AH National Qualifications 2025

X803/77/11

Statistics Paper 1

WEDNESDAY, 28 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.

[BLANK PAGE]

DO NOT WRITE ON THIS PAGE

Total marks — 30

Attempt ALL questions

1. An extract of a draft investigation 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

When listening to people talk, I find that I become very aware of when they repeatedly say ‘erm’ or ‘um’ between words and sentences. These are known as ‘filler sounds’ and other people have done research on why people use them and what they might indicate.

I found an online video of a university lecturer describing how they had to change their teaching to a hybrid model of both ‘live’ and ‘remote’ delivery, due to the Covid-19 Pandemic. It was in this video that I noticed the lecturer used lots of filler sounds, which I ultimately found to be quite distracting, and this prompted my investigation.

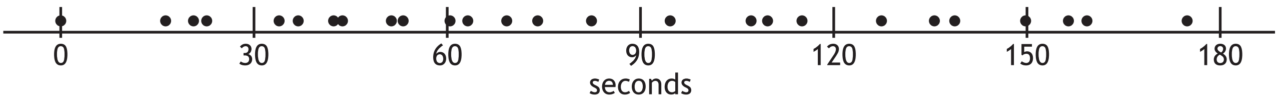
I watched the online video again and noted down every time the lecturer used a filler sound during their 14 minute talk. Interestingly, the very start of their talk began with a filler sound!

In this investigation I shall only explore the mean rate of the number of filler sounds spoken in a fixed time interval. I am interested to know whether this count could be modelled by a Poisson distribution.

Data

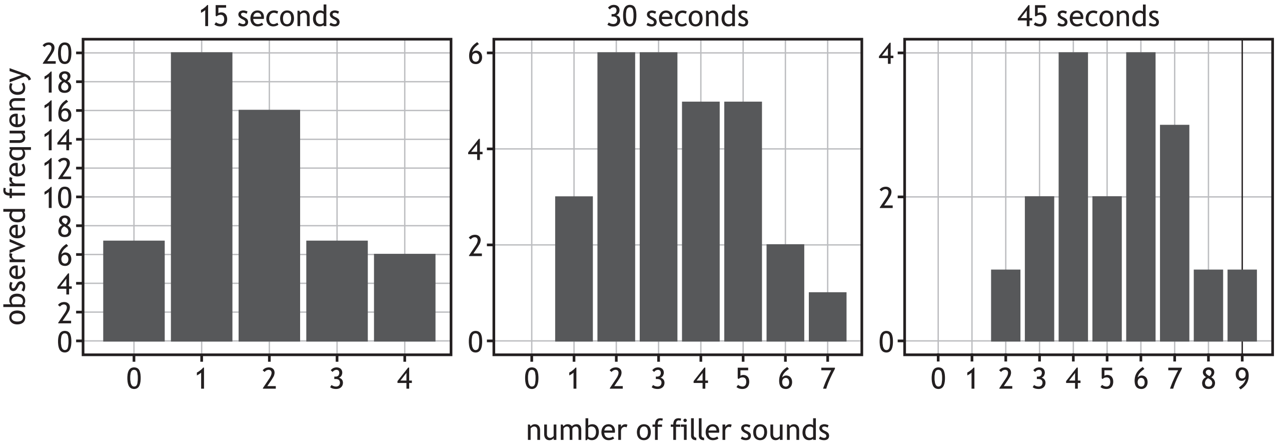
I recorded 97 instances of filler sounds during the 14 minute talk by the lecturer. A visualisation of when the filler sounds occurred during the first 3 minutes is shown in **Figure 1**. This shows that in the first 30 second interval, there were 4 filler sounds, then in the next 30 second interval there were 6 filler sounds, and so on.

Figure 1 Instances of filler sounds



I needed to divide the 14 minutes into smaller time intervals for my Poisson model and I considered time intervals of three different lengths. These are shown in **Figure 2**.

Figure 2 Distributions of number of filler sounds for time intervals of 15, 30 and 45 seconds



1. (continued)

I decided to proceed with using 30 second intervals and performed a chi-squared goodness-of-fit test.

Analysis

Defining X to be the number of filler sounds per 30 seconds, I summarised my recorded data in **Table 1** and calculated expected frequencies. However there were too many small values, so I had to combine categories, to give the values in **Table 2**.

Table 1

x_i	O_i	E_i
0	0	0.85
1	3	2.96
2	6	5.18
3	6	6.04
4	5	5.29
5	5	3.70
6	2	2.16
7	1	1.08

Table 2

x_i	O_i	E_i
0	0	0.85
1	3	2.96
2	6	5.18
3	6	6.04
4	5	5.29
5+	8	6.94

The data in **Table 2** generated a chi-squared statistic of 1.16 which has a p-value of 0.885. I was therefore able to conclude that the number of filler sounds did follow a Poisson distribution, using a 10% level of significance.

I was then curious to know whether this would hold true for another lecturer from another video. I repeated a similar analysis for a second lecturer who was talking on the same topic as the first lecturer for a similar length of time. Their mean rate of filler sounds per 30 seconds was 5.6, and it also seemed to follow a Poisson distribution.

I wanted to know whether this was statistically significantly more than the first lecturer. I considered performing either a two-sample z-test or t-test on the difference of their means, using the sample sizes of 28 for each lecturer, but I judged that a common assumption required for those tests was not likely to be satisfied.

Conclusion

To conclude, it appears that when some people include a noticeable number of filler sounds in their speech, the frequency of these filler sounds could be argued to follow a Poisson distribution, but this may not be the same rate for every person.

1. (continued)

Read lines 12 to 14.

(a) Clearly describe the three assumptions required to use the Poisson distribution as a model, in this context. **3**

Look at **Figure 2**.

(b) Describe an improvement to the graphs that should be made that would allow them to be more easily compared. **1**

Look at the first two columns of **Table 1**, labelled x_i and O_i.

(c) Calculate the mean rate for the Poisson distribution of X. **1**

Look at **Table 2**.

The researcher correctly used the rounded mean rate of 3.5 to calculate the expected frequencies.

For x_i corresponding to 5+, the expected frequency of 6.94 has been incorrectly obtained from the summation of the last three expected frequencies in **Table 1**.

(d) Calculate the expected frequency that the researcher should have obtained. **2**

Read lines 28 to 30.

(e) Determine the number of degrees of freedom that would have been used for this test, justifying your answer. **2**

Read lines 25 to 27 and look at **Table 2**.

(f) (i) Explain why the researcher should have also combined the categories for 0 and 1. **1**

(ii) After combining the categories for 0 and 1, calculate the correct value of the test statistic. **2**

(iii) State the null and alternative hypothesis for this test and state your conclusion using the correct test statistic at a 10% level of significance. **4**

Read lines 35 to 38 and look at **Figure 2**.

(g) State the common assumption that the researcher is referring to. **1**

[Turn over

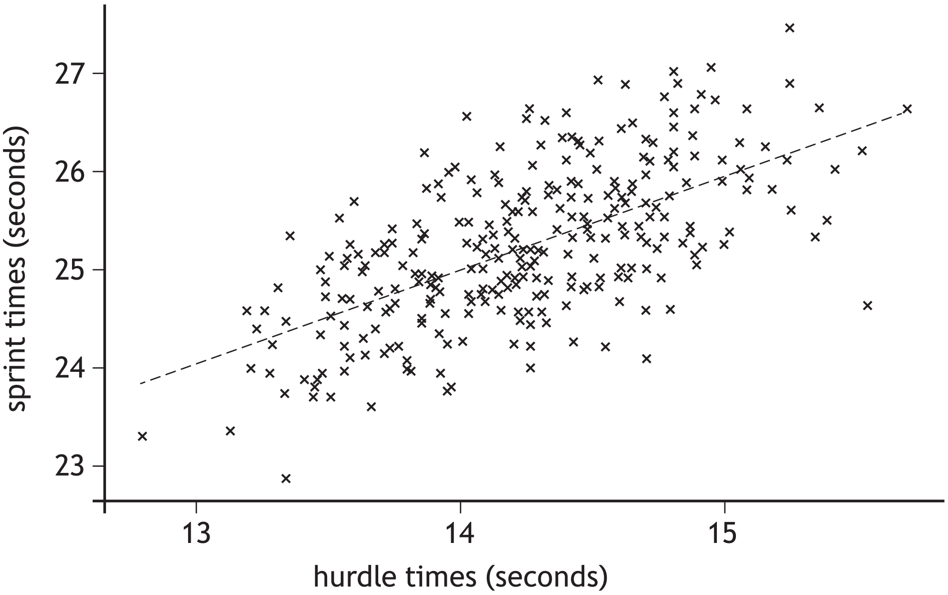
2. In athletics, the women's heptathlon involves athletes competing in seven different track and field events over two days. The relationship between an athlete's performance in the 200 metres sprint and the 100 metres hurdles was investigated. Data on the best performances of the top athletes during the 2018 season was sourced from the website of the International Association of Athletics Federations.

The data was used to investigate any relationship between the sprint times and hurdle times, both measured in seconds. The results from a hypothesis test conducted at the 5% significance level on the correlation are shown in **Output 1**. A scatterplot is shown in **Figure 1**, with the fitted least squares regression line shown as a dashed line.

Output 1

```
data:  sprint and hurdles
sample correlation coefficient, r = 0.6297338
t = 13.876, df = 293, p-value < 0.0001
null hypothesis: true correlation is equal to 0
alternative hypothesis: true correlation is not equal to 0
```

Figure 1



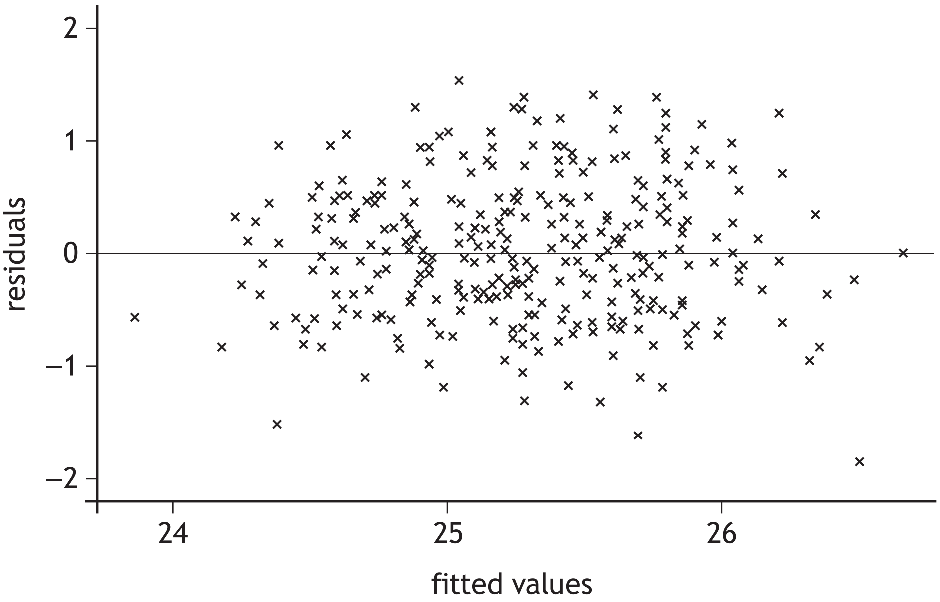
(a) Using **Output 1**, determine how many athletes' results were used in this analysis. **1**

(b) Write down the conclusion from the hypothesis test with reference to the context. **2**

To verify the assumptions required by the linear least squares regression model fitted in **Figure 1**, a residual plot is constructed and is shown in **Figure 2**.

2. **(continued)**

Figure 2



(c) With reference to the residual plot, comment on the validity of the two assumptions that are required about the location and spread of the residuals when using a linear model. **2**

During the 2019 season, the UK heptathlete Katerina Johnson-Thompson recorded her best hurdles time of 13.09 seconds.

Output 2 shows a prediction interval for her sprint time based upon her time for the hurdles, using the fitted model for the 2018 season that is assumed to be valid for the 2019 season. Two values have been deleted and replaced by `*****`.

Output 2

```
data:  hurdles and sprint
sprint = ***** + 0.9665 hurdles

variable   value
hurdles    13.09

fit        SE(fit)    99% PI
24.1366    0.6163     (22.5224, *****)
```

(d) Calculate the two missing values from **Output 2**. **2**

(e) State the further assumption needed about residuals, in order to use a prediction interval for estimation. **1**

(f) Assuming this assumption is met, write down what the prediction interval tells you about Katerina's sprint time. **1**

(g) Explain why the model in **Output 2** cannot be used to predict a hurdle time from a sprint time, and suggest what should be done to allow such a prediction to be made. **2**

2. (continued)

Output 3 shows a confidence interval for her sprint time based upon her time for the hurdles, using the same model from **Output 2**.

Output 3

```
variable   value
hurdles    13.09

fit        SE(fit)    99% CI
24.1366    0.6163     (23.9077, 24.3654)
```

Katerina Johnson-Thomson’s actual sprint time of 23.08 seconds was not captured by this confidence interval.

(h) Clearly explain what this confidence interval represents, and give a reason why it is not a concern that the time of 23.08 was not captured by this interval. **2**

[END OF QUESTION PAPER]