AH National Qualifications 2016

X703/77/11 Statistics

TUESDAY, 10 MAY

1:00 PM – 4:00 PM

Total marks — 100

Attempt ALL questions.

You may use a calculator.

Full credit will be given only to solutions which contain appropriate working.

State the units for your answer where appropriate.

Write your answers clearly in the answer booklet provided. In the answer booklet you must clearly identify the question number you are attempting.

Use **blue** or **black** ink.

Before leaving the examination room you must give your answer booklet to the Invigilator; if you do not, you may lose all the marks for this paper.

A booklet of Statistical Formulae and Tables is supplied for all candidates.

Total marks — 100

Attempt ALL questions

1. A biologist in Africa is investigating grazing patterns. The heights in a random sample of vegetation were measured to the nearest centimetre at various points on a line transect and are listed below.

26 15 28 23 25 15 16 20 22 27 17 30 48

(a) Draw a suitable statistical diagram to represent this sample. **1**

(b) Calculate the upper fence and comment on the measurement of 48 cm. **3**

2. A large number of knee operations is undertaken by 10 different surgeons.

Amongst these surgeons there is considerable variation in the number of operations they undertake. Some surgeons perform only a few and others perform hundreds.

A sample of 100 patients is to be taken to evaluate patient satisfaction using a quality of life questionnaire.

(a) If a simple random sample of 100 patients is taken, explain why this might be an unrepresentative sample. **1**

(b) Describe how a more representative sample might be taken, stating the type of sampling used. **2**

(c) Suggest another possible cause of variation in patient satisfaction other than the skill of the surgeon. **1**

3. The content weights of a random sample of 12 jars of honey yielded a mean of 147·8 g with standard deviation 2·379 g.

(a) Assuming that the weights are normally distributed, assess the evidence, at the 1% level of significance, that the population mean content weight is less than 150 g. **4**

(b) Explain why a z-test would be inappropriate in this situation. **1**

4. The ages x (years) of a small sample of adult gannets (a sea bird) on the Bass Rock were accurately determined by a biologist and are displayed below, along with summary statistics.

Ages of adult gannets on the Bass Rock

5	2 2 2 3 4 6 7 8 9
6	3 4 6 6 9
7	3 7
8	6
9	8
10	
11	
12	8
13	7

Key:

5|7 represents 5·7 years
$n = 20$
$\sum x = 142$ and $\sum x^2 = 1120{\cdot}16$

(a) (i) Comment on the shape of the distribution of this data. **1**

(ii) Calculate the mean and standard deviation for this sample. **1**

A second biologist accurately determines the ages of 20 different gannets from the same colony on the Bass Rock.

(b) Use the Central Limit Theorem to calculate the approximate probability that the mean age of the second biologist's sample is greater than 8·1 years. **4**

5. A popular gardening magazine wants to investigate the way in which to water tomato plants to give the biggest yield. Popular folklore says that tomatoes should be watered every day and under these conditions it is known that the variety "Gardener's Delight" gives a mean yield of 4·75 kg of tomatoes per plant.

An experiment to test if watering once a week changes the yield of Gardener's Delight produced the following yields (kg) of tomatoes on 9 plants.

4·00 4·50 4·75 4·88 4·99 5·10 5·00 5·25 5·63

(a) (i) Construct a 90% confidence interval for the population mean yield of tomatoes per plant, stating an assumption made. **5**

(ii) Explain whether or not this confidence interval supports the suggestion that watering once a week changes the yield. **2**

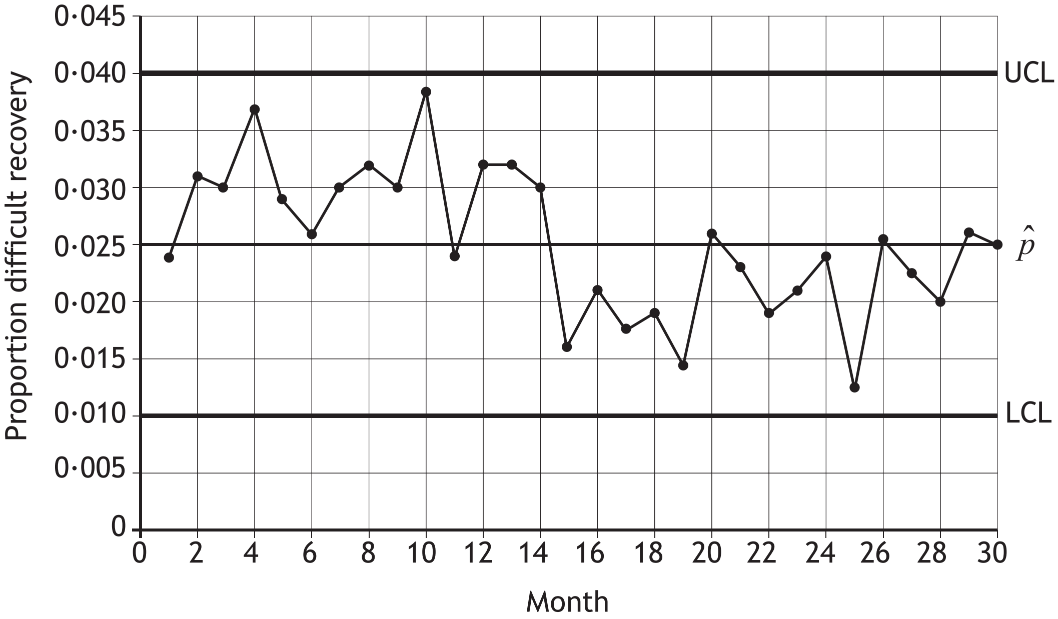
(b) Other than frequency of watering, suggest another factor which might contribute to the variation in tomato plant yields. **1**

MARKS

6. Working at Trondheim University Hospital in 2003, researchers Sigurd Fasting and Sven E. Gisvold used control charts to monitor the quality of anaesthesia processes. An adapted version of their work is considered below.

They investigated the percentage of surgical operations in which patients had a difficult recovery from general anaesthetic every month in the period from January 1997 (Month 1) to June 1999 (Month 30). The p-chart for the period of study, with 3-sigma limits, is shown below.

p-chart showing rate of difficult recovery from general anaesthesia



(a) Without further calculation, describe a feature of the control chart that indicates that the process is possibly influenced by special causes. **1**

Another possible indication of special cause variation is illustrated in the data from months 15 to 28 where 12 out of 14 consecutive data points lie below the centre line.

(b) Calculate the probability that exactly 12 out of 14 consecutive points lie either above or below the centre line, when a process is in control. **3**

Before the study was undertaken, changes had been made to anaesthetic practice in early 1998 and the data can be more sensibly viewed as two separate processes, before and after the change.

(c) State the numbers of the two months between which the change in practice was most likely to have been made. **1**

The estimated proportion of difficult recoveries, $\hat{p}$, illustrated on the centre line of the control chart, is 0·025 and the Upper Control Limit is 0·04.

(d) Show that the number of operations performed each month was 975. **3**

(e) Explain why it was appropriate to use the normal approximation to the binomial distribution in this study and calculate the approximate percentage of sample proportions which would be greater than 0·03. **4**

7. During a viral epidemic a doctor examines 150 people suffering from symptoms commonly associated with the virus. Of the 150 people examined, 90 are male of whom 40 actually have the virus. 10 of the examined females have the virus, the rest do not.

(a) Calculate the probability that an individual selected at random from this group is infected with the virus. **1**

(b) If 3 different people are selected at random without replacement from this group, what is the probability that all 3 have the disease? **2**

Of the people in this group with the virus 94% react positively to a clinical test to confirm the viral infection, as do 7% of the people without the virus.

(c) (i) Calculate the probability that a person selected at random reacts positively. **4**

(ii) Calculate the probability that a person selected at random has the virus given that he or she reacted positively. **3**

8. A new drug was trialled on 100 randomly chosen patients who had a particular disease and yielded a 75% recovery rate.

Another 100 randomly chosen patients with the disease were used as a control group, treated with an existing drug, giving a 65% recovery rate.

(a) Perform a z-test to determine whether there is any evidence that the new drug had a higher recovery rate. **7**

(b) For this particular type of drug trial explain why a two-tailed test might be more appropriate. **1**

9. In a particular school, the number of sporting injuries X occurring in a week can be modelled by the Poisson distribution with mean 4.

(a) State two assumptions that underlie the valid use of a Poisson model in this situation. **2**

(b) Calculate the probability that the number of sporting injuries in a week is more than two standard deviations above the mean. **3**

(c) Assuming that a school year is 38 weeks, use a suitable approximation to find the probability that there are fewer than 140 sporting injuries in a school year. **5**

(d) Considering the circumstance in which sporting injuries might occur, give a reason why the Poisson distribution might not be an appropriate model. **1**

[Turn over

10. Let X be a discrete random variable where $\mathrm{E}(X)=\frac{7}{4}$ and $\mathrm{V}(X)=\frac{3}{16}$.

(a) If X takes only the values 1 and 2, tabulate the probability distribution for X. **3**

The probability distribution of Y (independent of X) is

y	1	2
$p(y)$	$\frac{2}{5}$	$\frac{3}{5}$

(b) Calculate the mean and variance of Y. **2**

(c) Calculate the values of $\mathrm{E}(3X-Y)$ and $\mathrm{V}(3X-Y)$. **3**

11. Charles Darwin conducted a series of experimental studies on the reproductive biology of various plant species. His theory was that cross-pollination yielded offspring that were taller than those produced by self-pollination.

The data below come from a random sample of 12 pairs of plants. Each pair of plants was grown under identical conditions, but with one member of each pair, chosen randomly, to be bred via cross-pollination and the other by self-pollination.

A positive difference indicates a taller offspring by cross-pollination.

Cross	23·5	12·0	21·0	20·1	22·0	21·5	22·1	20·4	18·3	21·1	21·0	12·0
Self	17·4	20·4	20·0	20·1	20·0	18·6	18·6	15·3	16·5	18·0	18·0	18·0
Diff	6·1	–8·4	1·0	0·0	2·0	2·9	3·5	5·1	1·8	3·1	3·0	–6·0

(a) Assuming the differences are normally distributed, perform a test to assess the evidence from the data that the mean height for offspring bred via cross-pollination is greater than the mean height for offspring bred via self-pollination. **7**

(b) Without the assumption of normality state another test that might have been employed, together with its underlying assumption. Perform the test and comment on the conclusion with reference to that of the first test. **6**

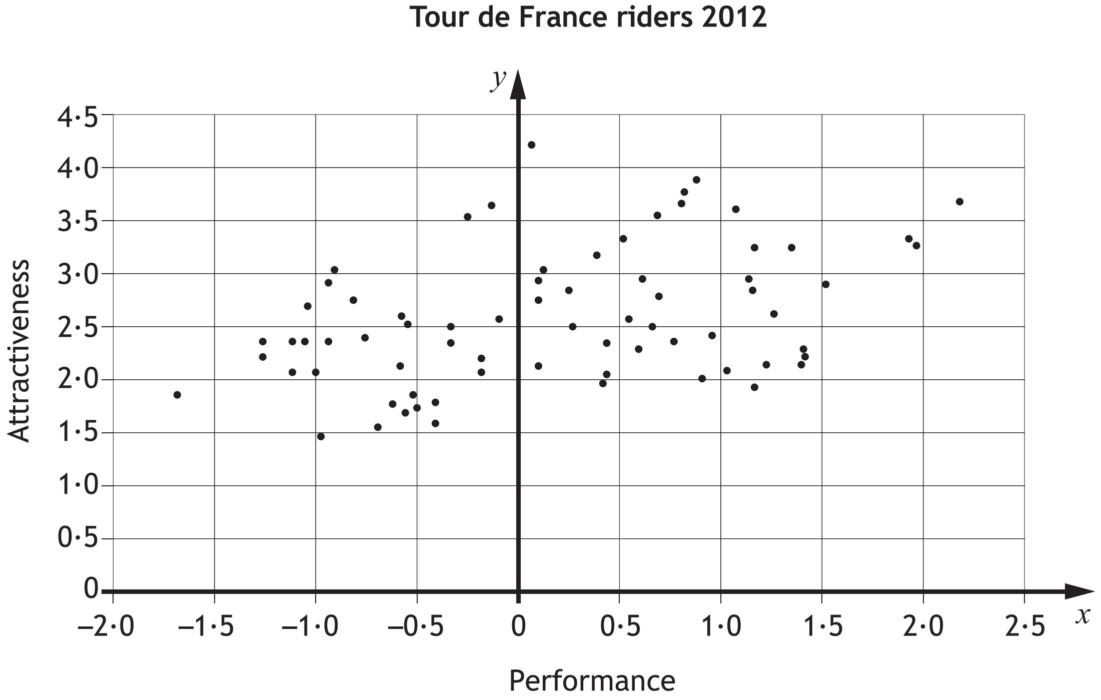
(c) Comment on the benefit of having a paired study design in this experiment. **1**

12. A study entitled "A relationship between attractiveness and performance in professional cyclists" was published in *Biology Letters* in January 2014 and a slightly modified version of the results is introduced below.

Participants in the study were shown head and shoulders portraits of a sample of 70 of the leading riders in the 2012 Tour de France cycle race. They were asked to then rate each rider, on a scale of 1 to 5 (with 5 the highest), in terms of attractiveness. Scores from all participants were combined to give a single value of attractiveness (y) for each rider.

A value for performance for each rider (x) was obtained by considering outcomes from different parts of the race, where higher values of x signify better performance.

A scatterplot of all 70 results is shown below.



(a) From the scatterplot, comment on the relationship between attractiveness and performance for the population of professional cyclists sampled. **1**

For the 70 results, the following statistics were obtained.

$$\sum x = 12 \cdot 0345 \quad \sum y = 179 \cdot 9185$$

$$S_{xx} = 58 \cdot 3111 \quad S_{yy} = 26 \cdot 1816 \quad S_{xy} = 15 \cdot 6348$$

(b) Calculate the product moment correlation coefficient of x and y.

State what this measures and comment on its value. **4**

[Turn over for next question

MARKS

12. (continued)

(c) Show that there is evidence at the 5% level that this correlation coefficient is significantly different from zero. **2**

(d) With reference to the test performed in (c) explain why relatively low correlations can be statistically significant. What would be your final conclusion to this study? **2**

(e) Give a reason why it might not be useful to fit a least squares regression line to this data. **1**

[END OF QUESTION PAPER]

ACKNOWLEDGEMENT

Question 12 – Information and graph is adapted from “A relationship between attractiveness and performance in professional cyclists” by Erik Postma, taken from *Biology Letters, February 2014, Vol 10, Issue 10.* Reproduced under The Royal Society open-access policy.