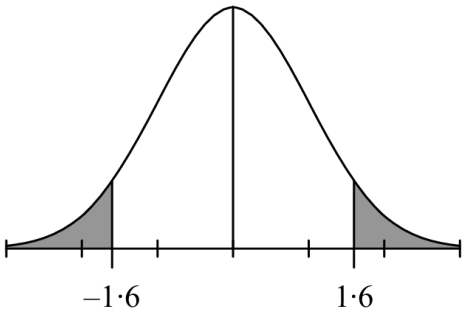
Question 1 (2013)

$$P(X \le 68) = P\left(Z \le \frac{68-60}{5}\right) = P(Z \le 1\cdot 6) = 0\cdot 9452$$

(ii) Find $P(52 \le X \le 68)$.

$$P(52 \le X \le 68) = P\left(\frac{52-60}{5} \le Z \le \frac{68-60}{5}\right)$$

$$= P(-1\cdot 6 \le Z \le 1\cdot 6)$$



$$P(Z \le -1\cdot 6) = P(Z \ge 1\cdot 6)$$

$$= 1 - P(Z \le 1\cdot 6)$$

$$= 1 - 0\cdot 9452 = 0\cdot 0548$$

$$P(-1\cdot 6 \le Z \le 1\cdot 6) = P(Z \le 1\cdot 6) - P(Z \le -1\cdot 6)$$

$$= 0\cdot 9452 - 0\cdot 0548 = 0\cdot 8904$$

OR

$$P(52 \le X \le 68) = P\left(\frac{52-60}{5} \le Z \le \frac{68-60}{5}\right)$$

$$= P(-1\cdot 6 \le Z \le 1\cdot 6)$$

$$= 1 - 2P(Z \ge 1\cdot 6)$$

$$= 1 - 2(1 - P(Z \le 1\cdot 6))$$

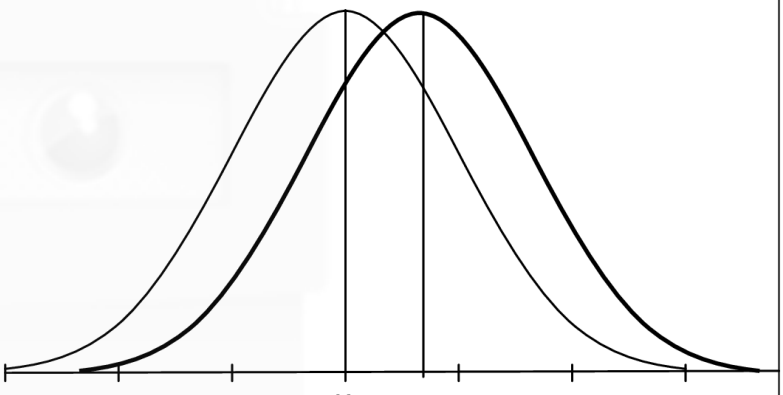
$$= 1 - 2(1 - 0\cdot 9452) = 1 - 2(0\cdot 0548) = 1 - 0\cdot 1096 = 0\cdot 8904$$

The effect, on plant growth, of each of the hormones is described. Sketch, on each diagram, a new distribution to show the effect of the hormone.

Hormone A

The effect of hormone A was to increase the height of all of the plants.

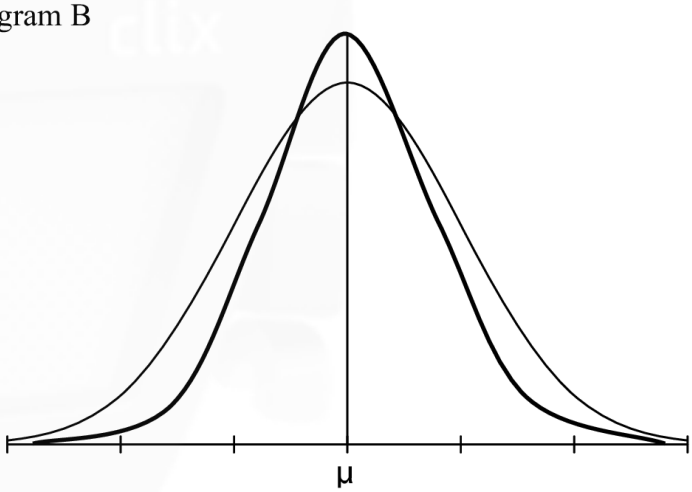
Diagram A



Hormone B

The effect of hormone B was to reduce the number of really small plants and the number of really tall plants. The mean was unchanged.

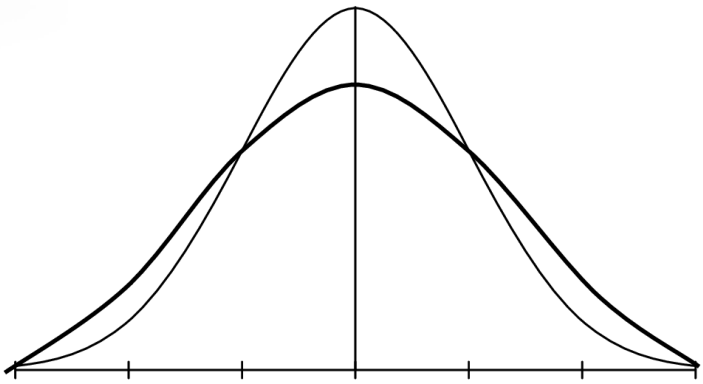
Diagram B



Hormone C

The effect of hormone C was to increase the number of small plants and the number of tall plants. The mean was unchanged.

Diagram C



Gender	Male: 479	Female: 521

Previously flown with *Go Fast Airlines*	Yes: 682	No: 318
Would fly again with *Go Fast Airlines*	Yes: 913	No: 87

Passenger Age	Mean age: 42 Median age: 31

Spend on in-flight facilities	Mean spend: €18·65 Median spend: €32·18

Was flight delayed	Yes	No	Don't Know
	231	748	21

Passenger satisfaction with overall service	Satisfied	Not satisfied	Don't Know
	664	238	98

(a) *Go Fast Airlines* used a **stratified random sample** to conduct the survey.

(i) Explain what is meant by a **stratified random sample**.

The population is divided into different subgroups which have common characteristics. Random samples are drawn from each subgroup according to their proportion of the population.

(ii) Write down 4 different passenger groups that the company might have included in their sample.

One solution:

Long haul economy class passengers.
Long haul business class passengers.
Long haul executive class passengers.
Short haul passengers

(b) **(i)** What is the probability that a passenger selected at random from this sample

- had their flight delayed

$$\frac{231}{1000} = 0 \cdot 231 \text{ or } \frac{231}{979} = 0 \cdot 236$$

- Was not satisfied with the overall service

$$\frac{238}{1000} = 0 \cdot 238 \text{ or } \frac{238}{902} = 0 \cdot 264$$

(ii) An employee suggests that the probability of selecting a passenger whose flight was delayed and who was not satisfied with the overall service should be equal to the product of the two probabilities in **(i)** above. Do you agree with the employee?

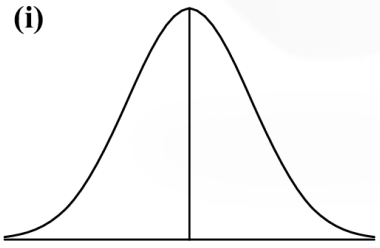
Answer: No.

Reason:

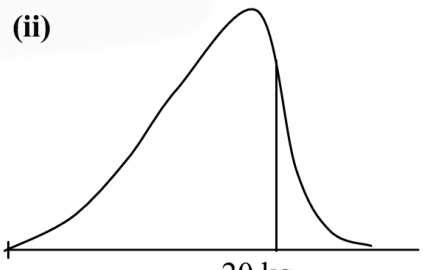
If it was this would imply that the events were independent but this is not likely since a passenger who had his flight delayed is likely to be not satisfied with the service.

(c) Which of the graphs below do you think is most likely to represent the distribution of the weights of passenger baggage?

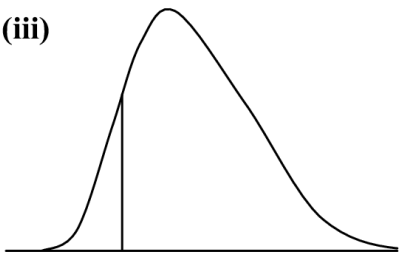
(i)



(ii)



(iii)

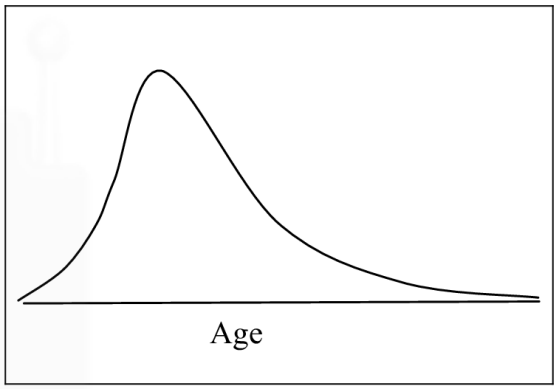


Answer: Graph (ii)

Reason:

A lot of the passengers are likely to have baggage with a weight of less than the maximum 20 kg.

(d) **(i)** Draw a sketch of the possible distribution of the ages of the passengers based on the data in the survey.



(ii) Explain your answer.

The median is less than the mean so the graph is skewed to the right.

(e) **(i)** The company repeatedly asserts that 70% of their customers are satisfied with their overall service Use an hypothesis test at the 5% level of significance to decide whether there is sufficient evidence to conclude that their claim is valid in May. State the null hypothesis and state your conclusion clearly.

Null Hypothesis: The satisfaction level is unchanged. $p = 0 \cdot 7$
The 95% margin of error for a sample of size 1000 is $\frac{1}{\sqrt{1000}} = 0 \cdot 0316$.
The recorded satisfaction level for May is $0 \cdot 664$.
This is outside the range $[0 \cdot 7 - 0 \cdot 0316 , 0 \cdot 7 + 0 \cdot 0316] = [0 \cdot 7316 , 0 \cdot 6684]$.
Reject the null hypothesis.
There is evidence to conclude that the company claim is not valid in May.

OR

Null Hypothesis: The satisfaction level is unchanged. $p = 0 \cdot 7$
The 95% margin of error for a sample of size 1000 is $\frac{1}{\sqrt{1000}} = 0 \cdot 0316$.
The 95% confidence interval for the population proportion is
$0 \cdot 664 - 0 \cdot 0316 < p < 0 \cdot 664 + 0 \cdot 0316 = 0 \cdot 6324 < p < 0 \cdot 6956$
$p = 0 \cdot 7$ is outside this range.
Reject the null hypothesis.
There is evidence to conclude that the company claim is not valid in May.

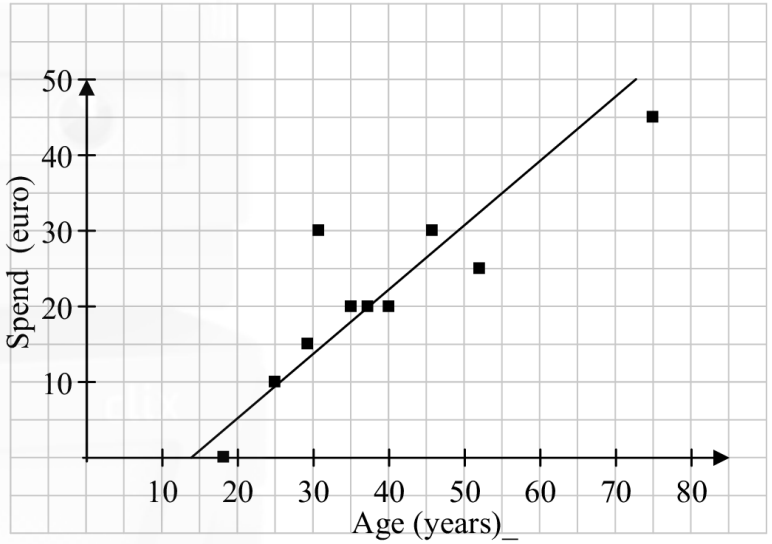
(ii) A manager of the airline says: "If we survey 2000 passengers from June on, we will half the margin of error in our surveys." Is the manager correct?

Answer: No.

Reason:
For a sample of size n, the margin of error is
$\frac{1}{\sqrt{n}}$. $\frac{1}{2}\frac{1}{\sqrt{1000}} \neq \frac{1}{\sqrt{2000}}$

or $0 \cdot 0158 \neq \cdot 022$

(i) Draw a scatter plot of the data.



(ii) Calculate the correlation coefficient between passenger age and in-flight spend.

$0 \cdot 88$

(iii) What can you conclude from the completed scatter plot and the correlation coefficient?

Older passengers tend to spend more.

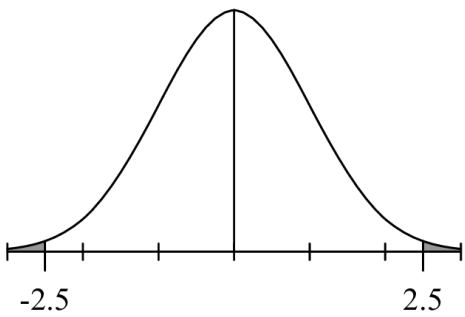
(iv) Sketch the line of best fit in the completed scatter plot above.

Question 3 (2012)

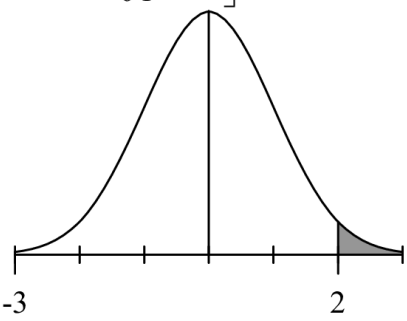
$$Z = \frac{20{\cdot}25 - 20}{0{\cdot}1} = 2{\cdot}5$$

$$P(|X - 20| > 0{\cdot}25) = P(|Z| > 2{\cdot}5)$$
$$= 2(1 - P(Z \leq 2.5))$$
$$= 2(1 - 0{\cdot}9938)$$
$$= 0{\cdot}0124$$

Answer = $10\,000 \times 0{\cdot}0124 = 124$.



$$P(X \le 1{\cdot}975) + P(X \ge 20{\cdot}25) = P\left[Z \le \frac{19{\cdot}75 - 20{\cdot}25}{0{\cdot}1}\right] + P\left[Z \ge \frac{20{\cdot}25 - 20{\cdot}05}{0{\cdot}1}\right]$$
$$= P(Z \le -3) + P(Z \ge 2)$$
$$= 1 - P(Z \le 3) + 1 - P(Z \le 2)$$
$$= 1 - 0{\cdot}9987 + 1 - 0{\cdot}9772$$
$$= 2 - 1{\cdot}9759$$
$$= 0{\cdot}0241$$



$\frac{0{\cdot}0241}{0{\cdot}0124} = 1{\cdot}9435\ldots \Rightarrow 94{\cdot}35\%$ increase

or increase: $0{\cdot}0241 - 0{\cdot}0124 = 0{\cdot}0117$

% Increase: $\left(\frac{0{\cdot}0117}{0{\cdot}0124}\right)100 = 94{\cdot}35\%$

Question 4 (2012)

(i) the arrears rates?

They've gone up a lot – they were mostly between 1 and 5 in 2009, and mostly between 5 and 15 in 2011.

(ii) the rates of interest being paid?

They've gone up a lot too – they were mostly between 2·3 and 4·1% in 2009, and mostly between 4 and 6% in 2011.

(iii) the relationship between the arrears rate and the interest rate?

There appears to be a stronger relationship 2011 than in 2009.

(b) What additional information would you need before you could estimate the median interest rate being paid by mortgage holders in September 2011?

You would need to know how many mortgage holders are represented by each point on the relevant diagram.

(c) Regarding the relationship between the arrears rate and the interest rate for September 2011, the authors of the report state: "The direction of causality … is important" and they go on to discuss this.

Explain what is meant by the "direction of causality" in this context.

It is a question of whether higher arrears rates cause interest rates to go up, or whether higher interest rates cause arrears rates to go up, (assuming there is a causal relationship at all).

(i) What is the probability that a property selected at random (from all those examined) will be in negative equity?
Give your answer correct to two decimal places.

$\frac{145414}{475136} = 0{\cdot}30604711 = 0{\cdot}31$ (to two decimal places)

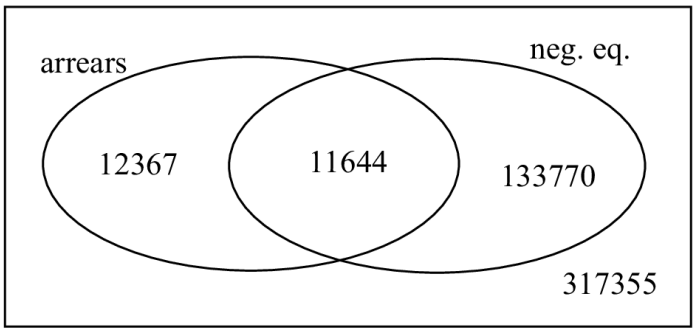
(ii) What is the probability that a property selected at random from all those in negative equity will also be in arrears?
Give your answer correct to two decimal places.

$\frac{11644}{145414} = 0{\cdot}08007482 = 0{\cdot}08$ (to two decimal places)

(iii) Find the probability that a property selected at random from all those in arrears will also be in negative equity.
Give your answer correct to two decimal places.

	arrears	¬arrears	total
neg. eq.	**11644**	133770	**145414**
¬neg. eq.	12367	**317355**	329722
total	24011	451125	**475136**

$\frac{11644}{24011} = 0{\cdot}4849 = 0{\cdot}48$ (to two decimal places)



$$P(A|N) = \frac{P(A \cap N)}{P(N)} \Rightarrow 0{\cdot}08007 = \frac{P(A \cap N)}{0{\cdot}30604} \Rightarrow P(A \cap N) = 0{\cdot}0245$$

But $P(A) = \frac{24011}{475136} = 0{\cdot}05053$

$$P(N|A) = \frac{P(N \cap A)}{P(A)} = \frac{0{\cdot}0245}{0{\cdot}05053} = 0{\cdot}4848 = 0{\cdot}48 \text{ (to two decimal places)}$$

Null hypothesis: proportion in negative equity unchanged: $p = 0{\cdot}31$.
Alternative hypothesis: it has changed: $p \neq 0{\cdot}31$.

95% margin of error for samples of size 2000 is $\frac{1}{\sqrt{2000}} \approx 0{\cdot}0224$

So, reject null hypothesis if observed proportion lies outside $0{\cdot}31 \pm 0{\cdot}0224$.

Observed proportion = $\frac{552}{2000} \approx 0{\cdot}276$.

$0{\cdot}276 \notin [0{\cdot}2876, 0{\cdot}3224]$

Outside margin of error, so reject null hypothesis.

The proportion in negative equity has changed.

OR

Null hypothesis: proportion in negative equity unchanged: $p = 0{\cdot}31$.

95% margin of error for samples of size 2000 is $\frac{1}{\sqrt{2000}} \approx 0{\cdot}0224$

Observed proportion = $\frac{552}{2000} \approx 0{\cdot}276$.

$\therefore$ the 95% confidence interval for the population proportion is:

$0{\cdot}276 - 0{\cdot}0224 < p < 0{\cdot}276 + 0{\cdot}0244$

$0{\cdot}2536 < p < 0{\cdot}2984$

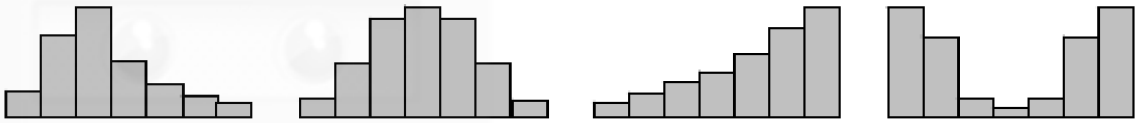
0·31 outside this range.

Therefore reject null hypothesis. Proportion in negative equity has changed.

Question 5 (2012)

Question 2 **(25 marks)**

(a) Complete the table below indicating whether a statement is correct (✓) or incorrect with respect to each data set.



A B C D

	A	B	C	D
The data are skewed to the left			✓	
The data are skewed to the right	✓			
The mean is equal to the median		✓		✓
The mean is greater than the median	✓			
There is a single mode	✓	✓	✓	

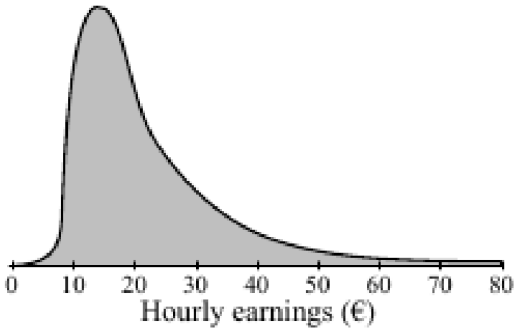
(b) Assume the four histograms are drawn on the same scale. State which of them has the largest standard deviation, and justify your answer.

Histogram D has the largest standard deviation.
Justification: The standard deviation measures how far the data are spread out from the mean. Since histogram D is symmetric, the mean is in the middle.
But we can also see that in histogram D, most of the data are spread out to the far right and left of the diagram, generally far from the mean.
Therefore, of the four histograms, D has the largest standard deviation.

Question 6 (2012)

(b) The distribution of the hourly earnings of all employees in Ireland in October 2009 is shown in the diagram. It can be seen that the distribution is positively skewed.

The mean is €22.05.
The median is €17.82.
The standard deviation is €10.64
The lower quartile is €12.80
The upper quartile is €26.05



(Source: adapted from: CSO. *National Employment Survey 2008 and 2009*)

(i) If six employees are selected at random from this population, what is the probability that exactly four of them had hourly earnings of more than €12.80?

Since €12.80 is the lower quartile, the probability of any one randomly selected person having hourly earnings of more than €12.80 is 0.75. If we repeat this selecting 6 times we are conducting a Bernouilli trial

$$P(r \text{ successes from } n \text{ trials}) = \binom{n}{r} p^r (1-p)^{n-r}$$

where p is the probability of a success. Let "success" be earning over €12.80. Then the probability of choosing 4 people who earn over €12.80 in 6 trials is

$$\begin{aligned} & \binom{6}{4} 0.75^4 (0.25)^2 \\ = \; & (15)(0.3164)(0.0625) \\ = \; & 0.296625 \end{aligned}$$

(ii) In a computer simulation, random samples of size 200 are repeatedly selected from this population and the mean of each sample is recorded. A thousand such sample means are recorded.

Describe the expected distribution of these sample means. Your description should refer to the shape of this distribution and to its mean and standard deviation.

The Law of Large Numbers says that these sample means will be normally distributed with mean = 22.05 and standard deviation = $\frac{10.64}{\sqrt{200}} = 0.752$

(iii) How many of the sample means would you expect to be greater than €23?

Let X be a normally distributed random variable with mean 22.05 and standard deviation 0.752

We will begin by finding the probability that $X \leq 23$. Since X is normally distributed with mean 22.05 and standard deviation 0.752, we will convert it into a Standard Normal Z distribution:

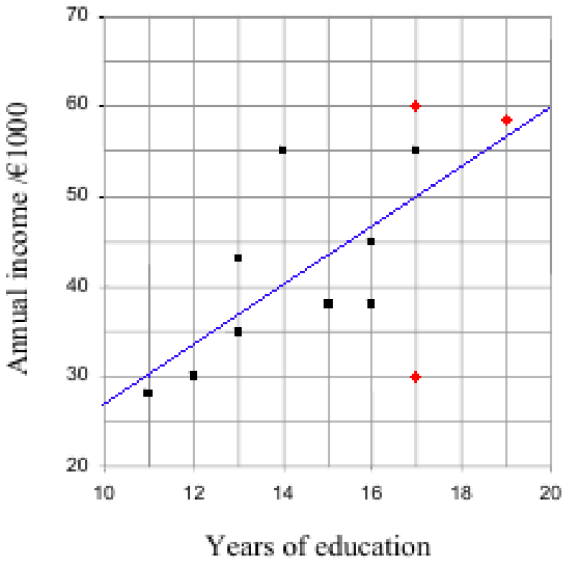
$$X \leq 23 \iff Z \leq \frac{23 - 22.05}{0.752} = 1.26$$

Note that the last figure is taken to two decimal places. From this calculation, we know that $P(X \leq 23) = P(Z \leq 1.26)$. From the z-tables, this probability works out to be 0.8762. This means that $P(X > 23) = 1 - 0.8762 = 0.1238$. A thousand sample means are calculated, so the expected number of sample means greater than €23 is $1000 \times 0.1238 \approx 124$.

Question 7 **(75 marks)**

(a) An economics students wants to find out whether the length of time people spend in education affects the income they earn. The student carries out a small study. Twelve adults are asked to state their annual income and the number of years they spent in full-time education. The data are given in the table below, and a partially completed scatter plot is given.

Years of education	Income /€1,000
11	28
12	30
13	35
13	43
14	55
15	38
16	45
16	38
17	55
17	60
17	30
19	58



(i) The last three rows of the data have not been included on the scatter plot. Insert them now.

See red points inserted in the scatter plot above.

(ii) Calculate the correlation coefficient.

The correlation coefficient may be calculated using your electronic calculator. This is sufficient for exam purposes.

For information, it is calculated by hand as follows:

Let $x_1, x_2, \ldots, x_{12}$ denote the values for "Years" and let $y_1, y_2, \ldots y_{12}$ denote the values for "Income". Then, the correlation coefficient is defined by

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{12} y_i\right)^2}}$$

where n is the number of data points. Consider the following table:

x_i	y_i	$x_i y_i$	$(x_i)^2$	$(y_i)^2$
11	28	308	121	784
12	30	360	144	900
13	35	455	169	1,125
13	43	559	169	1,849
14	55	770	196	3,025
15	38	570	225	1,444
16	45	720	256	2,025
16	38	608	256	1,444
17	55	935	289	3,025
17	60	1.020	289	3,600
17	30	510	289	900
19	58	1.102	361	3,364
180	515	7,917	2,764	23,585
$\sum_{i=1}^{12} x_i$	$\sum_{i=1}^{12} y_i$	$\sum_{i=1}^{12} x_i y_i$	$\sum_{i=1}^{12} (x_i)^2$	$\sum_{i=1}^{12} (y_i)^2$

Thus, our correlation coefficient becomes

$$\rho = \frac{12(7917) - (180)(515)}{\sqrt{12(2,764) - (180)^2}\sqrt{12(23,585) - (515)^2}} = 0.62324$$

(iii) What can you conclude from the scatter plot and the correlation coefficient?

The above scatter plot implies that there is a rough linear relationship between years in education and income. We can see that the data is increasing roughly together, although there are some outliers which do not follow the same pattern as the rest of data.

This agrees with the correlation coefficient. A value close to 0 represents no relationship between the data, and a value close to 1 represents a strong positive linear relationship. Our value ($\rho = 0.623$) means that there is a positive linear relationship, but not a very strong one.
We can conclude that there is a moderate positive correlation between number of years in education and annual income i.e. as one increases, the other tends to also increase.

(iv) Add the line of best fit to the scatter plot above.

See blue line above.

(v) Use the line of best fit to estimate the annual income of somebody who has spent 14 years in education.

€41,000

(vi) By taking suitable readings from your diagram, or otherwise, calculate the slope of the line of best fit.

Using the points $(14,41)$ and $(17,50)$ the slope of the line is $\dfrac{50-41}{17-14} = \dfrac{9}{3} = 3$.

(vii) Explain how to interpret the slope in this context.

The slope is the rise over the run. In this case, as we can write the slope of 3 as $\dfrac{3}{1}$, it implies that for every 1 extra year of education, there is a corresponding rise of 3 units of income. As the units of income are thousands of euro, then this implies a rise of 3000 euro in annual income per extra year of education.

(viii) The student collected the data using a telephone survey. Numbers were randomly chosen from the Dublin area telephone directory. The calls were made in the evenings, between 7 and 9pm. If there was no answer, or if the person who answered did not agree to participate, then another number was chosen at random.

List **three** possible problems regarding the sample and how it was collected that might make the result of the investigation unreliable. In each case, state clearly why the issue you mention could cause a problem.

Problem 1:

Calls were made in the evening. This can affect the likelihood of getting responses from some demographics versus other demographics e.g. office workers vs shift workers, or employed people vs unemployed people, or parents of young children vs others. This may make the sample unrepresentative and could cause bias in the data.

Problem 2:

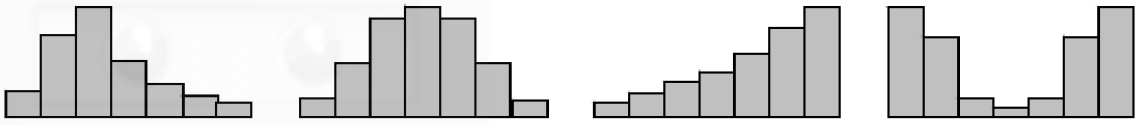
The telephone directory shows landline numbers only. Many younger people do not have a landline. This may make the sample unrepresentative, causing bias in the data..

Problem 3:

If the survey is intended to represent the whole country, then using the Dublin telephone directory will mean that the data is unrepresentative, causing bias.

Question 2 **(25 marks)**

(a) Complete the table below indicating whether a statement is correct (✓) or incorrect with respect to each data set.



A B C D

	A	B	C	D
The data are skewed to the left			✓	
The data are skewed to the right	✓			
The mean is equal to the median		✓		
The mean is greater than the median	✓			
There is a single mode	✓	✓	✓	

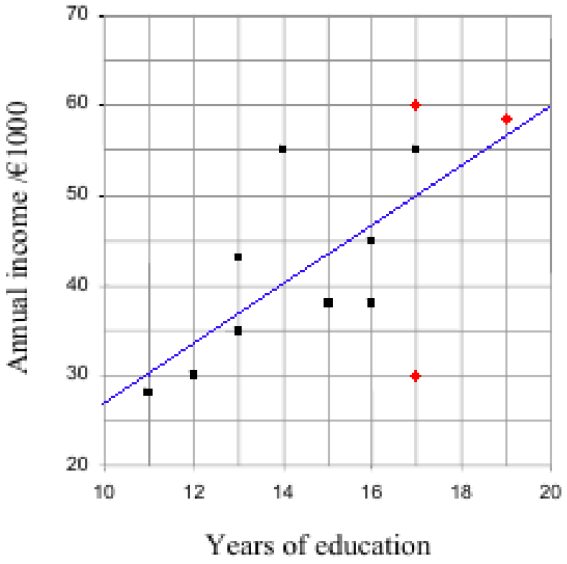
(b) Assume the four histograms are drawn on the same scale. State which of them has the largest standard deviation, and justify your answer.

Histogram D has the largest standard deviation.
Justification: The standard deviation measures how far the data are spread out from the mean. Since histogram D is symmetric, the mean is in the middle.
But we can also see that in histogram D, most of the data are spread out to the far right and left of the diagram, generally far from the mean.
Therefore, of the four histograms, D has the largest standard deviation.

Question 7 **(75 marks)**

(a) An economics students wants to find out whether the length of time people spend in education affects the income they earn. The student carries out a small study. Twelve adults are asked to state their annual income and the number of years they spent in full-time education. The data are given in the table below, and a partially completed scatter plot is given.

Years of education	Income /€1,000
11	28
12	30
13	35
13	43
14	55
15	38
16	45
16	38
17	55
17	60
17	30
19	58



(i) The last three rows of the data have not been included on the scatter plot. Insert them now.

See red points inserted in the scatter plot above.

(ii) Calculate the correlation coefficient.

The correlation coefficient may be calculated using your electronic calculator. This is sufficient for exam purposes.

For information, it is calculated by hand as follows:

Let $x_1, x_2, \ldots, x_{12}$ denote the values for "Years" and let $y_1, y_2, \ldots y_{12}$ denote the values for "Income". Then, the correlation coefficient is defined by

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n}(x_i)^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n}(y_i)^2 - \left(\sum_{i=1}^{12} y_i\right)^2}}$$

where n is the number of data points. Consider the following table:

x_i	y_i	$x_i y_i$	$(x_i)^2$	$(y_i)^2$
11	28	308	121	784
12	30	360	144	900
13	35	455	169	1,125
13	43	559	169	1,849
14	55	770	196	3,025
15	38	570	225	1,444
16	45	720	256	2,025
16	38	608	256	1,444
17	55	935	289	3,025
17	60	1.020	289	3,600
17	30	510	289	900
19	58	1.102	361	3,364
180	515	7,917	2,764	23,585
$\sum_{i=1}^{12} x_i$	$\sum_{i=1}^{12} y_i$	$\sum_{i=1}^{12} x_i y_i$	$\sum_{i=1}^{12} (x_i)^2$	$\sum_{i=1}^{12} (y_i)^2$

Thus, our correlation coefficient becomes

$$\rho = \frac{12(7917) - (180)(515)}{\sqrt{12(2,764) - (180)^2}\sqrt{12(23,585) - (515)^2}} = 0.62324$$

(iii) What can you conclude from the scatter plot and the correlation coefficient?

The above scatter plot implies that there is a rough linear relationship between years in education and income. We can see that the data is increasing roughly together, although there are some outliers which do not follow the same pattern as the rest of data.

This agrees with the correlation coefficient. A value close to 0 represents no relationship between the data, and a value close to 1 represents a strong positive linear relationship. Our value ($\rho = 0.623$) means that there is a positive linear relationship, but not a very strong one.
We can conclude that there is a moderate positive correlation between number of years in education and annual income i.e. as one increases, the other tends to also increase.

(iv) Add the line of best fit to the scatter plot above.

See blue line above.

(v) Use the line of best fit to estimate the annual income of somebody who has spent 14 years in education.

€41,000

(vi) By taking suitable readings from your diagram, or otherwise, calculate the slope of the line of best fit.

Using the points $(14,41)$ and $(17,50)$ the slope of the line is $\dfrac{50-41}{17-14} = \dfrac{9}{3} = 3$.

(vii) Explain how to interpret the slope in this context.

The slope is the rise over the run. In this case, as we can write the slope of 3 as $\dfrac{3}{1}$, it implies that for every 1 extra year of education, there is a corresponding rise of 3 units of income. As the units of income are thousands of euro, then this implies a rise of 3000 euro in annual income per extra year of education.

(viii) The student collected the data using a telephone survey. Numbers were randomly chosen from the Dublin area telephone directory. The calls were made in the evenings, between 7 and 9pm. If there was no answer, or if the person who answered did not agree to participate, then another number was chosen at random.

List **three** possible problems regarding the sample and how it was collected that might make the result of the investigation unreliable. In each case, state clearly why the issue you mention could cause a problem.

Problem 1:

Calls were made in the evening. This can affect the likelihood of getting responses from some demographics versus other demographics e.g. office workers vs shift workers, or employed people vs unemployed people, or parents of young children vs others. This may make the sample unrepresentative and could cause bias in the data.

Problem 2:

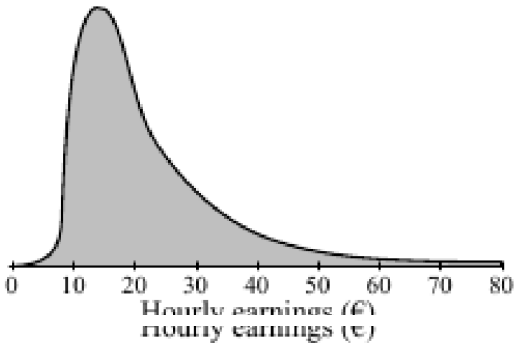
The telephone directory shows landline numbers only. Many younger people do not have a landline. This may make the sample unrepresentative, causing bias in the data..

Problem 3:

If the survey is intended to represent the whole country, then using the Dublin telephone directory will mean that the data is unrepresentative, causing bias.

Question 10 (2012)

The mean is €22.05.
The median is €17.82.
The standard deviation is €10.64
The lower quartile is €12.80
The upper quartile is €26.05



(i) If six employees are selected at random from this population, what is the probability that exactly four of them had hourly earnings of more than €12.80?

Since €12.80 is the lower quartile, the probability of any one randomly selected person having hourly earnings of more than €12.80 is 0.75. If we repeat this selecting 6 times we are conducting a Bernouilli trial

$$P(r \text{ successes from } n \text{ trials}) = \binom{n}{r} p^r (1-p)^{n-r}$$

where p is the probability of a success. Let "success" be earning over €12.80. Then the probability of choosing 4 people who earn over €12.80 in 6 trials is

$$\begin{aligned} & \binom{6}{4} 0.75^4 (0.25)^2 \\ = \; & (15)(0.3164)(0.0625) \\ = \; & 0.296625 \end{aligned}$$

(ii) In a computer simulation, random samples of size 200 are repeatedly selected from this population and the mean of each sample is recorded. A thousand such sample means are recorded.

Describe the expected distribution of these sample means. Your description should refer to the shape of this distribution and to its mean and standard deviation.

The Law of Large Numbers says that these sample means will be normally distributed with mean = 22.05 and standard deviation = $\dfrac{10.64}{\sqrt{200}} = 0.752$

(iii) How many of the sample means would you expect to be greater than €23?

Let X be a normally distributed random variable with mean 22.05 and standard deviation 0.752

We will begin by finding the probability that $X \leq 23$. Since X is normally distributed with mean 22.05 and standard deviation 0.752, we will convert it into a Standard Normal Z distribution:

$$X \leq 23 \iff Z \leq \frac{23 - 22.05}{0.752} = 1.26$$

Note that the last figure is taken to two decimal places. From this calculation, we know that $P(X \leq 23) = P(Z \leq 1.26)$. From the z-tables, this probability works out to be 0.8762. This means that $P(X > 23) = 1 - 0.8762 = 0.1238$. A thousand sample means are calculated, so the expected number of sample means greater than €23 is $1000 \times 0.1238 \approx 124$.

Question 11 (2011)

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

$$z_1 = \frac{14-20}{5} = -1{\cdot}2$$

$$z_2 = \frac{26-20}{2} = 1{\cdot}2$$

$$P(14 \le X \le 26) = P(-1{\cdot}2 \le z \le 1{\cdot}2)$$

$$\begin{aligned} P(-1{\cdot}2 \le z \le 1{\cdot}2) &= 1 - 2P(z > 1{\cdot}2) \\ &= 1 - 2[1 - P(z \le 1{\cdot}2)] \\ &= 2P(z \le 1{\cdot}2) - 1 \\ &= 0{\cdot}7698 \end{aligned}$$

Question 12 (2011)

A positive correlation between two variables does not mean that one is necessarily causing the other. For example, in a primary school there might be a correlation between reading ability and shoe size, but big feet don't make you read better and reading doesn't make your feet grow! In this case, both variables are connected to age – a 'confounding factor'

Question 13 (2011)

$$x = \frac{1}{5}(3x' + y') \text{ and } y = \frac{1}{5}(-2x' + y') \text{ and } A'\left(\frac{k}{8}, 0\right),\ B'(0, k).$$

$$\therefore A \text{ is } \left(\frac{3k}{40}, \frac{-2k}{40}\right) \text{ and } B \text{ is } \left(\frac{k}{5}, \frac{k}{5}\right).$$

$$\left|\angle A'OB'\right| = 90^\circ.$$

$$\text{Slope } OA = \frac{\frac{-2k}{40}}{\frac{3k}{40}} = -\frac{2}{3} \text{ and slope} OB = \frac{\frac{k}{5}}{\frac{k}{5}} = 1 \Rightarrow OA \text{ is not} \perp \text{to } OB.$$

$$\therefore \left|\angle AOB\right| \neq \left|\angle A'OB'\right|.$$

Blunders (-3)

B1 A or B incorrect
B2 Error in slope formula
B3 No conclusion or incorrect conclusion

Slips (-1)

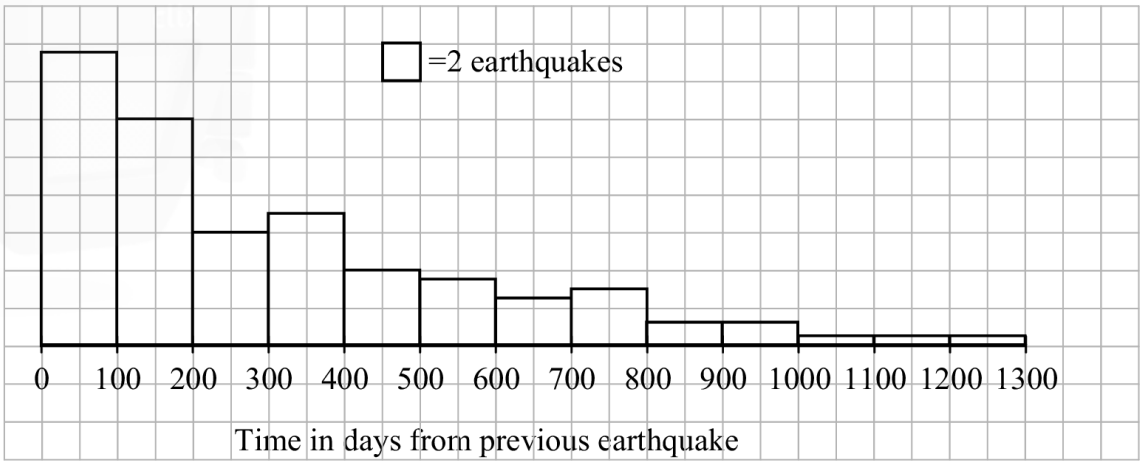
S1 Arithmetic errors

Attempts (2 marks)

A1 Effort to find A or B and stops
A2 Effort at finding angle other than required angle

Histogram.
First, divide up unequal intervals, and estimate an allocation of the data (optional step).

800 – 900	900 – 1000	1000 – 1100	1100 – 1200	1200 – 1300
2·5	2·5	1	1	1



The distribution is skewed to the right. (Or, e.g., there is a lot of data to the left, and it tails off to the right, etc.)

The median is approximately 220 days.

Because the distribution is not normal.

The best estimate is to assume the probability is as reflected in the proportion of such intervals in the historical data.

$$\frac{24}{115} \approx 0{\cdot}2$$

- They could have looked at the number of earthquakes each year, or some other interval of time (e.g. distribution of earthquakes per decade, per year, etc.)
- They could have redefined serious earthquakes as earthquakes greater than a certain magnitude; earthquakes in less populated areas are not included.
- The data set could have been broadened to include less serious earthquakes. This could result in a different pattern.

(i) Comment on the reporter's statement, using information from the diagram to support your answer, and suggest a more accurate statement.

The statement is too deterministic – strong earthquakes don't always cause tsunamis and weak ones sometimes do. A better statement would be "Strong earthquakes are more likely to cause tsunamis than weaker ones."

(ii) By taking suitable readings from the diagram, estimate the probability that an earthquake of magnitude between 6·5 and 7·0 will cause a tsunami.

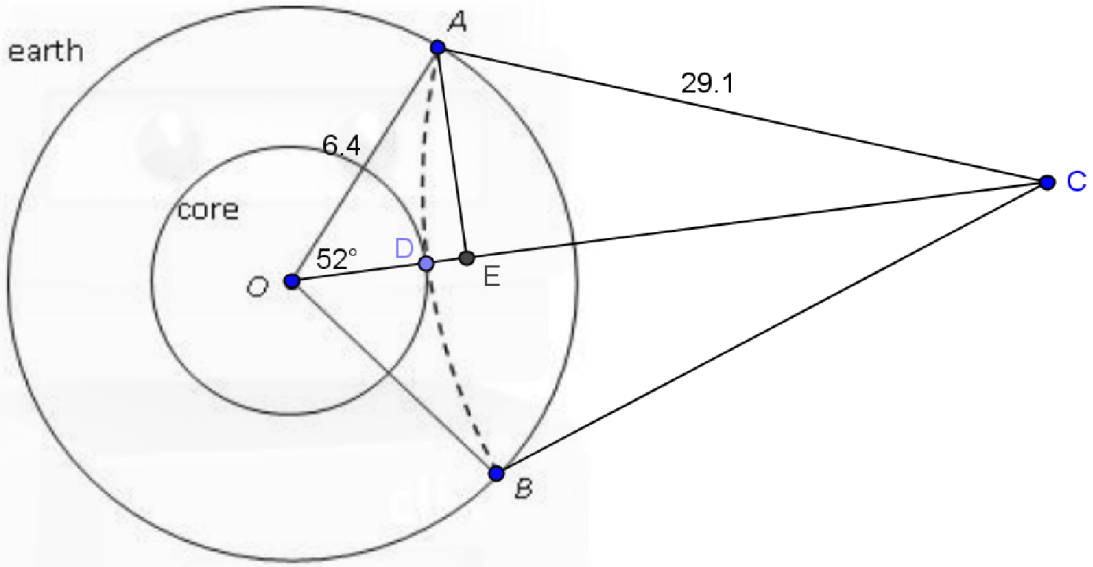
About 103 of these did and about 156 didn't. So probability is $\frac{103}{259} \approx 0{\cdot}4$

Tsunami : 142 + 60 + 8 =210

No tsunami: 139 + 36 + 7 = 182

$p \approx \frac{210}{392} \approx 0{\cdot}54$ *

$\binom{6}{4}(0{\cdot}54)^4(0{\cdot}46)^2 + \binom{6}{5}(0{\cdot}54)^5(0{\cdot}46) + (0{\cdot}54)^6 = 0{\cdot}421$



$|AE| = 6{\cdot}4 \sin 52 = 5{\cdot}04$

$|OE| = 6{\cdot}4 \cos 52 = 3{\cdot}94$

$|CE| = \sqrt{29{\cdot}1^2 - 5{\cdot}04^2} = 28{\cdot}66$

$|DE| = 29{\cdot}1 - 28{\cdot}66 = 0{\cdot}44$

$|OD| = 3{\cdot}94 - 0{\cdot}44 = 3{\cdot}5$ units (or 3500 km).

Or

Let $|OC| = x$. Then, in triangle OAC we have: $(29{\cdot}1)^2 = (6{\cdot}4)^2 + x^2 - 2(6{\cdot}4)x \cos 52°$

$x^2 - 7{\cdot}88x - 805{\cdot}85 = 0$

~~$x = -24{\cdot}7$~~ $\quad x = 32{\cdot}6$

$|OE| = 32{\cdot}6 - 29{\cdot}1 = 3{\cdot}5$ units (or 3500 km).

(a) From the diagram, estimate the correlation coefficient.

Answer: – 0·75

(b) Circle the *outlier* on the diagram and write down the person's age and maximum heart rate.

Age = 47 years

Max. heart rate = 137 bpm

(c) The line of best fit is shown on the diagram. Use the line of best fit to estimate the maximum heart rate of a 44-year-old person.

Answer: 176 bpm

(d) By taking suitable readings from the diagram, calculate the slope of the line of best fit.

Possible Readings

$(10, 200)$ and $(90, 144)$.

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$

$$m = \frac{144 - 200}{90 - 10} = -\frac{-56}{80} = -\frac{7}{10} \text{ or } m = -0{\cdot}7 .$$

(e) Find the equation of the line of best fit and write it in the form: $MHR = a - b \times (\text{age})$, where *MHR* is the maximum heart rate.

$$y - y_1 = m(x - x_1)$$
$$y - 200 = -0{\cdot}7(x - 10)$$
$$y = -0{\cdot}7x + 207$$
$$MHR = 207 - 0{\cdot}7 \times (age)$$

(f) The researchers compared their new rule for estimating maximum heart rate to an older rule. The older rule is: $MHR = 220 - \text{age}$. The two rules can give different estimates of a person's maximum heart rate. Describe how the level of agreement between the two rules varies according to the age of the person. Illustrate your answer with two examples.

For young adults the old rule gives a greater *MHR* than the new rule.

Adult aged 20

$MHR = 220 - 20 = 200$ bpm (Old rule)

$MHR = 207 - 0{\cdot}7(20) = 193$ bpm (New Rule)

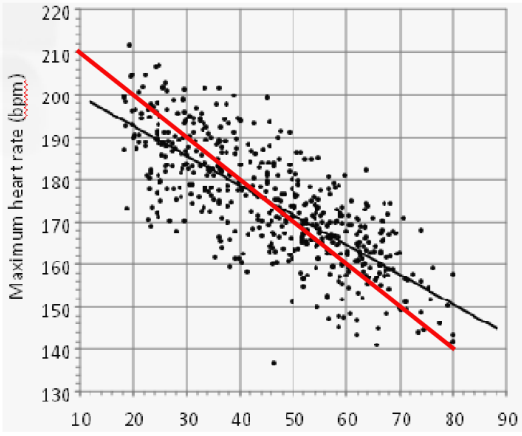
Towards middle age there is a greater agreement between the rules.

For older people the new rule gives a greater *MHR* than the old rule.

Adult aged 70

$MHR = 220 - 70 = 150$ bpm

$MHR = 207 - 0{\cdot}7(70) = 158$ bpm



(g) A particular exercise programme is based on the idea that a person will get most benefit by exercising at 75% of their estimated *MHR*. A 65-year-old man has been following this programme, using the old rule for estimating *MHR*. If he learns about the researchers' new rule for estimating *MHR*, how should he change what he is doing?

He should exercise a bit more intensely.

Using the old rule he exercises to 75% of $(220 - 65) = 116$ bpm.

Using the new rule he can exercise to 75% of $(207 - 0{\cdot}7 \times 65) = 121$ bpm.

(a) What is the probability that a randomly selected rod will be less than 39·7 mm in length?

$$P(X < 39{\cdot}7) = P\left(Z < \frac{39{\cdot}7 - 40}{0{\cdot}2}\right) = P(Z < -1{\cdot}5)$$
$$= P(z > 1{\cdot}5)$$
$$= 1 - P(Z \le 1{\cdot}5)$$
$$= 1 - 0{\cdot}9332$$
$$= 0{\cdot}0668$$

(b) Five rods are selected at random. What is the probability that at least two of them are less than 39·7 mm in length?

Binomial distribution with $n = 5$, $p = 0{\cdot}0668$, $q = 0{\cdot}9332$.

$$P(X \ge 2) = 1 - P(X < 2) = 1 - \left[P(X = 1) + P(X = 0)\right]$$
$$= 1 - \left[\binom{5}{1}(0{\cdot}0668)(0{\cdot}9332)^4 + \binom{5}{0}(0{\cdot}9332)^5\right]$$
$$= 0{\cdot}03895.$$

Or

$$P(X \ge 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$
$$= \binom{5}{2}(0{\cdot}0668)^2(0{\cdot}9332)^3 + \binom{5}{3}(0{\cdot}0668)^3(0{\cdot}9332)^2 + \binom{5}{4}(0{\cdot}0668)^4(.9332) + \binom{5}{5}(0{\cdot}0668)^5$$
$$= 0{\cdot}03895$$

$H_0 : \mu = 40$ mm (null hypothesis)
$H_1 : \mu \ne 40$ mm (alternative hypothesis)

$$\sigma_{\bar{x}} = \frac{0{\cdot}2}{\sqrt{10}} = 0{\cdot}0632456$$

Observed value of $\bar{x} = 39{\cdot}87$

$$\therefore \text{ Observed } z = \frac{39{\cdot}87 - 40}{0{\cdot}0632456} = -2{\cdot}055$$

The critical values for the test are $\pm 1{\cdot}96$

As $-2{\cdot}055 < -1{\cdot}96$, we reject the null hypothesis at the 5% level of significance and we conclude that the machine setting has become inaccurate.