HA060722368

Please check the examination details below before entering your candidate information

| Candidate surname | Other names |
|---|---|
| | |

Centre Number | Candidate Number

# Pearson Edexcel Level 3 GCE

## Wednesday 19 June 2024

Morning (Time: 2 hours)

Paper reference **9ST0/03**

# Statistics

## Advanced

## PAPER 3: Statistics in Practice

**You must have:**
Statistical formulae and tables booklet
Calculator

Total Marks

**Candidates may use any calculator allowed by Pearson regulations.**
**Calculators must not have retrievable mathematical formulae stored in them.**

## Instructions

- Use **black** ink or ball-point pen.
- If pencil is used for diagrams/sketches/graphs it must be dark (HB or B).
- **Fill in the boxes** at the top of this page with your name, centre number and candidate number.
- Answer **all** questions and ensure that your answers to parts of questions are clearly labelled.
- Answer the questions in the spaces provided
  – *there may be more space than you need.*
- You should show sufficient working to make your methods clear. Answers without working may not gain full credit.
- Unless otherwise stated, inexact answers should be given to three significant figures.
- Unless otherwise stated, statistical tests should be carried out at the 5% significance level.

## Information

- A booklet 'Statistical formulae and tables' is provided.
- There are 6 questions in this question paper. The total mark for this paper is 80.
- The marks for **each** question are shown in brackets
  – *use this as a guide as to how much time to spend on each question.*

## Advice

- Read each question carefully before you start to answer it.
- Try to answer every question.
- Check your answers if you have time at the end.
- If you change your mind about an answer, cross it out and put your new answer and any working underneath.

*Turn over* ▶

P75706A


F:1/1/1/1/

P 7 5 7 0 6 A 0 1 2 0

Pearson

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

**BLANK PAGE**

**Answer ALL questions. Write your answers in the spaces provided.**

**1** The stem and leaf diagram below shows the results of a test sat by 15 students.

| | |
|---|---|
| 4 | 5 |
| 5 | 1 8 8 |
| 6 | 1 3 7 |
| 7 | 1 1 3 5 5 9 |
| 8 | 3 |
| 9 | 1 |

**Key**

4 | 5 represents 45

A lower outlier is defined as a value < LQ – 1.5 IQR.

An upper outlier is defined as a value > UQ + 1.5 IQR.

Show that there are no outliers for the given data set.

**(6)**

using calculator: $Q_1 = 58$, $Q_3 = 75$ so IQR = 75 − 58 = 17

Lower boundary: $58 - 1.5 \times 17 = 32.5$

upper boundary: $75 + 1.5 \times 17 = 100.5$

since all data points are between 32.5 and 100.5, there are no outliers in this data set.

**(Total for Question 1 is 6 marks)**

2 Cats that have diabetes are treated with insulin. It is believed that insulin might cause a growth disease in cats.

Cats, that are being treated with insulin, routinely have their level of IGF-I, a growth hormone, measured to see if the insulin is causing any side effects.

A random sample of 23 cats with diabetes was obtained and these cats were put into three categories

- short-term, those who have received insulin for 31 days or less
- medium-term, those who have received insulin for 32 days to 14 months
- long-term, those who have received insulin for more than 14 months

Routine measurements of the levels of IGF-I (nanomoles per litre) in these cats are given in **Figure 1**

| IGF-I measurements | | |
|---|---|---|
| **Short-term** | **Medium-term** | **Long-term** |
| 51 | 74 | 189 |
| 49 | 52 | 163 |
| 30 | 49 | 158 |
| 28 | 48 | 144 |
| 26 | 45 | 142 |
| 25 | 40 | 109 |
| 24 | 38 | 88 |
| 23 | | 88 |
| Total 256 | 346 | 1081 |

1683

[Source: https://journals.sagepub.com/doi/pdf/10.1016/j.jfms.2004.01.002]

**Figure 1**

The data produced the following summary statistic

$$\sum x_i^2 = 182\,569$$

Viraj decides to carry out one-factor ANOVA to investigate whether there is any difference between the mean IGF-I measurement for the different lengths of time the cats received insulin.

(a) Complete Viraj's hypothesis test.

You should make any necessary assumptions.

**(10)**

$\Sigma\Sigma x^2 = 182569$ $T = 1683$ $n = 23$

$T_S = 256$ $T_M = 346$ $T_L = 1081$

$n_S = 8$ $n_M = 7$ $n_L = 8$

$$SS_T = \sum\sum x^2 - \frac{T^2}{n} = 182569 - \frac{1683^2}{23} = 59417.30435$$

$$SS_B = \sum \frac{T_i^2}{n_i} - \frac{T^2}{n} = \frac{256^2}{8} + \frac{346^2}{7} + \frac{1081^2}{8} - \frac{1683^2}{23} = 48212.71506$$

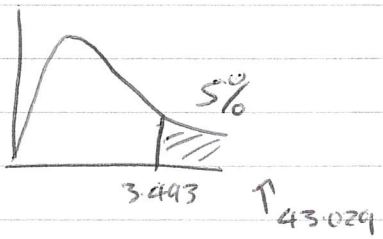
| | SS | $\nu$ | MS | F |
|---|---|---|---|---|
| Between Groups | 48212.71506 | 2 | 24106.35753 | 43.029 |
| Error | 11204.58929 | 20 | 560.2294645 | |
| Total | 59417.30435 | 22 | | |

$H_0$: $\mu_S = \mu_M = \mu_L$
$H_1$: At least 2 means are different

Use a one-way ANOVA at the 5% level using $\nu_1 = 2$ $\nu_2 = 20$

Test stat: 43.029

Critical region:




Result significant
Reject $H_0$
There is significant evidence to suggest a difference in at least two mean IGF-I levels between insulin durations.

**Question 2 continued**

(b) Suggest **two** blocking factors that could be used to improve the test in (a)

(2)

- Age of the cats
- Breed of the cats
- Gender of the cats

etc.

(c) State **two** assumptions required for the test in (a) to be valid.

(2)

The population of IGF-I levels between durations
- are normally distributed
- are independent of each other
- have the same population variance

(d) With reference to your answer in (c), give **one** reason why you would recommend that Viraj does **not** use one-factor ANOVA to investigate the data in Figure 1.

(1)

- The variances of the 3 durations may not be the same (e.g. variance of short-term = 128.57, variance of long-term = 1350.4)

**(Total for Question 2 is 15 marks)**

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

**BLANK PAGE**

3 Ewa was interested in the risk of Lyme disease in Poland. Ewa carried out an experiment to investigate the risk of Lyme disease in various locations. Lyme disease in humans is caused by a parasite called a 'tick'.

[Source: https://www.aaem.pl/pdf-71804-9030?filename=Risk%20of%20Lyme%20disease%20at.pdf]

Ewa compared two types of location.

- Timber acquisition, where wood was actively harvested
- Growing of forest, where forests were being grown for future harvesting

To compare the two different locations, the number of ticks in an area of 1 $m^2$ was recorded for a number of randomly selected areas in each location.

The number of ticks found in each of the randomly selected locations is shown in **Figure 2**

| rank | Number of ticks in 1 $m^2$ | | rank |
|---|---|---|---|
| | **Timber acquisition** | **Growing of forest** | |
| 9 | 15 | 16 | 10 |
| 7 | 11 | 12 | 8 |
| 6 | 8 | 7 | 5 |
| 2 | 3 | 5 | 4 |
| | | 4 | 3 |
| | | 1 | 1 |
| 24 | | Total | 31 |

**Figure 2**

(a) Explain why it would **not** be valid to analyse this data using a Wilcoxon signed-rank test.

(1)

The data are not paired and the two groups appear to be independent of each other.

Ewa says that a Wilcoxon rank-sum test should be used to investigate if there is a difference in average numbers of ticks between these two locations, rather than a *t*-test.

(b) Explain why Ewa might choose to use a Wilcoxon rank-sum test.

(1)

The number of ticks in each location may not be normally distributed.

**Question 3 continued**

(c) Conduct a Wilcoxon rank-sum test to investigate if there is a difference in average numbers of ticks per 1 $m^2$ between these two types of location.

**(8)**

Let X be the number of ticks in Timber acquisition and let Y be that for Grouting of forest.

$H_0$: The samples are taken from populations with identical distributions
$H_1$: $\eta_x - \eta_y \neq 0$

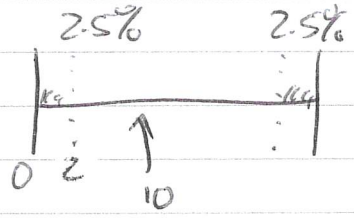
Use a 2-tailed Wilcoxon Rank-sum test at the 5% level using $n=4$, $m=6$

$T_x = 24$ $\quad$ $T_y = 31$

$U_x = 24 - \frac{4\times5}{2} = 14$ $\quad$ $U_y = 31 - \frac{6\times7}{2} = 10$

Test stat: $U=10$

Critical region:




Result not significant
Do not reject $H_0$
There is insufficient evidence to suggest, on average, the numbers of ticks per $1m^2$ differs between the two types of location.

**(Total for Question 3 is 10 marks)**

4 Steven uses a delivery company for his business. The company classifies packages as 'small' if they weigh up to a maximum of 2 kg

Steven sells pumpkins online.

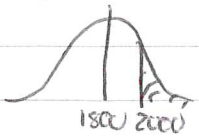
The weight of a packaged pumpkin, harvested after 20 weeks of growth, is assumed to be normally distributed with a mean of **1.8 kg** and a standard deviation of **100 g**

(a) Find the proportion of such packaged pumpkins that would be too heavy to be sent as 'small'.

**(1)**

Let $X$ be the weight of a packaged pumpkin after 20 weeks.

$X \sim N(1800, 100^2)$

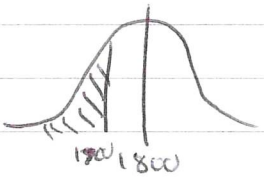



$P(X > 2000) = 0.0228$ (using calculator)

(b) Find the probability that, for a sample of 10 such packaged pumpkins, selected at random, their **mean** weight is below 1.7 kg

**(2)**

$\bar{X} \sim N(1800, \frac{100^2}{10})$




$P(\bar{x} < 1700) = 0.000783$ (using calculator)

**Question 4 continued**

Steven believes that if pumpkins are harvested after **19** weeks, then the mean weight of a packaged pumpkin will be lower than 1.75 kg

(c) Investigate Steven's belief, given that a sample of 10 packaged pumpkins harvested after 19 weeks is found to have a mean weight of 1.7 kg

You may assume that the standard deviation for the weights of these packaged pumpkins is also 100 g

**(5)**

Let $Y$ be the weight of a packaged pumpkin after 19 weeks

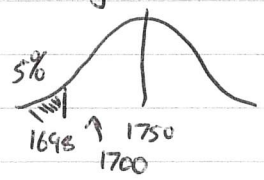
$H_0: \mu = 1750$

$H_1: \mu < 1750$

Use a 1-tailed z-test at the 5% level using

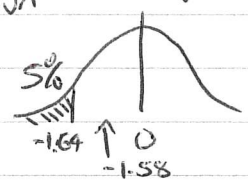
$\bar{Y} \sim N(1750, \frac{100^2}{10})$

TS: $\bar{y} = 1700$




$Z \sim N(0,1)$

TS: $\frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{1700 - 1750}{\frac{100}{\sqrt{10}}} = -1.58$




p-values

$P(\bar{Y} < 1700)$

$= P(Z < -1.58)$

$= 0.0569 > 5\%$

Result not significant. Do not reject $H_0$. Insufficient evidence to suggest the mean weight of a pumpkin after 19 weeks is lower than 1.7kg.

Steven obtained the sample of 10 pumpkins in part (c) from the corner of the field nearest to his house.

(d) By considering an assumption that had to be made about the sample in order for the investigation in (c) to be valid, comment on the validity of your conclusion in (c)

You should state the assumption made.

**(2)**

To use a z-test, the sample must be random.

Since Steve only collected pumpkins from 1 corner of the field, the sample is not random so the conclusion may not be valid.

**Question 4 continued**

(e) Define a parameter and a statistic.

Give an example of each in the context of this question.

(4)

A parameter is a numerical property of the population.
In this question, this could be the population mean weight of a packaged pumpkin after 20 weeks, 1.8 kg, or the population standard deviation of 100 g.

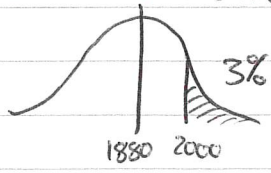
A statistic is a number calculated from a sample using no unknown parameters.
In this question, this could be the mean weight of a sample of pumpkins harvested after 19 weeks, 1.7 kg.

Steven also sells melons, whose weights may be assumed to be normally distributed, with a mean weight of 1880 g

(f) Given that 3% of the melons weigh more than 2 kg, determine the standard deviation of the weight of a melon.

(3)

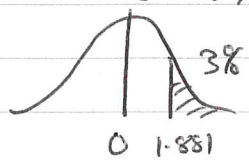
Let $M$ be the weight of a Melon. $M \sim N(1880, \sigma^2)$




Standardise: $Z = \frac{x-\mu}{\sigma}$ or $x = \mu + z\sigma$

$2000 = 1880 + z\sigma$

Sketch $Z \sim N(0,1)$




Solve $2000 = 1880 + 1.881\sigma$

$\sigma = 63.8$ g

**Question 4 continued**

A histogram of weights of melons is produced and is shown in **Figure 3**

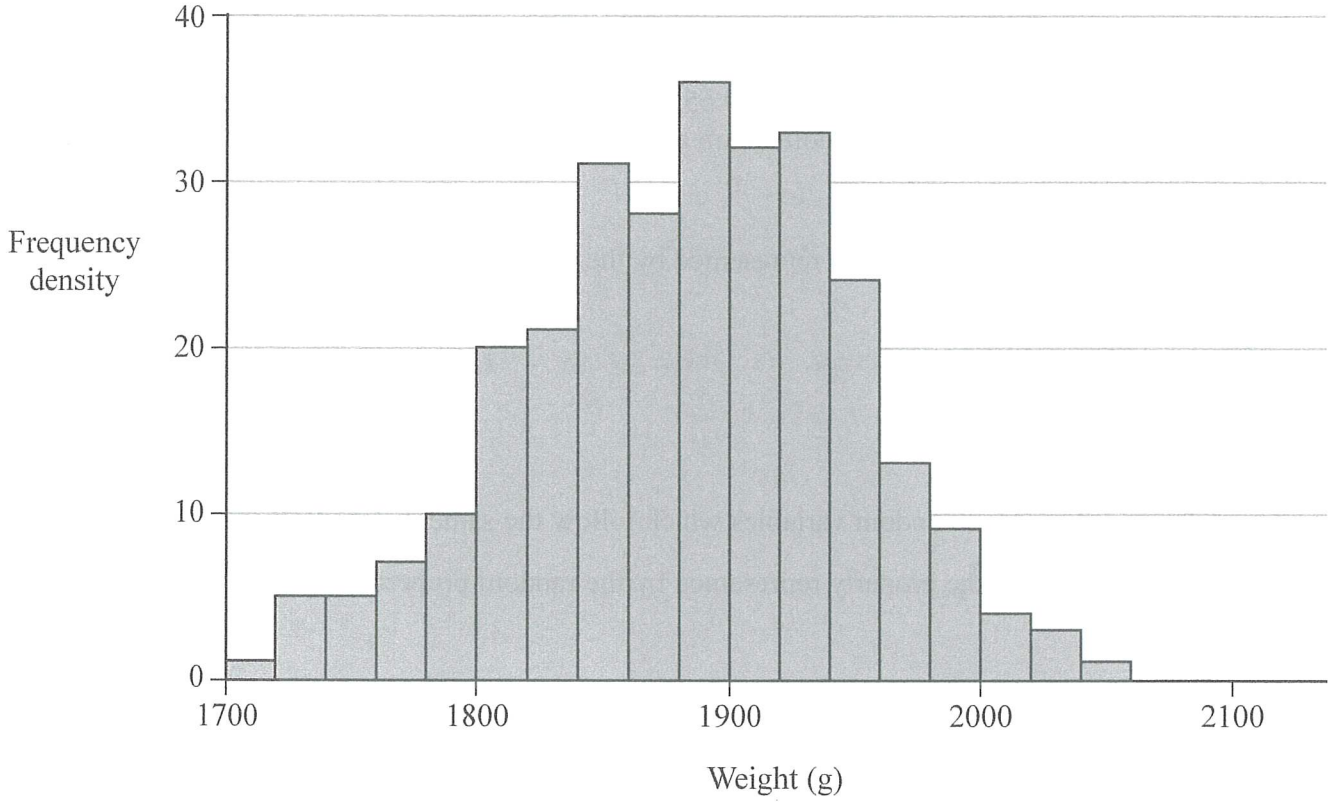



**Figure 3**

(g) Give **two** reasons why **Figure 3** suggests the normal distribution used in (f) is appropriate for modelling the weights of melons.

**(2)**

- Histogram looks like a bell-shaped curve
- Histogram looks symmetrical about 1880
- All observed data lies within 3 standard deviations of the mean (1688.6 and 2071.4)

etc.

**(Total for Question 4 is 19 marks)**

**5** Linus is interested in the US stock market, where it is possible to buy shares in companies.

Linus models the **change**, increase or decrease, in price of a share, on a randomly chosen day in 2023, as a normal distribution.

Using past data, Linus decides to use $X \sim N(0.65, 4.78)$ to model the change in price of a share in an electronics company, in dollars, on a randomly selected day.

[Source: www.nasdaq.com]

(a) Explain, in context, the property represented by the random variable $2X$

(1)

2X is double the change in price of a share in an electronics company on a random day

$X_1$ and $X_2$ are independent random variables which follow the same distribution as $X$

(b) Explain, in context, the property represented by the random property $X_1 + X_2$

(1)

$X_1 + X_2$ is the total change in price of a share in an electronics company on 2 randomly seleted days.

One working week for the stock market, consisting of Monday to Friday, is chosen at random from those in 2023

(c) Find the probability that, at the end of the chosen working week, the price of a share in the electronics company is at least five dollars more than it was at the start of that working week.

You may assume that changes in price are independent of day.

(3)

$E(X_1 + \ldots + X_5) = E(X_1) + \ldots + E(X_5)$

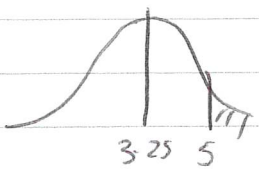
$= 5 \times 0.65 = 3.25$

$Var(X_1 + \ldots + X_5) = Var(X_1) + \ldots + Var(X_5)$

$= 5 \times 4.78 = 23.9$

So $X_1 + \ldots + X_5 \sim N(3.25, 23.9)$

$P(X_1 + \ldots + X_5 > 5) = 0.360$



**Question 5 continued**

Linus says the probability in part (c) is the probability that, during a randomly selected working week, the price of a share has changed by at least five dollars.

(d) Give **two** reasons why Linus is wrong.

(2)

- The probability is for an increase by at least 5 dollars, not a change
- The probability is for a share in that particular electronics company, not in general

The change in price of a share one day may impact the change in price of that share on the following day.

(e) Explain how this information impacts the assumption and the validity of the calculation in (c)

(2)

The calculation may not be valid since the change in price from day to day must be independent.

If the change on one day affects the change the next day, they are not independent.

Linus's estimates of the parameters were based on **only** data from January 2023

(f) Explain, with a reason, how Linus could improve the validity of his model.

(2)

- Extend the time frame to include all months in 2023 to account for seasonal changes in share prices
- Extend the time frame to include more years to account for time based variation like inflation.

etc.

**(Total for Question 5 is 11 marks)**

6 Josceline has a music player with 145 songs stored on it.

The shuffle function on Josceline's music player selects the next song to play by **randomly selecting** a song from all the songs stored on the music player.

Josceline's friend, Suji, has a different music player with 80 songs stored on it.

The shuffle function of Suji's music player creates a **list**, in **random order**, of all the songs stored on the music player, and then plays through this list. When it has played through all the songs, it repeats the process, creating a new random list.

(a) Identify the type of sampling that

(i) Josceline's shuffle function performs,

(1)

Unrestricted Random Sample

(ii) Suji's shuffle function performs.

(1)

Simple Random Sample

(b) Explain the difference between the two types of sampling identified in (a)(i) and (a)(ii)

(1)

Unrestricted random sampling allows repeats but simple random sampling does not all repeats.

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

DO NOT WRITE IN THIS AREA

**Question 6 continued**

(c) Explain how Josceline could use a random number generator to select a sample of 30 songs, from those on her music player.

(4)

- Number each song from 1 to 145
- Use a random number generator to generate numbers between 1 and 145 inclusive, ignoring repeats
- Stop when 30 different numbers are recorded
- Select the songs with these numbers.

(d) Find the probability that, when Josceline uses the shuffle function of her music player,

(i) the first and second songs played are the same,

(1)

$$\frac{1}{145}$$

(ii) the first four songs played are all different,

(2)

$$\frac{145 \times 144 \times 143 \times 142}{145 \times 145 \times 145 \times 145} = 0.959$$

(iii) the first 146 songs played are all different.

(1)

0

**Question 6 continued**

Suji claims that the shuffle function of her music player will never play the same song twice in a row.

(e) Explain why Suji is wrong.

(1)

At the end of the first list, a new list is generated. It is possible for the last song of the first list is the first song of the second list,

**Question 6 continued**

Of the 145 songs on Josceline's music player, 35 are by the same singer.

Josceline believes that, when using her shuffle function, songs by this singer come up more often than she would expect.

To investigate her belief, she counts the number of songs by this singer that appeared in the first 30 songs played with her shuffle function. She found that 11 of these songs were by the singer.

(f) Carry out a hypothesis test to investigate Josceline's belief.

**(5)**

expected chance $= \frac{35}{145} = \frac{7}{29}$

Let X be the number of songs by this singer to play.

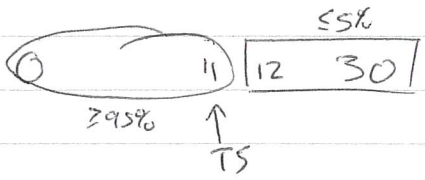
$H_0 : \pi = \frac{7}{29}$

$H_1 : \pi > \frac{7}{29}$

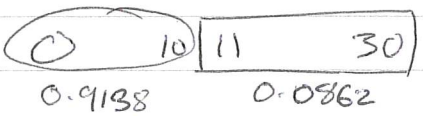
Use a 1-tailed proportion test at the 5% level using $B(30, \frac{7}{29})$

Test stat : 11

Critical region




or p-value




$P(x \leq 10) = 0.9138 < 0.95$

$P(x \leq 11) = 0.9604 > 0.95$

$0.0862 > 5\%$

Result not significant

Do not reject $H_0$

There is insufficient evidence to suggest the probability of a song by this singer appearing is more than expected.

**Question 6 continued**

Suji wants to carry out the same hypothesis test used in (f), using the first 30 songs played with her shuffle function.

(g) Explain why Suji's test would not be valid.

(2)

Since Suji's shuffle function is without replacement, the probability of songs by an artist appearing changes each time, so a Binomial distribution would not be valid.

**(Total for Question 6 is 19 marks)**

**TOTAL FOR PAPER IS 80 MARKS**

DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA DO NOT WRITE IN THIS AREA