AH National Qualifications 2024

X803/77/12

Statistics Paper 2

WEDNESDAY, 1 MAY

1:30 PM – 4:15 PM

Total marks — 90

Attempt ALL questions.

You may use a calculator.

To earn full marks you must show your working in your answers.

State the units for your answer where appropriate.

Write your answers clearly in the spaces provided in the answer booklet. The size of the space provided for an answer is not an indication of how much to write. You do not need to use all the space.

Additional space for answers is provided at the end of the answer booklet. If you use this space 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.

SQA ©

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

MARKS

Total marks — 90

Attempt ALL questions

1. The numbers of wild birds visiting a garden feeding station during the same one hour time period were recorded on fourteen consecutive days, and are given below.

11 27 45 63 65 70 77 79 87 88 90 95 102 130

Calculate the upper and lower fences and hence identify any possible outliers, giving a reason for your answer. **3**

2. A commuter has a choice of two routes to work, route A or route B.

The probability of choosing route A is 0.2. The probability of being late for work if choosing route A is 0.65 and the corresponding probability for route B is 0.12.

(a) Calculate the probability that the commuter is late for work on any given day. **2**

(b) Given that the commuter is late for work, calculate the probability that they went via route B. **3**

3. A local council has modernised its bin lorries. They report that 88% of bins are now completely emptied on the first time of lifting.

(a) In a particular street there are 12 bins. Calculate the probability that exactly 9 bins are completely emptied on the first lift, stating the distribution used. **2**

(b) On a second street there are 48 bins. Using an appropriate approximation, determine the probability that more than 75% of the bins are completely emptied on the first lift. **4**

4. The grades attained by a group of students from a Mathematics examination are below.

Grade	A	B	C	D	E
Number of students	185	170	197	163	155

Perform a test at the 10% level of significance to assess the evidence that the attainment grade frequencies follow the uniform distribution, U(5). **6**

[Turn over

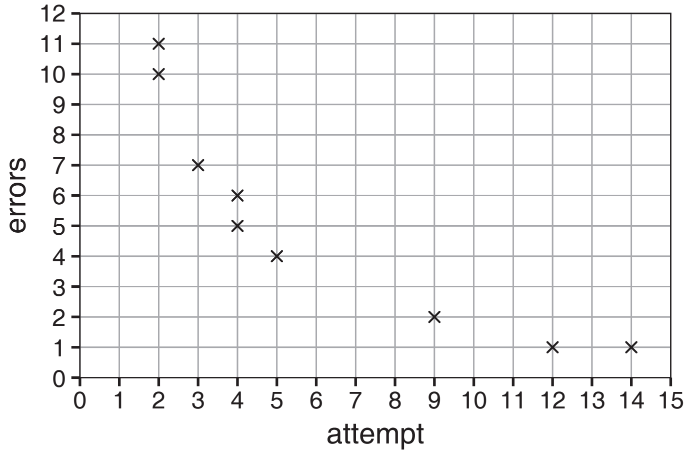
5. A psychology experiment was conducted in 1940 by Robert Tryon that examined the genetic differences in maze-learning ability in rats. He tested the proficiency of successive generations of rats at completing a maze by recording how many times they went down a blind alley (considered here to be an error) before finding the maze exit.

A researcher replicated some of Tryon's work with just one generation of rats.

Rats started at the maze entry point and each attempt ended when they either found the maze exit, or they returned to the entry point. Once a rat found the maze exit, their participation in the experiment was over, the number of errors they made along the way was recorded, and another rat began their attempts.

If a rat incorrectly left the maze by the entry point, it was recorded as a failed attempt, and the rat was given another attempt to find the maze exit.

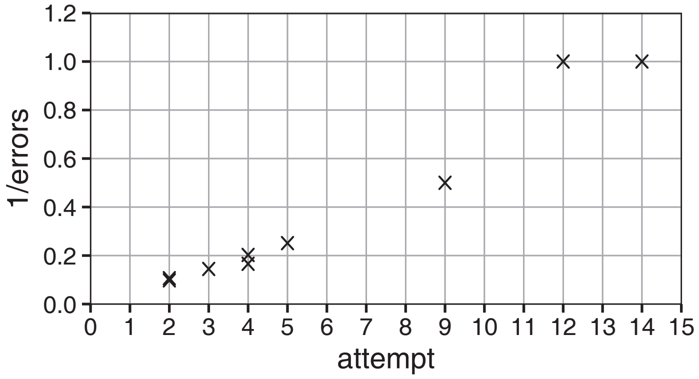
Data was recorded for the 9 rats that found the maze exit, and is shown below. For example, the rat that took 14 attempts to find the maze exit, only made one error on that 14th attempt.



(a) Describe the relationship between the number of errors, y, and the number of attempts, x. **2**

The data was transformed by calculating $\frac{1}{\text{errors}}$, to give the scatterplot and summary statistics shown on the next page.

5. (continued)



$$n = 9 \quad \sum x = 55 \quad \sum x^2 = 495 \quad S_{xx} = 158.89 \quad s = 0.078$$

$$\sum \frac{1}{y} = 3.45 \quad \sum \frac{1}{y^2} = 2.42 \quad \sum x\frac{1}{y} = 34.03$$

(b) (i) Calculate the least squares regression equation for the transformed data. **4**

The researcher was interested in knowing the expected number of errors that would be made by a single rat that completed the maze on its seventh attempt.

(ii) Calculate the appropriate 95% interval, either prediction or confidence, that would help answer the researcher's question. **5**

6. A shop has a large number of unsold Christmas crackers from last December and plans to sell them this year if they believe that 75% are working properly. A random sample of 20 is taken and tested, resulting in 14 working properly, meaning they make a snapping sound when pulled.

(a) Calculate a 99% confidence interval for the proportion of Christmas crackers that work properly, stating two assumptions that you have made. **5**

(b) Comment, with reason, on the belief that 75% of the unsold Christmas crackers work properly. **2**

[Turn over

7. A random variable X has the distribution $X \sim \text{U}(8)$.

A second random variable Y is defined as $Y = 3X - 2$.

Find the mean and standard deviation of Y. **5**

8. A house owner decides to grow the same type of plant from seed in two different areas of a large garden. They wish to find out if there is an area in the garden where the type of plant grows higher. After the plants have grown for two months, a random sample is taken from each area and the following statistics are obtained for plant height (cm).

	n	$\bar{x}$	s
Area 1	9	38.7	3.42
Area 2	11	36.9	3.51

Perform a hypothesis test to determine if there is significant evidence that the mean height of this type of plant is greater in Area 1, stating the underlying assumptions of the test used. **9**

9. A rally car is being prepared for a desert rally of length 6000 miles. A specific component on the car is known to last for a mean of 2400 miles. The driver fits a new component before the start of the rally.

(a) Calculate the parameter of the Poisson distribution that models the number of failures of this specific component during the rally, and state an assumption behind using this distribution. **2**

(b) Calculate the probability that the specific component does not fail during the rally. **2**

The driver carries spares of the specific component to replace any that fail during the rally, in order to keep going.

(c) Determine the minimum number of spares of the specific component the driver should carry to be at least 90% sure of having enough spares to complete the rally. **2**

10. A random variable is normally distributed with standard deviation 2.9.

A 90% confidence interval for the mean is to be constructed, with overall width less than 1.4.

Calculate the minimum sample size that should be taken. **5**

11. A university course collected information regarding the health and fitness of a random sample of 100 first year undergraduates currently enrolled at the university. The results for the first six participants are shown in the table below.

Age (years)	Height (m)	Mass (kg)	BMI (kg/m^2)	Pulse rate (bpm)	Activity level	Smoker
18	1.87	73.2	20.9	63	active	yes
19	1.85	75.4	22.0	73	inactive	no
18	1.83	88.0	26.3	65	moderate	no
20	1.76	70.1	22.6	78	moderate	no
18	1.80	77.6	24.0	60	inactive	yes
19	1.67	61.2	21.9	73	active	no

The definitions for each variable are as follows:

Age (years): the age in complete years at their last birthday.

Height (m) and Mass (kg): measured by an electronic gauge for consistency.

BMI (kg/m^2): Body Mass Index, calculated by $\text{BMI} = \frac{\text{Mass}}{\text{Height}^2}$.

Pulse Rate (bpm): count in beats per minute.

Activity level: ‘inactive’ is defined as doing physical activity for less than 30 minutes per week; ‘moderate’ is defined as doing 30 to 60 minutes of physical activity per week; and ‘active’ is defined as doing physical activity for over 60 minutes per week.

Smoker: ‘yes’ if the participant currently smokes or vapes; ‘no’ if they do not.

(a) (i) Name the variables from the table which measure discrete data. **1**

(ii) If a hypothesis test was performed to determine whether there was a difference in the median heights of smokers and non-smokers, state the assumption that would need to be met. **1**

(b) Name the type of data required for a chi-squared test for association and hence name the variables from the table that could be used for such a test. **2**

(c) State a statistic that would be appropriate to calculate when investigating any relationship between pulse rate and BMI. **1**

[Turn over

12. A Scottish beekeeper distributes boxes containing jars of honey. The mass (grams) of each jar is N(69,6). The mass (grams) of honey in one jar is N(453,16).

If the mass of the contents of a box containing 48 jars of honey exceeds 25 kilograms, then health and safety regulations state that two people are required to lift the box.

(a) Stating any assumption required, calculate the probability that two people are required to lift the box. **6**

The committee of the local Farmer's Market is considering regulating the sale of jars of honey and they expect that the total mass (grams) of a jar should be modelled by $N(522,5^2)$.

The committee took a random sample of 10 jars from the Scottish beekeeper which had a mean mass of 527.5 grams.

(b) Conduct an appropriate parametric test with a 1% significance level to determine whether the mean mass equals 522 grams, stating one assumption required. **6**

13. The Quality Control Manager of a fizzy drink company monitors the sugar content in their drinks by taking a random sample of five one-litre bottles every hour during the hours of daily operation. The results are plotted on an x-bar chart with a historical mean 102 grams of sugar and standard deviation 0.13 grams in every litre of fizzy drink.

(a) Calculate the 2-sigma limits for the x-bar chart. **2**

One day the machine, from where the samples were taken, broke down and so repairs were made overnight. The first two sample means of sugar content obtained the next morning were 101.86 and 101.89.

(b) If the 1-sigma limits are 101.94 and 102.06 and the 3-sigma limits are 101.83 and 102.17, determine the range that the third sample mean must be within to ensure that the process is in control. **3**

14. The Central Limit Theorem is often used when performing statistical inference on populations from large, representative random samples that are at least size 20. These samples of independent observations typically come from populations that have an unknown distribution shape, but that have either known parameter values or ones where the estimates from the sample produce approximations that are good.

(a) Without using statistical notation, write down:

(i) the theoretical distribution of the sample mean resulting from the Central Limit Theorem **1**

(ii) the parameters of this distribution, in terms of the population parameters. **2**

It is widely agreed that the birth weights of full-term pregnancy babies have a normal distribution. In a study to determine the population distribution's mean and standard deviation, a group of researchers visited their nearest hospital maternity ward and recorded the birth weights of the first 25 babies that were delivered.

The Central Limit Theorem is either not necessary or not appropriate to be used in this study. Give two reasons why this is the case, by making reference to:

(b) (i) the distribution that the study was investigating **1**

(ii) the study design, or its underlying assumptions. **1**

[END OF QUESTION PAPER]

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE