AH National Qualifications 2017

# X703/77/11 Statistics

WEDNESDAY, 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.

You may refer to the Statistics Advanced Higher Statistical Formulae and Tables.

MARKS

**Total marks — 100**

**Attempt ALL questions**

**1.** A charity in Africa rescues orphaned baby elephants with the aim of returning them to the wild. After it has been rescued, the probability of an orphaned elephant surviving the first month is 0·6. The probability of then reaching adulthood is 0·7. Having reached adulthood, the probability of being returned to the wild is 0·9.

(a) (i) Calculate the probability that an orphaned elephant is successfully returned to the wild.

(ii) Calculate the probability that an orphaned elephant reaches adulthood, but is not successfully returned to the wild. **3**

(b) What proportion of rescued orphaned elephants, which do not reach adulthood, die within the first month? **2**

**2.** The discrete random variable $X$ is such that $\mathrm{P}(X = x) = c$, for constant $c$, and $x = 20, 30, 40, 50, 60$.

(a) Obtain the probability distribution of $X$. **1**

(b) (i) Calculate the expectation and variance of $X$.

(ii) Calculate $\mathrm{P}(X > \mu)$. **3**

**3.** Last year it was claimed that pianists were awarded lower marks than violinists for their performance in the exams offered by a particular Music Exam Board.

(a) (i) Give a reason why obtaining a sample of such exam marks might not be possible.

(ii) Assuming that this problem has been overcome, you are asked to obtain a random sample of around 15% of pianists from the 20 centres in which the Exam Board operates. Outline how you would obtain such a sample, given that there is a strict budget which means that you must keep travelling expenses to an absolute minimum. **4**

A random sample of 16 pianists obtained a mean mark of 77 with standard deviation 10, while another random sample of 14 violinists obtained a mean of 82 with standard deviation 8.

(b) Perform a $t$-test to assess the evidence for the claim. **6**

MARKS

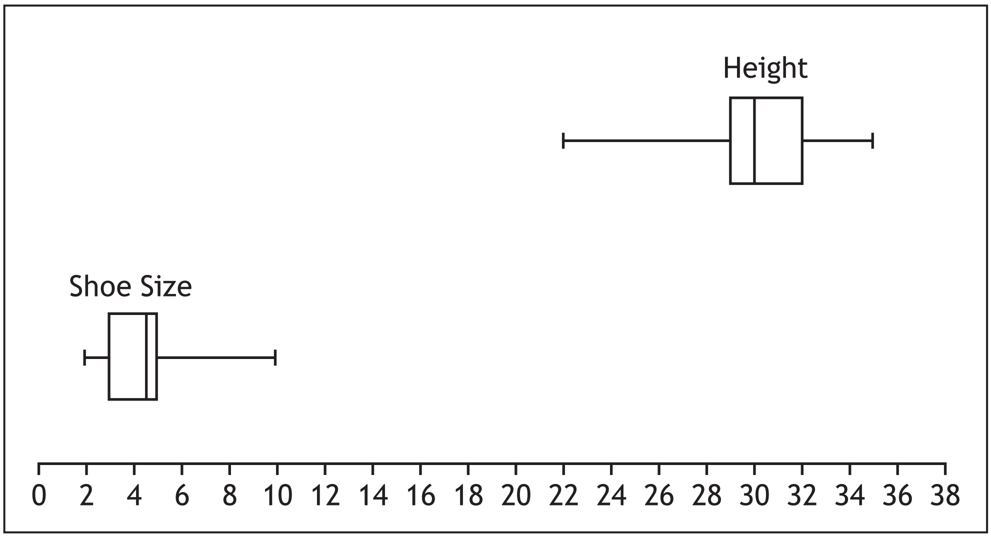
**4.** In a day care nursery group, children's heights (inches) were measured:

22 27 27 29 29 30 30 30 30 32 32 33 34 35

(a) Calculate the fences and identify any outliers. **3**

The corresponding shoe size of the children was measured and the boxplot below illustrates both sets of data.

**Nursery Group Shoe Size and Height**




(b) (i) Comment on the appropriateness of this plot for examining the relationship between height and shoe size.

(ii) Suggest an appropriate diagram to explore any relationship between shoe size and height and what statistical analysis could then be performed. **3**

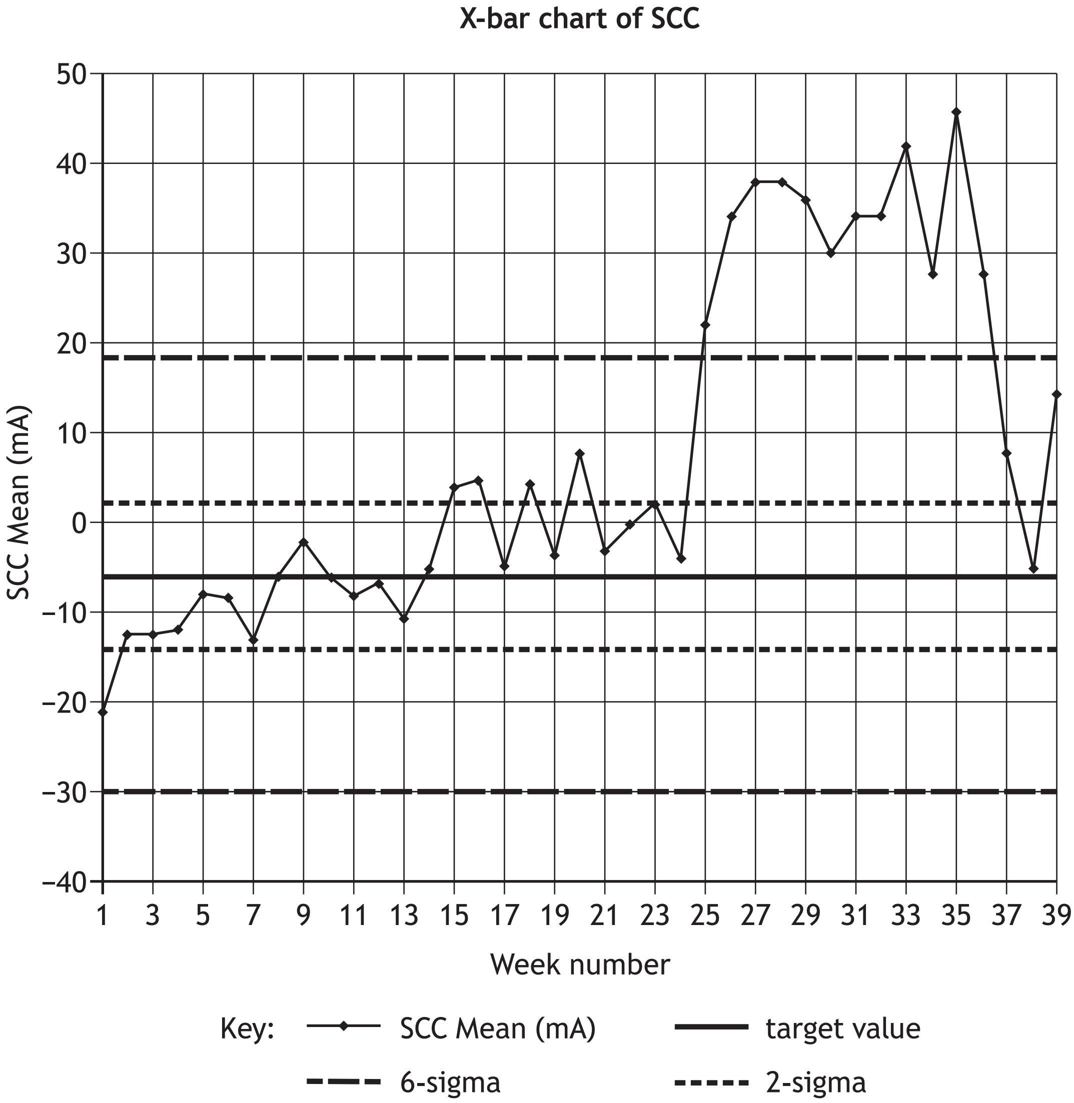
**[Turn over**

5. Radiation therapy is used in the treatment of some illnesses and requires a precise and consistent supply of X-ray radiation. Researchers in a School of Medicine studied an X-ray producing system which had suffered total equipment failure.

The precision and consistency of the X-ray radiation is controlled by Steering Coil Currents (SCC), measured in milliamps. Daily SCC measurements were grouped together to give a weekly mean value ($n=5$) for the 39 weeks up to and including the week of the total equipment failure. By retrospectively investigating this data, researchers were wanting to see if it might be possible to detect changes in SCC prior to total equipment failure.

The weekly mean values of SCC were used to produce the control chart shown below.

As it is extremely expensive to shut down and repair this type of system, control limits were set at 6-sigma, as opposed to 3-sigma, above and below the researchers' target value. In this study, these 6-sigma limits are referred to as high alarm limits.



MARKS

**5. (continued)**

As population parameters were unavailable, the data for the first 20 weeks were used to establish the control chart target value and 6-sigma limits.

For these data $\bar{x} = -6 \cdot 1$ and $s = 9 \cdot 0561$.

(a) (i) Calculate the high alarm limits.

(ii) Identify the week in which an SCC mean value first occurs outside the high alarm limits. **2**

Researchers also considered two low alarm conditions:

- 9 points in a row, on the same side of the centre line
- 2 out of 3 consecutive points falling beyond the same 2-sigma limit.

(b) (i) Calculate the probability of 9 points in a row, on the same side of the centre line, when the process is in control.

(ii) Identify the week numbers in which each of the low alarm conditions first occurs. **3**

**6.** A ski lift linking two ski resorts in Austria is used to transport 130 skiers and their personal equipment from one valley to another. The weights (kg) of skiers may be assumed to be N(75, 16) and the weights of their personal equipment may be assumed to be N(10, 2). The maximum total weight the ski lift is licensed to carry is 11 200 kg.

Stating any assumption required, calculate the probability that this total weight is exceeded. **6**

**[Turn over**

MARKS

**7.** In California there has been a significant water shortage for decades and biologists are concerned about the effect that this water shortage might have on the structure of forests, and in particular the number of large trees (> 61 cm diameter at breast height).

As the forests being investigated cover such vast areas it is not practical to count all the trees. To enable the calculation of a value for the mean number of large trees per hectare for a whole forest, the number of such trees is counted within some sample hectares. The vast areas to be surveyed dictate that the number of sample hectares is relatively small.

Data from 18 randomly sampled hectares in the 1930s yielded a mean of 7·46 trees per hectare with standard deviation 1·46.

(a) Calculate a 95% confidence interval for the mean number of large trees per hectare from the 1930s data. **3**

A recent survey from the 2000s yielded a mean of 5·87 trees per hectare.

(b) (i) By considering this recent mean and the 1930s confidence interval, comment on the suggestion that the number of large trees might be in decline.

(ii) Explain whether or not this statistical analysis validates the biologists' concern that water shortage has had an effect on the number of large trees per hectare. **3**

**8.** For each of the given distributions, obtain the probability that the value of the random variable lies no more than one standard deviation from the mean.

(a) $W \sim \text{Po}(4)$ **3**

(b) $X \sim \text{N}(4, 4)$ **2**

(c) $Y \sim \text{U}(6, 10)$ **3**

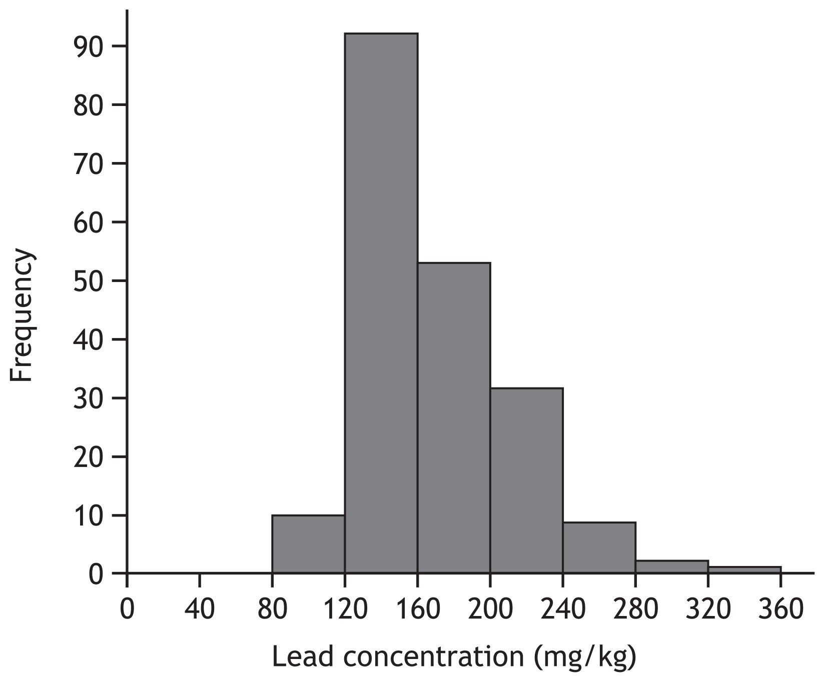
**9.** Environmental Scientists working in Fujian Province in China were concerned about the amount of lead found in the blood of children living in rural villages situated near lead mines and processing plants. As part of their study, they collected soil samples from villages suspected of having contaminated soil, and also from villages known to be contamination free. For each soil sample, the concentration of lead (mg/kg) was determined. The lead concentration in a single sample is denoted by the continuous random variable $X$.

During data collection, each village was divided into square grids of size 50 metres by 50 metres and a soil sample was taken from the centre of each grid.

(a) Suggest a problem that might occur when using this sampling strategy and what might be done to overcome this problem. **2**

To determine the background levels of naturally occurring lead in soil, the scientists collected data from many villages in Fujian province known to be contamination free. The concentration, $X$ mg/kg, of lead in soil in single sample grid squares collected from these villages is shown in the histogram below, with $\mu = 165 \cdot 6$ and $\sigma = 23 \cdot 1$ mg/kg.

**Concentration of Lead**




(b) State the distribution of the sample means, $\bar{X}$, with $n = 25$, expected from the parent distribution illustrated above and justify your answer. **3**

One small village (referred to as village A) close to a lead mine was suspected of having some lead contamination. 25 random samples were collected from this village and a mean value of 174·5 mg/kg was calculated.

(c) Carry out an appropriate statistical test, at the 5% level of significance, to determine if the mean soil lead level in village A is greater than 165·6 mg/kg. **5**

**[Turn over**

MARKS

**10.** A researcher is studying woodland rodents as hosts for parasite transmission. The study involves capturing, examining, marking and releasing rodents on a number of sites in the Loch Lomond basin in the West of Scotland. The theoretical chance of a recapture (capturing a rodent that has previously been marked and released), determined from previous studies, is 20%.

(a) At one site the researcher captures 20 individuals. What is the probability that exactly 3 are recaptures? **2**

(b) At a second site the researcher captures 45 individuals. Using an appropriate approximation, determine the probability that between 5 and 10 (inclusive) are recaptures. **4**

The following contingency table summarises the whole data set.

| | | Sex of Rodent | |
|---|---|---|---|
| | | Male | Female |
| Recapture Status | Recaptured | 58 | 51 |
| | Not Recaptured | 255 | 182 |

(c) By conducting a hypothesis test, determine if there is any association between sex and recapture status. **6**

**11.** A local council is proposing to construct a new bypass and plans to hold a referendum to decide whether or not to proceed. A local newspaper wishes to seek public opinion and commissions two companies to survey residents who live in the council area on this issue.

(a) Company A takes a random sample of 100 residents and finds that 61% are in favour of the bypass. Construct an approximate 99% confidence interval for the proportion of all residents who are in favour of the bypass, suggesting why it is only an approximate interval. **4**

(b) Company B finds that 58% of its random sample are in favour of the bypass and claims to be 99% confident that the majority of all the residents are in favour. Calculate the minimum size of sample that has been taken in order to justify this claim. **3**

MARKS

**12.** The Offwell Woodland and Wildlife Trust undertook a project to restore a former woodland site, surrounded by existing trees and of approximate size 1200 square metres, back to a heathland habitat at its Woodland Education Centre in East Devon.

Part of the site was divided into two strips. The strips are managed in different ways, with strip 1 being cut in spring and strip 2 being cut in the autumn. The effect on bluebell distribution of the different cutting times was investigated.

Within each strip, twenty quadrats (a wooden frame with an area of one square metre) were placed at random, and within each quadrat the percentage bluebell cover was estimated and the data shown in the stem-and-leaf diagram below.

**Percentage Bluebell Cover**

| Strip 1 | | Strip 2 |
|---:|:---:|:---|
| 8 8 | 0 | |
| 5 5 5 5 0 0 | 1 | 3 5 5 5 5 |
| 5 5 5 5 0 0 0 0 0 | 2 | 0 0 0 4 5 7 |
| | 3 | |
| 0 | 4 | 0 5 5 9 |
| | 5 | 5 |
| 5 | 6 | 0 7 |
| 5 | 7 | 0 |
| | 8 | 0 |

4|7 represents 47%

$n = 20$ $\qquad$ $m = 20$

(a) By considering the stem-and-leaf diagram, explain why it might be appropriate to conduct a Mann-Whitney test on this data. **2**

The rank sum of the data from Strip 1 is 480.

(b) (i) Conduct the Mann-Whitney test to determine if there is evidence, at the 10% level, that there is a difference in the median percentage bluebell cover between Strips 1 and 2.

(ii) How would your conclusion have been different if the test was performed at the 5% level of significance? Comment on this difference in the context of the biological study.

(iii) How might the precision of the results for percentage cover in a strip be improved? **8**

[Turn over

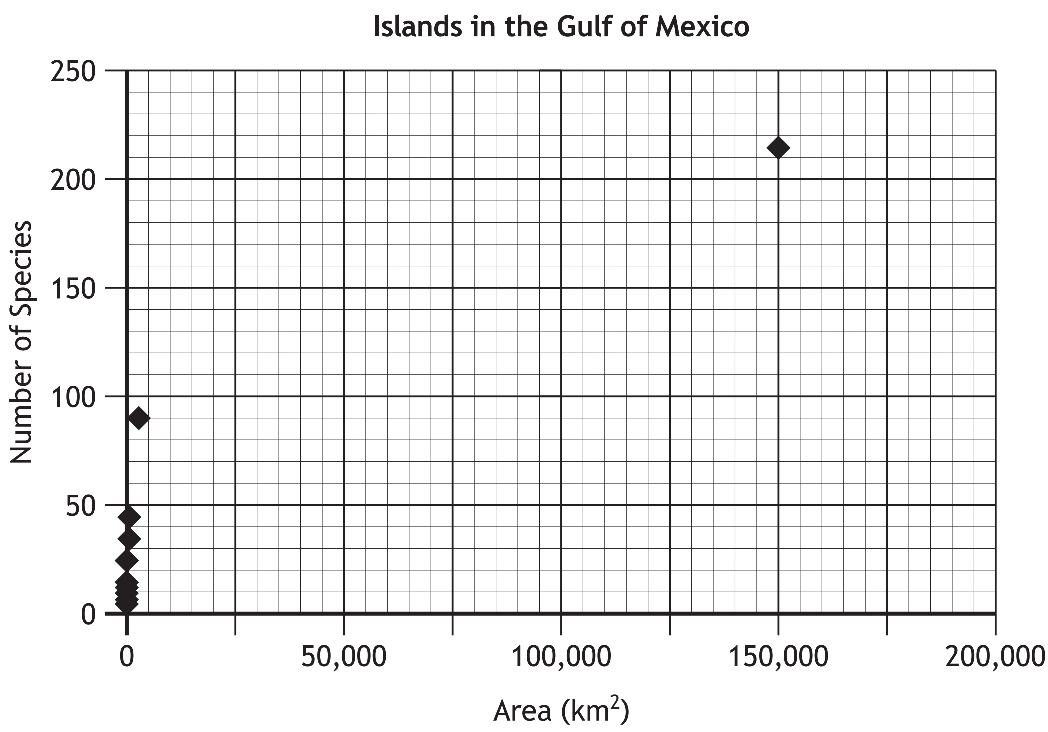
**13.** A biogeographical study in the islands of the Gulf of Mexico by Browne and Peck (1996) investigated the relationship between island area and the number of species of Cerambycidae (long-horned beetles) on the island.

Although Florida and South Florida are not strictly speaking islands, they were considered as such by Browne and Peck.

The data collected for the 13 islands is recorded below.

| **Island** | **Area ($km^2$)** | **Number of species** |
|---|---|---|
| Florida | 149 913 | 213 |
| South Florida | 5080 | 91 |
| Key Largo | 55·1 | 44 |
| Matecumbe Key | 4·3 | 16 |
| Fat Deer Key | 3·7 | 12 |
| Key Vaca | 2·9 | 15 |
| No Name Key | 3·1 | 16 |
| Big Pine Key | 17·1 | 24 |
| Big Torch Key | 2·3 | 16 |
| Cudjoe Key | 9·2 | 8 |
| Sugarloaf Key | 10·2 | 10 |
| Key West | 11·9 | 24 |
| Dry Tortugas | 0·9 | 3 |

The scatterplot obtained from this data is shown below.




(a) Make two comments on this scatterplot. **2**

**13. (continued)**

The researchers transformed the data by taking $\log_{10}$ of both variables and this produced the figures below. A scatterplot of this transformed data showed a clear indication of a linear relationship.

| **Island** | $\log_{10}$ **(Area)** $(x)$ | $\log_{10}$ **(Number of species)** $(y)$ |
|---|---|---|
| Florida | 5·1758 | 2·3283 |
| South Florida | 3·7058 | 1·9590 |
| Key Largo | 1·7411 | 1·6434 |
| Matecumbe Key | 0·6334 | 1·2041 |
| Fat Deer Key | 0·5682 | 1·0791 |
| Key Vaca | 0·4623 | 1·1760 |
| No Name Key | 0·4913 | 1·2041 |
| Big Pine Key | 1·2329 | 1·3802 |
| Big Torch Key | 0·3617 | 1·2041 |
| Cudjoe Key | 0·9637 | 0·9030 |
| Sugarloaf Key | 1·0086 | 1·0000 |
| Key West | 1·0755 | 1·3802 |
| Dry Tortugas | −0·0457 | 0·4771 |

Taking $x$ as $\log_{10}$ (Area) and $y$ as $\log_{10}$ (Number of species), the following statistics are obtained:

$$\sum x = 17{\cdot}375,\ S_{xx} = 26{\cdot}2676,\ S_{yy} = 2{\cdot}6395,\ S_{xy} = 7{\cdot}5234$$

and the equation of the least squares regression line of $y$ on $x$ is found to be

$$y = 0{\cdot}9202 + 0{\cdot}2864x.$$

(b) Calculate the coefficient of determination for the transformed data and comment on its value in the context of this biogeographical study. **2**

Researchers were unable to visit an island called Tepui Key, but were able to calculate its area to be 350 km$^2$, which gives an $x$-value of $\log_{10} 350 = 2{\cdot}5441$.

(c) In this individual case, calculate, with 90% confidence, the maximum and minimum number of species on Tepui Key that the researchers might have expected to find. **7**

(NB: $\log_{10} p = q$ can be written as $10^q = p$)

**[END OF QUESTION PAPER]**

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE