National
Qualifications
2023 MODIFIED

X803/77/11

Statistics
Paper 1

FRIDAY, 19 MAY
9:00 AM – 10:00 AM

Total marks — 30

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.

MARKS

Total marks — 30

Attempt ALL questions

1. In an effort to understand the nutritional properties of food items sold by coffee shop chains in the UK, a random sample of food items was taken from a randomly chosen national coffee shop chain. Nutritional information such as the fat and calorie content of these food items were obtained. Each food item was also classified as either a 'bakery' item or a 'non-bakery' item.

Using these data, **Figure 1** and **Output 1** display graphical and numerical summaries of the observed fat values (in grams) and calorie values (in kilocalories) of the bakery and non-bakery items separately.

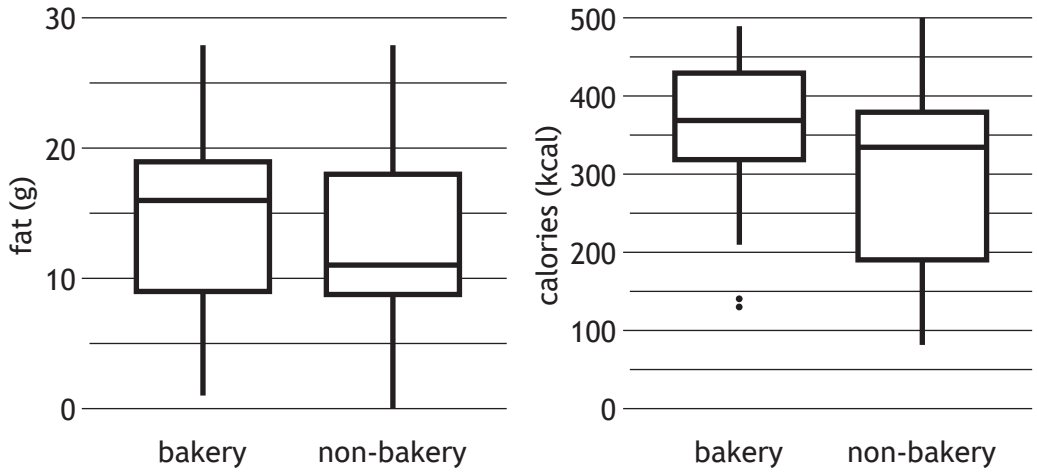

Figure 1 Boxplots of fat and calorie values by type of item (bakery and non-bakery).

Output 1

```
Bakery items (n=41)
       fat              calories
Min.    : 1.00    Min.    :130.0
1st Qu. : 9.00    1st Qu. :320.0
Median  :16.00    Median  :370.0
Mean    :14.56    Mean    :368.8
3rd Qu. :19.00    3rd Qu. :430.0
Max.    :28.00    Max.    :490.0
------------------------------
Non-bakery items (n=36)
       fat              calories
Min.    : 0.00    Min.    : 80.0
1st Qu. : 8.75    1st Qu. :190.0
Median  :11.00    Median  :335.0
Mean    :12.86    Mean    :304.7
3rd Qu. :18.00    3rd Qu. :380.0
Max.    :28.00    Max.    :500.0
```

(a) Comment on what **Figure 1** reveals about the location and spread of the distributions of fat contents of bakery and non-bakery items in this coffee chain. **2**

(b) State the criteria used to identify outliers. **1**

1. (continued) **MARKS**

(c) Using the information in **Output 1**, show that the two values identified as outliers in **Figure 1** are indeed outliers. **2**

(d) Suggest a valid reason for the removal of the values identified as outliers in **Figure 1**. **1**

Assuming that the distributions of fat values and calorie values were normally distributed, and that this sample was representative of all food items sold by the coffee shop chain, statistical inference was performed to compare the nutritional properties of all the bakery and non-bakery items in this coffee shop chain. The results are shown in **Output 2** and **Output 3**. One value has been deleted and replaced by `*****`.

Output 2

```
Two sample t-test comparing fat content of bakery and
non-bakery items
t = 1.0496, df = 75, p-value = *****
alternative hypothesis: true difference in means is not equal
to 0
sample estimates:
mean in group Bakery      mean in group Non-bakery
      14.56098                  12.86111
```

Output 3

```
Two sample t-test comparing calorie content of bakery and
non-bakery items
t = 2.7767, df = 75, p-value = 0.006931
alternative hypothesis: true difference in means is not equal
to 0
sample estimates:
mean in group Bakery      mean in group Non-bakery
      368.7805                  304.7222
```

(e) Determine the information that was not included in **Output 1** that would help validate the use of the t-tests reported in **Output 2** and **Output 3**. **1**

(f) Write down the null hypothesis being tested in **Output 2** and the null hypothesis being tested in **Output 3**. **1**

(g) Using **Table 3** in the Statistical Formulae and Tables booklet provided, approximate the p-value (to 4 decimal places) shown as `*****` in **Output 2**. **2**

(h) Referring to **Output 3**, interpret the p-value and comment on whether there would be an impact to your mean calorie intake if you consumed either bakery or non-bakery items from the coffee shop chain. **2**

Given that the coffee shop chain was selected at random from all the UK coffee shop chains, the data analysed above can be used to draw inferences about all the coffee shop chains in the UK.

(i) Name the sampling method that this scenario corresponds to, and describe how the sampling method was conducted. **3**

2. An extract of a report's first draft by a student of an agricultural college is given below.

It is known to contain some flaws and questionable methodology.

Read it and then answer the questions that follow.

Introduction

Field corn, which is also called maize, is an important crop for animal feed. In all crops, as the crop density increases, the crop yield also increases until a certain point. However, due to the availability of environmental resources such as water and nutrients, competition between the plants outstrips the supply of these resources. This inhibits plant growth and the crop yield starts to decrease.

The aim of this study is to investigate the relationship between the crop density of one type of field corn that is planted on college grounds and the resulting crop yield in order to maximise the crop yield for the college's animals. The crop density is measured by counting the number of plants per unit area, whilst the crop yield is the total weight of the harvested corn which is then converted into grams per unit area.

Method

The field chosen for this study had consistent environmental conditions across it such as soil type, pH levels, drainage, altitude, and aspect (the compass direction the field faces); and had no crops grown on it the previous year. The field was split into 39 non-overlapping plots. Each plot was assigned a number from 1 to 39 and each of 13 different crop densities of field corn was allocated three of these plots through random selection. Each plot was planted on the same day with their specific crop density of corn plants, and the yield from all plots in the field were obtained on the same day of harvest. When fertiliser was applied, it was spread on the same day and was spread evenly across all plots. After the plots were harvested later in the year, stratified sampling was applied to choose one crop yield from each of the 13 different crop densities. The crop yield from these plots was weighed before being converted into a gram per square metre measurement.

Data and analysis

The plot number for each crop density, together with that crop density and corresponding crop yield after harvest is shown in **Figure 1**.

Plot number	9	31	14	1	24	32	26	38	12	17	22	6	36
Crop density	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8
Crop yield	538	585	622	676	721	763	806	861	920	960	1019	1074	1137

Figure 1 Plot number, crop density (plants/m^2) and crop yield results (g/m^2)

A scatterplot of the crop density and crop yield data was generated and is shown in **Figure 2**.

2. (continued)

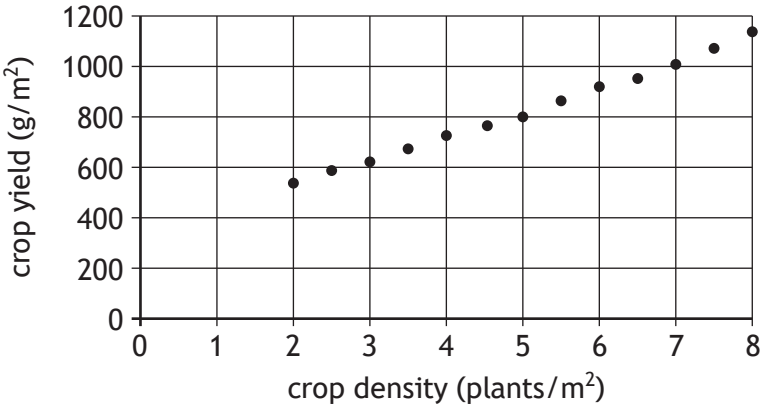

Figure 2 Scatterplot of crop density and crop yield

A least squares linear regression model was fitted to the data, giving the equation $y = 98.626x + 328.56$ where x is the crop density and y is the crop yield.

A residual plot for the crop yield data was then generated to give the graph in **Figure 3**.

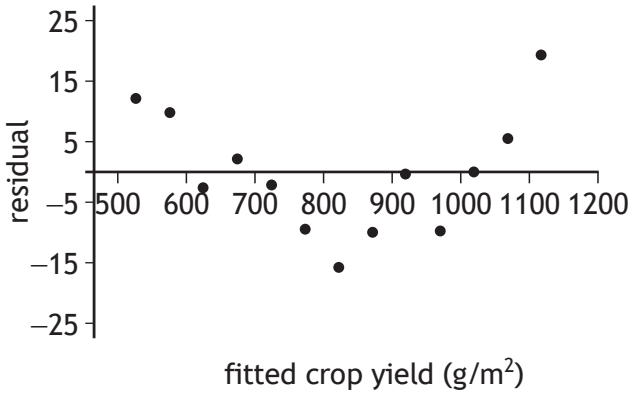

Figure 3 Residual plot for crop yield

35 After looking at the residual plot, a transformation was applied to the original data by taking the square root of the crop yield to give the following data in **Figure 4** below.

Crop density	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8
$\sqrt{\text{crop yield}}$	23.2	24.2	24.9	26.0	26.9	27.6	28.4	29.3	30.3	31.0	31.9	32.8	33.7

Figure 4 Crop density (plants/m^2) and $\sqrt{\text{crop yield}}$ $\left(\sqrt{\text{g}}\,/\,\text{m}\right)$ data, rounded to 1 d.p.

The scatterplot for the data in **Figure 4** was generated (not shown here) and again suggested a linear relationship. The following statistics were therefore calculated for crop density (x) and $\sqrt{\text{crop yield}}$ $\left(\sqrt{y}\right)$ using the exact values for $\sqrt{\text{crop yield}}$:

40 $\sum x = 65 \quad \sum \sqrt{y} = 370.2569 \quad S_{xx} = 45.5 \quad S_{\sqrt{y}\sqrt{y}} = 136.6022 \quad S_{x\sqrt{y}} = 78.8165$

2. (continued)

A new least squares linear regression model was fitted to this transformed data. The residual plot for this new model is shown in **Figure 5**.

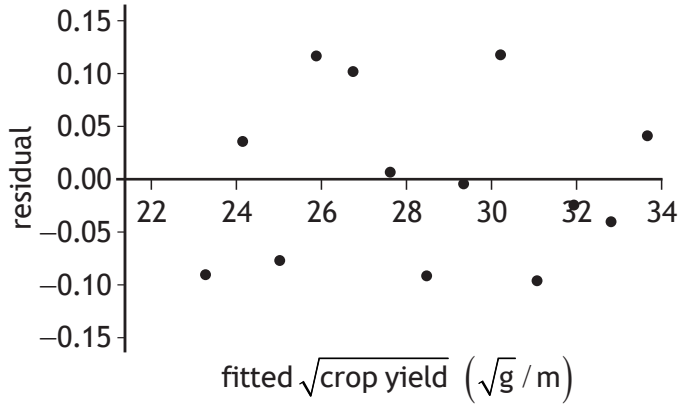

Figure 5 Residual plot for $\sqrt{\text{crop yield}}$

Conclusion

Based on the residual plots, the linear model for crop yield and crop density is a good fit for all crop densities for this type of field corn in the college's field.

Study the scatter plot in **Figure 2** and the residual plot in **Figure 3**.

(a) (i) Explain why the scatter plot in **Figure 2** suggests that a linear model would be suitable for the data. **1**

(ii) State how you would calculate a residual and explain what a residual measures. **2**

(iii) Fully explain why the residual plot in **Figure 3** suggests that the linear model that was fitted is not suitable. **2**

Review the data in **Figure 4**, lines 38 to 42 and the residual plot in **Figure 5** where the square root transformation was applied and the resulting data was analysed.

(b) (i) Using the statistics from line 40, calculate the correct least squares regression line of $\sqrt{y}$ on x. **3**

(ii) Using your least squares regression line from (b) (i), calculate the residual value for a crop density of 3.5 plants/m^2. Using the copy of **Figure 5** in the answer booklet, circle the point that relates to this fitted value and residual value. **3**

MARKS

2. (continued)

In order to identify the maximum crop yield, the college student wanted to know the crop density at which the crop yield might start to decrease.

(c) (i) Explain why it would be bad practice to use the least squares regression line for crop densities greater than 8 plants per square metre. **1**

(ii) Give a reason why the least squares regression line would not help the college student with the aim of the study. **1**

Read the concluding statement in lines 44 to 45.

(d) Re-write an improved concluding sentence in terms of what the investigation between crop density and crop yield actually found. **2**

[END OF QUESTION PAPER]

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE