, $h$(m) | 1.35–1.37 | 1.38–1.40 | 1.41–1.43 | | |
| Frequency of $h$ | 19 | 17 | 5 | | |

It is suggested that a suitable model for these data is N(1.32, 0.0016).

**a** Conduct a goodness of fit test using a $2\frac{1}{2}\%$ significance level.

**b** Estimate values of $\mu$ and $\sigma$ using the observed values and using these conduct a goodness of fit test using a $2\frac{1}{2}\%$ significance level.

**c** Select the best model and use it to tell the outfitter how many of each size should be ordered per year if the normal annual sales are 1200.

**Solution:**

**a** $H_0$: The data can be modelled by N(1.32, 0.0016).

$H_1$: The data cannot be modelled by N(1.32, 0.0016).

Here the variance $= 0.0016$ so the standard deviation $= \sqrt{0.0016} = 0.04$

i.e. $\sigma = 0.04$

| Class $A$ | $b$ | $z = \dfrac{b-\mu}{\sigma}$ (rounded values) | $P(a \le X \le b)$ | $E$ |
|---|---|---|---|---|
| | $\le 1.225$ | $-2.38$ | $0.0009$ | $0.108$ |
| 1.225 | 1.255 | $-1.63$ | $0.0516 - 0.0009 = 0.0507$ | $6.084$ |
| 1.255 | 1.285 | $-0.88$ | $0.1894 - 0.0516 = 0.1378$ | $16.536$ |
| 1.285 | 1.315 | $-0.13$ | $0.4483 - 0.1894 = 0.2589$ | $31.068$ |
| 1.315 | 1.345 | $0.63$ | $0.7357 - 0.4483 = 0.2874$ | $31.488$ |
| 1.345 | 1.375 | $1.38$ | $0.9162 - 0.7357 = 0.1805$ | $21.660$ |
| 1.375 | 1.405 | $2.13$ | $0.9834 - 0.9162 = 0.0672$ | $8.064$ |
| $\ge 1.405$ | | | $1 - 0.9834 = 0.0166$ | $1.992$ |

Pool to get $E$ to be 5 or more.

Same here.

Total frequency $= 120$ ← This is 1 restriction.

| Class | 1.25 or less | 1.26–1.28 | 1.29–1.31 | 1.32–1.34 | 1.35–1.37 | 1.38 or more |
|---|---|---|---|---|---|---|
| $O$ | 18 | 18 | 23 | 20 | 19 | 22 |
| $E$ | 6.192 | 16.536 | 31.068 | 31.488 | 21.660 | 10.056 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(18-6.192)^2}{6.192} + \frac{(18-16.536)^2}{16.536} + \cdots + \frac{(22-10.056)^2}{10.056}$$

$$\text{t.s. } \left(\chi^2\right) = 43.4$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$\chi^2$ values in the range 43.35 to 44.4 accepted.

Number of classes after pooling.

Degrees of freedom $= 6 - 1$ ← Number of restrictions.

$= 5$

critical value $= \chi_2^5\left(2\frac{1}{2}\%\right) = 12.833$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude the data cannot be modelled by N(1.32, 0.0016).

**b**

| Mid-interval value | 1.21 | 1.24 | 1.27 | 1.30 | 1.33 | 1.36 | 1.39 | 1.42 |
|---|---|---|---|---|---|---|---|---|
| Frequency | 9 | 9 | 18 | 23 | 20 | 19 | 17 | 5 |

Mean $= 1.3165 = \mu$

standard derivation $= s = 0.0569 \simeq \sigma$

Total frequency $= 120$

This gives 3 restrictions in total.

$H_0$: The data can be modelled by $N(1.3165, 0.0569^2)$.

$H_1$: The data cannot be modelled by $N(1.3165, 0.0569^2)$.

| Class $a$ | Class $b$ | $z = \frac{b-\mu}{\sigma}$ (rounded values) | $P(a \le X \le b)$ | $E$ |
|---|---|---|---|---|
| | $\le 1.225$ | $-1.61$ | 0.0537 | 6.444 |
| 1.225 | 1.255 | $-1.08$ | $0.1401 - 0.0537 = 0.0864$ | 10.368 |
| 1.255 | 1.285 | $-0.55$ | $0.2912 - 0.1401 = 0.1511$ | 18.132 |
| 1.285 | 1.315 | $-0.03$ | $0.4880 - 0.2912 = 0.1968$ | 23.616 |
| 1.315 | 1.345 | 0.50 | $0.6915 - 0.4880 = 0.2035$ | 24.420 |
| 1.345 | 1.375 | 1.03 | $0.8485 - 0.6915 = 0.1570$ | 18.840 |
| 1.375 | 1.405 | 1.56 | $0.9406 - 0.8485 = 0.0921$ | 11.052 |
| $\ge 1.405$ | | | $1 - 0.9406 = 0.0594$ | 7.128 |

| Class | 1.22 or less | 1.23–1.25 | 1.26–1.28 | 1.29–1.31 | 1.32–1.34 | 1.35–1.37 | 1.38–1.40 | 1.41 or more |
|---|---|---|---|---|---|---|---|---|
| $O$ | 9 | 9 | 18 | 23 | 20 | 19 | 17 | 5 |
| $E$ | 6.444 | 10.368 | 18.132 | 23.616 | 24.420 | 18.840 | 11.052 | 7.128 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(9-6.444)^2}{6.444} + \frac{(9-10.368)^2}{10.368} + \cdots + \frac{(5-7.128)^2}{7.128}$$

$$\text{t.s. } \left(\chi^2\right) = 5.85$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$\chi^2$ can be in range 5.78 to 5.85.

Degrees of freedom $= 8 - 3$
$= 5$

critical value $= \chi_5^2 \left(2\frac{1}{2}\%\right) = 12.832$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by $N(1.3165, 0.0569^2)$.

**c** The outfitter should use the model in part **b** because it is based on the experimental evidence.

If normal annual sales are 1200 (this is 10 times what is in the model so multiply the $E$'s by 10) s/he should order

*Size 1* $(6.444 + 10.368) \times 10$ i.e. 168
*Size 2* $(18.132 + 23.616) \times 10$ i.e. 417
*Size 3* $(24.420 + 18.840) \times 10$ i.e. 433
*Size 4* $(11.052 + 7.128) \times 10$ i.e. 182

Those will vary according to the answers in part **b**.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 6**

**Question:**

A hamster breeder is studying the weight of adult hamsters. Each hamster from a random sample of 50 hamsters is weighed and the results, given to the nearest g, are recorded in the following table.

| **Weight (g)** | 85–94 | 95–99 | 100–104 | 105–109 | 110–119 |
|---|---|---|---|---|---|
| **Frequency** | 6 | 9 | 17 | 14 | 4 |

**a** Show that an estimate of the mean weight of the hamsters is 102 g.
The breeder proposes that the weight of an adult hamster, in g, should be modelled by the random variable $W$ having a normal distribution with standard deviation 6. The breeder fits a normal distribution and obtains the following expected frequencies.

| $W$ | $W \le 94.5$ | $94.5 < W \le 99.5$ | $99.5 < W \le 104.5$ | $104.5 < W \le 109.5$ | $W > 109.5$ |
|---|---|---|---|---|---|
| Expected frequency | $r$ | 11.64 | $s$ | 11.64 | $t$ |

**b** Find the values of $r$, $s$ and $t$.

**c** Stating your hypotheses clearly, test at the 10% level of significance whether or not a normal distribution with a standard deviation of 6 is a suitable model for $W$.

**Solution:**

**a**

| Mid-interval value | 89.5 | 97 | 102 | 107 | 114.5 |
|---|---|---|---|---|---|
| Frequency | 6 | 9 | 17 | 14 | 4 |

$$\text{mean} = \frac{[6\times89.5+9\times97+17\times102+14\times107+4\times114.5]}{[6+9+17+14+4]}$$

$\text{mean} = 102$ ← This is 1 restriction in part c.

**b** Total frequency $= 50$ ← This is 1 restriction in part c.

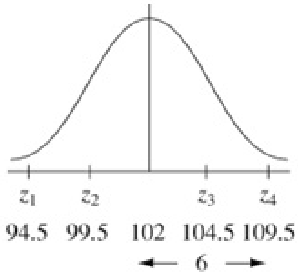
$\sigma = 6$ given ← This is *not* a restriction since an estimate of it isn't calculated.

$$z_1 = \frac{(94.5-102)}{6} = -1.25$$

$$z_2 = \frac{(99.5-102)}{6} = -0.42$$

$$z_3 = \frac{(104.5-102)}{6} = 0.42$$

$$z_4 = \frac{(109.5-102)}{6} = 1.25$$




$t = \mathrm{E}(W > 109.5) = 50\times[1-\Phi(1.25)] = 50\times[1-0.8944] = 5.28$

so $t = 5.28$

and from symmetry (because the $Z$-scores are symmetrical)

$r = 5.28$

and $s = 50-[5.28+11.64+11.64+5.28]$

so $s = 16.16$

**c** $\mathrm{H}_0$: The data can be modelled by $\mathrm{N}(102, 6^2)$.

$\mathrm{H}_1$: The data cannot be modelled by $\mathrm{N}(102, 6^2)$.

| Class | $W \le 94.5$ | $94.5 < W \le 99.5$ | $99.5 < W < 104.5$ | $104.5 < W \le 109.5$ | $W > 109.5$ |
|---|---|---|---|---|---|
| $O$ | 6 | 9 | 17 | 14 | 4 |
| $E$ | 5.28 | 11.64 | 16.16 | 11.64 | 5.28 |

$$\text{Test statistic } (\chi^2) = \frac{(6-5.28)^2}{5.28} + \frac{(9-11.64)^2}{11.64} + \frac{(17-16.16)^2}{16.16} + \frac{(14-11.64)^2}{11.64} + \frac{(4-5.28)^2}{5.28}$$

$\text{t.s. } (\chi^2) = 1.53$

Number of classes.

Degrees of freedom $= 5-2 = 3$

Number of restrictions.

Critical value $= \chi_3^2\,(10\%) = 6.251$

$\chi^2 <$ c.v. so accept $\mathrm{H}_0$.

Conclude no reason to doubt the data could be modelled by $\mathrm{N}(102, 6^2)$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise C, Question 7**

**Question:**

A hamster breeder is studying the weight of adult hamsters. Each hamster from a random sample of 50 hamsters is weighed and the results, given to the nearest g, are recorded in the following table.

| ***Weight (g)*** | 85–94 | 95–99 | 100–104 | 105–109 | 110–119 |
|---|---|---|---|---|---|
| ***Frequency*** | 6 | 9 | 17 | 14 | 4 |

**a** Show that an estimate of the mean weight of the hamsters is 102 g.
The breeder proposes that the weight of an adult hamster, in g, should be modelled by the random variable $W$ having a normal distribution with standard deviation 6. The breeder fits a normal distribution and obtains the following expected frequencies.

| $W$ | $W \le 94.5$ | $94.5 < W \le 99.5$ | $99.5 < W \le 104.5$ | $104.5 < W \le 109.5$ | $W > 109.5$ |
|---|---|---|---|---|---|
| Expected frequency | $r$ | 11.64 | $s$ | 11.64 | $t$ |

**b** Find the values of $r$, $s$ and $t$.

**c** Stating your hypotheses clearly, test at the 10% level of significance whether or not a normal distribution with a standard deviation of 6 is a suitable model for $W$.

**Solution:**

**a**

| Mid-interval value | 89.5 | 97 | 102 | 107 | 114.5 |
|---|---|---|---|---|---|
| Frequency | 6 | 9 | 17 | 14 | 4 |

$$\text{mean} = \frac{[6\times89.5+9\times97+17\times102+14\times107+4\times114.5]}{[6+9+17+14+4]}$$

$\text{mean} = 102$ ← This is 1 restriction in part c.

**b** Total frequency $= 50$ ← This is 1 restriction in part c.

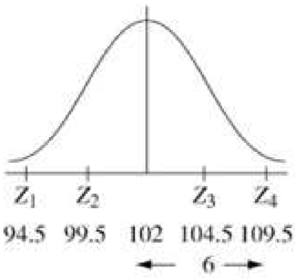
$\sigma = 6$ given ← This is *not* a restriction since an estimate of it isn't calculated.

$$Z_1 = \frac{(94.5-102)}{6} = -1.25$$

$$Z_2 = \frac{(99.5-102)}{6} = -0.42$$

$$Z_3 = \frac{(104.5-102)}{6} = 0.42$$

$$Z_4 = \frac{(109.5-102)}{6} = 1.25$$




$t = \mathrm{E}(W > 109.5) = 50\times[1-\Phi(1.25)] = 50\times[1-0.8944] = 5.28$

so $t = 5.28$

and from symmetry (because the $Z$-scores are symmetrical)

$r = 5.28$

and $s = 50 - [5.28 + 11.64 + 11.64 + 5.28]$

so $s = 16.16$

**c** $\mathrm{H}_0$: The data are from $\mathrm{N}(102, 6^2)$.

$\mathrm{H}_1$: The data aren't from $\mathrm{N}(102, 6^2)$.

| Class | $W \le 94.5$ | $94.5 < W \le 99.5$ | $99.5 < W < 104.5$ | $104.5 < W \le 109.5$ | $W > 109.5$ |
|---|---|---|---|---|---|
| $O$ | 6 | 9 | 17 | 14 | 4 |
| $E$ | 5.28 | 11.64 | 16.16 | 11.64 | 5.28 |

$$\text{Test statistic} = \frac{(6-5.28)^2}{5.28} + \frac{(9-11.64)^2}{11.64} + \frac{(17-16.16)^2}{16.16} + \frac{(14-11.64)^2}{11.64} + \frac{(4-5.28)^2}{5.28}$$

$$\text{t.s.} = 1.53$$

Number of classes.

Degrees of freedom $= 5 - 2 = 3$ ← Number of restrictions.

Critical value $= \chi_3^2(10\%) = 6.251$

t.s. < c.v. so accept $\mathrm{H}_0$.

Conclude no reason to doubt the data could be from $\mathrm{N}(102, 6^2)$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 1**

**Question:**

When analysing the results of a $3 \times 2$ contingency table it was found that

$$\sum_{i=1}^{6} \frac{\left(O_i - E_i\right)^2}{E_i} = 2.38$$

Write down the number of degrees of freedom and the critical value appropriate to these data in order to carry out a $\chi^2$ test of significance at the 5% level. ***E***

**Solution:**

$\nu = 2, \quad \chi_2^2\,(5\%) = 5.99$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 2**

**Question:**

Three different types of locality were studied to see if the ownership, or non-ownership, of a television was or was not related to the locality. $\sum \frac{(O_i - E_i)^2}{E_i}$ was evaluated and found to be 13.1. Using a 5% level of significance, carry out a suitable test and state your conclusion.

**Solution:**

$H_0$ : Ownership is not related to locality.
$H_1$: Ownership is related to locality.
Test statistic $\chi^2 \quad = 13.1$
Degrees of freedom $= (3-1) \times (2-1) = 2$
critical value $= \chi_2^2\,(5\%) = 5.991$
$\chi^2 >$ c.v. so reject $H_0$.
Conclude there is evidence that ownership of a television is related to the locality.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 3**

**Question:**

In a college, three different groups of students sit the same examination. The results of the examination are classified as Credit, Pass or Fail. In order to test whether or not there is a difference between the groups with respect to the proportions of students in the three grades, the statistic $\sum \frac{(O_i - E_i)^2}{E_i}$ is evaluated and found to be equal to 10.28.

**a** Explain why there are 4 degrees of freedom in this situation.
**b** Using a 5% level of significance, carry out the test and state your conclusions.

**Solution:**

**a** Degrees of freedom $= (3-1) \times (3-1) = 4$

**b** $H_0$: There is no association between the group and the results.

$H_1$: There is an association between the group and the results.

Test statistic $= 10.28$

critical value $= \chi_4^2\,(5\%) = 9.488$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between the groups and the exam results they achieve.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 4**

**Question:**

The grades of 200 students in both Mathematics and English were studied with the following results.

| | | English grades | | |
|---|---|---|---|---|
| | | *A* | *B* | *C* |
| Maths grades | *A* | 17 | 28 | 18 |
| | *B* | 38 | 45 | 16 |
| | *C* | 12 | 12 | 14 |

Using a 0.05 significance level, test these results to see if there is an association between English and Mathematics results. State your conclusions.

**Solution:**

$H_0$: There is no association between Mathematics and English results.

$H_1$: There is an association between Mathematics and English results.

*Observed frequencies* →

| | | *English grades* | | | |
|---|---|---|---|---|---|
| | | *A* | *B* | *C* | Total |
| *Maths grades* | *A* | 17 | 28 | 18 | 63 |
| | *B* | 38 | 45 | 16 | 99 |
| | *C* | 12 | 12 | 14 | 38 |
| | Total | 67 | 85 | 48 | 200 |

Expected frequency Maths $A$ and English $A$

$$= \frac{63 \times 67}{200} = 21.105$$

*Expected frequencies* →

| | | *English Grades* | | |
|---|---|---|---|---|
| | | *A* | *B* | *C* |
| *Maths grades* | *A* | 21.105 | 26.775 | 15.120 |
| | *B* | 33.165 | 42.075 | 23.760 |
| | *C* | 12.730 | 16.150 | 9.120 |

$$\text{Test statistic } (\chi^2) = \frac{(17-21.105)^2}{21.105} + \frac{(28-26.775)^2}{26.775} + \cdots + \frac{(14-9.120)^2}{9.120}$$

$$\text{t.s. } (\chi^2) = 8.57 \text{ (or 8.56)}$$

Degrees of freedom $= (3-1)\times(3-1) = 4$

> Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi^2_4(5\%) = 9.488$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no evidence of an association between Mathematics and English results.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 5**

**Question:**

The number of trains on time, and the number of trains that were late was observed at three different London stations. The results were:

| | | Observed frequency | |
|---|---|---|---|
| | | On time | Late |
| Station | *A* | 26 | 14 |
| | *B* | 30 | 10 |
| | *C* | 44 | 26 |

Using the $\chi^2$ statistic and a significance test at the 5% level, decide if there is any association between station and lateness.

**Solution:**

$H_0$ : There is no association between station and lateness.
$H_1$ : There is an association between station and lateness.

Observed frequencies

| | | *On Time* | *Late* | *Total* |
|---|---|---|---|---|
| *Station* | *A* | 26 | 14 | 40 |
| | *B* | 30 | 10 | 40 |
| | *C* | 44 | 26 | 70 |
| Total | | 100 | 50 | 150 |

Expected frequency 'On time' and 'station *A*'

$$= \frac{40 \times 100}{150} = 26.666\ldots \text{ or } 26\frac{2}{3}$$

Expected frequencies

| | | *On Time* | *Late* |
|---|---|---|---|
| *Station* | *A* | 26.666 | 13.333 |
| | *B* | 26.666 | 13.333 |
| | *C* | 46.666 | 23.333 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(26-26.666)^2}{26.666} + \frac{(14-13.333)^2}{13.333} + \cdots + \frac{(26-23.333)^2}{23.333}$$

$$\text{t.s.}\left(\chi^2\right) = 1.76$$

$$\text{Degrees of freedom} = (3-1)\times(2-1)$$
$$= 2$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value = $\chi_2^2\,(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no evidence of an association between station and lateness.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 6**

**Question:**

In addition to being classed into grades A, B, C, D and E, 200 students are classified as male or female and their results summarised in a contingency table. Assuming all expected values are 5 or more, the statistic $\sum \frac{(O_i - E_i)^2}{E_i}$ was 14.27.

Stating your hypotheses and using a 1% significance level, investigate whether or not gender and grade are associated.

**Solution:**

$H_0$ : There is no association between gender and grade.

$H_1$: There is an association between gender and grade.

Test statistic $\chi^2 = 14.27$

Degrees of freedom $= (5-1)\times(2-1) = 4$

critical value $= \chi_4^2\,(1\%) = 13.277$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between the gender of the student and the grade achieved.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 7**

**Question:**

In a random sample of 60 articles made in factory A, 8 were defective. In factory B, 6 out of 40 similar articles were defective.

**a** Draw up a contingency table.

**b** Test at the 0.05 significance level the hypothesis that quality was independent of the factory involved.

**Solution:**

**a**

Observed frequencies →

| | *Factory* | | |
|---|---|---|---|
| | *A* | *B* | Total |
| *OK* | 52 | 34 | 86 |
| *Defective* | 8 | 6 | 14 |
| Total | 60 | 40 | 100 |

**b** $H_0$ : quality is independent of factory.

$H_1$: quality isn't independent of factory.

Expected frequency 'OK' and '*A*' $= \frac{86 \times 60}{100} = 51.6$

Expected frequencies →

| | *A* | *B* |
|---|---|---|
| *OK* | 51.6 | 34.4 |
| *Defective* | 8.4 | 5.6 |

$$\text{Test statistic } (\chi^2) = \frac{(52-51.6)^2}{51.6} + \frac{(34-34.4)^2}{34.4} + \cdots + \frac{(6-5.6)^2}{5.6}$$
$$= 0.0554$$

Degrees of freedom $= (2-1)\times(2-1) = 1$

> Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_1^2\,(5\%) = 3.841$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt that quality is independent of factory.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 8**

**Question:**

During an influenza epidemic, 15 boys and 8 girls became ill out of a class of 22 boys and 28 girls. Assuming that this group may be treated as a random sample of the age group, test at the 5% significance level the hypothesis that there is no connection between gender and susceptibility to influenza.

**Solution:**

*Observed frequencies* →

| | *Boys* | *Girls* | Total |
|---|---|---|---|
| *Flu* | 15 | 8 | 23 |
| *No Flu* | 7 | 20 | 27 |
| Total | 22 | 28 | 50 |

$H_0$ : There is no association between gender and susceptibility to influenza.

$H_1$: There is an association between gender and susceptibility to influenza.

Expected frequency 'Boy' and 'Flu' $= \dfrac{23 \times 22}{50} = 10.12$

*Expected frequencies* →

| | *Boys* | *Girls* |
|---|---|---|
| *Flu* | 10.12 | 12.88 |
| *No Flu* | 11.88 | 15.12 |

Test statistic $\left(\chi^2\right) = \dfrac{(15-10.12)^2}{10.12} + \dfrac{(8-12.88)^2}{12.88} + \cdots + \dfrac{(20-15.12)^2}{15.12}$

t.s. $\left(\chi^2\right) = 7.78$

Degrees of freedom $= (2-1)\times(2-1) = 1$

> Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_1^2\,(5\%) = 5.991$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between gender and susceptibility to influenza.

More boys were observed to catch flu than would be expected and fewer girls were observed to catch flu than would be expected so it appears boys are more susceptible to flu.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Goodness of fit and contingency tables

**Exercise D, Question 9**

**Question:**

In a study of marine organisms, a biologist collected specimens from three beaches and counted the number of males and females in each sample, with the following results.

| | | Beach | | |
|---|---|---|---|---|
| | | *A* | *B* | *C* |
| *Gender* | *Male* | 46 | 80 | 40 |
| | *Female* | 54 | 120 | 160 |

Using a significance level of 5%, test these results to see if there is any difference between the beaches with regard to the numbers of male and female organisms.

**Solution:**

$H_0$: There is no association between gender and beach.
$H_1$: There is an association between gender and beach.

Observed frequencies →

| | | Beach | | | |
|---|---|---|---|---|---|
| | | *A* | *B* | *C* | Total |
| *Gender* | *Male* | 46 | 80 | 40 | 166 |
| | *Female* | 54 | 120 | 160 | 334 |
| | Total | 100 | 200 | 200 | 500 |

Expected frequency 'male' and '*A*' $= \frac{166 \times 100}{500} = 33.2$

Expected frequencies →

| | *A* | *B* | *C* |
|---|---|---|---|
| *Male* | 33.2 | 66.4 | 66.4 |
| *Female* | 66.8 | 133.6 | 133.6 |

Test statistic $(\chi^2) = \frac{(46-33.2)^2}{33.2} + \frac{(80-66.4)^2}{66.4} + \cdots + \frac{(160-133.6)^2}{133.6}$

t.s. $(\chi^2) = 27.3$

Degrees of freedom $= (2-1)\times(3-1) = 2$

> Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi^2_2(5\%) = 5.991$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between the gender of the organism and the beach on which it is found.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise D, Question 10**

**Question:**

A research worker studying the ages of adults and the number of credit cards they possess obtained the results shown below;

| | | *Number of cards* | |
|---|---|---|---|
| | | $\leq 3$ | $> 3$ |
| *Age* | $< 30$ | 74 | 20 |
| | $\geq 30$ | 50 | 35 |

Use the $\chi^2$ statistic and a significance test at the 5% level to decide whether or not there is an association between age and number of credit cards possessed. ***E***

**Solution:**

$H_0$ : There is no association between age and number of credit cards.

$H_1$: There is an association between age and number of credit cards.

Observed Frequencies →

| | | *Number of cards* | | |
|---|---|---|---|---|
| | | $\leq 3$ | $> 3$ | Total |
| *Age* | $< 30$ | 74 | 20 | 94 |
| | $\geq 30$ | 50 | 35 | 85 |
| | Total | 124 | 55 | 179 |

Expected frequency '$< 30$' and '$\leq 3$' $= \dfrac{94 \times 124}{179} = 65.117$

Expected Frequencies →

| | $\leq 3$ | $> 3$ |
|---|---|---|
| $< 30$ | 65.117 | 28.883 |
| $\geq 30$ | 58.883 | 26.117 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(74-65.117)^2}{65.117} + \frac{(20-28.883)^2}{28.883} + \ldots + \frac{(35-26.117)^2}{26.117}$$
$$= 8.31 (\text{or } 8.30 \text{ direct from calculator})$$

Degrees of freedom $= (2-1)\times(2-1) = 1$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_1^2 (5\%) = 3.841$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between age and number of credit cards possessed.

More people over 30 were observed to carry 3 or more cards than would be expected and more people under 30 were observed to carry fewer than 3 credit cards than would be expected.

Older people are more likely to possess more credit cards than are younger people.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 1**

**Question:**

The random variable $Y$ has a $\chi^2$ distribution with 10 degrees of freedom. Find $y$ such that $P(Y < y) = 0.99$. ***E***

**Solution:**

23.209

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 2**

**Question:**

The random variable $X$ has a chi-squared distribution with 8 degrees of freedom. Find $x$ such that $\mathrm{P}(X > x) = 0.05$. ***E***

**Solution:**

15.507

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 3**

**Question:**

As part of an investigation into visits to a Health Centre a $5\times3$ contingency table was constructed. A $\chi^2$ test of significance at the 5% level is to be carried out on the table. Write down the number of degrees of freedom and the critical region appropriate to this test. ***E***

**Solution:**

$$\text{Degrees of freedom} = (5-1)\times(3-1)$$
$$= 8$$

Critical region is $\chi^2 > 15.507$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 4**

**Question:**

Data are collected in the form of a $4 \times 4$ contingency table.
To carry out a $\chi^2$ test of significance one of the rows was amalgamated with another row and the resulting value of $\sum \frac{(O-E)^2}{E}$ calculated.
Write down the number of degrees of freedom and the critical value of $\chi^2$ appropriate to this test assuming a 5% significance level. ***E***

**Solution:**

Amalgamation gives a $3 \times 4$ contingency table

$$\begin{aligned} \text{Degrees of freedom } &= (3-1)\times(4-1) \\ &= 6 \\ \text{critical value } &= \chi^2_6 \text{ } (5\%) \\ &= 12.592 \end{aligned}$$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 5**

**Question:**

A new drug to treat the common cold was used with a randomly selected group of 100 volunteers. Each was given the drug and their health was monitored to see if they caught a cold. A randomly selected control group of 100 volunteers was treated with a dummy pill. The results are shown in the table below.

| | Cold | No cold |
|---|---|---|
| Drug | 34 | 66 |
| Dummy Pill | 45 | 55 |

Using a 5% significant level, test whether or not the chance of catching a cold is affected by taking the new drug. State your hypotheses carefully. ***E***

**Solution:**

$H_0$ : There is no association between catching a cold and taking the new drug.
$H_1$: There is an association between catching a cold and taking the new drug.

*Observed frequencies* →

| | *Cold* | *No cold* | Total |
|---|---|---|---|
| *Drug* | 34 | 66 | 100 |
| *Dummy* | 45 | 55 | 100 |
| Total | 79 | 121 | 200 |

Expected frequency 'Drug' and 'cold' $= \dfrac{100 \times 79}{200} = 39.5$

*Expected Frequencies* →

| | *Cold* | *No cold* |
|---|---|---|
| *Drug* | 39.5 | 60.5 |
| *Dummy* | 39.5 | 60.5 |

$$\text{Test statistic } \left(\chi^2\right) = \frac{(34-39.5)^2}{39.5} + \frac{(66-60.5)^2}{60.5} + \frac{(45-39.5)^2}{39.5} + \frac{(55-60.5)^2}{60.5}$$

$$\text{t.s. } \left(\chi^2\right) = 2.53$$

Degrees of freedom $= (2-1) \times (2-1) = 1$

> Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

critical value $= \chi_1^2\,(5\%) = 3.841$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no evidence that there is an association between catching a cold and taking the new drug. It appears taking the new drug doesn't affect the chance of a person catching a cold.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**

**Exercise E, Question 6**

**Question:**

Breakdowns on a certain stretch of motorway were recorded each day for 80 consecutive days. The results are summarised in the table below.

| Number of breakdowns | 0 | 1 | 2 | > 2 |
|---|---|---|---|---|
| Frequency | 38 | 32 | 10 | 0 |

It is suggested that the number of breakdowns per day can be modelled by a Poisson distribution.

Using a 5% significant level, test whether or not the Poisson distribution is a suitable model for these data. State your hypotheses clearly. ***E***

**Solution:**

$H_0$ : The data can be modelled by a Poisson distribution.
$H_1$: The data cannot be modelled by a Poisson distribution.

Total frequency = 80 ← This is 1 restriction (total frequencies agree.)

mean = $\bar{x}$ = 0.65 ← This is 1 restriction (calculation of estimate of parameter.)

Get this directly from your calculator.

$$E(0) = 80 \times P(0) = 80 \times \left[ e^{-0.65} \times \frac{0.65^0}{0!} \right] = 41.764$$

$$E(1) = 80 \times P(1) = 80 \times \left[ e^{-0.65} \times \frac{0.65^1}{1!} \right] = 27.146$$

$$E(2) = 80 \times P(2) = 80 \times \left[ e^{-0.65} \times \frac{0.65^2}{2!} \right] = 8.823$$

Pool these classes to get all the $E$'s to be 5 or more.

$$E(3 \text{ or more}) = 80 - [41.764 + 27.146 + 8.823] = 2.267$$

| Class | 0 | 1 | 2 or more |
|---|---|---|---|
| $O$ | 38 | 32 | 10 |
| $E$ | 41.764 | 27.146 | 11.09 |

$$\text{Test statistic } (\chi^2) = \frac{(38-41.764)^2}{41.764} + \frac{(32-27.146)^2}{27.146} + \frac{(10-11.09)^2}{11.09}$$

$$\text{t.s. } (\chi^2) = 1.31$$

Number of classes after pooling.

Degrees of freedom = 3 − 2 = 1

Number of restrictions.

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Critical value = $\chi_1^2$ (5%) = 3.841

$\chi^2$ < c.v. so accept $H_0$.

Conclude there is no reason to doubt that the data can be modelled by a Poisson distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 7**

**Question:**

A survey in a college was commissioned to investigate whether ot not there was any association between gender and passing a driving test. A group of 50 males and 50 females were asked whether they passed or failed their driving test at the first attempt. All the students asked had taken the test. The results were as follows.

| | **Pass** | **Fail** |
|---|---|---|
| **Male** | 23 | 27 |
| **Female** | 32 | 18 |

Stating your hypotheses clearly test, at the 10% level, whether or not there is any evidence of an association between gender and passing a driving test at the first attempt. ***E***

**Solution:**

$H_0$ : There is no association between gender and passing a driving test 'first time'.

$H_1$: There is an association between gender and passing a driving test 'first time'.

*Observed frequencies* →

| | *Pass* | *Fail* | Total |
|---|---|---|---|
| *Male* | 23 | 27 | 50 |
| *Female* | 32 | 18 | 50 |
| Total | 55 | 45 | 100 |

Expected frequency 'Male' and 'pass' $= \dfrac{50 \times 55}{100} = 27.5$

*Expected frequencies* →

| | *Pass* | *Fail* |
|---|---|---|
| *Male* | 27.5 | 22.5 |
| *Female* | 27.5 | 22.5 |

Test statistic $\left(\chi^2\right) = \dfrac{(23-27.5)^2}{27.5} + \dfrac{(27-22.5)^2}{22.5} + \dfrac{(32-27.5)^2}{27.5} + \dfrac{(18-22.5)^2}{22.5}$

t.s. $\left(\chi^2\right) = 3.27$

Degrees of freedom $= (2-1)\times(2-1) = 1$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Critical value $= \chi_1^2\,(10\%) = 2.706$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between gender and passing a driving test 'first time'.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 8**

**Question:**

Successful contestants in a TV game show were allowed to select from one of five boxes, four of which contained prizes, and one of which contained nothing. The boxes were numbered 1 to 5, and, when the show had run for 100 weeks, the choices made by the contestants were analysed with the following results:

| Box number | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| **Frequency** | 20 | 16 | 25 | 18 | 21 |

**a** Explain why these data could possibly be modelled by a discrete uniform distribution.

**b** Using a significance level of 5%, test to see if the discrete uniform distribution is a good model in this particular case.

**Solution:**

**a** We would expect each box to have an equal chance of being opened. The box numbers are discrete values.

**b** $H_0$: The data can be modelled by a discrete uniform distribution.

$H_1$: The data cannot be modelled by a discrete uniform distribution.

Total frequency $= 100$

This is 1 restriction (total frequencies agree.)

There are 5 boxes so expected frequency for each $= \frac{100}{5} = 20$

| Box | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $O$ | 20 | 16 | 25 | 18 | 21 |
| $E$ | 20 | 20 | 20 | 20 | 20 |

Test statistic $(\chi^2) = \frac{(20-20)^2}{20} + \frac{(16-20)^2}{20} + \frac{(25-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(21-20)^2}{20}$

t.s. $(\chi^2) = 2.3$

Degrees of freedom $= 5 - 1 = 4$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Critical value $= \chi_4^2(5\%) = 9.488$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude there is no reason to doubt that the data could be modelled by a discrete uniform distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 9**

**Question:**

A pesticide was tested by applying it in the form of a spray to 50 samples of 5 flies. The numbers of dead flies after 1 hour were then counted with the following results:

| Number of dead flies | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| Frequency | 1 | 1 | 5 | 11 | 24 | 8 |

**a** Calculate the probability that a fly dies when sprayed.
**b** Using a significance level of 5%, test to see if these data could be modelled by a binomial distribution.

**Solution:**

**a** Total number of dead flies $= 0\times1+1\times1+2\times5+3\times11+4\times24+5\times8$
$= 180$

Total number of flies sprayed $= 50\times5$
$= 250$

$\therefore$ $P$ (fly dies when sprayed) $= \frac{180}{250} = 0.72$

This is 1 restriction in part **b** (calculation of estimate of parameter).

**b** $H_0$: The data can be modelled by B(5, 0.72).
$H_1$: The data cannot be modelled by B(5, 0.72).

This is 1 restriction (total frequencies agree).

Total frequency = 50

$$\left.\begin{aligned} E(0) &= 50\times P(0) = 50\times\left[\binom{5}{0}\times0.72^0\times0.28^5\right] = 0.0861\\ E(1) &= 50\times P(1) = 50\times\left[\binom{5}{1}\times0.72^1\times0.28^4\right] = 1.1064\\ E(2) &= 50\times P(2) = 50\times\left[\binom{5}{2}\times0.72^2\times0.28^3\right] = 5.6900\end{aligned}\right\}$$

Pool these classes to get all $E$'s to be 5 or more.

Similarly $E(3) = 14.6313$, $E(4) = 18.8117$, $E(5) = 9.6746$

| Number of dead flies | 2 or fewer | 3 | 4 | 5 |
|---|---|---|---|---|
| $O$ | 7 | 11 | 24 | 8 |
| $E$ | 6.8825 | 14.6313 | 18.8117 | 9.6746 |

Test statistic $(\chi^2) = \frac{(7-6.8825)^2}{6.8825} + \frac{(11-14.6313)^2}{14.6313} + \frac{(24-18.8117)^2}{18.8117} + \frac{(8-9.6746)^2}{9.6746}$

t.s. $(\chi^2) = 2.62$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Number of classes after pooling.

Degrees of freedom $= 4 - 2 = 2$

Number of restrictions.

Critical value $= \chi_2^2\,(5\%) = 5.991$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude we have no reason to doubt that B(5, 0.72) is a good model for these data.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 10**

**Question:**

The number of accidents per week at a certain road junction was monitored for four years. The results obtained are summarised in the table.

| **Number of accidents** | 0 | 1 | 2 | > 2 |
|---|---|---|---|---|
| **Number of weeks** | 112 | 56 | 40 | 0 |

Using a 5% level of significance, carry out a $\chi^2$ test of the hypothesis that the number of accidents per week has a Poisson distribution. ***E***

**Solution:**

$H_0$ : A Poisson distribution is a good model.
$H_1$: A Poisson distribution is not a good model.

Total frequency $= 112 + 56 + 40 = 208$

This is 1 restriction (total frequencies agree).

mean = 0.654

This is 1 restriction (calculation of estimate of parameter).

$$E(0) = 208 \times P(0) = 208 \times \left[ e^{-0.654} \times \frac{0.654^0}{0!} \right] = 108.15$$

$$E(1) = 208 \times P(1) = 208 \times \left[ e^{-0.654} \times \frac{0.654^1}{1!} \right] = 70.73$$

$$E(2) = 208 \times P(2) = 208 \times \left[ e^{-0.654} \times \frac{0.654^2}{2!} \right] = 23.13$$

$$E(\text{more than } 2) = 208 - [108.15 + 70.73 + 23.13] = 5.99$$

| Number of accidents | 0 | 1 | 2 | 2 or more |
|---|---|---|---|---|
| $O$ | 112 | 56 | 40 | 0 |
| $E$ | 108.15 | 70.73 | 23.13 | 5.99 |

$$\text{Test statistic } \left( \chi^2 \right) = \frac{(112 - 108.15)^2}{108.15} + \frac{(56 - 70.73)^2}{70.73} + \frac{(40 - 23.13)^2}{23.13} + \frac{(0 - 5.99)^2}{5.99}$$

$$\text{t.s. } \left( \chi^2 \right) = 21.5$$

Number of classes.

Degrees of freedom $= 4 - 2 = 2$

Number of restrictions.

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

Critical value $= \chi_2^2 (5\%) = 5.991$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence the data cannot be modelled by a Poisson distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 11**

**Question:**

A tensile test is carried out on 100 steel bars which are uniform in section. The distances from the mid-points of the bars at which they fracture is recorded with the following results.

| *Distance* | $0 < d \le 10$ | $10 < d \le 20$ | $20 < d \le 30$ | $30 < d \le 40$ | $40 < d \le 50$ | $50 < d \le 60$ |
|---|---|---|---|---|---|---|
| *Frequency* | 15 | 17 | 18 | 20 | 12 | 18 |

Test at the 0.05 significance level if these data can be modelled by a continuous uniform distribution.

**Solution:**

$H_0$: The data can be modelled by a continuous uniform distribution.

$H_1$: The data cannot be modelled by a continuous uniform distribution.

Total frequency = 100

This is 1 restriction (total frequencies agree).

Expected frequency for each class $= \frac{100}{6} = 16.667 \quad (\text{or } 16\frac{2}{3})$

$$\therefore \text{Test statistic } (\chi^2) = \frac{(15-16.667)^2}{16.667} + \frac{(17-16.667)^2}{16.667} + \cdots + \frac{(18-16.667)^2}{16.667}$$

Values found for $\chi^2$ will vary according to the rounding you do. If possible work to the degree of accuracy of your calculation. Never work to any less than the degree of accuracy in the question (at least one more place is a good idea). A range of answers that encompass sensible degrees of accuracy would be accepted in an examination.

$$\text{t.s. } (\chi^2) = 2.36$$

$$\text{Degrees of freedom} = 6 - 1 = 5$$

$$\text{Critical value} = \chi_5^2\,(5\%) = 11.070$$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt the data could be modelled by a continuous uniform distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 12**

**Question:**

Samples of stones were taken at two points on a beach which were 1 mile apart. The rock types of the stones were found and classified as igneous, sedimentary or other types, with the following results.

| | | Site | |
|---|---|---|---|
| | | *A* | *B* |
| Rock type | *Igneous* | 30 | 10 |
| | *Sedimentary* | 55 | 35 |
| | *Other* | 15 | 15 |

Use a 5% significance level to see if the rocks at both sites come from the same population.

**Solution:**

$H_0$ : There is no association between site and type of rock found.
$H_1$: There is an association between site and type of rock found.

*Observed frequencies*

| | | Site | | |
|---|---|---|---|---|
| | | *A* | *B* | Total |
| *Rock Type* | *Igneous* | 30 | 10 | 40 |
| | *Sedimentary* | 55 | 35 | 90 |
| | *Other* | 15 | 15 | 30 |
| | Total | 100 | 60 | 160 |

Expected frequency '*Igneous*' and '*A*' $= \dfrac{40 \times 100}{160} = 25$

*Expected frequencies*

| | *A* | *B* |
|---|---|---|
| *Igneous* | 25 | 15 |
| *Sedimentary* | 56.25 | 33.75 |
| *Other* | 18.75 | 11.25 |

$$\text{Test statistic } (\chi^2) = \frac{(30-25)^2}{25} + \frac{(10-15)^2}{15} + \cdots + \frac{(15-11.25)^2}{11.25}$$

$$\text{t.s. } (\chi^2) = 4.74$$

$$\text{Degrees of freedom } = (3-1)\times(2-1) = 2$$

$$\text{Critical value } = \chi_2^2\,(5\%) = 5.991$$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no evidence of an association between site and type of rock found so the rocks at both sites could come from the same population.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 13**

**Question:**

A small shop sells a particular item at a fairly steady yearly rate. When looking at the weekly sales it was found that the number sold varied. The results for the 50 weeks the shop was open were as shown in the table.

| **Weekly sales** | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | $> 8$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Frequency** | 0 | 4 | 7 | 8 | 10 | 6 | 7 | 4 | 4 | 0 |

**a** Find the mean number of sales per week.

**b** Using a significance level of 5%, test to see if these can be modelled by a Poisson distribution.

**Solution:**

**a** mean $= 4.28$

Get this directly from your calculator.

**b** $H_0$ : The data can be modelled by Po (4.28).

$H_1$ : The data cannot be modelled by Po (4.28).

Total frequency $= 50$

This is 1 restriction (total frequencies agree).

$\lambda = 4.28$

This is 1 restriction (calculation of estimate of parameter).

$$\left.\begin{aligned} E(0) &= 50\times P(0) = 50\times\left[e^{-4.28}\times\frac{4.28^0}{0!}\right] = 0.6921 \\ E(1) &= 50\times P(1) = 50\times\left[e^{-4.28}\times\frac{4.28^1}{1!}\right] = 2.9623 \\ E(2) &= 50\times P(2) = 50\times\left[e^{-4.28}\times\frac{4.28^2}{2!}\right] = 6.3394 \end{aligned}\right\}$$

Pool these classes to get all the $E$'s to be 5 or more.

Similarly, $E(3) = 9.0442, E(4) = 9.6773, E(5) = 8.2838$

$E(6) = 5.9091, E(7) = 3.6130, E(8) = 1.9329,$

$E(8 \text{ or more}) = 50 - [0.6921 + 2.9623 + \ldots + 1.9329] = 1.5459$

The last 3 classes must be pooled to get all the $E$'s to be 5 or more.

| Weekly sales | 2 or fewer | 3 | 4 | 5 | 6 | 7 or more |
|---|---|---|---|---|---|---|
| $O$ | 11 | 8 | 10 | 6 | 7 | 8 |
| $E$ | 9.9938 | 9.0442 | 9.6773 | 8.2838 | 5.9091 | 7.0918 |

$$\text{Test statistic } (\chi^2) = \frac{(11-9.9938)^2}{9.9938} + \frac{(8-9.0442)^2}{9.0442} + \cdots + \frac{(8-7.0918)^2}{7.0918}$$

$$\text{t.s. } (\chi^2) = 1.18$$

Number of classes after pooling.

$$\text{Degrees of freedom} = 6 - 2$$
$$= 4$$

Number of restrictions.

$$\text{Critical value} = \chi_4^2\,(5\%) = 9.488$$

$\chi^2 <$ c.v. so accept $H_0$.

Conclude no reason to doubt that Po (4.28) could be a good model for the data.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Goodness of fit and contingency tables**
**Exercise E, Question 14**

**Question:**

A study was done of how many students in a college were left-handed and how many were right-handed. As well as left- or right-handedness the gender of each person was also recorded with the following results

| | Left handed | Right handed |
|---|---|---|
| **Male** | 100 | 600 |
| **Female** | 80 | 800 |

Use a significance test at the 0.05 level to see if there is an association between gender and left- and right-handedness.

**Solution:**

$H_0$: There is no association between gender and left- and right-handedness.
$H_1$: There is an association between gender and left- and right-handedness.

*Observed frequencies* →

| | *Left-handed* | *Right-handed* | Total |
|---|---|---|---|
| *Male* | 100 | 600 | 700 |
| *Female* | 80 | 800 | 880 |
| Total | 180 | 1400 | 1580 |

Expected frequency 'Male' and 'Left-handed' $= \dfrac{700 \times 180}{1580}$
$= 79.747$

*Expected frequencies* →

| | *Left-handed* | *Right-handed* |
|---|---|---|
| *Male* | 79.747 | 620.253 |
| *Female* | 100.253 | 779.747 |

Test statistic $(\chi^2) = \dfrac{(100-79.747)^2}{79.747} + \dfrac{(600-620.253)^2}{620.253} + \cdots + \dfrac{(800-779.747)^2}{779.747}$
$= 10.4 \quad$ (3 s.f.)

Degrees of freedom $= (2-1)\times(2-1) = 1$

Critical value $= \chi_1^2\ (5\%) = 3.841$

$\chi^2 >$ c.v. so reject $H_0$.

Conclude there is evidence of an association between gender and left- and right-handedness in this population.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**

**Exercise A, Question 1**

**Question:**

For each of the data sets of ranks given below, calculate the Spearman's rank correlation coefficient and interpret the result.

**a**

| | | | | | | |
|---|---|---|---|---|---|---|
| $r_x$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $r_y$ | 3 | 2 | 1 | 5 | 4 | 6 |

**b**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $r_y$ | 2 | 1 | 4 | 3 | 5 | 8 | 7 | 9 | 6 | 10 |

**c**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $r_x$ | 5 | 2 | 6 | 1 | 4 | 3 | 7 | 8 |
| $r_y$ | 5 | 6 | 3 | 8 | 7 | 4 | 2 | 1 |

**Solution:**

**a**

| $r_x$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $r_y$ | 3 | 2 | 1 | 5 | 4 | 6 |
| $d$ | −2 | 0 | 2 | −1 | 1 | 0 |
| $d^2$ | 4 | 0 | 4 | 1 | 1 | 0 |

$\sum d^2 = 10$

$$r_s = 1 - \frac{6 \times 10}{6(6^2 - 1)}$$

$$r_s = 0.714$$

i.e. positive correlation between the pairs of ranks.

This value is between weak and strong positive correlation between the pairs of ranks.

**b**

| $r_x$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_y$ | 2 | 1 | 4 | 3 | 5 | 8 | 7 | 9 | 6 | 10 |
| $d$ | −1 | 1 | −1 | 1 | 0 | −2 | 0 | −1 | 3 | 0 |
| $d^2$ | 1 | 1 | 1 | 1 | 0 | 4 | 0 | 1 | 9 | 0 |

$\sum d^2 = 18$

$$r_s = 1 - \frac{6 \times 18}{10(10^2 - 1)}$$

$$r_s = 0.891$$

i.e. fairly strong positive correlation between the pairs of ranks of $x$ and $y$.

**c**

| $r_x$ | 5 | 2 | 6 | 1 | 4 | 3 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $r_y$ | 5 | 6 | 3 | 8 | 7 | 4 | 2 | 1 |
| $d$ | 0 | −4 | 3 | −7 | −3 | −1 | 5 | 7 |
| $d^2$ | 0 | 16 | 9 | 49 | 9 | 1 | 25 | 49 |

$\sum d^2 = 158$

$$r_s = 1 - \frac{6 \times 158}{8(8^2 - 1)}$$

$$r_s = 0.881$$

i.e. fairly strong negative correlation between the pairs of ranks of $x$ and $y$.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 2**

**Question:**

The number of goals scored by football teams and their positions in the league were recorded as follows for the top 12 teams.

| Team | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Goals | 49 | 44 | 43 | 36 | 40 | 39 | 29 | 21 | 28 | 30 | 33 | 26 |
| League position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |

**a** Find $\Sigma d^2$.
**b** Calculate Spearman's rank correlation coefficient for these data. What conclusions can be drawn from this result?

**Solution:**

| Goals | 49 | 44 | 43 | 36 | 40 | 39 | 29 | 21 | 28 | 30 | 33 | 26 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| League position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| $r_G$ | 1 | 2 | 3 | 6 | 4 | 5 | 9 | 12 | 10 | 8 | 7 | 11 |
| $r_L$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| $d$ | 0 | 0 | 0 | 2 | −1 | −1 | 2 | 4 | 1 | −2 | −4 | −1 |
| $d^2$ | 0 | 0 | 0 | 4 | 1 | 1 | 4 | 16 | 1 | 4 | 16 | 1 |

The League position is their rank in the League $(r_L)$.

$\sum d^2 = 48$

$$r_s = 1 - \frac{6 \times 48}{12(12^2 - 1)}$$

$$r_s = 0.832$$

This shows fairly strong positive correlation between the pairs of ranks.
The more goals a team scores the higher their league position is likely to be.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 3**

**Question:**

A sample of a class's statistics projects was taken, and the projects were assessed by two teachers independently. Each teacher decided their rank order with the following results.

| Project | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Teacher A | 5 | 8 | 1 | 6 | 2 | 7 | 3 | 4 |
| Teacher B | 7 | 4 | 3 | 1 | 6 | 8 | 2 | 5 |

**a** Find $\Sigma d^2$.
**b** Calculate the rank correlation coefficient and state any conclusions you draw from it.

**Solution:**

**a**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $r_A$ | 5 | 8 | 1 | 6 | 2 | 7 | 3 | 4 |
| $r_B$ | 7 | 4 | 3 | 1 | 6 | 8 | 2 | 5 |
| $d$ | −2 | 4 | −2 | 5 | −4 | −1 | 1 | −1 |
| $d^2$ | 4 | 16 | 4 | 25 | 16 | 1 | 1 | 1 |

$$\sum d^2 = 68$$

**b** $r_s = 1 - \dfrac{6 \times 68}{8(8^2 - 1)}$

$r_s = 0.190$

There is virtually no correlation between the pairs of ranks awarded by the two teachers. They appear to be judging the projects using different criteria.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 4**

**Question:**

A veterinary surgeon and a trainee veterinary surgeon both rank a small herd of cows for quality. Their rankings are shown below.

| Cow | A | D | F | E | B | C | H | J |
|---|---|---|---|---|---|---|---|---|
| Qualified vet | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Trainee vet | 1 | 2 | 5 | 6 | 4 | 3 | 8 | 7 |

Find the rank correlation coefficient for these data, and comment on the experience of the trainee vet.

**Solution:**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $r_Q$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $r_T$ | 1 | 2 | 5 | 6 | 4 | 3 | 8 | 7 |
| $d$ | 0 | 0 | −2 | −2 | 1 | 3 | −1 | 1 |
| $d^2$ | 0 | 0 | 4 | 4 | 1 | 9 | 1 | 1 |

$$\sum d^2 = 20$$

$$r_s = 1 - \frac{6 \times 20}{8(8^2 - 1)}$$

$$r_s = 0.762$$

There is fairly strong positive correlation between the pairs of ranks. This suggests the trainee vet is rating the cows for quality in a similar way to the qualified vet.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 5**

**Question:**

Two adjudicators at an ice dance skating competition award marks as follows.

| Competitor | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Judge 1 | 7.8 | 6.6 | 7.3 | 7.4 | 8.4 | 6.5 | 8.9 | 8.5 | 6.7 | 7.7 |
| Judge 2 | 8.1 | 6.8 | 8.2 | 7.5 | 8.0 | 6.7 | 8.5 | 8.3 | 6.6 | 7.8 |

**a** Explain why you would use Spearman's rank correlation coefficient in this case.

**b** Calculate the rank correlation coefficient $r_s$, and comment on how well the judges agree.

**Solution:**

**a** The marks are discrete values within a very restricted scale.
They are also subjective judgements, not measurements.

**b**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| J1 | 7.8 | 6.6 | 7.3 | 7.4 | 8.4 | 6.5 | 8.9 | 8.5 | 6.7 | 7.7 |
| J2 | 8.1 | 6.8 | 8.2 | 7.5 | 8.0 | 6.7 | 8.5 | 8.3 | 6.6 | 7.8 |
| $r_1$ | 4 | 9 | 7 | 6 | 3 | 10 | 1 | 2 | 8 | 5 |
| $r_2$ | 4 | 8 | 3 | 7 | 5 | 9 | 1 | 2 | 10 | 6 |
| $d$ | 0 | 1 | 4 | −1 | −2 | 1 | 0 | 0 | −2 | −1 |
| $d^2$ | 0 | 1 | 16 | 1 | 4 | 1 | 0 | 0 | 4 | 1 |

$$\sum d^2 = 28$$

$$r_s = 1 - \frac{6 \times 28}{10(10^2 - 1)}$$

$$r_s = 0.830$$

This shows a fairly strong positive correlation between the pairs of ranks of the marks awarded by the two judges so it appears they are judging the ice dances using similar criteria and with similar standards.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 6**

**Question:**

**a** A teacher believes that he can predict the positions in which his students will finish in an A-Level examination. When the results were out he wished to compare his predictions with the actual results. Which correlation test should he use and why?

**b** The table shows predicted and actual orders.

| Student | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Predicted, $p$ | 2 | 4 | 1 | 3 | 8 | 6 | 9 | 5 | 10 | 7 |
| Actual, $a$ | 3 | 4 | 2 | 8 | 1 | 6 | 7 | 9 | 10 | 5 |

Calculate Spearman's rank correlation coefficient $r_s$ between $a$ and $p$. Comment on the result.

**Solution:**

**a** The teacher should use Spearman's rank correlation coefficient because the data being used are ranks (i.e. he is concerned with order).

**b**

| $p$ | 2 | 4 | 1 | 3 | 8 | 6 | 9 | 5 | 10 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|
| $a$ | 3 | 4 | 2 | 8 | 1 | 6 | 7 | 9 | 10 | 5 |
| $d$ | −1 | 0 | −1 | −5 | 7 | 0 | 2 | −4 | 0 | 2 |
| $d^2$ | 1 | 0 | 1 | 25 | 49 | 0 | 4 | 16 | 0 | 4 |

$$\sum d^2 = 100$$

$$r_s = 1 - \frac{6 \times 100}{10(10^2 - 1)}$$

$$r_s = 0.394$$

This shows weak positive correlation between the orders. The teacher has very limited ability to predict the positions in which the students will finish. It doesn't appear to be very reliable.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise A, Question 7**

**Question:**

A doctor assessed the lung damage suffered by a number of his patients who smoked, and asked each one 'For how many years have you smoked?' with the following results.

| Patient | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| Number of years smoked | 15 | 22 | 25 | 28 | 30 | 31 | 42 |
| Lung damage grade | 30 | 50 | 55 | 35 | 40 | 42 | 58 |

Calculate Spearman's rank correlation coefficient $r_s$ and comment on the result. Give your value of $\Sigma d^2$.

**Solution:**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Years smoked ($y$) | 15 | 22 | 25 | 28 | 30 | 31 | 42 |
| Lung damage grade ($g$) | 30 | 50 | 55 | 35 | 40 | 42 | 58 |
| $r_y$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $r_g$ | 1 | 5 | 6 | 2 | 3 | 4 | 7 |
| $d$ | 0 | −3 | −3 | 2 | 2 | 2 | 0 |
| $d^2$ | 0 | 9 | 9 | 4 | 4 | 4 | 0 |

$$\sum d^2 = 30$$

$$r_s = 1 - \frac{6 \times 30}{7(7^2 - 1)}$$

$$r_s = 0.464$$

This shows weak positive correlation between the pairs of ranks. There is some association between the lung damage grade and the number of years a person has smoked for – it is likely the lung damage grade is also likely to depend on other factors since the correlation is not very strong.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 1**

**Question:**

A product-moment correlation coefficient of 0.3275 was obtained from a sample of 40 pairs of values. Test whether or not this value shows evidence of correlation.

**a** at the 0.05 level (use a two-tailed test),

**b** at the 0.02 level (use a two-tailed test).

**Solution:**

**a** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}\right\}$ 2-tail $\alpha = 0.05$

test statistic $= 0.3275$

critical values $= \pm 0.3120$

t.s. > c.v. since $0.3275 > 0.3120$

reject $H_0$

Conclude there is evidence of correlation at the 5% used of significance.

**b** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}\right\}$ 2-tail $\alpha = 0.02$

test statistic $= 0.3275$

critical values $= \pm 0.3665$

t.s. < c.v. since $0.3275 < 0.3665$

accept $H_0$

Conclude no evidence of correlation at the 2% level of significance.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 2**

**Question:**

**a** Calculate the product-moment correlation coefficient for the following data, giving values for $S_{xx}, S_{yy}$ and $S_{xy}$.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x$ | 2 | 3 | 4 | 4 | 5 | 5 | 6 |
| $y$ | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

**b** Test, for these data, the null hypothesis that there is no correlation between $x$ and $y$. Use a 1% significance level.
State any assumptions you have made.

**Solution:**

**a**

> In an exam do not set up a table of values – you will run out of time! Get the values directly from your calculator.

$\sum x = 29, \sum x^2 = 131, \sum y = 28, \sum y^2 = 140, \sum xy = 99, n = 7$

$$S_{xy} = 99 - \frac{29 \times 28}{7} = -17$$

$$S_{xx} = 131 - \frac{(29)^2}{7} = 10.857$$

$$S_{yy} = 140 - \frac{(28)^2}{7} = 28$$

$$\therefore r = \frac{-17}{\sqrt{(10.857 \times 28)}}$$

$$r = -0.975$$

**b** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}\right\}$ 2-tail $\alpha = 0.01$

test statistic $= r = -0.975$
critical values $= \pm 0.8745$
t.s. < c.v. since $-0.975 < -0.8745$
so reject $H_0$
Conclude there is evidence of correlation between $x$ and $y$.
Assumption: $x$ and $y$ are both normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 3**

**Question:**

The ages $X$ (years) and heights $Y$ (cm) of 11 members of a football team were recorded and the following statistics were used to summarise the results.

$$\Sigma X = 168, \Sigma Y = 1275, \Sigma XY = 20\,704, \Sigma X^2 = 2585$$
$$\Sigma Y^2 = 320\,019$$

**a** Calculate the product-moment correlation coefficient for these data.

**b** Test the assertion that height and weight are positively correlated by using a suitable test. State your conclusion in words and any assumptions you have made. (Use a 5% level of significance.)

**Solution:**

**a** $$r = \frac{\left[20\,704 - \frac{168 \times 1275}{11}\right]}{\sqrt{\left[\left(2585 - \frac{168^2}{11}\right)\left(320\,019 - \frac{1275^2}{11}\right)\right]}}$$

$$\boxed{r = \frac{s_{xy}}{s_{xx}s_{yy}}}$$

$r = 0.677$

**b** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{array}\right\}$ 1-tail $\alpha = 0.05$

test statistic $= 0.677$

critical values $= 0.5214$

t.s. > c.v. so reject $H_0$

Conclude there is evidence of positive correlation between the age and height of members of a football team – the older the player, the taller they tend to be.

Assumption: both the ages and the heights of the players are normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 4**

**Question:**

**a** Explain briefly your understanding of the term 'correlation'. Describe how you used, or could have used, correlation in a project or in class work.

**b** Twelve students sat two Biology tests, one theoretical the other practical. Their marks are shown below.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Marks in theoretical test ($t$)** | 5 | 9 | 7 | 11 | 20 | 4 | 6 | 17 | 12 | 10 | 15 | 16 |
| **Marks in practical test ($p$)** | 6 | 8 | 9 | 13 | 20 | 9 | 8 | 17 | 14 | 8 | 17 | 18 |

Find to 3 significant figures,

**i** the value of $S_{tp}$

**ii** the product-moment correlation coefficient.

**c** Use a 0.05 significance level and a suitable test to check the statement that 'students who do well in theoretical Biology also do well in practical Biology tests'.

**Solution:**

**a** The product-moment coefficient of correlation is the measure of the strength of the linear link between two variables.
You could use it to investigate whether there is correlation between the age of a lichen and its diameter, for example.

**b**

| $t$ | 5 | 9 | 7 | 11 | 20 | 4 | 6 | 17 | 12 | 10 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | 6 | 8 | 9 | 13 | 20 | 9 | 8 | 17 | 14 | 8 | 17 | 18 |

Enter these data into your calculator to obtain the following results directly – *don't* complete a table of results in an exam as you will run out of time!

$$\sum t = 132, \sum t^2 = 1742, \sum p = 147, \sum p^2 = 2057, \sum tp = 1872, n = 12$$

**i** $S_{tp} = 1872 - \dfrac{132 \times 147}{12}$

$S_{tp} = 255$

**ii** $r = 0.935$

Get this value directly from your calculator.

**c** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{array}\right\}$ 1-tail $\alpha = 0.05$

test statistic $= 0.935$

critical values $= 0.4973$

t.s. > c.v. so reject $H_0$.

Conclude there is positive correlation between theoretical biology and practical biology marks – this implies that students who do well in theoretical biology tests also tend to do well in practical biology tests.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 5**

**Question:**

The following table shows the marks attained by 8 students in English and Mathematics tests.

| Student | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| English | 25 | 18 | 32 | 27 | 21 | 35 | 28 | 30 |
| Mathematics | 16 | 11 | 20 | 17 | 15 | 26 | 32 | 20 |

**a** Calculate the product-moment correlation coefficient.
A teacher thinks that the population correlation coefficient between the marks is likely to be zero.
**b** Test the teacher's idea at the 5% level of significance.

**Solution:**

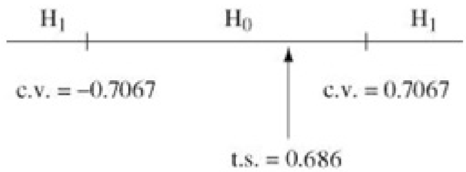
**a** $r = 0.68556\ldots$

In the exam get this directly from your calculator. If you set up a table of results you are likely to run out of time.

so $r = 0.686$ (3 s.f.)

**b**

$$\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}\right\} \quad \text{2-tail} \quad \alpha = 0.05$$




test statistic $= 0.686$

critical values $= \pm 0.7067$

at upper tail, t.s. < c.v. since $0.686 < 0.7067$ so accept $H_0$.

Conclude there is evidence the p.m.c.c. could be zero so the teacher's theory is supported.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise B, Question 6**

**Question:**

A small company decided to import fine Chinese porcelain. They believed that in the long term this would prove to be an increasingly profitable arrangement with profits increasing proportionally to sales. Over the next 6 years their sales and profits were as shown in the table below.

| Year | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 |
|---|---|---|---|---|---|---|
| Sales in thousands | 165 | 165 | 170 | 178 | 178 | 175 |
| Profits in £1000 | 65 | 72 | 75 | 76 | 80 | 83 |

Using a 1% significance level test to see if there is any evidence that the company's beliefs were correct, and that profit and sales were positively correlated.

**Solution:**

$r = 0.793$ ← Get this value directly from your calculator. *Don't* set up a table of values as you will be likely to run out of time in the exam if you do!

$$\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{array}\right\} \text{ 1-tail } \alpha = 0.01$$

test statistic $= 0.793$

critical values $= 0.8822$

t.s. < c.v. so accept $H_0$.

Conclude there is insufficient evidence at the 1% significance level to support the company's belief.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 1**

**Question:**

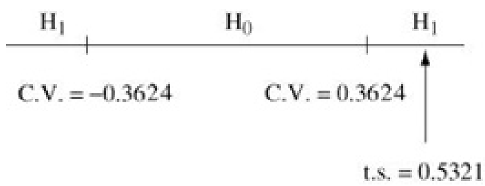
A Spearman's rank correlation obtained from the fuel consumption of a selection of 30 cars and their engine sizes gave a rank correlation coefficient $r_s = 0.5321$. Investigate whether or not the fuel consumption is related to the engine size. State your null and alternative hypotheses. (Use a 5% level of significance.)

**Solution:**

$H_0: \rho_s = 0$ There is no association between fuel consumption and engine size.
$H_1: \rho_s \neq 0$ There is an association between fuel consumption and engine size.
2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.5321$
$n = 30$ so critical value $= \pm 0.3624$
t.s. > c.v. in the upper tail
since $0.5321 > 0.3624$
so reject $H_0$.




Conclude there is evidence of an association between fuel consumption and engine size.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 2**

**Question:**

For one of the activities at a gymnastics competition, 8 gymnasts were awarded marks out of 10 for each of artistic performance and technical ability. The results were as follows.

| Gymnast | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Technical ability | 8.5 | 8.6 | 9.5 | 7.5 | 6.8 | 9.1 | 9.4 | 9.2 |
| Artistic performance | 6.2 | 7.5 | 8.2 | 6.7 | 6.0 | 7.2 | 8.0 | 9.1 |

The value of the product-moment correlation coefficient for these data is 0.774.

**a** Stating your hypotheses clearly and using a 1% level of significance, test for evidence of a positive association between technical ability and artistic performance. Interpret this value.

**b** Calculate the value of the rank correlation coefficient for these data.

**c** Stating your hypotheses clearly and using a 1% level of significance, interpret this coefficient.

**d** Explain why the rank correlation coefficient might be the better one to use with these data.

**Solution:**

**a** $H_0: \rho = 0$
$H_1: \rho > 0$ $\Big\}$ $\boxed{\text{1-tail } \alpha = 0.01}$

Test statistic $= r = 0.774$
$n = 8$ critical value $= 0.7887$
t.s. < c.v. in upper tail test
so accept $H_0$.
Conclude there is insufficient evidence of positive correlation between technical ability and artistic performance at the 1% significance level.

**b**

| $T$ | 8.5 | 8.6 | 9.5 | 7.5 | 6.8 | 9.1 | 9.4 | 9.2 |
|---|---|---|---|---|---|---|---|---|
| $A$ | 6.2 | 7.5 | 8.2 | 6.7 | 6.0 | 7.2 | 8.0 | 9.1 |
| $r_T$ | 6 | 5 | 1 | 7 | 8 | 4 | 2 | 3 |
| $r_A$ | 7 | 4 | 2 | 6 | 8 | 5 | 3 | 1 |
| $d$ | −1 | 1 | −1 | 1 | 0 | −1 | −1 | 2 |
| $d^2$ | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 4 |

$\sum d^2 = 10$

$$r_s = 1 - \frac{6 \times 10}{8(8^2 - 1)}$$

$r_s = 0.881$

**c** $H_0: \rho_s = 0$ There is no association between technical ability and artistic performance.
$H_1: \rho_s > 0$ There is positive association between technical ability and artistic performance.

1-tail $\alpha = 0.01$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.881$
$n = 8$ critical value $= 0.8333$
upper tail test, t.s. > c.v.
so reject $H_0$

Conclude there is evidence of a positive association between technical ability and artistic performance. Gymnasts who are better in their technical ability also appear to be better in their artistic performance.

**d** The data are discrete results in a limited range. They are judgments, not measurements. It is also unlikely that these scores will both be normally distributed.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 3**

**Question:**

Two judges ranked 8 ice skaters in a competition according to the table below.

| Skater / Judge | i | ii | iii | iv | v | vi | vii | viii |
|---|---|---|---|---|---|---|---|---|
| A | 2 | 5 | 3 | 7 | 8 | 1 | 4 | 6 |
| B | 3 | 2 | 6 | 5 | 7 | 4 | 1 | 8 |

**a** Evaluate Spearman's rank correlation coefficient between the ranks of the two judges.

**b** Use a suitable test, at the 5% level of significance. Interpret your findings to investigate for evidence of positive association between the rankings of the judges. ***E***

**Solution:**

**a**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $r_A$ | 2 | 5 | 3 | 7 | 8 | 1 | 4 | 6 |
| $r_B$ | 3 | 2 | 6 | 5 | 7 | 4 | 1 | 8 |
| $d$ | −1 | 3 | −3 | 2 | 1 | −3 | 3 | −2 |
| $d^2$ | 1 | 9 | 9 | 4 | 1 | 9 | 9 | 4 |

$\sum d^2 = 46$

$$r_s = 1 - \frac{6 \times 46}{8(8^2 - 1)}$$

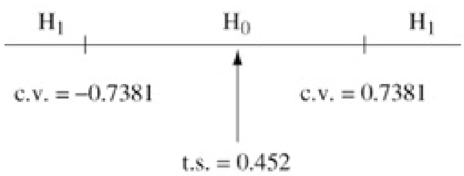
$$r_s = 0.452$$

**b** $H_0: \rho_s = 0$ There is no association between the rankings of the 2 judges.
$H_1: \rho_s \neq 0$ There is an association between the rankings of the 2 judges.

2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.452$
Critical values $= \pm 0.7381$
t.s. < c.v. in the upper tail
since $0.452 < 0.7381$
so accept $H_0$.




Conclude there is no association between the rankings awarded by the 2 judges. They appear to be using different criteria in their judgements.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**

**Exercise C, Question 4**

**Question:**

Each of the teams in a school hockey league had the total number of goals scored by them and against them recorded, with the following results.

| Team | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| Goals for | 39 | 40 | 28 | 27 | 26 | 30 | 42 |
| Goals against | 22 | 28 | 27 | 42 | 24 | 38 | 23 |

Investigate whether there is any correlation between the goals for and those against by using Spearman's rank correlation coefficient. Use a suitable test at the 1% level to investigate the statement, 'A team that scores a lot of goals concedes very few goals'.

**Solution:**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $F$ | 39 | 40 | 28 | 27 | 26 | 30 | 42 |
| $A$ | 22 | 28 | 27 | 42 | 24 | 38 | 23 |
| $r_F$ | 3 | 2 | 5 | 6 | 7 | 4 | 1 |
| $r_A$ | 7 | 3 | 4 | 1 | 5 | 2 | 6 |
| $d$ | −4 | −1 | 1 | 5 | 2 | 2 | −5 |
| $d^2$ | 16 | 1 | 1 | 25 | 4 | 4 | 25 |

$\sum d^2 = 76$

$$r_s = 1 - \frac{6 \times 76}{7(7^2 - 1)}$$

$r_s = -0.357$

$H_0: \rho_s = 0$ There is no association between 'goals for' and 'goals against'.

$H_1: \rho_s < 0$ There is negative association between 'goals for' and 'goals against'.

1-tail $\alpha = 0.01$

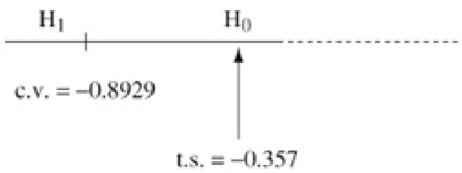
You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = -0.357$

critical value $= -0.8929$

Lower tail test where t.s. > c.v.

so accept $H_0$.




Conclude there is insufficient evidence to support the statement.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 5**

**Question:**

The weekly takings and weekly profits for six branch shops of a company are set out below.

| Shop | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Takings (£) | 400 | 6200 | 3600 | 5100 | 5000 | 3800 |
| Profits (£) | 400 | 1100 | 450 | 750 | 800 | 500 |

**a** Calculate the coefficient of rank correlation $r_s$ between the takings and profit.

**b** It is assumed that profits and takings will be positively correlated. Using a suitable hypothesis test (stating the null and alternative hypotheses) test this assertion at the 5% level of significance.

**Solution:**

**a**

| | | | | | | |
|---|---|---|---|---|---|---|
| $T$ | 400 | 6200 | 3600 | 5100 | 5000 | 3800 |
| $P$ | 400 | 1100 | 450 | 750 | 800 | 500 |
| $r_T$ | 6 | 1 | 5 | 2 | 3 | 4 |
| $r_P$ | 6 | 1 | 5 | 3 | 2 | 4 |
| $d$ | 0 | 0 | 0 | −1 | 1 | 0 |
| $d^2$ | 0 | 0 | 0 | 1 | 1 | 0 |

$\sum d^2 = 2$

$$r_s = 1 - \frac{6 \times 2}{6(6^2 - 1)}$$

$$r_s = 0.943$$

**b** $H_0: \rho_s = 0$ There is no association between takings and profits.
$H_1: \rho_s > 0$ There is positive association between takings and profits.

1-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

upper tail test where t.s. > c.v.
so reject $H_0$.
Conclude there is evidence to support the statement.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 6**

**Question:**

The rankings of 12 students in Mathematics and Music were as follows.

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mathematics** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| **Music** | 6 | 4 | 2 | 3 | 1 | 7 | 5 | 9 | 10 | 8 | 11 | 12 |

**a** Calculate the coefficient of rank correlation $r_s$. [Show your value of $\Sigma d^2$.]

**b** Test the assertion that there is no correlation between these subjects. State the null and alternative hypotheses used. Use a 5% significance level.

**Solution:**

**a**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_{\text{math}}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| $r_{\text{mus}}$ | 6 | 4 | 2 | 3 | 1 | 7 | 5 | 9 | 10 | 8 | 11 | 12 |
| $d$ | −5 | −2 | 1 | 1 | 4 | −1 | 2 | −1 | −1 | 2 | 0 | 0 |
| $d^2$ | 25 | 4 | 1 | 1 | 16 | 1 | 4 | 1 | 1 | 4 | 0 | 0 |

$$\sum d^2 = 58$$

$$r_s = 1 - \frac{6 \times 58}{12(12^2 - 1)}$$

$$r_s = 0.797$$

**b** $\left.\begin{array}{l}\text{H}_0: \rho_s = 0 \text{ There is no correlation between the rankings in Mathematics and Music.} \\ \text{H}_1: \rho_s \neq 0 \text{ There is correlation between the rankings in Mathematics and Music.}\end{array}\right\}$ 2-tail $\alpha = 0.05$

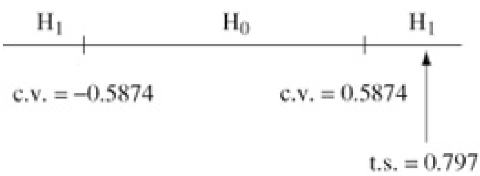
You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= 0.797$

critical values $= \pm 0.5874$

t.s. > c.v. in upper tail, since $0.797 > 0.5874$

so reject $\text{H}_0$.




Conclude there is evidence of an association between the rankings in Mathematics and Music.

Here it appears that students who do well in Mathematics are also likely to do well in Music.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 7**

**Question:**

**a** A child is asked to place 10 objects in order and gives the ordering

A C H F B D G E J I

The correct ordering is

A B C D E F G H I J

Find a coefficient of rank correlation between the child's ordering and the correct ordering.

**b** Use a 5% significance level and test whether there is an association between the child's order and the correct ordering. Draw conclusions about this result. ***E***

**Solution:**

**a**

| Child's order | A | C | H | F | B | D | G | E | J | I |
|---|---|---|---|---|---|---|---|---|---|---|
| Correct order | A | B | C | D | E | F | G | H | I | J |
| $r_{child}$ | 1 | 3 | 8 | 6 | 2 | 4 | 7 | 5 | 10 | 9 |
| $r_{correct}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $d$ | 0 | 1 | 5 | 2 | −3 | −2 | 0 | −3 | 1 | −1 |
| $d^2$ | 0 | 1 | 25 | 4 | 9 | 4 | 0 | 9 | 1 | 1 |

$$\sum d^2 = 54$$

$$r_s = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$r_s = 0.673$$

**b** $\left.\begin{array}{l} H_0 : \rho_s = 0 \text{ There is no association between the child's ordering and the correct ordering.} \\ H_1 : \rho_s \neq 0 \text{ There is an association between the child's ordering and the correct ordering.} \end{array}\right\}$ 2-tail $\alpha = 0.05$

| |
|---|
| You should always put your hypotheses in terms of $\rho$ first. |

Test statistic $= 0.673$

critical values $= 0.5636$

t.s. > c.v. so reject $H_0$.

Conclude there is evidence of a positive association between the child's ordering and the correct ordering so the child is showing some ability to perform the task.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise C, Question 8**

**Question:**

The crop of a root vegetable was measured over six consecutive years, the years being ranked for wetness. The results are given in the table below.

| Year | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Crop (10 000 tons) | 62 | 73 | 52 | 77 | 63 | 61 |
| Rank of wetness | 5 | 4 | 1 | 6 | 3 | 2 |

Calculate, to 3 decimal places, a Spearman's rank correlation coefficient for these data. Test the assertion that crop and wetness are not correlated. (Use a 5% level of significance).

**Solution:**

| $C$ | 62 | 73 | 52 | 77 | 63 | 61 |
|---|---|---|---|---|---|---|
| $r_c$ | 4 | 2 | 6 | 1 | 3 | 5 |
| $r_w$ | 5 | 4 | 1 | 6 | 3 | 2 |
| $d$ | −1 | −2 | 5 | −5 | 0 | 3 |
| $d^2$ | 1 | 4 | 25 | 25 | 0 | 9 |

$\sum d^2 = 64$

$$r_s = 1 - \frac{6 \times 64}{6(6^2 - 1)} = -0.82857\ldots$$

$r_s = -0.829$

$H_0$: $\rho_s = 0$ There is no correlation between the ranks of crop yield and wetness.
$H_1$: $\rho_s \neq 0$ There is correlation between the ranks of crop yield and wetness.

2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= -0.829$

critical value $= 0.8857$

In the lower tail, t.s. > c.v. since $-0.829 > 0.8857$
so accept $H_0$.
Conclude at the 1% significance level there is insufficient evidence of correlation between the ranks of crop yield and wetness. (More samples need to be taken as the coefficient is quite close to the critical value).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 1**

**Question:**

**a** Two judges at a cat show place the 10 entries in the following rank orders.

| Cat | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| First judge | 4 | 6 | 1 | 2 | 5 | 3 | 10 | 9 | 8 | 7 |
| Second judge | 2 | 9 | 3 | 1 | 7 | 4 | 6 | 8 | 5 | 10 |

Find a coefficient of rank correlation between the two rankings and, using the tables provided, comment on the extent of the agreement between the two judges.

**b** Explain briefly the role of the null and alternative hypotheses in a test of significance. ***E***

**Solution:**

**a**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_F$ | 4 | 6 | 1 | 2 | 5 | 3 | 10 | 9 | 8 | 7 |
| $r_S$ | 2 | 9 | 3 | 1 | 7 | 4 | 6 | 8 | 5 | 10 |
| $d$ | 2 | −3 | −2 | 1 | −2 | −1 | 4 | 1 | 3 | −3 |
| $d^2$ | 4 | 9 | 4 | 1 | 4 | 1 | 16 | 1 | 9 | 9 |

$$\sum d^2 = 58$$

$$r_s = 1 - \frac{6 \times 58}{10(10^2 - 1)} = 0.64848\ldots$$

$$r_s = 0.648$$

We are looking for *positive* correlation between the rankings awarded by the two judges. There is clear evidence of a positive correlation at the 10% and 5% level(s) of significance since our test statistic $r_s = 0.648$ is bigger than the critical value 0.5636. There is insufficient evidence of positive correlation at the 1% significance level where the critical value 0.7455 is greater than the test statistic.

**b** The null hypothesis is what we assume to be true unless proved otherwise and the alternative hypothesis is what we conclude is happening if we reject the null hypothesis. The null hypothesis is only rejected if it is true with a probability equal to the significance level of the test.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 2**

**Question:**

**a** Explain briefly the conditions under which you would measure association using a rank correlation coefficient.

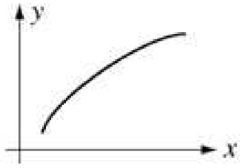
**b** Nine applicants for places at a college were interviewed by two tutors. Each tutor ranked the applicants in order of merit. The rankings are shown below.

| Applicant | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Tutor 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Tutor 2 | 1 | 3 | 5 | 4 | 2 | 7 | 9 | 8 | 6 |

Investigate the extent of the agreement between the two tutors. ***E***

**Solution:**

**a** Where there is a link between the variables $x$ and $y$ but it isn't linear; it could be




for example.

Where the results you have are rankings already where items have been put in order of preference or judgements, or where alphabetical grades have been awarded − here the data won't be from a bivariate normal distribution.

**b**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $r_{T_1}$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $r_{T_2}$ | 1 | 3 | 5 | 4 | 2 | 7 | 9 | 8 | 6 |
| $d$ | 0 | −1 | −2 | 0 | 3 | −1 | −2 | 0 | 3 |
| $d^2$ | 0 | 1 | 4 | 0 | 9 | 1 | 4 | 0 | 9 |

$\sum d^2 = 28$

$$r_s = 1 - \frac{6 \times 28}{9(9^2 - 1)}$$

$$r_s = 0.7666\ldots = 0.767$$

$\left.\begin{array}{l}\text{H}_0: \rho_s = 0 \text{ There is no correlation between the ranks awarded by the two tutors.}\\ \text{H}_1: \rho_s > 0 \text{ There is positive correlation between the ranks awarded by the two tutors.}\end{array}\right\}$ 1-tail

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.767$

This is greater than the critical value of 0.6833 at the $2\frac{1}{2}\%$ significance level so we would reject $\text{H}_0$ and we would conclude there is evidence of agreement between the two tutors at the $2\frac{1}{2}\%$ significance level.

At the 1% significance level the test statistic and critical value are very close so it is inconclusive at this level of significance.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 3**

**Question:**

In a ski-jump contest each competitor made two jumps. The order of merit for the 10 competitors who completed both jumps are shown.

| Ski-jumper | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| First jump | 2 | 9 | 7 | 4 | 10 | 8 | 6 | 5 | 1 | 3 |
| Second jump | 4 | 10 | 5 | 1 | 8 | 9 | 2 | 7 | 3 | 6 |

**a** Calculate, to 2 decimal places, a rank correlation coefficient for the performance of the ski-jumpers in the two jumps.

**b** Using a 5% significance, and quoting from the table of critical values, investigate whether there is positive association between the two jumps. State your null and alternative hypotheses clearly. ***E***

**Solution:**

**a**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_{1st}$ | 2 | 9 | 7 | 4 | 10 | 8 | 6 | 5 | 1 | 3 |
| $r_{2nd}$ | 4 | 10 | 5 | 1 | 8 | 9 | 2 | 7 | 3 | 6 |
| $d$ | −2 | −1 | 2 | 3 | 2 | −1 | 4 | −2 | −2 | −3 |
| $d^2$ | 4 | 1 | 4 | 9 | 4 | 1 | 16 | 4 | 4 | 9 |

$$\sum d^2 = 56$$

$$r_s = 1 - \frac{6 \times 56}{10(10^2 - 1)}$$

$$r_s = 0.6606$$

$$r_s = 0.66 \text{ (2 d.p.)}$$

**b** $H_0: \rho_s = 0$ There is no correlation between the order of merit for the two jumps.
$H_1: \rho_s > 0$ There is positive correlation between the order of merit for the two jumps.

1-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= 0.66$

Critical value $= 0.5636$

t.s. > c.v. so reject $H_0$.

Conclude there is evidence of positive correlation between the order of merit for the two jumps – jumpers who did well in the first jump are also likely to do well in the second jump.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 4**

**Question:**

An expert on porcelain is asked to place seven china bowls in date order of manufacture, assigning the rank 1 to the oldest bowl. The actual dates of manufacture and the order given by the expert are shown below.

| Bowl | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| Date of manufacture | 1920 | 1857 | 1710 | 1896 | 1810 | 1690 | 1780 |
| Order given by expert | 7 | 3 | 4 | 6 | 2 | 1 | 5 |

**a** Find, to 3 decimal places, the Spearman's rank correlation coefficient between the order of manufacture and the order given by the expert.

**b** Refer to the table of critical values to comment on your results. State clearly the null hypothesis being tested. ***E***

**Solution:**

**a**

| Date of manufacture | 1920 | 1857 | 1710 | 1896 | 1810 | 1690 | 1780 |
|---|---|---|---|---|---|---|---|
| Date rank | 7 | 5 | 2 | 6 | 4 | 1 | 3 |
| Expert rank | 7 | 3 | 4 | 6 | 2 | 1 | 5 |
| $d$ | 0 | 2 | −2 | 0 | 2 | 0 | −2 |
| $d^2$ | 0 | 4 | 4 | 0 | 4 | 0 | 4 |

$\sum d^2 = 16$

$$r_s = 1 - \frac{6 \times 16}{7(7^2 - 1)}$$

$r_s = 0.714$ (3 d.p.)

**b** $H_0: \rho_s = 0$ There is no association between the pairs of ranks.
$H_1: \rho_s > 0$ There is positive association between the pairs of ranks, i.e. they agree with each other. } 1-tail

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= 0.714\ldots$

At the 5% significance level the critical value $= 0.7143$

t.s. > c.v. so reject $H_0$.

Conclude at the 5% level of significance there is minimal evidence that the expert is ranking the porcelain in date order correctly.

At the $2\frac{1}{2}\%$ significance level the critical value is 0.7857 which is bigger than the test statistic meaning $H_0$ would be accepted. So at the $2\frac{1}{2}\%$ significance level there is insufficient evidence of agreement.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 5**

**Question:**

A small bus company provides a service for a small town and some neighbouring villages. In a study of their service a random sample of 20 journeys was taken and the distances $x$, in kilometres, and journey times $t$, in minutes, were recorded. The average distance was 4.535 km and the average journey time was 15.15 minutes.

**a** Using $\Sigma x^2 = 493.77, \Sigma t^2 = 4897, \Sigma xt = 1433.8$, calculate the product-moment correlation coefficient for these data.

**b** Stating your hypotheses clearly test, at the 5% level, whether or not there is evidence of a positive correlation between journey time and distance.

**c** State any assumptions that have to be made to justify the test in **b**. ***E***

**Solution:**

**a** $n = 20$

$\bar{x} = \frac{\sum x}{n}$ gives $4.535 = \frac{\sum x}{20}$

so $\sum x = 20 \times 4.535 = 90.7$

$\bar{t} = \frac{\sum t}{n}$ gives $15.15 = \frac{\sum t}{20}$

so $\sum t = 20 \times 15.15 = 303$

$$\therefore r = \frac{\left[1433.8 - \frac{(90.7)(303)}{20}\right]}{\sqrt{\left(493.77 - \frac{90.7^2}{20}\right)\left(4897 - \frac{303^2}{20}\right)}}$$

$r = 0.375$

**b** $\left.\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{array}\right\}$ 1-tail $\alpha = 0.05$

Test statistic = 0.375
critical value = 0.3783
t.s. < c.v. since $0.375 < 0.3783$
so accept $H_0$.

Conclude there is insufficient evidence of positive correlation between distance and time at the 5% level of significance.

**c** Both distance and journey time are normally distributed.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 6**

**Question:**

A group of students scored the following marks in their Statistics and Geography examinations.

| Student | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Statistics | 64 | 71 | 49 | 38 | 72 | 55 | 54 | 68 |
| Geography | 55 | 50 | 51 | 47 | 65 | 45 | 39 | 82 |

**a** Find the value of the Spearman's rank correlation coefficient between the marks of these students.

**b** Stating your hypotheses and using a 5% level of significance, test whether marks in Statistics and marks in Geography are associated.

**Solution:**

**a**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $S$ | 64 | 71 | 49 | 38 | 72 | 55 | 54 | 68 |
| $G$ | 55 | 50 | 51 | 47 | 65 | 45 | 39 | 82 |
| $r_S$ | 4 | 2 | 7 | 8 | 1 | 5 | 6 | 3 |
| $r_G$ | 3 | 5 | 4 | 6 | 2 | 7 | 8 | 1 |
| $d$ | 1 | –3 | 3 | 2 | –1 | –2 | –2 | 2 |
| $d^2$ | 1 | 9 | 9 | 4 | 1 | 4 | 4 | 4 |

$$\sum d^2 = 36$$

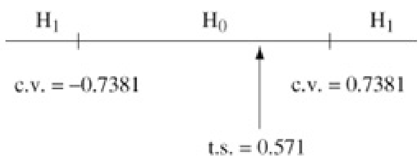
$$r_s = 1 - \frac{6 \times 36}{8(8^2 - 1)}$$

$$r_s = 0.571$$

**b** $\left.\begin{array}{l} \text{H}_0: \ \rho_s = 0 \text{ There is no association between marks in Statistics and Geography.} \\ \text{H}_1: \ \rho_s \neq 0 \text{ There is an association between marks in Statistics and Geography.} \end{array}\right\}$ 2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.571$
critical value $= \pm 0.7381$
In the upper tail t.s. < c.v.
so accept $\text{H}_0$.




Conclude there is no evidence of an association between marks students score in Statistics and Geography, i.e. students who are good at one of these subjects aren't necessarily going to do well in the other subject.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 7**

**Question:**

An international study of female literacy investigated whether there was any correlation between the life expectancy of females and the percentage of adult females who were literate. A random sample of 8 countries was taken and the following data were collected.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Life expectancy (years)** | 49 | 76 | 69 | 71 | 50 | 64 | 78 | 74 |
| **Literacy (%)** | 25 | 88 | 80 | 62 | 37 | 86 | 89 | 67 |

**a** Evaluate Spearman's rank correlation coefficient for these data.

**b** Stating your hypotheses clearly test, at the 5% level of significance, whether or not there is evidence of a correlation between the rankings of literacy and life expectancy for females.

**c** Give one reason why Spearman's rank correlation coefficient and not the product-moment correlation coefficient has been used in this case. ***E***

**Solution:**

**a**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| $LE$ | 49 | 76 | 69 | 71 | 50 | 64 | 78 | 74 |
| $Lit$ | 25 | 88 | 80 | 62 | 37 | 86 | 89 | 67 |
| $r_{LE}$ | 8 | 2 | 5 | 4 | 7 | 6 | 1 | 3 |
| $r_{Lit}$ | 8 | 2 | 4 | 6 | 7 | 3 | 1 | 5 |
| $d$ | 0 | 0 | 1 | −2 | 0 | 3 | 0 | −2 |
| $d^2$ | 0 | 0 | 1 | 4 | 0 | 9 | 0 | 4 |

$\sum d^2 = 18$

$$r_s = 1 - \frac{6 \times 18}{8(8^2 - 1)}$$

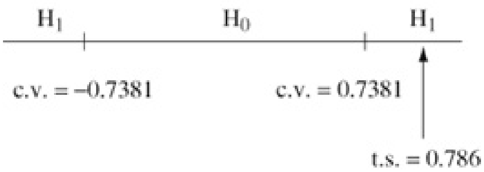
$r_s = 0.786$

**b** $H_0: \rho_s = 0$ There is no correlation between the rankings of literacy and life expectancy for females.
$H_1: \rho_s \neq 0$ There is correlation between the rankings of literacy and life expectancy for females.

2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.786$
critical value $= \pm 0.7381$
In upper tail t.s. > c.v.
so reject $H_0$.




Conclude there is evidence of correlation between the rankings of literacy and life expectancy for women. It appears that where a higher percentage of women are literate, they appear to have a higher life expectancy.

**c** We cannot assume that both life expectancy and percentage of literary are both normally distributed.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Regression and correlation

**Exercise D, Question 8**

**Question:**

Six friesian cows were ranked in order of merit at an agricultural show by the official judge and by a student vet.
The ranks were as follows:

| | | | | | | |
|---|---|---|---|---|---|---|
| **Official judge** | 1 | 2 | 3 | 4 | 5 | 6 |
| **Student vet** | 1 | 5 | 4 | 2 | 6 | 3 |

**a** Calculate Spearman's rank correlation coefficient between these rankings.

**b** Investigate whether or not there was agreement between the rankings of the judge and the student.
State clearly your hypotheses, and carry out an appropriate one-tailed significance test at the 5% level. ***E***

**Solution:**

**a**

| | | | | | | |
|---|---|---|---|---|---|---|
| $r_{OJ}$ | 1 | 2 | 3 | 4 | 5 | 6 |
| $r_{SV}$ | 1 | 5 | 4 | 2 | 6 | 3 |
| $d$ | 0 | −3 | −1 | 2 | −1 | 3 |
| $d^2$ | 0 | 9 | 1 | 4 | 1 | 9 |

$\sum d^2 = 24$

$$r_s = 1 - \frac{6 \times 24}{6(6^2 - 1)}$$

$r_s = 0.314$

**b** $\mathrm{H}_0: \rho_s = 0$ There is no correlation between the rankings of the official judge and student vet.
$\mathrm{H}_1: \rho_s > 0$ There is positive correlation between the rankings of the official judge and student vet.

1-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.314$

critical value $= 0.8286$

t.s. < c.v. so accept $\mathrm{H}_0$.

Conclude there is insufficient evidence of agreement between the rankings of the official judge and the student vet. They appear to be ranking using different criteria to each other.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 9**

**Question:**

As part of a survey in a particular profession, age, $x$ years, and salary, £ $y$ thousands, were recorded.
The values of $x$ and $y$ for a randomly selected sample of ten members of the profession are as follows:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 30 | 52 | 38 | 48 | 56 | 44 | 41 | 25 | 32 | 27 |
| y | 22 | 38 | 40 | 34 | 35 | 32 | 28 | 27 | 29 | 41 |

**a** Calculate, to 3 decimal places, the product-moment correlation coefficient between age and salary.
**b** State two conditions under which it might be appropriate to use Spearman's rank correlation coefficient.
**c** Calculate, to 3 decimal places, the Spearman's rank correlation coefficient between age and salary.
It is suggested that there is no correlation between age and salary.
**d** Set up appropriate null and alternative hypotheses and carry out an appropriate test. (Use a 5% significance level.) ***E***

**Solution:**

**a** $r = 0.340$

Get this value directly from your calculator in an examination. Don't set up a table of values – you will run out of time if you do.

**b** When both sets of data aren't from normal distributions. When at least one set of data is given as grades (letters) *or* ranking of preference or size.

**c**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $x$ | 30 | 52 | 38 | 48 | 56 | 44 | 41 | 25 | 32 | 27 |
| $y$ | 22 | 38 | 40 | 34 | 35 | 32 | 28 | 27 | 29 | 41 |
| $r_x$ | 8 | 2 | 6 | 3 | 1 | 4 | 5 | 10 | 7 | 9 |
| $r_y$ | 10 | 3 | 2 | 5 | 4 | 6 | 8 | 9 | 7 | 1 |
| $d$ | −2 | −1 | 4 | −2 | −3 | −2 | −3 | 1 | 0 | 8 |
| $d^2$ | 4 | 1 | 16 | 4 | 9 | 4 | 9 | 1 | 0 | 64 |

$\sum d^2 = 112$

$$r_s = 1 - \frac{6 \times 112}{10(10^2 - 1)}$$

$r_s = 0.321$ (3 d.p.)

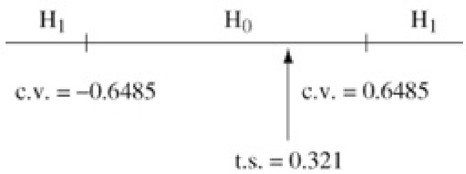
**d** $H_0$: $\rho_s = 0$ There is no correlation between the rankings of age and salary.
$H_1$: $\rho_s \neq 0$ There is correlation between the rankings of age and salary.

2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Here the test on Spearman's coefficient should be used because it is unlikely that salary and age are both normally distributed.

Test statistic $= r_s = 0.321$
critical value $= \pm 0.6485$
In upper tail, t.s. < c.v.
Since $0.321 < +0.6485$
accept $H_0$.




Conclude no evidence of correlation between the rankings of salary and age. This means that the older a person is, in the profession, it doesn't mean they will earn more.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 10**

**Question:**

A machine hire company kept records of the age, $x$ months, and the maintenance costs, £$y$, of one type of machine. The following table summarises the data for a random sample of 10 machines.

| Machine | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| Age, $x$ | 63 | 12 | 34 | 81 | 51 | 14 | 45 | 74 | 24 | 89 |
| Maintenance costs, $y$ | 111 | 25 | 41 | 181 | 64 | 21 | 51 | 145 | 43 | 241 |

**a** Calculate, to 3 decimal places, the product-moment correlation coefficient. (You may use $\sum x^2 = 30\,625, \sum y^2 = 135\,481, \sum xy = 62\,412$.)

**b** Calculate, to 3 decimal places, the Spearman's rank correlation coefficient.

**c** For a different type of machine similar data were collected. From a large population of such machines a random sample of 10 was taken and the Spearman's rank correlation coefficient, based on $\sum d^2 = 36$, was 0.782.

Using a 5% level of significance and quoting from the tables of critical values, interpret this rank correlation coefficient. Use a two-tailed test and state clearly your null and alternative hypotheses. ***E***

**Solution:**

**a** $r = 0.937$

Get this directly from your calculator in the examination. There is no need to set up a table of values; you will run out of time if you do.

**b**

| $x$ | 63 | 12 | 34 | 81 | 51 | 14 | 45 | 74 | 24 | 89 |
|---|---|---|---|---|---|---|---|---|---|---|
| $y$ | 111 | 25 | 41 | 181 | 64 | 21 | 51 | 145 | 43 | 241 |
| $r_x$ | 4 | 10 | 7 | 2 | 5 | 9 | 6 | 3 | 8 | 1 |
| $r_y$ | 4 | 9 | 8 | 2 | 5 | 10 | 6 | 3 | 7 | 1 |
| $d$ | 0 | 1 | −1 | 0 | 0 | −1 | 0 | 0 | 1 | 0 |
| $d^2$ | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |

$\sum d^2 = 4$

$$r_s = 1 - \frac{6 \times 4}{10(10^2 - 1)}$$

$r_s = 0.976$ (3 d.p.)

**c** $H_0: \rho_s = 0$ There is no association between age and maintenance costs.
$H_1: \rho_s \neq 0$ There is an association between age and maintenance costs.

$\Big\}$ 2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.782$
$n = 10$, critical value(s) $= \pm 0.6485$
In upper tail, t.s. > c.v. since $0.782 > +0.6485$ so reject $H_0$.

Conclude there is evidence of an association between age and maintenance costs.
It appears that older machines cost more to maintain.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**

**Exercise D, Question 11**

**Question:**

The data below show the height above sea level, $x$ metres, and the temperature, $y°$ C, at 7.00 a.m., on the same day in summer at nine places in Europe.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Height, $x$(m)** | 1400 | 400 | 280 | 790 | 390 | 590 | 540 | 1250 | 680 |
| **Temperature, $y$ (°C)** | 6 | 15 | 18 | 10 | 16 | 14 | 13 | 7 | 13 |

**a** The product-moment correlation coefficient is $-0.975$. Test this at the 5% significance level. Interpret your result in context.

On the same day the number of hours of sunshine was recorded and Spearman's rank correlation between hours of sunshine and temperature, based on $\sum d^2 = 28$, was 0.767.

**b** Stating your hypotheses and using a 5% two-tailed test, interpret this rank correlation coefficient. ***E***

**Solution:**

**a** $H_0 : \rho = 0$
$H_1 : \rho < 0$ 1-tail $\alpha = 0.05$

Test statistic $= -0.975$

$n = 9$ critical value $= -0.5822$

Lower tail test, t.s. < c.v. since $-0.975 < -0.5822$ reject $H_0$.

Conclude there is evidence of negative correlation. There is evidence that the greater the height above sea level, the lower the temperature at 7.00 a.m. is likely to be.

**b** $H_0 :\ \rho_s = 0$ There is no association between hours of sunshine and temperature.
$H_1 :\ \rho_s \neq 0$ There is an association between hours of sunshine and temperature. 2-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.767$

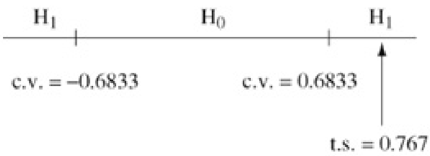
$n = 9$ critical value $= \pm 0.6833$

In upper tail

t.s. > c.v. since

$0.767 > +0.6833$

so reject $H_0$.




Conclude there is evidence of an association between hours of sunshine and temperature.

The more hours of sunshine the warmer the temperature.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

### Regression and correlation

**Exercise D, Question 12**

**Question:**

a Explain briefly, referring to your project work if you wish, the conditions under which you would measure association by using a rank correlation coefficient rather than a product-moment coefficient.

b At an agricultural show 10 Shetland sheep were ranked by a qualified judge and by a trainee judge. Their rankings are shown in the table.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Qualified judge** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Trainee judge** | 1 | 2 | 5 | 6 | 7 | 8 | 10 | 4 | 3 | 9 |

Calculate a rank correlation coefficient for these data.

c Using a suitable table and a 5% significance level, state your conclusions as to whether there is some degree of agreement between the two sets of ranks. ***E***

**Solution:**

a You use a rank correlation coefficient if at least one of the sets of data isn't from a normal distribution, or if at least one of the sets of data is a letter grading or an order of preference. It is also used if there is a non-linear association between the variables.

b

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $r_Q$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| $r_T$ | 1 | 2 | 5 | 6 | 7 | 8 | 10 | 4 | 3 | 9 |
| $d$ | 0 | 0 | −2 | −2 | −2 | −2 | −3 | 4 | 6 | 1 |
| $d^2$ | 0 | 0 | 4 | 4 | 4 | 4 | 9 | 16 | 36 | 1 |

$$\sum d^2 = 78$$

$$r_s = 1 - \frac{6 \times 78}{10(10^2 - 1)}$$

$$r_s = 0.527$$

c $\left.\begin{array}{l}\text{H}_0: \rho_s = 0 \text{ There is no correlation between the rankings of the qualified and trainee judges.} \\ \text{H}_1: \rho_s > 0 \text{ There is positive correlation between the rankings of the qualified and trainee judges.}\end{array}\right\}$ 1-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.527$

critical value $= 0.5636$

t.s. < c.v. since $0.527 < 0.5636$. These two values are very close.

Accept $\text{H}_0$.

Conclude there is insufficient evidence of agreement between the rankings awarded by the qualified and trainee judges at the 5% level of significance.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 13**

**Question:**

a Explain briefly the use of a null hypothesis and a level of significance in statistical work.

b The positions in a league table of 8 rugby clubs at the end of a season are shown, together with the average attendance (in hundreds) at home matches during the season.

| Club | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Average attendance | 30 | 32 | 12 | 19 | 27 | 18 | 15 | 25 |

Calculate the coefficient of rank correlation between position in the league and home attendance. Comment on your results. ***E***

**Solution:**

a The null hypothesis is what is assumed to be true unless proved otherwise. (The alternative hypothesis tells you what is likely to be happening if the null hypothesis is rejected.) The level of significance tells you the probability of rejecting the null hypothesis if it is true. The null hypothesis is only rejected in favour of the alternative hypothesis if by doing so the probability of being wrong is less than or equal to the significance level.

b

| Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Attendance (hundreds) | 30 | 32 | 12 | 19 | 27 | 18 | 15 | 25 |
| $r_p$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| $r_a$ | 2 | 1 | 8 | 5 | 3 | 6 | 7 | 4 |

Keep these rows together in the table to make it easy to find the $d$ values.

| $d$ | −1 | 1 | −5 | −1 | 2 | 0 | 0 | 4 |
|---|---|---|---|---|---|---|---|---|
| $d^2$ | 1 | 1 | 25 | 1 | 4 | 0 | 0 | 16 |

$$\sum d^2 = 48$$

$$r_s = 1 - \frac{6 \times 48}{8(8^2 - 1)}$$

$$r_s = 0.429$$

There is weak positive correlation between the ranking of attendance and position in the league. Being higher in the league doesn't necessarily mean the attendance will be higher – it may help though.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Regression and correlation**
**Exercise D, Question 14**

**Question:**

The ages, in months, and the weights, in kg, of a random sample of nine babies are shown in the table below.

| Baby | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Age ($x$) | 1 | 2 | 2 | 3 | 3 | 3 | 4 | 4 | 5 |
| Weight ($y$) | 4.4 | 5.2 | 5.8 | 6.4 | 6.7 | 7.2 | 7.6 | 7.9 | 8.4 |

**a** The product-moment correlation coefficient between weight and age for these babies was found to be 0.972. By testing for positive correlation at the 5% significance level interpret this value.

**b** A boy who does not know the weights or ages of these babies is asked to list them, by guesswork, in order of increasing weight. He puts them in the order

A C E B G D I F H

Obtain, to 3 decimal places, a rank correlation coefficient between the boy's order and the true weight order.

**c** Referring to the tables and using a 5% significance level, investigate for any agreement between the boy's order and the weight order. Discuss any conclusions you draw from you results.

**Solution:**

**a** $\left.\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho > 0 \end{aligned}\right\}$ 1-tail $\alpha = 0.05$

Test statistic $= r = 0.972$

$n = 9$ critical value $= 0.5822$

*upper tail* t.s. > c.v. since $0.972 > 0.5822$

so reject $H_0$.

Conclude there is evidence of a positive association between age and weight. This means the older a baby is, the heavier it is likely to be.

**b**

| Weight | 4.4 | 5.2 | 5.8 | 6.4 | 6.7 | 7.2 | 7.6 | 7.9 | 8.4 |
|---|---|---|---|---|---|---|---|---|---|
| Boy's order | A | C | E | B | G | D | I | F | H |
| $r_w$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $r_b$ | 1 | 3 | 5 | 2 | 7 | 4 | 9 | 6 | 8 |
| $d$ | 0 | −1 | −2 | 2 | −2 | 2 | −2 | 2 | 1 |
| $d^2$ | 0 | 1 | 4 | 4 | 4 | 4 | 4 | 4 | 1 |

$\sum d^2 = 26$

$$r_s = 1 - \frac{6 \times 26}{9(9^2 - 1)}$$

$r_s = 0.783$ (3 d.p.)

**c** $H_0$: $\rho_s = 0$ There is no association between the boy's order and the true weight order.
$H_1$: $\rho_s > 0$ There is positive association between the boy's order and the true weight order.
1-tail $\alpha = 0.05$

You should always put your hypotheses in terms of $\rho$ first.

Test statistic $= r_s = 0.783$

critical value $= 0.6000$

upper tail test where t.s. > c.v. since $0.783 > 0.6000$ so reject $H_0$.

Conclude there is evidence of positive association between the boy's order and the actual weight of the babies.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 1**

**Question:**

The time, in minutes, it takes Robert to complete the puzzle in his morning newspaper each day is normally distributed with mean 18 and standard deviation 3. After taking a holiday, Robert records the times taken to complete a random sample of 15 puzzles and he finds that the mean time is 16.5 minutes. You may assume that the holiday has not changed the standard deviation of times taken to complete the puzzle.
Stating your hypotheses clearly test, at the 5% level of significance, whether or not there has been a reduction in the mean time Robert takes to complete the puzzle. ***E***

**Solution:**

$H_0 : \mu = 18, H_1 : \mu < 18$

Remember to identify which is $H_0$ and which is $H_1$. This is a one-tail test since we are only interested in whether the time taken to solve the puzzle has reduced. You must use the correct parameter $(\mu)$.

$$z = \frac{(16.5 - 18)}{\left(\frac{3}{\sqrt{15}}\right)} = -1.9364\ldots$$

Using $z = \dfrac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$.

5% one tail c.v. is $z = -1.6449$

Use the percentage point table and quote the figure in full.

significant *or* reject $H_0$ *or* in critical region.

$-1.9364 < -1.6449$

There is evidence that the (mean) time to complete the puzzles has reduced

*or* Robert is getting faster (at doing the puzzles)

State your conclusion in the context of the question.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 2**

**Question:**

In a trial of diet $A$ a random sample of 80 participants were asked to record their weight loss, $x$ kg, after their first week of using the diet. The results are summarised by

$$\sum x = 361.6 \quad \text{and} \quad \sum x^2 = 1753.95.$$

**a** Find unbiased estimates for the mean and variance of weight lost after the first week of using diet $A$.

The designers of diet $A$ believe it can achieve a greater mean weight loss after the first week than a standard diet $B$. A random sample of 60 people used diet $B$. After the first week they had achieved a mean weight loss of 4.06 kg, with an unbiased estimate of variance of weight loss of $2.50\,\text{kg}^2$.

**b** Test, at the 5% level of significance, whether or not the mean weight loss after the first week using diet $A$ is greater than that using diet $B$. State your hypotheses clearly.

**c** Explain the significance of the Central Limit Theorem to the test in part **b**.

**d** State an assumption you have made in carrying out the test in part **b**. ***E***

**Solution:**

**a** $\bar{x} = \frac{361.6}{80} = 4.52$

$\hat{\sigma}^2 = s^2 = \frac{1753.95 - 80 \times \bar{x}^2}{79} = 1.5128$

or $\hat{\sigma}^2 = s^2 = \frac{80}{79} \times \left( \frac{1753.95}{80} - \bar{x}^2 \right) = 1.5128$

Using $\frac{\sum x^2 - n\bar{x}^2}{n-1}$ or $\frac{n}{n-1}\left( \frac{\sum x^2}{n} - \bar{x}^2 \right)$.

**b** $H_0: \mu_A = \mu_B \quad H_1: \mu_A > \mu_B$

This is a difference of means test. When stating hypotheses you must make it clear which mean is greater when it is one-tailed test.

$z = \frac{4.52 - 4.06}{\sqrt{\frac{1.5128}{80} + \frac{2.50}{60}}} = \left( \frac{0.46}{\sqrt{0.060576}} \right)$

$= 1.8689$ or $-1.8689$ if $B - A$ was used.

Using $z = \frac{(\bar{A} - \bar{B}) - (\mu_A - \mu_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$.

One tail c.v. is $z = 1.6449$

$1.87 > 1.6449$ so reject $H_0$.

Use the percentage point table and quote the figure in full.

There is evidence that diet $A$ is better than diet $B$ *or* evidence that (mean) weight lost in first week using diet $A$ is more than with $B$.

State your conclusion in the context of the question.

**c** CLT enables you to assume that $\bar{A}$ and $\bar{B}$ are normally distributed since both samples are large.

**d** Assumed $\sigma_A^2 = s_A^2$ and $\sigma_B^2 = s_B^2$

Variance must be known to use the test. Remember $\sigma^2$ is the population variance and $s^2$ is an unbiased estimator of the population variance.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 3**

**Question:**

A random sample of the daily sales (in £s) of a small company is taken and, using tables of the normal distribution, a 99% confidence interval for the mean daily sales is found to be

(123.5,154.7).

Find a 95% confidence interval for the mean daily sales of the company. ***E***

**Solution:**

$$123.5 = \bar{x} - 2.5758 \times \frac{\sigma}{\sqrt{n}} \quad ①$$

$$154.7 = \bar{x} + 2.5758 \times \frac{\sigma}{\sqrt{n}} \quad ②$$

99% confidence interval so each tail is 0.05. Use the percentage point table and quote the figure in full. C.I. $\bar{x} \pm 2.5758 \times \frac{\sigma}{\sqrt{n}}$

$$\bar{x} = \frac{1}{2}(123.5 + 154.7) = 139.1$$

Add equations ①and ② to find $\bar{x}$ or calculate the mean of the given limits.

$$2.5758 \times \frac{\sigma}{\sqrt{n}} = 154.7 - 139.1$$

$$= 15.6$$

$$\frac{\sigma}{\sqrt{n}} = \frac{15.6}{2.5758}$$

Substitute $\bar{x}$ into equation ① or ② to find $\frac{\sigma}{\sqrt{n}}$.

$$\text{So 95\% C.I.} = 139.1 \pm 1.9600 \times \frac{15.6}{2.5758}$$

$$= (127.22\ldots, 150.97\ldots)$$

95% confidence interval so each tail is 0.025. Substitute in $\bar{x}$ and $\frac{\sigma}{\sqrt{n}}$.

$$= (127, 151)$$

Answers should be given to at least 3 significant figures.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 4**

**Question:**

A set of scaffolding poles come in two sizes, long and short. The length $L$ of a long pole has the normal distribution $N(19.7, 0.5^2)$. The length $S$ of a short pole has the normal distribution $N(4.9, 0.2^2)$. The random variables $L$ and $S$ are independent. A long pole and a short pole are selected at random.

**a** Find the probability that the length of the long pole is more than 4 times the length of the short pole.

Four short poles are selected at random and placed end to end in a row. The random variable $T$ represents the length of the row.

**b** Find the distribution of $T$.

**c** Find $P(|L-T| < 0.1)$. ***E***

**Solution:**

**a** Let $X = L - 4S$ then
$E(X) = 19.7 - 4 \times 4.9 = 0.1$

Use $E(aX - bY) = aE(X) - bE(X)$.

$$\begin{aligned}\text{Var}(X) &= \text{Var}(L) + 4^2\text{Var}(S)\\ &= 0.5^2 + 16 \times 0.2^2\\ &= 0.89\end{aligned}$$

Use
$\text{Var}(aX - bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$.

$$\begin{aligned}P(X > 0) &= P\left(Z > \frac{0 - 0.1}{\sqrt{0.89}}\right)\\ &= P(Z > -0.10599\ldots)\\ &= (0.542 \text{ to } 0.544)\end{aligned}$$

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

Look $z$ up in the tables or use a graphical calculator.

This is the accepted range of values for the answer.

**b** $T = S_1 + S_2 + S_3 + S_4$

$$\begin{aligned}E(T) &= 4 \times 4.9\\ &= 19.6\\ \text{Var}(T) &= 4 \times 0.2^2\\ &= 0.16 \text{ or } 0.4^2\end{aligned}$$

Since we are adding we use
$E(X_1 + X_2) = E(X_1) + E(X_2)$ and
$\text{Var}(X_1 + X_2) = \text{Var}(X_1) + \text{Var}(X_2)$.

**c** Let $Y = L - T$

$$\begin{aligned}E(Y) &= E(L) - E(T)\\ &= 0.1\end{aligned}$$

Using $E(X - Y) = E(X) - E(Y)$.

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(L) + \text{Var}(T)\\ &= 0.5^2 + 0.4^2\\ &= 0.41\end{aligned}$$

Using $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$.

$$\begin{aligned}P(-0.1 < Y < 0.1) &= P\left(Z < \frac{0.1 - 0.1}{\sqrt{0.41}}\right) - P\left(Z < \frac{0.1 - 0.1}{\sqrt{0.41}}\right)\\ &= P(Z < 0) - P(Z < -0.31)\\ &= 0.5 - (1 - 0.6217)\\ &= 0.1217 \text{ (tables)}\\ &\quad \text{or } 0.1226 \ldots \text{ (calc)}\end{aligned}$$

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 5**

**Question:**

Describe one advantage and one disadvantage of

**a** quota sampling,

**b** simple random sampling. ***E***

**Solution:**

**a** *Advantages*: Any one of

- does not require the existence of:
  a sampling frame
  a population list
- field work can be done quickly as representative sample can be achieved with a small sample size
- costs kept to a minimum (cheaply)
- administration relatively easy
- non-response not an issue

This is a book work. You need to learn the advantages and disadvantages of all the sampling methods.

*Disadvantages*: Any one of

- not possible to estimate sampling errors
- interviewer choice and may not be able to judge easily/may lead to bias
- non-response not recorded
- non-random process

**b** *Advantages*: any one of

- random process so possible to estimate sampling errors
- free from bias

*Disadvantages*: any one of

- not suitable when sample size is large
- sampling frame required which may not exist or may be difficult to construct for a large population

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 6**

**Question:**

A report on the health and nutrition of a population stated that the mean height of three-year-old children is 90 cm and the standard deviation is 5 cm. A sample of 100 three-year-old children was chosen from the population.

**a** Write down the approximate distribution of the sample mean height. Give a reason for your answer.

**b** Hence find the probability that the sample mean height is at least 91 cm. ***E***

**Solution:**

**a** $\bar{X} \sim \mathrm{N}\left(90, \frac{5^2}{100}\right)$ i.e. N (90, 0.25)

Application of Central Limit Theorem (as sample large)

**b** $\mathrm{P}(\bar{X} \geq 91) = 1 - \mathrm{P}\left(Z < \frac{91-90}{0.25}\right)$

$= 1 - \mathrm{P}(Z < 2)$

$= 1 - 0.9772$

$= 0.0228$

Using $z = \frac{x-\mu}{\frac{\sigma}{\sqrt{n}}}$.

You do not interpolate so round your $z$ value to 2 decimal places and use the tables or use a graphical calculator.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 7**

**Question:**

A machine produces metal containers. The weights of the containers are normally distributed. A random sample of 10 containers from the production line was weighed, to the nearest 0.1 kg, and gave the following results

49.7, 50.3, 51.0, 49.5, 49.9
50.1, 50.2, 50.0, 49.6, 49.7.

**a** Find unbiased estimates of the mean and variance of the weights of the population of metal containers.

The machine is set to produce metal containers whose weights have a population standard deviation of 0.5 kg.

**b** Estimate the limits between which 95% of the weights of metal containers lie.

**c** Determine the 99% confidence interval for the mean weight of metal containers. ***E***

**Solution:**

**a** $\bar{X} = \frac{500}{10} = 50$

$s^2 = \frac{25\,001.74 - 10 \times 50^2}{9}$

$= 0.193$

Using $\frac{\sum x^2 - n\bar{x}^2}{n-1}$ or $\frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right)$.

**b** Limits are $50 \pm 1.96 \times 0.5$
$= (49.02, 50.98)$

**c** Confidence interval is

$(50 - 2.5758 \times \frac{0.5}{\sqrt{10}}, 50 + 2.5758 \times \frac{0.5}{\sqrt{10}})$
$= (49.593, 50.407)$
$= (49.6, 50.4)$

99% confidence interval so each tail is 0.005. Use the percentage point table and quote the figure in full.
C.I.: $\bar{x} \pm 2.5758 \times \frac{\sigma}{\sqrt{n}}$

Final answers should be given to at least 3 significant figures.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 8**

**Question:**

A school has 15 classes and a sixth form. In each class there are 30 students. In the sixth form there are 150 students. There are equal numbers of boys and girls in each class. There are equal numbers of boys and girls in the sixth form. The headteacher wishes to obtain the opinions of the students about school uniforms.
Explain how the headteacher would take a stratified sample of size 40. ***E***

**Solution:**

Total in school $= (15 \times 30) + 150 = 600$

Need a random sample of $\frac{30}{600} \times 40$

$= 2$ from each of the 15 classes

and a random sample of $\frac{150}{600} \times 40$

$= 10$ from sixth form;

To describe a stratified sample your need to
1) work out how many people to take from each strata,
2) explain how to collect the sample from each strata.

Label the boys in each class from 1–15 and the girls from 1–15.
Use random numbers to select 1 girl and 1 boy from each class.
Label the boys in the sixth form from 1–75 and the girls from 1–75.
Use random numbers to select 5 different boys and 5 different girls from the sixth form.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 9**

**Question:**

A workshop makes two types of electrical resistor.
The resistance, $X$ ohms, of resistors of Type A is such that $X \sim N(20, 4)$.
The resistance, $Y$ ohms, of resistors of Type B is such that $Y \sim N(10, 0.84)$.
When a resistor of each type is connected into a circuit, the resistance $R$ ohms of the circuit is given by $R = X + Y$ where $X$ and $Y$ are independent.
Find

**a** $E(R)$,

**b** $Var(R)$,

**c** $P(28.9 < R < 32.64)$. ***E***

**Solution:**

**a** $E(R) = 20 + 10 = 30$

Using $E(X+Y) = E(X) + E(Y)$.

**b** $Var(R) = 4 + 0.84$
$= 4.84$

Using $Var(X+Y) = Var(X) + Var(Y)$.

**c** $R \sim N(30, 4.84)$

$P(28.9 < R < 32.64) = P(R < 32.64) - P(R < 28.9)$

$$= P\left(Z < \frac{32.64 - 30}{2.2}\right) - P\left(Z < \frac{28.9 - 30}{2.2}\right)$$

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

$= P(Z < 1.2) - P(Z < -0.5)$
$= 0.8849 - (1 - 0.6915)$
$= 0.8849 - 0.3085$
$= 0.5764$

Look $z$ up in the tables or use a graphical calculator.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 10**

**Question:**

The drying times of paint can be assumed to be normally distributed. A paint manufacturer paints 10 test areas with a new paint. The following drying times, to the nearest minute, were recorded.

82, 98, 140, 110, 90, 125, 150, 130, 70, 110.

**a** Calculate unbiased estimates for the mean and the variance of the population of drying times of this paint.

Given that the population standard deviation is 25,

**b** find a 95% confidence interval for the mean drying time of this paint.

Fifteen similar sets of tests are done and the 95% confidence interval is determined for each set.

**c** Estimate the expected number of these 15 intervals that will enclose the true value of the population mean $\mu$. ***E***

**Solution:**

**a** $\hat{\mu} = \bar{x} = \dfrac{82+98+140+110+90+125+150+130+70+110}{10}$

$= 110.5$

$s^2 = \dfrac{128\,153 - 10 \times 110.5^2}{9}$

$= 672.28$

Using $\dfrac{\sum x^2 - n\bar{x}^2}{n-1}$ or $\dfrac{n}{n-1}\left(\dfrac{\sum x^2}{n} - \bar{x}^2\right)$.

**b** 95% confidence limits are

$110.5 \pm 1.96 \times \dfrac{25}{\sqrt{10}}$

$= (95.005, 125.995)$

$= (95.0, 126)$

95% confidence interval so each tail is 0.025. Use the percentage point table and quote the figure in full.

C.I.: $\bar{x} \pm 1.9600 \times \dfrac{\sigma}{\sqrt{n}}$

Answers should be given to at least 3 significant figures.

**c** Number of intervals $= \dfrac{95}{100} \times 15 = 14.25$

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 11**

**Question:**

Some biologists were studying a large group of wading birds. A random sample of 36 were measured and the wing length, $x$ mm, of each wading bird was recorded. The results are summarised as follows.

$\sum x = 6046, \sum x^2 = 1\ 06\ 338.$

**a** Calculate unbiased estimates of the mean and the variance of the wing lengths of these birds.

Given that the standard deviation of the wing lengths of this particular type of bird is actually 5.1 mm,

**b** find a 99% confidence interval for the mean wing length of the birds from this group. ***E***

**Solution:**

**a** $\bar{x} = \left(\frac{6046}{36} = \right) 167.94\ldots$

$s^2 = \frac{1\ 016\ 338 - 36 \times \bar{x}^2}{35}$

$= 27.0$

Using $\frac{\sum x^2 - n\bar{x}^2}{n-1}$ or $\frac{n}{n-1}\left(\frac{\sum x^2}{n} - \bar{x}^2\right)$.

**b** 99% confidence interval is: $\bar{x} \pm 2.5758 \times \frac{5.1}{\sqrt{36}}$

99% confidence interval so each tail is 0.005. Use the percentage point table and quote the figure in full.

C.I.: $\bar{x} \pm 2.5758 \times \frac{\sigma}{\sqrt{n}}$

$= 167.94 \pm 2.5758 \times \frac{5.1}{\sqrt{36}}$

$= (165.75, 170.13)$

$= (166, 170)$

Answers should be given to at least 3 significant figures.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 12**

**Question:**

The weights of adult men are normally distributed with a mean of 84 kg and a standard deviation of 11 kg.

**a** Find the probability that the total weight of 4 randomly chosen adult men is less than 350 kg.

The weights of adult women are normally distributed with a mean of 62 kg and a standard deviation of 10 kg.

**b** Find the probability that the weight of a randomly chosen adult man is less than one and a half times the weight of a randomly chosen adult woman. ***E***

**Solution:**

**a** $X = M_1 + M_2 + M_3 + M_4$

$\mathrm{E}(X) = 4 \times 84$
$= 336$

Since we are adding we use $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$ and $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$.

$\mathrm{Var}(X) = 4 \times 11^2$
$= 484 \text{ or } 22^2$

$X \sim \mathrm{N}(336, 22^2)$

$\mathrm{P}(X < 350) = \mathrm{P}\left(Z < \frac{350 - 336}{22}\right)$

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

$= \mathrm{P}(Z < 0.64)$
$= 0.738 \text{ or } 0.739$

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

**b** $M \sim \mathrm{N}(84, 121)$ and $W \sim \mathrm{N}(62, 100)$

Let $Y = M - 1.5W$

$\mathrm{E}(Y) = 84 - 1.5 \times 62 = -9$

Using $\mathrm{E}(X_1 - bX_2) = \mathrm{E}(X_1) - b\mathrm{E}(X_2)$ and $\mathrm{Var}(X_1 - bX_2) = \mathrm{Var}(X_1) + b^2\mathrm{Var}(X_2)$.

$\mathrm{Var}(Y) = \mathrm{Var}(M) + 1.5^2\mathrm{Var}(W)$
$= 11^2 + 1.5^2 \times 10^2$
$= 346$

$\mathrm{P}(Y < 0) = \mathrm{P}\left(Z < \frac{0 - (-9)}{\sqrt{346}}\right)$

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

$= \mathrm{P}(Z < 0.48)$
$= 0.684 \sim 0.686$

You do not interpolate so round your $z$ value to 2 decimal places and use the tables or use a graphical calculator.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 13**

**Question:**

A researcher is hired by a cleaning company to survey the opinions of employees on a proposed pension scheme. The company employs 55 managers and 495 cleaners.
To collect data the researcher decides to give a questionnaire to the first 50 cleaners to leave at the end of the day.

**a** Give 2 reasons why this method is likely to produce biased results.

**b** Explain briefly how the researcher could select a sample of 50 employees using

   **i** a systematic sample,

   **ii** a stratified sample.

Using the random number tables in the formulae book, and starting with the top left hand corner (8) and working across, 50 random numbers between 1 and 550 inclusive were selected. The first two suitable numbers are 384 and 100.

**c** Find the next two suitable numbers. ***E***

**Solution:**

**a** Only cleaners – no managers i.e. not all types.
*or* not a random sample; 1st 50 may be in same shift/group/share same views.

**b** **i** Label employees (1–550) or obtain an ordered list.
Select first using random numbers (from 1–11).
Then select every 11th person from the list
e.g. if person 8 is selected then the sample is 8, 19, 30, 41, ... 547

**ii** Label managers (1–55) and cleaners (1–495)
Use random numbers to select...

**iii** ...5 managers and 45 cleaners

**c** 390, 372 (They must be in this order.)

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 14**

**Question:**

A sociologist is studying how much junk food teenagers eat. A random sample of 100 female teenagers and an independent random sample of 200 male teenagers were asked to estimate what their weekly expenditure on junk food was. The results are summarised below.

| | $n$ | mean | s.d. |
|---|---|---|---|
| **Female teenagers** | 100 | £5.48 | £3.62 |
| **Male teenagers** | 200 | £6.86 | £4.51 |

**a** Using a 5% significance level, test whether or not there is a difference in the mean amounts spent on junk food by male teenagers and female teenagers. State your hypotheses clearly.

**b** Explain briefly the importance of the Central Limit Theorem in this problem. ***E***

**Solution:**

**a** $H_0: \mu_F = \mu_M \quad H_1: \mu_F \neq \mu_M$

This is a difference of means test.

$$z = \frac{6.86 - 5.48}{\sqrt{\frac{4.51^2}{200} + \frac{3.62^2}{100}}}$$
$$= 2.860\ldots$$

Using $z = \dfrac{(\bar{A} - \bar{B}) - (\mu_A - \mu_B)}{\sqrt{\dfrac{\sigma_A^2}{n_A} - \dfrac{\sigma_B^2}{n_B}}}$.

2-tail 5% critical value $(\pm)$ 1.96

2.860 > +1.900

Use the percentage point table and quote the figure in full.

Significant result or reject the null hypothesis.
There is evidence of a difference in the (mean) amount spent on junk food by male and female teenagers.

State your conclusion in the context of the question.

**b** CLT enables us to assume $\bar{F}$ and $\bar{M}$ are normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 15**

**Question:**

**a** State two reasons why stratified sampling might be chosen as a method of sampling when carrying out a statistical survey.
**b** State one advantage and one disadvantage of quota sampling. ***E***

**Solution:**

**a** Population divides into mutually exclusive/distinct groups/strata.
Its results will best reflect those of the population since the sample structure reflects that of the population.

**b** *Advantages*: Any one of
- enables fieldwork to be done quickly
- costs kept to a minimum
- administration is relatively easy

*Disadvantages*: Any one of
- non-random so not possible to estimate sampling errors
- subject to possible interviewer bias
- non-response not recorded

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 16**

**Question:**

A sample of size 5 is taken from a population that is normally distributed with mean 10 and standard deviation 3. Find the probability that the sample mean lies between 7 and 10. ***E***

**Solution:**

$X \sim N(10, 3^2) \therefore \bar{X} \sim N(10, \frac{9}{5})$

$$\begin{aligned} P(7 \leq \bar{X} \leq 10) &= P\left(\frac{7-10}{\sqrt{\frac{9}{5}}} < Z < 0\right) \\ &= P(-2.236 < Z < 0) \\ &= \Phi(0) - \{1 - \Phi(2.24)\} \\ &= 0.4875 \end{aligned}$$

Using $z = \dfrac{x - \mu}{\frac{\sigma}{\sqrt{n}}}$

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 17**

**Question:**

A computer company repairs large numbers of PCs and wants to estimate the mean time to repair a particular fault. Five repairs are chosen at random from the company's records and the times taken, in seconds, are

205 310 405 195 320.

**a** Calculate unbiased estimates of the mean and the variance of the population of repair times from which this sample has been taken.

It is known from previous results that the standard deviation of the repair time for this fault is 100 seconds. The company manager wants to ensure that there is a probability of at least 0.95 that the estimate of the population mean lies within 20 seconds of its true value.

**b** Find the minimum sample size required. ***E***

**Solution:**

**a** Let $X$ represent repair time

$\therefore \sum x = 1435 \therefore \bar{x} = \dfrac{1435}{5} = 287$

Using $\dfrac{\sum x^2 - n\bar{x}^2}{n-1}$ or $\dfrac{n}{n-1}\left(\dfrac{\sum x^2}{n} - \bar{x}^2\right)$.

$\sum x^2 = 442\,575$

$\therefore s^2 = \dfrac{442\,575 - 5 \times 287^2}{4}$

$= 7682.5$

**b** $\mathrm{P}(|\mu - \hat{\mu}|) < 20 = 0.95$

The repair time is between 80 and 120. 95% confidence interval so each tail is 0.025. Use the percentage point table and quote the figure in full.
C.I.: $\bar{x} \pm 1.96 \times \dfrac{\sigma}{\sqrt{n}}$.

$\therefore 1.96 \times \dfrac{\sigma}{\sqrt{n}} = 20$

$\therefore n = \dfrac{1.96^2\sigma^2}{20^2} = \dfrac{1.96^2\sigma^2}{400} = \dfrac{1.96^2 \times 100^2}{400} = 96.04$

$\therefore$ Sample size $(\geq)$ 97 required

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 18**

**Question:**

The random variable $D$ is defined as

$$D = A - 3B + 4C$$

where $A \sim \mathrm{N}(5, 2^2)$, $B \sim \mathrm{N}(7, 3^2)$ and $C \sim \mathrm{N}(9, 4^2)$, and $A$, $B$ and $C$ are independent.

**a** Find $\mathrm{P}(D < 44)$.

The random variables $B_1$, $B_2$ and $B_3$ are independent and each has the same distribution as $B$. The random variable $X$ is defined as

$$X = A - \sum_{i=1}^{3} B_i + 4C.$$

**b** Find $\mathrm{P}(X > 0)$. ***E***

**Solution:**

**a** $\mathrm{E}(D) = \mathrm{E}(A) - 3\mathrm{E}(B) + 4\mathrm{E}(C)$

$= 5 - 3 \times 7 + 4 \times 9$

$= 20$

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

$\mathrm{Var}(D) = \mathrm{Var}(A) + 9\mathrm{Var}(B) + 16\mathrm{Var}(C)$

$= 2^2 + 9 \times 3^2 + 16 \times 4^2$

$= 341$

Using $\mathrm{Var}(aX \pm bY) = a^2\mathrm{Var}(X) \pm b^2\mathrm{Var}(Y)$.

$\mathrm{P}(D < 44) = \mathrm{P}\left(z < \dfrac{44 - 20}{\sqrt{341}}\right)$

Using $z = \dfrac{x - \mu}{\sigma}$.

$= \mathrm{P}(z < 1.30)$

$= 0.9032$

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

**b** $\mathrm{E}(X) = 20$

$\mathrm{Var}(X) = \mathrm{Var}(A) + 3\mathrm{Var}(B) + 16\,\mathrm{Var}(C)$

$= 2^2 + 3 \times 3^2 + 16 \times 4^2$

$C = 287$

$X = A - \sum_{i=1}^{3} B_i + 4C$
$= A - (B_1 + B_2 + B_3) + 4C$

$\mathrm{P}(X > 0) = \mathrm{P}\left(z > \dfrac{-20}{\sqrt{287}}\right)$

Using $z = \dfrac{x - \mu}{\sigma}$.

$= \mathrm{P}(z > -1.18)$

$= 0.8810$

$= 0.881$ (3 s.f.)

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 1**
**Exercise A, Question 19**

**Question:**

A manufacturer produces two flavours of soft drink, cola and lemonade. The weights, $C$ and $L$, in grams, of randomly selected cola and lemonade cans are such that $C \sim \mathrm{N}(350, 8)$ and $L \sim \mathrm{N}(345, 17)$.

**a** Find the probability that the weights of two randomly selected cans of cola will differ by more than 6 g.

One can of each flavour is selected at random.

**b** Find the probability that the can of cola weighs more than the can of lemonade.

Cans are delivered to shops in boxes of 24 cans. The weights of empty boxes are normally distributed with mean 100 g and standard deviation 2 g.

**c** Find the probability that a full box of cola cans weighs between 8.51 kg and 8.52 kg.

**d** State an assumption you made in your calculation in part **c**. ***E***

**Solution:**

**a** Let $W = C_1 - C_2$

$\mathrm{E}(W) = 350 - 350 = 0$

$\mathrm{Var}(W) = 8 + 8$
$= 16$

$\therefore W \sim \mathrm{N}(0, 16)$

Using $\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y)$

Using $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

$\therefore \mathrm{P}(|W| > 6) = 2\mathrm{P}(W > 6)$

$= 2 \times \mathrm{P}\left(Z > \frac{6-0}{\sqrt{16}}\right)$

$= 2 \times \mathrm{P}(Z > 1.5)$

$= 2 \times (1 - 0.9332)$

$= 0.1336$

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

**b** Let $W = C - L$

$\mathrm{E}(W) = 350 - 345 = 5$

$\mathrm{Var}(W) = 8 + 17$
$= 25$

$\therefore W \sim \mathrm{N}(5, 25)$

$\mathrm{P}(W > 0) = \mathrm{P}\left(Z > \frac{-5}{\sqrt{25}}\right)$

$= \mathrm{P}(Z < -1)$

$= 0.8413$

Using $\mathrm{E}(X - Y) = \mathrm{E}(X) - \mathrm{E}(Y)$

Using $\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

If you have used $W = L - C$ then

$\mathrm{P}(W < 0) = \mathrm{P}\left(Z < \frac{(0-(-5))}{\sqrt{25}}\right)$

$= \mathrm{P}(Z < 1)$.

**c** Let $W = C_1 + \ldots + C_{24} + B$

A full box consists of 24 cans and a box.

$\therefore \mathrm{E}(W) = 24 \times 350 + 100 = 8500$

$\mathrm{Var}(W) = 24 \times 8 + 2^2 = 196$

Since we are adding we use $\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2)$ and $\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$

$\mathrm{P}(8510 \le W \le 8520) = \mathrm{P}\left(\frac{8510-8500}{\sqrt{196}} \le Z \le \frac{8520-8500}{\sqrt{196}}\right)$

$= \mathrm{P}(0.71\ldots \le Z \le 1.43\ldots)$

$= 0.9236 - 0.7611$

$= 0.1625$

(or 0.16096 directly from calculator)

Units must be consistent so change them all to grams and use $z = \frac{x-\mu}{\sigma}$.

You do not interpolate so round your $z$ value to 2 decimal places and use the tables, or use a graphical calculator.

**d** All random variables are independent and normally distributed.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 1**

**Question:**

During a village show, two judges, $P$ and $Q$, had to award a mark out of 30 to some flower displays. The marks they awarded to a random sample of 8 displays were as follows:

| Display | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| Judge P | 25 | 19 | 21 | 23 | 28 | 17 | 16 | 20 |
| Judge Q | 20 | 9 | 21 | 13 | 17 | 14 | 11 | 15 |

**a** Calculate Spearman's rank correlation coefficient for the marks awarded by the two judges.

After the show, one competitor complained about the judges. She claimed that there was no positive correlation between their marks.

**b** Stating your hypotheses clearly, test whether or not this sample provides support for the competitor's claim. Use a 5% level of significance. ***E***

**Solution:**

**a**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| $P$ Rank | 2 | 6 | 4 | 3 | 1 | 7 | 8 | 5 |
| $Q$ Rank | 2 | 8 | 1 | 6 | 3 | 5 | 7 | 4 |
| $d$ | 0 | −2 | 3 | −3 | −2 | 2 | 1 | 1 |
| $d^2$ | 0 | 4 | 9 | 9 | 4 | 4 | 1 | 1 |

Remember to rank the data. It does not matter whether you rank from highest to lowest or vice versa as long as you do the same for both judges.

The sum of your $d$'s should be zero. The $d^2$ should all be positive.

$\sum d^2 = 32$

$$r_s = 1 - \frac{6 \times 32}{8 \times (8^2 - 1)}$$

Using $r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$ from the formula book.

$$= \frac{13}{21} \text{ or } 0.619$$

If you give your answer as a decimal it should be given to 3 significant figures.

**b** $H_0 : \rho = 0 \quad H_1 : \rho > 0$

Make sure your hypotheses are clearly written using the symbol $\rho$. This is a one-tail test so only interested if positive i.e.

$r_s$ 1-tail 5% critical value is 0.6429

Look up the value under 0.05 in the table for Spearman's. Quote the figure in full.

$0.619 < 0.6429$ so accept $H_0$ or not significant.
So insufficient evidence of a positive correlation between judges
or
competitor's claim is justified.

Draw a conclusion in the context of the question.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**

**Exercise A, Question 2**

**Question:**

The Director of Studies at a large college believed that students' grades in Mathematics were independent of their grades in English. She examined the results of a random group of candidates who had studied both subjects and she recorded the number of candidates in each of the 6 categories shown.

| | Mathematics grade A or B | Mathematics grade C or D | Mathematics grade E or U |
|---|---|---|---|
| English grade A or B | 25 | 25 | 10 |
| English grade C to U | 5 | 30 | 15 |

a Stating your hypotheses clearly, test the Director's belief using a 10% level of significance. You must show each step of your working.

The Head of English suggested that the Director was losing accuracy by combining the English grades C to U in one row. He suggested that the Director should split the English grades into two rows, grades C or D and grades E or U as for Mathematics.

b State why this might lead to problems in performing the test. ***E***

**Solution:**

**a** $H_0$ : Mathematics grades are independent of English grades

*or*

no association between Mathematics grades and English grades.

$H_1$: Mathematics and English grades are dependent.

*or*

There is an association between Mathematics grades and English grades.

For a contingency table.
For $H_0$ you should use the words 'no association' or 'independent'.
For $H_1$ you should use the words 'is an association' or 'dependent'.

*Expected frequencies*

| | $M_{A,B}$ | $M_{C\text{ or }D}$ | $M_{E,U}$ |
|---|---|---|---|
| $E_{A,B}$ | 16.364 | 30 | 13.636 |
| $E_{C\text{ to }U}$ | 13.636 | 25 | 11.364 |

Expected frequency 'Maths A or B' and 'English A or B' $\dfrac{60\times30}{110}=16.364$

Show the working for at least one calculation of an expected value.

$$\text{Test statistic} = \sum\frac{(O_i-E_i)^2}{E_i}$$

$$=\frac{(25-16.364)^2}{16.364}+\frac{(25-30)^2}{25}+\cdots+\frac{(15-11.364)^2}{11.364}$$

t.s. $=13.994$

Degrees of freedom $=(2-1)\times(3-1)=2$

Critical value $=X_2^2\,(10\%)=4.605$

t.s. > c.v. since $13.994>4.605$

so reject $H_0$.

Conclude there is evidence of an association between Mathematics and English grades.

**b** May have some expected frequencies $<5$ (and hence need to pool rows/columns).

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 3**

**Question:**

A quality control manager regularly samples 20 items from a production line and records the number of defective items $x$. The results of 100 such samples are given in Table 1 below.

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 or more |
|---|---|---|---|---|---|---|---|---|
| **Frequency** | 17 | 31 | 19 | 14 | 9 | 7 | 3 | 0 |

**Table 1**

**a** Estimate the proportion of defective items from the production line.
The manager claimed that the number of defective items in a sample of 20 can be modelled by a binomial distribution. He used the answer in part **a** to calculate the expected frequencies given in Table 2.

| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 or more |
|---|---|---|---|---|---|---|---|---|
| **Expected frequency** | 12.2 | 27.0 | $r$ | 19.0 | $s$ | 3.2 | 0.9 | 0.2 |

**Table 2**

**b** Find the value of $r$ and the value of $s$ giving your answers to 1 decimal place.
**c** Stating your hypotheses clearly, use a 5% level of significance to test the manager's claim.
**d** Explain what the analysis in part **c** tells the manager about the occurrence of defective items from this production line. ***E***

**Solution:**

**a** mean number of defectives $= \dfrac{0\times17+1\times31+\cdots+6\times3}{17+31+\cdots+3} = \dfrac{200}{100} = 2$

This is 1 restriction in part c.

Total frequency $= 17+31+\cdots+3 = 100$

This is 1 restriction in part c.

In binomial distribution with $n = 20$
mean $= np$ so $2 = 20\times p$
$\therefore p = 0.1$

**b** $r = 100\times\binom{20}{2}(0.1)^2(0.9)^{18}$
$= 28.517$
$= 28.5$ (1 d.p.)
$s = 100 - 91 = 9.0$ (1 d.p.)

You could use the tables to work this out $100\times(0.6769-0.3917)$.

The total of the expected frequencies is the same as the total of the observed frequencies. Here it is 100.

**c** $H_0$: B(20, 0.1) is a good/suitable model/fit
$H_1$: B(20, 0.1) is *not* a suitable model

| $x$ | 0 | 1 | 2 | 3 | $\geq 4$ |
|---|---|---|---|---|---|
| $O_i$ | 17 | 31 | 19 | 14 | 19 |
| $E_i$ | 12.2 | 27.0 | 28.5 | 19.0 | 13.3 |
| $\frac{(O-E)^2}{E}$ | 1.889 | 0.593 | 3.167 | 1.316 | 2.443 |

The classes for 4, 5, 6 and 7 or more have been combined. This is so that all the expected frequencies are greater than 5.

Test statistic $= \sum\dfrac{(O-E)^2}{E} = 9.41$

or

$$\sum\frac{O_i^2}{E_i} - N = \frac{17^2}{12.2} + \frac{31^2}{27} + \ldots + \frac{19^2}{13.3} - 100$$
$$= 9.41$$

It is often easier to use the formula $\sum\dfrac{O_i^2}{E_i} - N$.

$\nu = 5 - 2 = 3$

Degrees of freedom = (number of cells after pooling) − (restrictions)

critical value $= \chi_3^2\,(5\%) = 7.815$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

t.s. > c.v. so reject $H_0$.
(significant result) binomial distribution is not a suitable model

**d** Defective items do not occur independently **or** not with constant probability.

Since the binomial does not fit then the laws for a binomial distribution can not be true.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 4**

**Question:**

The table below shows the price of an ice cream and the distance of the shop where it was purchased from a particular tourist attraction.

| Shop | Distance from tourist attraction (m) | Price (£) |
|---|---|---|
| *A* | 50 | 1.75 |
| *B* | 175 | 1.20 |
| *C* | 270 | 2.00 |
| *D* | 375 | 1.05 |
| *E* | 425 | 0.95 |
| *F* | 580 | 1.25 |
| *G* | 710 | 0.80 |
| *H* | 790 | 0.75 |
| *I* | 890 | 1.00 |
| *J* | 980 | 0.85 |

**a** Find, to 3 decimal places, the Spearman rank correlation coefficient between the distance of the shop from the tourist attraction and the price of an ice cream.

**b** Stating your hypotheses clearly and using a 5% one-tailed test, interpret your rank correlation coefficient. ***E***

**Solution:**

| Shop | Distance | Price | $d$ | $d^2$ |
|---|---|---|---|---|
| $A$ | 1 | 9 | −8 | 64 |
| $B$ | 2 | 7 | −5 | 25 |
| $C$ | 3 | 10 | −7 | 49 |
| $D$ | 4 | 6 | −2 | 4 |
| $E$ | 5 | 4 | 1 | 1 |
| $F$ | 6 | 8 | −2 | 4 |
| $G$ | 7 | 2 | 5 | 25 |
| $H$ | 8 | 1 | 7 | 49 |
| $I$ | 9 | 5 | 4 | 16 |
| $J$ | 10 | 3 | 7 | 49 |

Remember to rank the data. It does not matter whether you rank from highest to lowest or vice versa as long as you do the same for both distance and price.

The sum of your $d$'s should be zero. The $d^2$ should all be positive.

$\sum d^2 = 286$

**a** $r_s = 1 - \dfrac{6 \times 286}{10(100-1)}$

Using $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ from the formula book.

$= -\dfrac{11}{15}$ or $-0.733$

If you give your answer as a decimal it should be given to 3 significant figures.

**b** $H_0 : \rho = 0$

$H_1 : \rho < 0$

Make sure your hypotheses are clearly written using the parameter $\rho$.

test statistic $= -0.733$

c.v. $= -0.5636$

t.s. < c.v. since $-0.733 < -0.5636$

Look up the value under 0.05 in the table for Spearman's. Quote the figure in full.

Reject $H_0$, evidence there is a significant *negative* correlation between the rank of the price of an ice cream and the rank of the distance from a tourist attraction. i.e. the further from a tourist attraction you travel the less you are likely to pay for an ice cream.

Draw a conclusion in the context of the question.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 5**

**Question:**

Five coins were tossed 100 times and the number of heads recorded. The results are shown in the table below.

| Number of heads | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| **Frequency** | 6 | 18 | 29 | 34 | 10 | 3 |

**a** Suggest a suitable distribution to model the number of heads when five unbiased coins are tossed.

**b** Test, at the 10% level of significance, whether or not the five coins are unbiased. State your hypotheses clearly. ***E***

**Solution:**

**a** B(5, 0.5)

**b** $H_0$: B(5, 0.5) is a suitable model (good fit)
$H_1$: B(5, 0.5) is a not a suitable model (not a good fit)

Total frequency $= 6+18+29+34+10+3 = 100$

This is 1 restriction.

Expected value for no heads $= 100 \times 0.5^5 = 3.125$

Expected value for 1 head $= 100 \times \binom{5}{1} 0.5^5 = 15.625$

Show the working for at least one calculation of an expected value.

Expected value for 2 heads $= 100 \times \binom{5}{2} 0.5^5 = 31.25$

Expected value for 3 heads $= 100 \times \binom{5}{3} 0.5^5 = 31.25$

Write down all the expected values.

Expected value for 4 heads $= 100 \times \binom{5}{4} 0.5^5 = 15.625$

Expected value for 5 heads $= 100 \times \binom{5}{5} 0.5^5 = 3.125$

| | $O$ | $E$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|---|
| 0 or 1 | 24 | 18.75 | 1.47 |
| 2 | 29 | 31.25 | 0.162 |
| 3 | 34 | 31.25 | 0.242 |
| 4 or 5 | 13 | 18.75 | 1.76 |

test statistic $= \sum \frac{(O-E)^2}{E} = 3.64$

or

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$$\sum \frac{O_i^2}{E_i} - N = \frac{24^2}{18.75} + \frac{29^2}{31.25} + \frac{34^2}{31.25} + \frac{13^2}{18.75} - 100$$
$$= 3.64$$

$\nu = 4 - 1 = 3$

Degrees of freedom = (number of cells after pooling) − 1 since the parameter $p$ is known.

$\chi_3^2(10\%) = 6.251$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

t.s. < c.v. since $3.64 < 6.251$
Insufficient evidence to reject $H_0$.
B(5, 0.5) is a suitable model.
No evidence that coins are biased.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 6**

**Question:**

People over the age of 65 are offered an annual flu injection. A health official took a random sample from a list of patients who were over 65. She recorded their gender and whether or not the offer of an annual flu injection was accepted or rejected. The results are summarised below.

| Gender | Accepted | Rejected |
|---|---|---|
| **Male** | 170 | 110 |
| **Female** | 280 | 140 |

Using a 5% significance level, test whether or not there is an association between gender and acceptance or rejection of an annual flu injection. State your hypotheses clearly. ***E***

**Solution:**

$H_0$ : No association between gender and acceptance
or
gender is independent of acceptance
$H_1$: There is an association between gender and acceptance
or
gender is not independent of acceptance

For a contingency table
For $H_0$ you should use the words 'no association' or 'independent'.
For $H_1$ you should use the words 'is an association' or 'dependent'.

'male' and 'accepted' expected frequency $= \frac{450 \times 280}{700} = 180$

Show the working for at least one calculation of an expected value.

Expected

| *Expected (obs)* | *Accept* | *Not accept* | Total |
|---|---|---|---|
| *Males* | 180 (170) | 100 (110) | 280 |
| *Females* | 270 (280) | 150 (140) | 420 |
| Totals | 450 | 250 | 700 |

Write down all the expected values.

| $O$ | $E$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|
| 170 | 180 | 0.5556 |
| 110 | 100 | 1.0000 |
| 280 | 270 | 0.3704 |
| 140 | 150 | 0.6667 |

The formula $\sum \frac{(O_i - E_i)^2}{E_i}$ is in the formula book. Write down at least two of the calculations.

$$\sum \frac{(O-E)^2}{E} = 2.59$$

$$\sum \frac{O_i^2}{E_i} - N = \frac{170^2}{180} + \frac{110^2}{100} + \ldots + \frac{140^2}{150} - 700$$

$$= 2.59$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$\nu = (2-1) \times (2-1) = 1$

Contingency table therefore degrees of freedom $= (c-1)(r-1)$.

$\chi_1^2 (5\%) = 3.841$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

$3.841 > 2.59$ . There is insufficient evidence to reject $H_0$
There is no association between a person's gender and their acceptance of the offer of a flu jab.

Draw a conclusion in the context of the question.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 7**

**Question:**

An area of grass was sampled by placing a $1\text{m} \times 1\text{m}$ square randomly in 100 places. The numbers of daisies in each of the squares were counted. It was decided that the resulting data could be modelled by a Poisson distribution with mean 2. The expected frequencies were calculated using the model.
The following table shows the observed and expected frequencies.

| Number of daisies | Observed frequency | Expected frequency |
|---|---|---|
| 0 | 8 | 13.53 |
| 1 | 32 | 27.07 |
| 2 | 27 | $r$ |
| 3 | 18 | $s$ |
| 4 | 10 | 9.02 |
| 5 | 3 | 3.61 |
| 6 | 1 | 1.20 |
| 7 | 0 | 0.34 |
| $\geq 8$ | 1 | $t$ |

**a** Find values for $r$, $s$ and $t$.

**b** Using a 5% significance level, test whether or not this Poisson model is suitable. State your hypotheses clearly.

An alternative test might have been to estimate the population mean by using the data given.

**c** Explain how this would have affected the test. ***E***

**Solution:**

**a** $r = 100 \times (0.6767 - 0.4060) = 27.07$

$s = 100 \times (0.8571 - 0.6767) = 18.04$

The total of the observed values is 100.

$t = 100 - [13.53 + 27.07 + 27.07 + 18.04$
$+ 9.02 + 3.61 + 1.20 + 0.34]$

$t = 0.12$

$t$ is worked out using the total of the expected frequencies is 100.

**b** $H_0$: A Poisson model Po(2) is a suitable model.

$H_1$: A Poisson model Po(2) is *not* a suitable model.

| Number of daisies | Observed | Expected |
|---|---|---|
| 0 | 8 | 13.53 |
| 1 | 32 | 27.07 |
| 2 | 27 | 27.07 |
| 3 | 18 | 18.04 |
| 4 | 10 | 9.02 |
| $\geq 5$ | 5 | 5.27 |

The classes for 5, 6, 7 and $\geq 8$ have been combined. This is so that all the expected frequencies are greater than 5.

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(8 - 13.53)^2}{13.53} + \frac{(32 - 27.07)^2}{27.07} + \ldots + \frac{(5 - 5.27)^2}{5.27}$$

$= 3.28$ (awrt)

The formula $\sum \frac{(O_i - E_i)^2}{E_i}$ is in the formula book. Write down at least two of the calculations.

or

$$\sum \frac{O_i^2}{E_i} - N = \frac{8^2}{13.53} + \frac{32^2}{27.07} + \cdots + \frac{5^2}{5.27} - 100$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$= 3.28$ (awrt)

$\nu = 6 - 1 = 5$

Degrees of freedom = number of cells − 1 since the parameter $\lambda$ is known.

$\chi_5^2 (5\%) = 11.070$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

$3.28 < 11.070$ There is insufficient evidence to reject $H_0$.

Po(2) is a suitable model.

**c** The mean must be calculated and then $\lambda =$ mean. The expected values, and hence $\sum \frac{(O - E)^2}{E}$ would be different, and the degrees of freedom would be 1 less, also changing the critical value.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 8**

**Question:**

The numbers of deaths from pneumoconiosis and lung cancer in a developing country are given in the table.

| Age group (years) | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70 and over |
|---|---|---|---|---|---|---|
| Deaths from pneumoconiosis (1000*s*) | 12.5 | 5.9 | 18.5 | 19.4 | 31.2 | 31.0 |
| Deaths from lung cancer (1000*s*) | 3.7 | 9.0 | 10.2 | 19.0 | 13.0 | 18.0 |

The correlation between the number of deaths in the different age groups for each disease is to be investigated.

**a** Give **one** reason why Spearman's rank correlation coefficient should be used.

**b** Calculate Spearman's rank correlation coefficient for these data.

**c** Use a suitable test, at the 5% significance level, to interpret your result. State your hypotheses clearly. ***E***

**Solution:**

**a** The variables cannot be assumed to be normally distributed.

**b**

| | 20–29 | 30–39 | 40–49 | 50–59 | 60–69 | 70+ |
|---|---|---|---|---|---|---|
| Rank $x$ | 5 | 6 | 4 | 3 | 1 | 2 |
| Rank $y$ | 6 | 5 | 4 | 1 | 3 | 2 |
| $d$ | –1 | 1 | 0 | 2 | –2 | 0 |
| $d^2$ | 1 | 1 | 0 | 4 | 4 | 0 |

Remember to rank the data. It does not matter whether you rank from highest to lowest or vice versa as long as you do the same for both.

The sum of your $d$'s should be zero. The $d^2$ should all be positive.

$\sum d^2 = 10$

$$r_s = 1 - \frac{6 \times 10}{6(36-1)}$$

Using $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ from the formula book.

$$= \frac{5}{7} \text{ or } 0.714$$

If you give your answer as a decimal it should be given to 3 significant figures.

**c** $\mathrm{H}_0 : \rho = 0$

$\mathrm{H}_1 : \rho \neq 0$ (or $\rho_s > 0$)

Make sure your hypotheses are clearly written using the parameter $\rho$.

$n = 6 \Rightarrow 5\%$ critical value $= 0.8857$
(or 0.8286)
$0.714 < 0.8857$ (or $0.714 < 0.8826$)

Look up the value under 0.05 in the table for Spearman's. Quote the figure in full.

No evidence to reject $\mathrm{H}_0$. No evidence of a positive correlation between the rates of deaths from pneumoconiosis and lung cancer.

Draw a conclusion in the context of the question.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 9**

**Question:**

Students in a mixed sixth form college are classified as taking courses in either arts, science or humanities. A random sample of students from the college gave the following results.

| | | Course | | |
|---|---|---|---|---|
| | | Arts | Science | Humanities |
| Gender | Boy | 30 | 50 | 35 |
| | Girl | 40 | 20 | 42 |

Showing your working clearly, test, at the 1% level of significance, whether or not there is an association between gender and the type of course taken. State your hypotheses clearly. ***E***

**Solution:**

$H_0$: There is no association between course and gender or course is independent of gender

$H_1$: There is an association between course and gender or course is dependent on gender

For a contingency table
For $H_0$ you should use the words 'no association' or 'independent'.
For $H_1$ you should use the words 'is an association' or 'dependent'.

| | *Arts* | *Science* | *Hums* | Total |
|---|---|---|---|---|
| *Boy* | 30 | 50 | 35 | 115 |
| *Girl* | 40 | 20 | 42 | 102 |
| Total | 70 | 70 | 77 | 217 |

Expected frequency 'boy' and 'arts'

$$= \frac{115 \times 70}{217} = 37.0967\ldots$$

Show the working for at least one calculation of an expected value.

| Expected (Obs) | A | S | H |
|---|---|---|---|
| Boy | 37.1 (30) | 37.1 (50) | 40.8 (35) |
| Girl | 32.9 (40) | 32.9 (20) | 36.2 (42) |

Write down all the expected values.

$$\sum \frac{(O-E)^2}{E} = \frac{(30-37.1)^2}{37.1} + \frac{(40-32.9)^2}{32.9} + \ldots + \frac{(42-36.2)^2}{36.2}$$
$$= 1.358 + 4.485 + 0.824 + 1.532 + 5.058 + 0.929 = 14.18$$

The formula $\sum \frac{(O_i - E_i)^2}{E_i}$ is in the formula book. Write down at least two of the calculations.

$$\left[ \text{or } \sum \frac{O^2}{E} - N = \frac{30^2}{37.1} + \frac{40^2}{32.9} + \ldots + \frac{42^2}{36.2} - 217 \right]$$
$$= 14.2 \text{ (3 s.f.)}$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$v = (3-1)(2-1) = 2$

Contingency table therefore degrees of freedom $= (c-1)(r-1)$.

$\chi_2^2$ (1%) critical value is 9.210

Look up the value under 0.01 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

14.18 > 9.210
Significant result or reject null hypothesis.

There is evidence of an association between course taken and gender.

Draw a conclusion in the context of the question.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 10**

**Question:**

The product-moment correlation coefficient is denoted by $r$ and Spearman's rank correlation coefficient is denoted by $r_s$.

**a** Sketch separate scatter diagrams, with five points on each diagram, to show
  **i** $r = 1$,
  **ii** $r_s = -1$ but $r > -1$.

Two judges rank seven collie dogs in a competition. The collie dogs are labelled A to G and the rankings are as follows.

| **Rank** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **Judge 1** | A | C | D | B | E | F | G |
| **Judge 2** | A | B | D | C | E | G | F |

**b** **i** Calculate Spearman's rank correlation coefficient for these data.
  **ii** Stating your hypotheses clearly, test, at the 5% level of significance, whether or not the judges are generally in agreement. ***E***

**Solution:**

**a** **i**

**ii**

**b** **i**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $r_1$ | 1 | 4 | 2 | 3 | 5 | 6 | 7 |
| $r_2$ | 1 | 2 | 4 | 3 | 5 | 7 | 6 |
| $d$ | 0 | 2 | −2 | 0 | 0 | −1 | 1 |
| $d^2$ | 0 | 4 | 4 | 0 | 0 | 1 | 1 |

Although the data is ranked it is easiest to rewrite it in a familiar form.

The sum of your $d$'s should be zero. The $d^2$ should all be positive.

$\sum d^2 = 10$

$$r_s = 1 - \frac{6 \times 10}{7(49-1)}$$

$$= \frac{23}{28} \text{ or } 0.821 \text{ (3 s.f.)}$$

Using $r_s = 1 - \dfrac{6\sum d^2}{n(n^2-1)}$ from the formula book.

If you give your answer as a decimal it should be given to 3 significant figures.

**ii** $H_0 : \rho_s = 0 \quad H_1 : \rho_s > 0$

Make sure your hypotheses are clearly written using the parameter $\rho_s$. This time you are testing if in agreement therefore you are testing if positively correlated.

test statistic $= r_s = 0.821$

critical value is 0.7143

Look up the value under 0.05 in the table for Spearman's. Quote the figure in full.

$0.821 > 0.7143$ so significant result or reject null hypothesis.

Draw a conclusion in the context of the question.

There is evidence of a (positive) correlation between the ranks awarded by the judges *or* the judges agree.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 11**

**Question:**

Ten cuttings were taken from each of 100 randomly selected garden plants. The numbers of cuttings that did not grow were recorded.
The results are as follows.

| Number of cuttings which did not grow | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8, 9 or 10 |
|---|---|---|---|---|---|---|---|---|---|
| Frequency | 11 | 21 | 30 | 20 | 12 | 3 | 2 | 1 | 0 |

**a** Show that the probability of a randomly selected cutting, from this sample, not growing is 0.223.

A gardener believes that a binomial distribution might provide a good model for the number of cuttings, out of 10, that do not grow.
He uses a binomial distribution, with the probability 0.2 of a cutting not growing. The calculated expected frequencies are as follows.

| Number of cuttings which did not grow | 0 | 1 | 2 | 3 | 4 | 5 or more |
|---|---|---|---|---|---|---|
| Expected frequency | $r$ | 26.84 | $s$ | 20.13 | 8.81 | $t$ |

**b** Find the values of $r$, $s$ and $t$.

**c** State clearly the hypotheses required to test whether or not this binomial distribution is a suitable model for these data.

The test statistic for the test is 4.17 and the number of degrees of freedom used is 4.

**d** Explain fully why there are 4 degrees of freedom.

**e** Stating clearly the critical value used, carry out the test using a 5% level of significance. ***E***

**Solution:**

**a** $\text{mean} = \dfrac{0\times11+1\times21+2\times30+\ldots+7\times1}{11+21+30+\ldots+1}$

$\text{mean} = \dfrac{223}{100} = 2.23$

In binomial, $n = 10$, mean $= np$, $2.23 = 10\times p$

so $p = 0.223$

**b** $r = (0.8)^{10}\times100 = 10.7374 = 10.74$ (2 d.p.)

You could use the tables to work these out. But you will need to use $p = 0.2$ so it is easier to do them this way.

$s = \binom{10}{2}(0.8)^8\times(0.2)^2\times100 = 30.198\ldots$

$= 30.20$ (2 d.p.)

$t = 100-[r+s+26.84+20.13+8.81]$

$= 3.28$

The total of the expected frequencies is the same as the total of the observed frequencies. Here it is 100.

**c** $H_0$: B(10, 0.2) is a suitable model for these data.

$H_1$: B(10, 0.2) is *not* a suitable model for these data.

**d** Since $t < 5$, the last two groups are combined and $\nu = 5-1 = 4$

Since there are then 5 cells and the parameter $p$ is given

**e** Critical value $\chi_4^2(5\%) = 9.488$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

$4.17 < 9.488$ so not significant or do not reject null hypothesis.

The binomial distribution with $p = 0.2$ is a suitable model for the number of cuttings that do not grow.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 12**

**Question:**

A researcher carried out a survey of three treatments for a fruit tree disease.

| | No action | Remove diseased branches | Spray with chemicals |
|---|---|---|---|
| **Tree died within 1 year** | 10 | 5 | 6 |
| **Tree survived for 1–4 years** | 5 | 9 | 7 |
| **Tree survived beyond 4 years** | 5 | 6 | 7 |

Test, at the 5% level of significance, whether or not there is any association between the treatment of the trees and their survival. State your hypotheses and conclusion clearly.

**Solution:**

$H_0$: There is no association between treatment and survival *or* treatment is independent of survival
$H_1$: There is association between treatment and survival *or* treatment is dependent on survival

For a contingency table
For $H_0$ you should use the words 'no association' or 'independent'.
For $H_1$ you should use the words 'is an association' or 'dependent'.

No action and tree died within 1 year

expected frequency $= \frac{20 \times 21}{60} = 7$

Show the working for at least one calculation of an expected value.

| Expected (Obs) | No action | Remove diseased branches | Spray with chemicals | Totals |
|---|---|---|---|---|
| Tree died within 1 year | 7(10) | 7(5) | 7(6) | 21 |
| Survived 1–4 years | 7(5) | 7(9) | 7(7) | 21 |
| Survived > 4 years | 6(5) | 6(6) | 6(7) | 18 |
| Totals | 20 | 20 | 20 | 60 |

Write down all the expected values

$$\sum \frac{(O-E)^2}{E} = \frac{9}{7}+\frac{4}{7}+\frac{1}{7}+\frac{4}{7}+\frac{4}{7}+0+\frac{1}{6}+0+\frac{1}{6}$$
$$= 3.4761\ldots$$

The formula $\sum \frac{(O_i - E_i)^2}{E_i}$ is in the formula book. Write down at least two of the calculations.

or

$$\sum \frac{O_i^2}{E_i} - N = \frac{10^2}{7}+\frac{5^2}{7}+\ldots+\frac{6^2}{7}-60$$
$$= 3.47619\ldots$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$\nu = (3-1)\times(3-1) = 4$

Contingency table therefore degrees of freedom $= (c-1)(r-1)$.

Critical value $\chi_4^2\ (5\%) = 9.488$

or CR: $\chi^2 > 9.488$

$3.47619 < 9.488$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

(or since 3.47619... is *not* in the critical region (i.e. $< 9.488$) there is insufficient evidence to reject $H_0$.

Draw a conclusion in the context of the question.

There is no evidence of association between treatment and length of survival.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 13**

**Question:**

Over a period of time, researchers took 10 blood samples from one patient with a blood disease. For each sample, they measured the levels of serum magnesium, $s$ mg/dl, in the blood and the corresponding level of the disease protein, $d$ mg/dl. The results are shown in the table.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **s** | 1.2 | 1.9 | 3.2 | 3.9 | 2.5 | 4.5 | 5.7 | 4.0 | 1.1 | 5.9 |
| **d** | 3.8 | 7.0 | 11.0 | 12.0 | 9.0 | 12.0 | 13.5 | 12.2 | 2.0 | 13.9 |

[Use $\sum s^2 = 141.51, \sum d^2 = 1081.74$ and $\sum sd = 386.32$]

**a** Draw a scatter diagram to represent these data.

**b** State what is measured by the product-moment correlation coefficient.

**c** Calculate $S_{ss}$, $S_{dd}$ and $S_{sd}$.

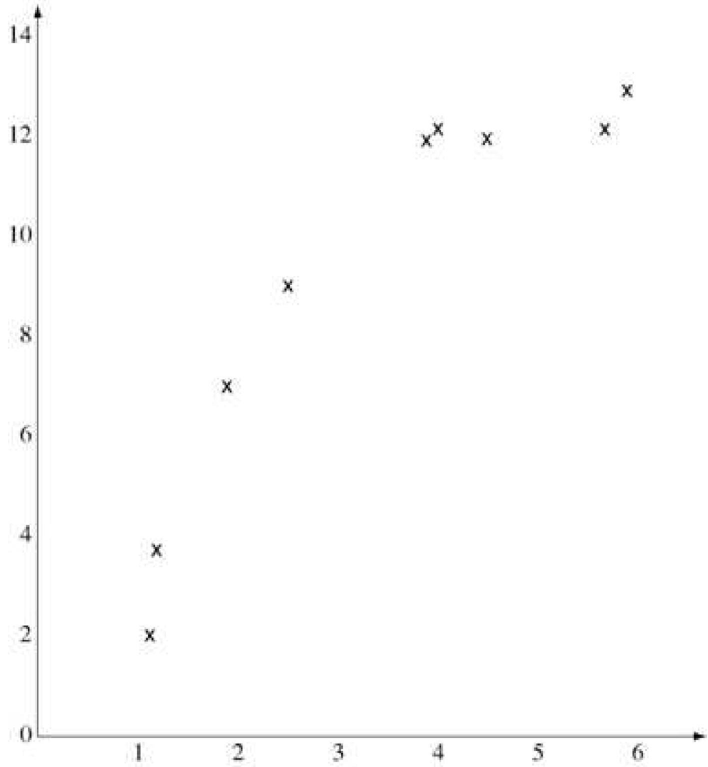
**d** Calculate the value of the product-moment correlation coefficient $r$ between $s$ and $d$.

**e** Stating your hypotheses clearly, test, at the 1% significance level, whether or not the correlation coefficient is greater than zero.

**f** With reference to your scatter diagram, comment on your result in part **e**. ***E***

**Solution:**

**a**




**b** The strength of the linear link between two variables.

**c** $S_{ss} = 141.51 - \dfrac{33.9^2}{10} = 26.589; S_{dd} = 152.444; S_{sd} = 59.524$

**d** $r = \dfrac{59.524}{\sqrt{152.444 \times 26.589}}$
$= 0.93494\ldots$

The formulae for these are under S1 in the formula book.

Make sure your hypotheses are clearly written using the parameter $\rho$.

**e** $\mathrm{H}_0 : \rho = 0; \mathrm{H}_1 : \rho > 0$
test statistic $= r = 0.935$
Critical value at $1\% = 0.7155$
$0.935 > 0.7155$

Look up the value under 0.01 in the table for product-moment coefficient. Quote the figure in full.

so reject $\mathrm{H}_0$ : levels of serum and disease are positively correlated.

Draw a conclusion in the context of the question.

**f** Linear correlation significant but scatter diagram looks non-linear.

The product-moment correlation coefficient should not be used here since the association/relationship is not linear.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 14**

**Question:**

The number of times per day a computer fails and has to be restarted is recorded for 200 days. The results are summarised in the table.

| Number of restarts | Frequency |
|---|---|
| 0 | 99 |
| 1 | 65 |
| 2 | 22 |
| 3 | 12 |
| 4 | 2 |

Test whether or not a Poisson model is suitable to represent the number of restarts per day. Use a 5% level of significance and state your hypothesis clearly. ***E***

**Solution:**

$H_0$: Poisson distribution is a suitable model

$H_1$: Poisson distribution is not a suitable model

$$\hat{\lambda} = \frac{(0 \times 99) + (1 \times 65) + \ldots + (4 \times 2)}{200} = \frac{153}{200} = 0.765$$

As $\lambda$ is not given you must work it out.

Total frequency = 200

This is 1 restriction.

Expected frequency for $(X = 2) = \frac{0.765^2 \, e^{-0.765}}{2} \times 200$
$= 27.23250$

Show the working for at least one calculation of an expected value.

Number of restarts gives

| $X$ | Observed frequency | Expected frequency | |
|---|---|---|---|
| 0 | 99 | 93.06678... | |
| 1 | 65 | 71.19604... | |
| 2 | 22 | 27.23250... | |
| 3 | 12 } 14 | 6.94428... } | 8.50468 |
| $\geq 4$ | 2 } | 1.56040... } | |

Write down all the expected values.

Combine the classes for 3 and $\geq 4$. This is so that all the expected frequencies are greater than 5.

$$\sum \frac{(O_i - E_i)^2}{E_i} = \frac{(99 - 93.07\ldots)^2}{93.066\ldots} + \ldots + \frac{(14 - 8.505)^2}{8.505}$$
$$= 5.47$$

or

$$\sum \frac{O_i^2}{E_i} - N = \frac{99^2}{93.066\ldots} + \ldots + \frac{14^2}{8.505} - 200$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$\nu = 4 - 2 = 2$

Degrees of freedom = number of cells − 2 since the parameter $\lambda$ is unknown. It had to be estimated from the information given.

Critical value $\chi_2^2(5\%) = 5.991$

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

test statistic < c.v. so 5.47 is not in the critical region so accept $H_0$.

Number of computer failures per day can be modelled by a Poisson distribution.

# Solutionbank S3
## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 15**

**Question:**

A research worker studying colour preference and the age of a random sample of 50 children obtained the results shown below.

| Age in years | Red | Blue | Totals |
|---|---|---|---|
| 4 | 12 | 6 | 18 |
| 8 | 10 | 7 | 17 |
| 12 | 6 | 9 | 15 |
| Totals | 28 | 22 | 50 |

Using a 5% significance level, carry out a test to decide whether or not there is an association between age and colour preference. State your hypotheses clearly. ***E***

**Solution:**

$H_0$ : No association between age and colour preference (they are independent)
$H_1$: Association between age and colour preference (they are not independent)

For a contingency table
For $H_0$ you should use the words 'no association' or 'independent'.
For $H_1$ you should use the words 'is an association' or 'dependent'.

'4' and 'red'

expected frequency $y = \frac{18 \times 28}{50} = 10.08$

Show the working for at least one calculation of an expected value.

| $O$ | $E$ | $\frac{(O-E)^2}{E}$ |
|---|---|---|
| 12 | 10.08 | 0.3657... |
| 6 | 7.92 | 0.4654... |
| 10 | 9.52 | 0.0242... |
| 7 | 7.48 | 0.0308... |
| 6 | 8.4 | 0.6857... |
| 9 | 6.6 | 0.8727... |

Write down all the expected values.

test statistic $= \sum \frac{(O_i - E_i)^2}{E_i} = 2.4446..$

or

$$\sum \frac{O_i^2}{E_i} - N = \frac{12^2}{10.08} + .. + \frac{9^2}{6.6\ldots} - 500$$
$$= 2.4446\ldots$$

It is often easier to use the formula $\sum \frac{O_i^2}{E_i} - N$.

$v = (3-1) \times (2-1) = 2$

Contingency table therefore degrees of freedom $= (c-1)(r-1)$.

critical value $= \chi_2^2\ (5\%) = 5.991$

(or CR: $\chi^2 > 5.991$)

Look up the value under 0.05 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

$2.4446 < 5.991$
so insufficient evidence to reject $H_0$.
No association between age and colour preference.

# Solutionbank S3

## Edexcel AS and A Level Modular Mathematics

**Review Exercise 2**
**Exercise A, Question 16**

**Question:**

A manufacture claims that the batteries used in his mobile phones have a mean lifetime of 360 hours and a standard deviation of 20 hours, when the phone is left on standby. To test this claim 100 phones were left on standby until the batteries ran flat. The lifetime $t$ hours of the batteries was recorded.
The results are as follows.

| **t** | 300– | 320– | 340– | 350– | 360– | 370– | 380– | 400– |
|---|---|---|---|---|---|---|---|---|
| **Frequency** | 1 | 9 | 28 | 20 | 16 | 18 | 7 | 1 |

A researcher believes that a normal distribution might provide a good model for the lifetime of the batteries
She calculated the expected frequencies as follows using the distribution $\mathrm{N} \sim (360, 20)$.

| $t$ | < 320 | 320– | 340– | 355– | 365– | 370– | 380– | 400– |
|---|---|---|---|---|---|---|---|---|
| Expected frequency | 2.28 | 13.59 | 24.26 | $r$ | $s$ | 14.98 | 13.59 | 2.28 |

**a** Find the values of $r$ and $s$.
**b** Stating clearly your hypotheses, test, at the 1% level of significance, whether or not this normal distribution is a suitable model for these data.

**Solution:**

**a** $\text{P}(355 < T < 365) = P\left(z < \dfrac{365-360}{20}\right) - P\left(z < \dfrac{355-360}{20}\right)$
$= P(z < 0.25) - P(z < -0.25)$
$= 0.5987 - (1 - 0.5987)$
$= 0.1974$

Using $z = \dfrac{x-\mu}{\sigma}$.

$r = 0.1974 \times 100$
$= 19.74$
$s = 100 - 2.28 - 13.59 - 24.26 - 19.74 - 14.98 - 13.59 - 2.28$
$= 9.28$

You could use the normal distribution to work out the expected value. This is quicker.

**b** $\text{H}_0 : \text{N} \sim (360, 20)$ is a suitable model.
$\text{H}_1 : \text{N} \sim (360, 20)$ is not a suitable model.

| $t$ | < 340 | 340– | 355– | 365– | 370– | 380– |
|---|---|---|---|---|---|---|
| Observed frequency | 10 | 28 | 20 | 16 | 18 | 8 |
| Expected frequency | 15.87 | 24.26 | 19.74 | 9.28 | 14.98 | 15.87 |

test statistic $= \sum \dfrac{(O_i - E_i)^2}{E_i} = \dfrac{(10-15.87)^2}{15.87} + \ldots + \dfrac{(8-15.87)^2}{15.87}$
$= 12.13$

or

$\sum \dfrac{O_i^2}{E_i} - N = \dfrac{10^2}{15.87} + .. + \dfrac{8^2}{15.87} - 100$

It is often easier to use the formula $\sum \dfrac{O_i^2}{E_i} - N$.

$\nu = 6 - 1 = 5$

Degrees of freedom = number of cells − 1 since $\mu$ and $\sigma$ are given.

Critical value $\chi_5^2 (1\%) = 15.086$

Look up the value under 0.01 in the percentage points of the $\chi^2$ distribution. Quote the figure in full.

$12.13 < 15.086$ so accept $\text{H}_0$.

The distribution can be modelled by a $\text{N} \sim (360, 20)$.