AH National Qualifications 2024

X803/77/11

Statistics Paper 1

WEDNESDAY, 1 MAY

12:00 NOON – 1:00 PM

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.

SQA ©

X8037711

A/HTP

[OPEN OUT]

DO NOT WRITE ON THIS PAGE

Total marks — 30

Attempt ALL questions

1. An extract of a draft report by a researcher is given below.

It is known to contain some flaws and questionable methodology.

Read it and then answer the questions that follow.

Introduction

The British Board of Film Classification awards an age rating for every film that is released in cinemas in the UK. Since around 2010, online film streaming companies have been increasing their share of the film viewing market and this may be having an effect on the age ratings of films that are released for the viewing public. This report investigates whether there has been any change in the age ratings of UK films over recent years.

Method

A data set of the release year and age rating of films was required. In addition, it was known that age ratings have changed over the years with the introduction of categories such as '18' and removal of categories such as 'X'. Therefore a degree of data cleaning was required in order to create a dataset using the current film age ratings of 18, 15 and the 'family friendly' age ratings of 12, Parental Guidance (PG) and Universal (U).

The intended population was all feature films released in the UK between 1980 and 2019 inclusive, and a sample was taken from a leading online film database. The database listed over 22 000 films worldwide for this time period. However, without advanced programming skills, analysis of this full data set was not possible. I therefore chose to obtain a random sample of worldwide films from every year, from 1980 to 2019 inclusive. This involved searching in the database for films released each year and then randomly selecting films from the pages of results. This was done for each individual year, one at a time. Due to time constraints, I opted for a sample size of 5 per year, giving 50 films per decade and therefore a total of 200 films in all.

Analysis

Table 1 shows the sample data of age ratings, grouped by decade.

Output 1 shows the result of running a chi-squared test for association on **Table 1**.

Table 1

	Age rating					
Decade	**U**	**PG**	**12**	**15**	**18**	**Total**
1980s	3	8	1	21	17	50
1990s	2	7	5	20	16	50
2000s	2	5	8	31	4	50
2010s	2	3	14	26	5	50

Output 1

```
Number of cases in table: 200
Test for independence:
Chisq = 32.71, df = 12, p-value = 0.001075
```

1. (continued)

This first analysis suggests even at the 1% level that the age rating of films is not independent of the decade, which means that the proportion(s) of films in at least one of the age categories must have changed over time.

From looking at the data in **Table 1**, it would appear that the films with age ratings 12 and 18 have trends that are moving in opposite directions. With the growth of online film streaming from 2010, it was decided to look at the proportions of films with each of these age ratings from before and after 2010 to determine if there had indeed been a change. **Table 2** summarises the subset of data to be analysed.

Table 2

	Age rating		
Decade(s)	**12**	**18**	**Total all films**
1980s to 2000s	14	37	150
2010s	14	5	50

From **Table 2**, looking at the age rating 12 films, the z-test statistic for whether there had been an increase in proportion of films between the two time periods was 3.294. For the age rating 18 films, the z-test statistic for whether there had been a decrease in proportion of films between the two time periods was -2.205. With the 5% critical values being ± 1.64, we have evidence to reject the null hypothesis in both cases.

<u>Conclusion</u>

This report has demonstrated that the numbers of films in every age rating category have changed over recent decades. The proportion of films that were age rated 12 has increased since 2010 whilst the proportion of films that were age rated 18 has decreased since 2010. This shows that there is an increase in the number of ‘family friendly’ films being released for viewing in cinemas and ultimately available from online film streaming services.

MARKS

1. (continued)

(a) Read lines 13 to 15.

Write down one concern that you have regarding the choice of the online film database for the issue being investigated. **1**

(b) (i) Read lines 16 to 21.

Explain why the random sampling method used would not be classed as stratified random sampling. **1**

(ii) Briefly explain how stratified random sampling of 1% of the population could have been performed to obtain such a sample. **2**

(c) (i) Look at **Output 1**.

State what the null and alternative hypotheses would be for this test. **1**

(ii) Describe the check which should have been performed before proceeding with calculating the chi-squared test statistic. **1**

(iii) If the check had not been satisfied, describe a process that might be carried out and give the resulting number of degrees of freedom that would arise from that process. **2**

(d) (i) Read lines 34 to 36.

Show the calculation that generates the z-test statistic of -2.205. **3**

(ii) Describe the check which should have been performed before proceeding with the hypothesis test for the population proportions, and comment on the consequences of this check not being satisfied. **2**

(e) Identify the two parts of the conclusion that are not supported by the report and write down what is incorrect about them. **2**

[Turn over

2. A clinical trial was conducted with 29 patients with a particular virus to assess the effectiveness of a new treatment.

Each patient was randomly allocated to receive a standard treatment (‘standard’) or a new treatment (‘new’). All patients were monitored for 3 months.

The primary outcome was each patient’s viral load after 3 months. Viral load is the number of particles of the virus in a blood sample, measured as the number of copies of the virus per millilitre of blood. A lower viral load is a better outcome than a higher viral load. These data are shown in **Table 1**.

Table 1 Viral loads after 3 months for 29 patients randomly allocated to either a standard treatment or a new treatment

Standard	New
7500	410
8000	250
2000	800
550	1400
1250	7900
1000	7400
2250	1020
6800	6000
3400	920
6300	1420
9100	2700
970	4200
1040	5200
670	4100
400	

 MARKS

Figure 1 shows histograms of the viral loads of the two treatments separately. Results from an analysis of the data using statistical software are shown in the outputs below.

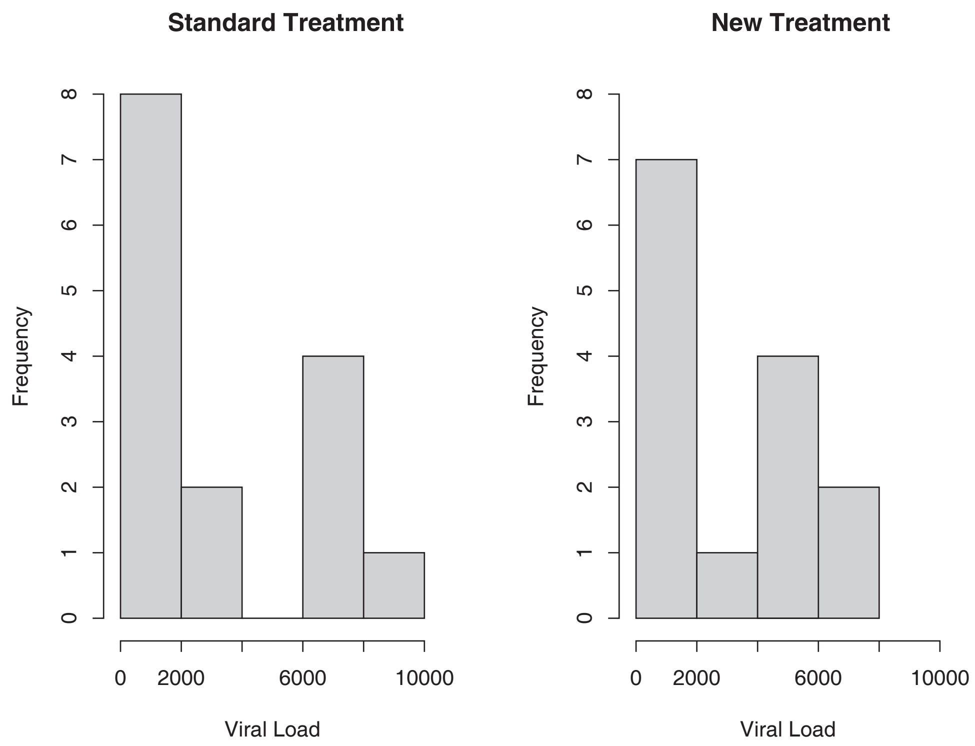

Figure 1 Histograms of the viral loads of the two treatments

It is proposed that a two sample z-test could compare the central locations of the distributions of the viral loads of the standard and new treatments using the data in **Table 1** which is displayed in **Figure 1**.

(a) State two reasons why it would not be appropriate to use a two sample z-test on these data. **2**

Output 1 presents the truncated results of a Mann-Whitney test applied to the data in **Table 1**.

Output 1

```
Mann Whitney test
data: Viral Load by Treatment
W = 205
alternative hypothesis: true location shift is not equal to 0
```

(b) (i) Write down the hypotheses that **Output 1** can be used to test. **1**

(ii) Without performing any calculations, clearly describe how the test statistic W in **Output 1** was calculated. **3**

(iii) State the critical value of W that would be used with **Output 1** to perform the test at the 5% significance level. Hence, state and interpret the conclusion of the test at this level. **3**

[Turn over

2. **(continued)**

The 15 patients on the standard treatment remained on the standard treatment beyond the clinical trial and were monitored for a further 3 months to see if there was a difference in their viral loads.

Their viral loads at 3 months (ie the end of the clinical trial) and 6 months are shown in **Table 2** and **Output 2** presents the results of a Wilcoxon Signed Rank test applied to these data. In **Output 2**, one value has been deleted and replaced by *****.

Table 2 Viral loads for patients on the standard treatment at 3 months and again at 6 months

Patient ID	Month 3 viral load	Month 6 viral load
1	7500	7480
2	8000	8010
3	2000	2010
4	550	560
5	1250	1210
6	1000	980
7	2250	2180
8	6800	6790
9	3400	3370
10	6300	6260
11	9100	9060
12	970	960
13	1040	990
14	670	640
15	400	430

Output 2

```
Wilcoxon signed rank test
data: Month.3.Viral.Load minus Month.6.Viral.Load
differences (in order of Patient ID):
20 −10 −10 −10 40 20 70 10 30 40 40 10 50 30 −30
ranks (in order of Patient ID):
***** 3 3 3 12 6.5 15 3 9 12 12 3 14 9 9
W = 18
null hypothesis: true location shift is equal to 0
alternative hypothesis: true location shift is not equal to 0
```

MARKS

2. (continued)

(c) Determine the missing value in **Output 2**, indicated by *****. **1**

(d) Using **Output 2**, calculate the rank sums for both negative and positive differences and justify the value of the test statistic W. **2**

By consulting Table 7 on *page 15* in the Statistical Formulae and Tables booklet, it is possible to identify the smallest interval that would contain the p-value of the test statistic W in **Output 2**.

(e) (i) State the maximum and minimum values of this interval for the p-value. **1**

(ii) Hence, or otherwise, complete this hypothesis test at the 5% significance level, and write an appropriate conclusion in context. **2**

[END OF QUESTION PAPER]

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE