2018

QUESTION ONE (8 marks)

An investigation was carried out by a clothing retailer regarding the impact of advertising expenditure on the total value of sales over two different types of outlet; rural and urban. The following data were obtained:

Table 1: Advertising and Sales Data

Advertising Expenditure (A) ($00)	Rural Sales (S) ($000)	Advertising Expenditure (A) ($00)	Urban Sales (S) ($000)
0	2.6	1	2.8
2	3.5	2	3.9
2	2.9	2	4.1
3	3.9	3	3.9
3	3.8	3	3.8
3	3.6	3	4.2
4	4.7	4	4.3
6	5.0	6	5.1
7	4.0	7	5.2
7	5.0	12	0.5
8	3.5	8	5.5
9	4.1	9	5.8
10	4.2	10	6.1
9	4.0	9	5.7
8	3.7	8	5.3
7	4.1	7	5.4
11	4.0	11	6.2
8	3.9	9	5.7
9	3.8	8	5.3
8	3.8	8	5.2

Figures 1 and 2 below show a scatterplot for the data for each type of outlet.

Figure 1: Rural Sales

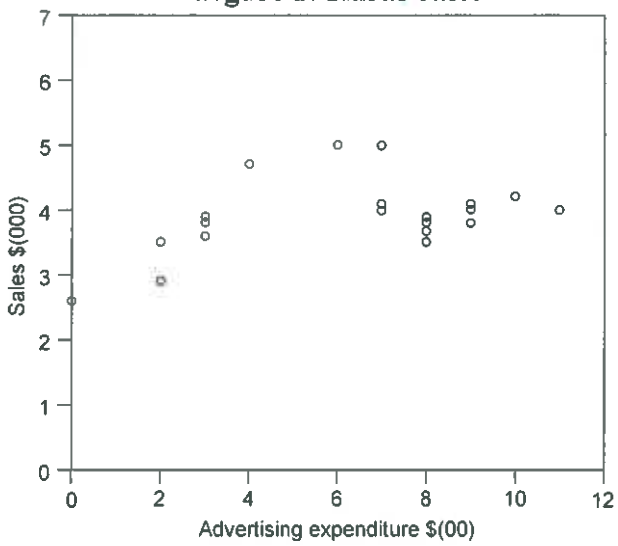
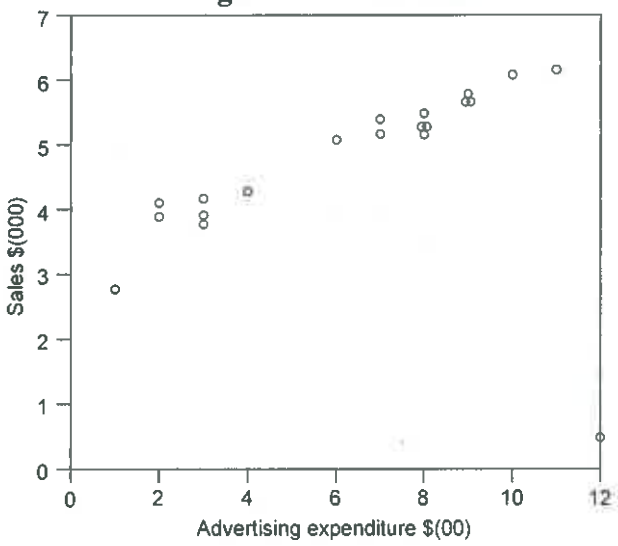


Figure 2: Urban Sales



The following are possible models to fit to these data.

$S = 0.0825A + 3.3955$

$S = -0.0356A^2 + 0.4783A + 2.6392$

$S = 0.1263A + 3.8788$

$S = 0.29A + 3.12$

(a) Describe the relationship between the advertising expenditure, A, and sales, S, for each type of outlet.

(b) By selecting an appropriate model for each type of outlet, obtain a sales prediction for an advertising expenditure of $500, one for urban outlets and one for rural outlets.

(c) Discuss any reservations there may be with using your sales predictions in (b).

(d) Suggest two further variables that would possibly influence sales.

Describe the expected relationship between each of your suggested variables and sales.

(e) It was claimed that increasing the advertising expenditure to $1400 for each type of outlet would not lead to any increase in sales.

Discuss this claim using the given data.

2018

QUESTION THREE (8 marks)

(a) A random sample of 400 articles published during 2017 was taken from a news website to investigate the nature of the words used in the headlines.

(i) The headlines for these articles were analysed using sentiment analysis. Sentiment analysis assigns a score on a continuous scale between 0 and 1 that measures the overall sentiment or emotion of a piece of text. A sentiment score of 0 indicates a completely negative sentiment, and a sentiment score of 1 indicates a completely positive sentiment.

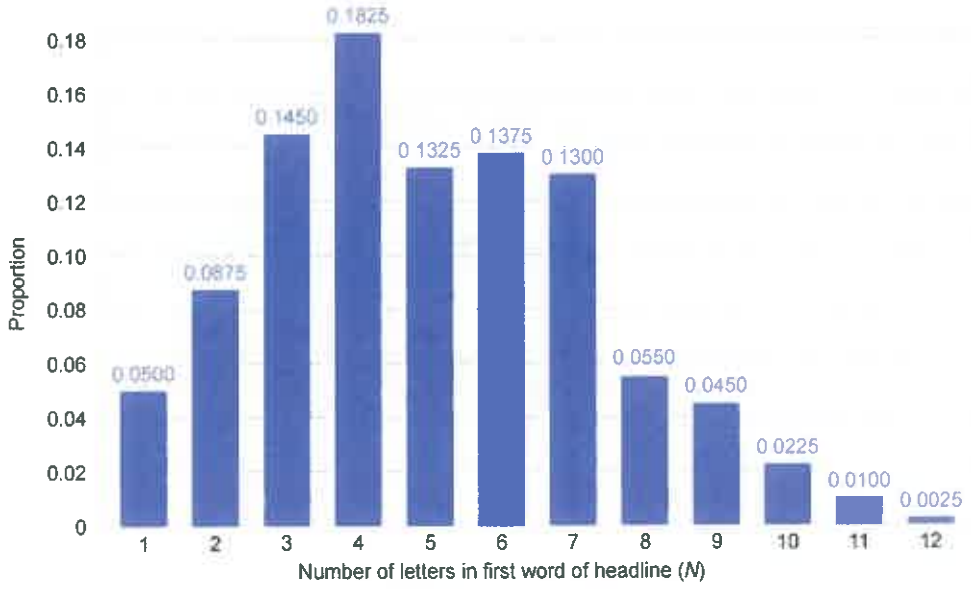
The mean sentiment score of the headlines was found to be 0.538, and the standard deviation 0.287. 74.9% of the headlines had a sentiment score of less than 0.8.

Explain why a normal distribution may not be a good model for the sentiment scores of all headlines for articles published on this news website.

Support your answer with at least one calculation.

(ii) The distribution of the first word length of the first word of each headline of these 400 articles is shown in the following graph.

Figure 4: Lengths of First Word of Each Headline



Let the random variable N be the number of letters in the first word of each headline in the sampled articles.

Investigate whether a Poisson distribution would be a good model for $N - 1$.

Support your answer with statistical reasoning and calculations.

(b) Some news websites display how many minutes a reader is expected to take to read an article. A random sample of 200 news articles was taken from each of two different news websites, Website A and Website B, that displayed expected reading times. For each article, the reading time displayed and the number of words for each article were used to calculate the reading speed in words per minute for the article. The table below shows an excerpt of the sample data from Website A.

Table 4: Sample Data Excerpt from Website A

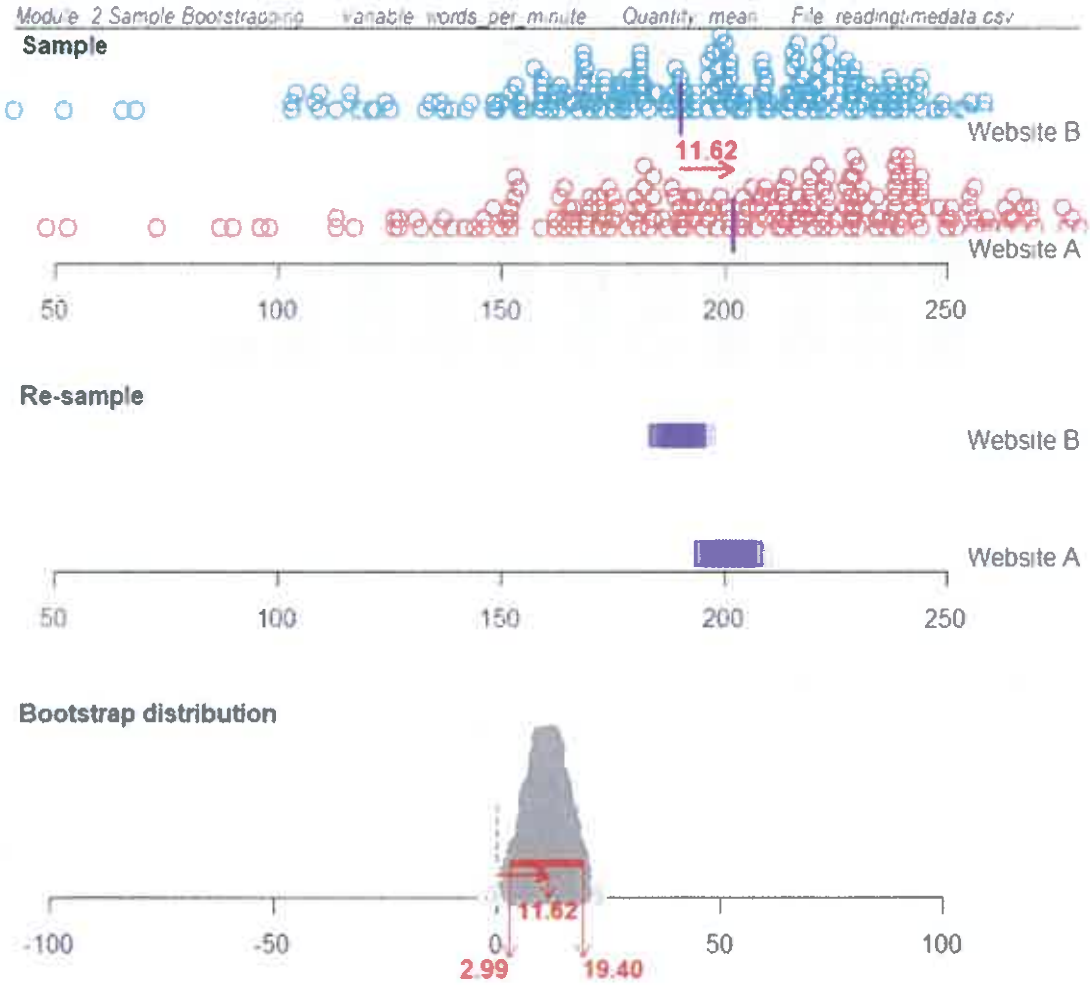
Website	Article	Number of words	Reading time in minutes	Reading speed (words per minute)
A	1	684	3.6	190
A	2	299	1.3	230
A	3	187	1.0	187
...	...	...	...	...

Figure 5 below gives a summary of the reading speeds for the sample from each website and the bootstrap distribution of the difference between the means of the reading speeds for the two websites.

What can be concluded about the mean reading speeds of news websites A and B?

Use Figure 5 to support your answer.

Figure 5: Reading Speeds and Bootstrap Distribution of Difference between Means



2018

QUESTION FOUR (8 marks)

(a) The developer of a particular pregnancy test claims that 98% of women who are pregnant will test positive for pregnancy using the test, and only 4% of women who are not pregnant will test positive for pregnancy using the test (referred to as a "false positive").

(i) Explain why the proportion of women using this test who are actually pregnant when the test is positive for pregnancy is not necessarily 98%.

Support your answer with statistical reasoning and calculations.

(ii) A study was conducted to investigate the accuracy of the pregnancy test. The study found that:

- 94% of the women who were pregnant tested positive for pregnancy
- 81% of the women who were not pregnant tested negative for pregnancy
- 44 of the 55 women who tested positive for pregnancy were pregnant

What proportion of the women in this study were pregnant?

Support your answer with statistical reasoning and calculations.

(b) Read the following excerpts from report on *Attitudes and Behaviour towards Alcohol Survey (ABAS) 2013/14 to 2015/16: Attitudes to drinking in pregnancy.*

Purpose
This report presents descriptive results about New Zealanders' attitudes to drinking alcohol during pregnancy. Results from the survey are used to inform the planning and development of alcohol activities, policies, and programmes that aim to reduce alcohol-related harm in New Zealand.

Method
For each survey in 2013/14, 2014/15 and 2015/16, approximately 4000 people aged 15 years and over were surveyed over four months (November, December, January, and February). Households were stratified into telephone directory regions. A random sample of telephone numbers was generated from all number ranges found in the White Pages using a Random Digit Dialling (RDD) approach. The mode of the interview was Computer-Assisted Telephone Interviewing (CATI).

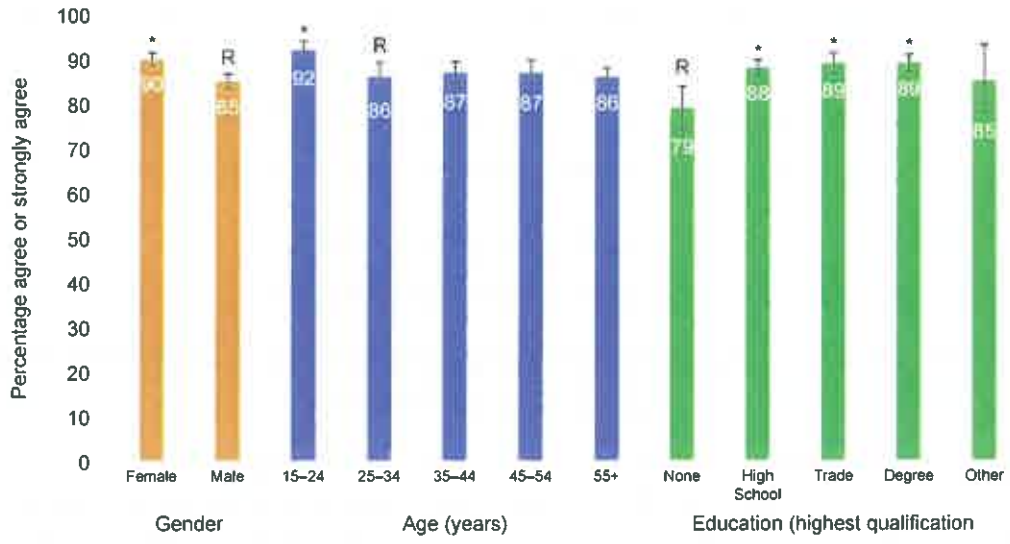
Analysis
This report presents the analysis of five questions from ABAS that assessed New Zealanders' attitudes towards drinking in pregnancy. Responses to these attitude statements were on a five-point scale of 'strongly agree', 'agree, 'neither agree nor disagree', 'disagree' and 'strongly disagree'. The data have been weighted (adjusted) so that the sample reflects the makeup of the New Zealand population at the last Census (2013). Results are presented as weighted estimates with error bars representing the 95% confidence intervals. The confidence level for comparing estimates by sub-group was set at 95%.

Drinking in pregnancy is OK
General attitude to drinking in pregnancy was assessed by asking all respondents to indicate their level of agreement with the statement 'During pregnancy drinking small amounts of alcohol is OK'. Overall, 84% [95% CI: 82, 85] of respondents disagreed with this statement in 2015/16. This question was also asked of respondents in the 2013/14 and 2014/15 surveys. Overall, there were no significant changes in level of disagreement across the three survey years.

Encourage others to stop drinking if pregnant
To assess the level of support from others to encourage pregnant women not to drink, all respondents were asked to indicate their level of agreement with the statement 'I would

encourage a friend or family member to stop drinking completely if she was pregnant'. Overall, 88% [95% CI: 86, 89] of respondents agreed with this statement. As shown in Figure 6, agreement was higher among: females (compared with males), 15 to 24-year-olds (compared with 25 to 34-year-olds) and those with a formal qualification (compared with no formal qualifications).

Figure 6: Percentage of Respondents who agreed with the statement 'I would encourage a friend or family member to stop drinking conpletely if she was pregnant' in 2015/16, by gender, age, and education level



Base = All respondents (ABAS 2015 / 2016)
* Significantly different from the reference group (indicated with 'R')

(i) Describe two strengths in respect to the method used for this survey.

(ii) Describe two potential non-sampling errors for this survey.

(iii) In the paragraph titled **Drinking in pregnancy is OK**, explain what was meant by the expression "no significant changes".

(iv) Figure 6 has error bars (vertical lines) to represent 95% confidence intervals.

Give two possible reasons why the error bars for the "Other" sub-group are longer than those for the "Degree" subgroup.

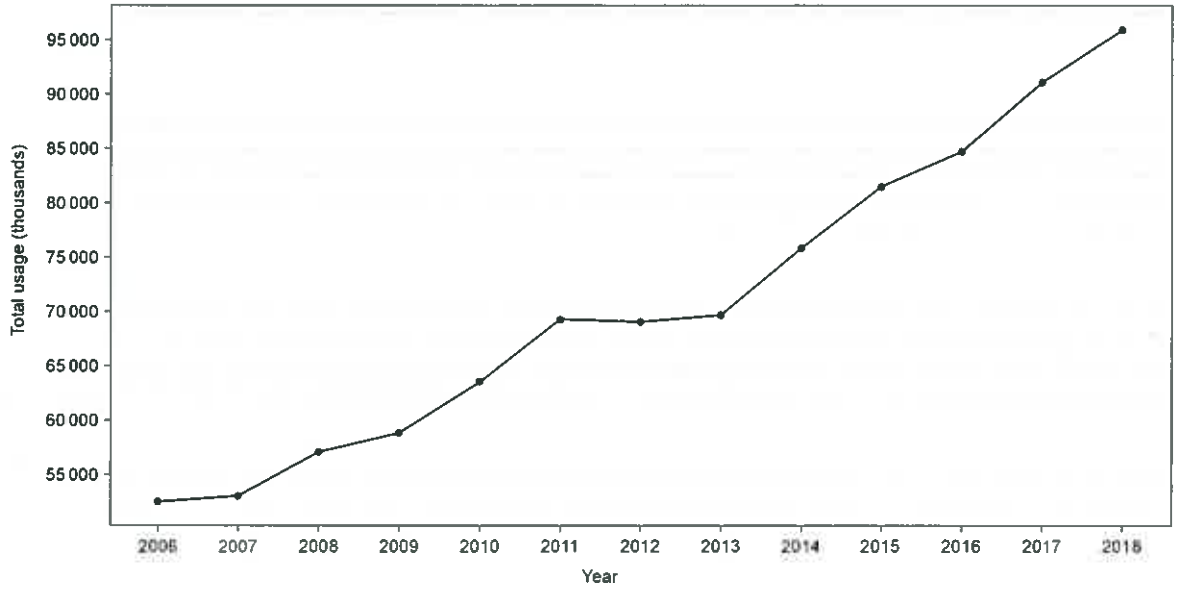
2019

QUESTION TWO

Data was obtained from Auckland Transport on the usage of buses, ferries, and trains per quarter for the years 2006 to 2018. Usage was measured in thousands of trips made.

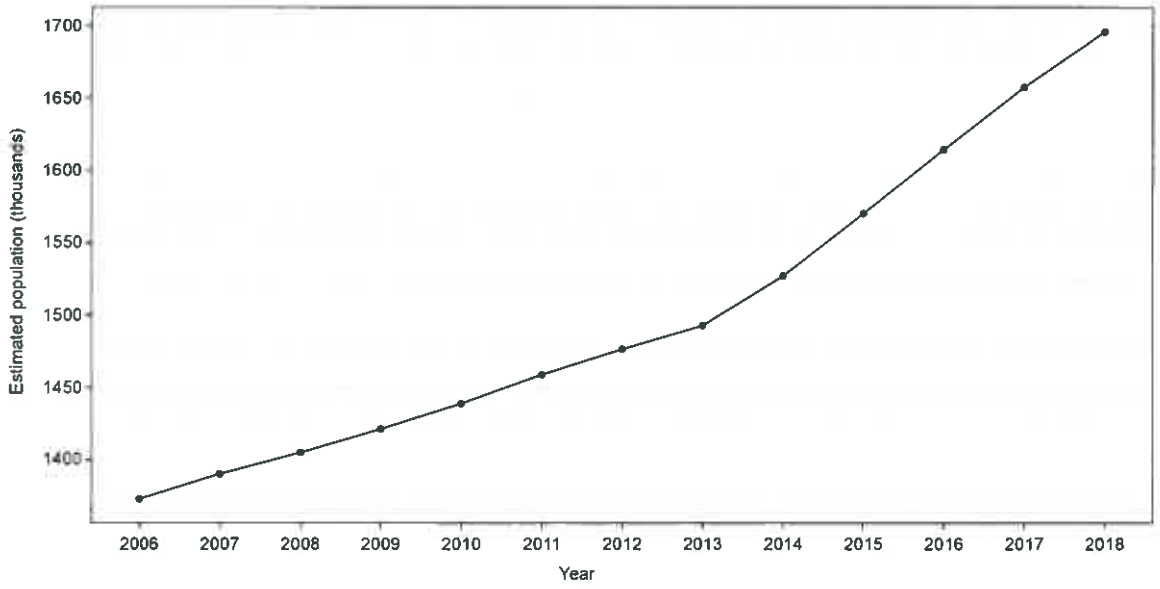
(a) Figure 5 displays the data on total usage of buses, ferries, and trains per year.

Figure 5: Total usage in thousands, 2006–2018



Data was also obtained from Statistics New Zealand on the estimated population of the Auckland Region for the years 2006 to 2018. Figure 6 displays this data.

Figure 6: Estimated population of Auckland region, 2006–2018

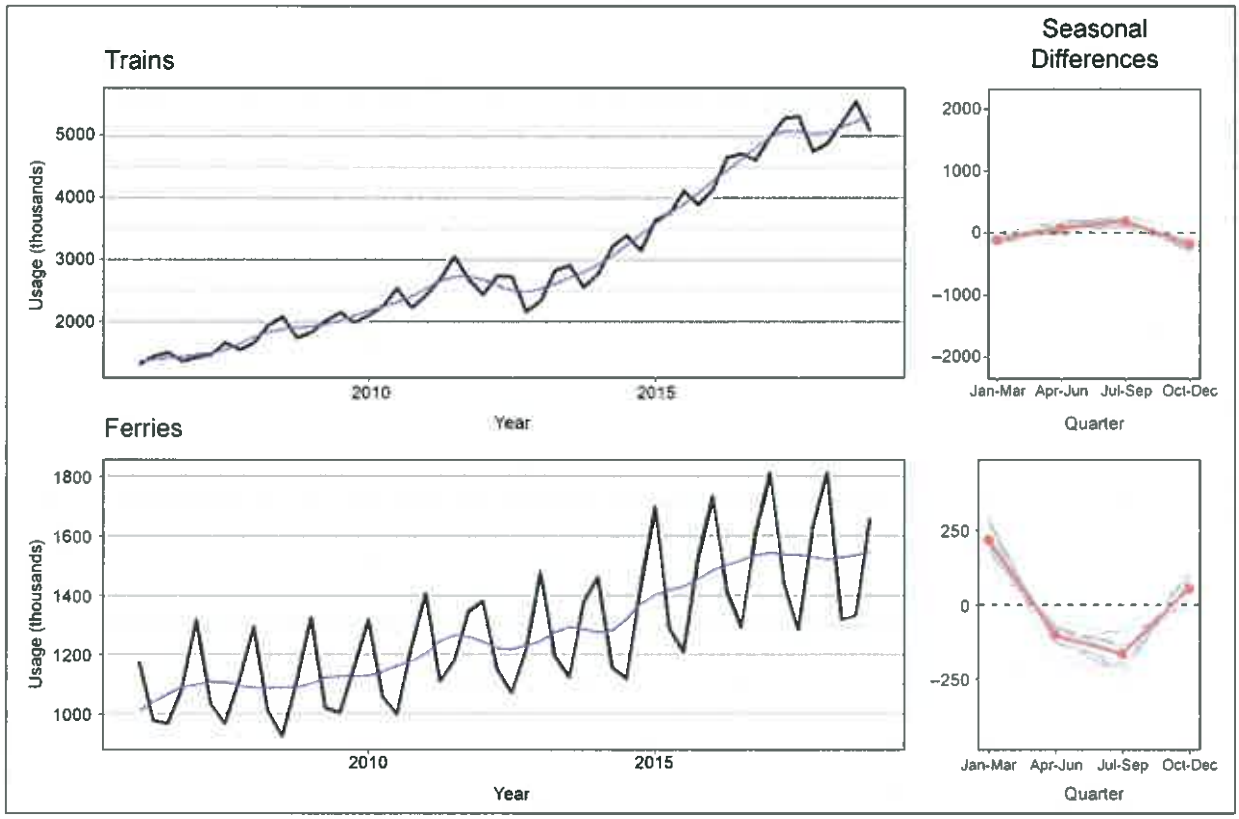


It was claimed that the rate of increase in total usage of Auckland Transport was similar to the rate of increase for the estimated population of the Auckland region.

Using Figures 5 and 6, investigate if this claim is justified.

(b) Figure 7 displays the raw data for the usage of trains and ferries for the years 2006 to 2018, with smoothed trend curves shown in blue. Figure 7 also displays the seasonal differences, with their average (mean) shown in red.

Figure 7: Usage of trains and ferries, 2006–2018



Write two short paragraphs comparing the features of the data for the usage of trains and ferries over the period 2006 to 2018.

(c) Two different Holt-Winters Additive models were used to obtain forecasts for the usage of **buses** per quarter for the years 2019 and 2020.

Figure 8 shows the raw and fitted data for the years 2006 to 2018, and forecasts produced from Model 1.

Figure 9 shows the raw and fitted data for the years 2016 to 2018, and forecasts produced from Model 2.

Both figures also include the forecast tables.

Figure 8: Model 1

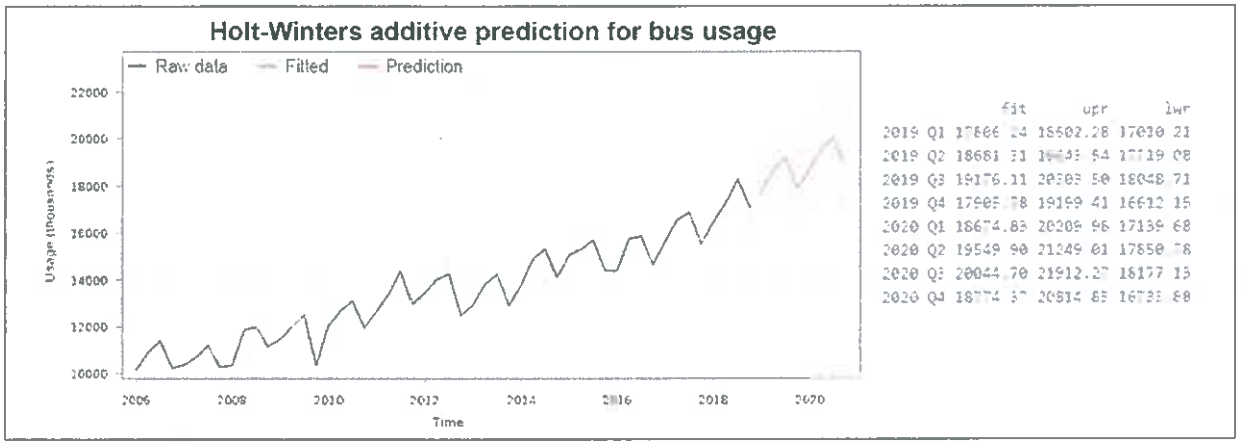
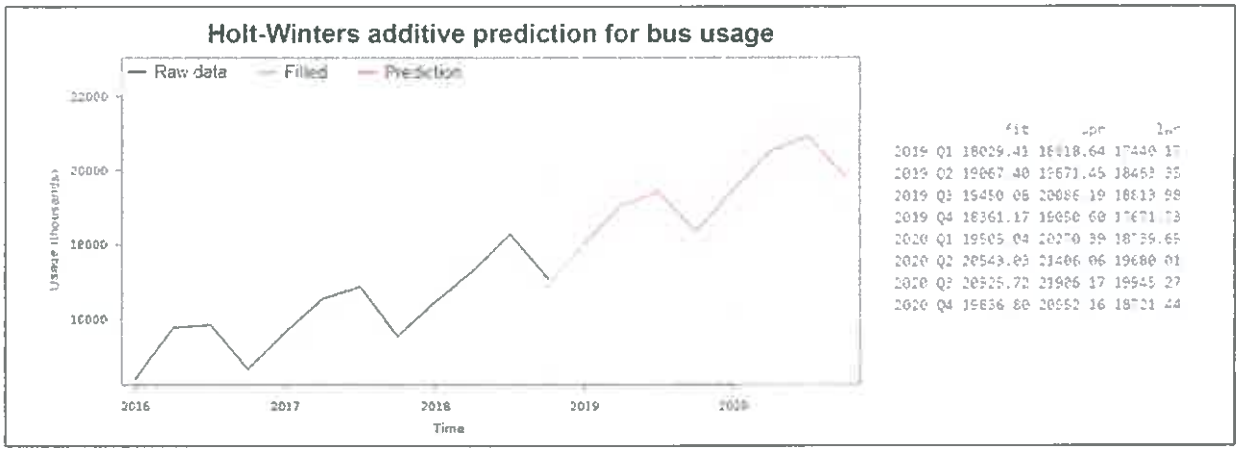


Figure 9: Model 2



(i) Compare the forecasts from Models 1 and 2 for the second quarter of 2020.

(ii) Give reasons for the similarities or differences between the forecasts for the second quarter of 2020 in (i) in terms of the different models.

2019

3. (b) An experiment was carried out to investigate if university students could be encouraged to walk or cycle to university more often. The participants were 136 university students who were members of an environment club, all of whom sometimes, but not always, walked or cycled to university. These students were asked to install an app on their phones that requested them to record each evening whether they had walked or cycled to university that day.

The students were randomly allocated into one of two groups of 68. Students in one group were sent daily messages from the app either encouraging them to walk or cycle to university the next day or providing information about the benefits of walking or cycling. Students in the other group were not sent daily messages. After three months, the data collected through the app was used to determine how many days each student had walked or cycled to university.

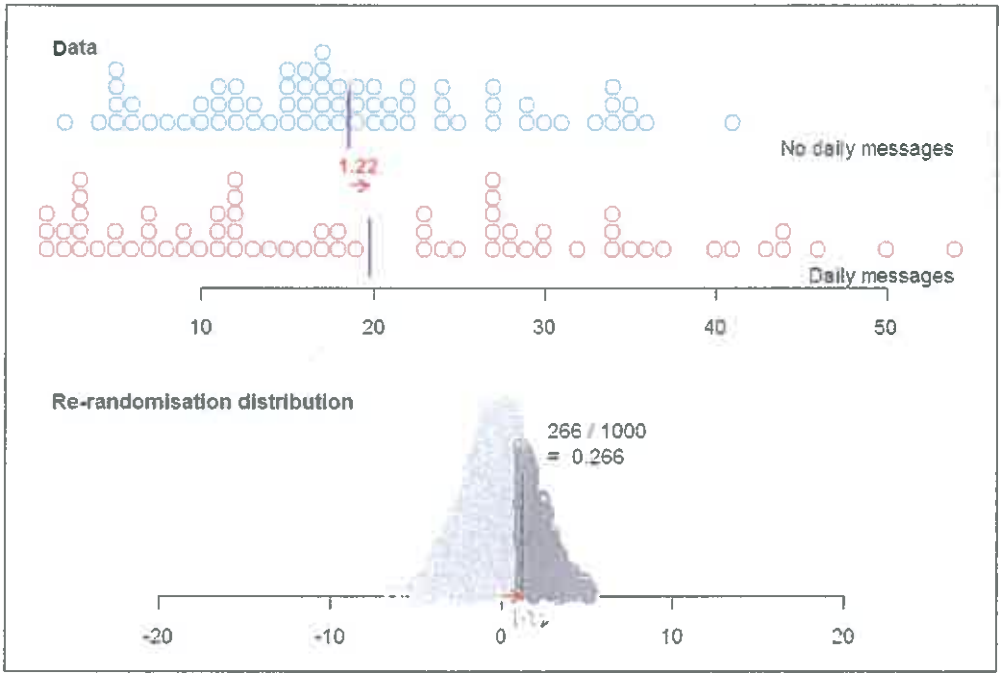
(i) Write a short paragraph that summarises the design of the experiment using appropriate statistical terminology.

A randomisation test was carried out using the difference between the mean number of days participants had cycled or walked to university for the two groups. Table 1 gives some statistics and Figure 12 gives some output from this test.

Table 1: Summary of number of days cycled or walked by group

	Min	Lower quartile	Median	Upper quartile	Max	Mean	Standard deviation	Group size
Daily messages	1	7.75	17	29.25	54	19.72	14.07	68
No daily messages	2	12.00	17	24.00	41	18.50	9.20	68

Figure 12: Summary of number of days cycled or walked, and re-randomisation distribution



(ii) Interpret the randomisation test output and explain why the result could have been expected in this context.

(iii) Give TWO questions you would want answered about the study, data, or participants before accepting the result in (ii).

For each of the questions, explain why it would be important to know each answer.

2019

QUESTION FOUR

Read the following report.

What do passengers do during travel time?

This study reports on 812 adult passengers in Wellington, New Zealand. The aim of this study was to assess the frequency of passenger activities during bus and train travel using structured observations of passengers in a sample of bus and train routes and times in the Wellington area.

Bus and train routes selected were short (20-minute) or long (up to 2-hour) distances, downtown and suburban routes, encompassing wealthier and poorer areas, and included routes where passengers had a clear choice of bus or train mode. Both morning (before 9.00 a.m.) and evening (3.00 p.m. to 6.30 p.m.) peak commuting times were included for observations, as were several night and midday times. Public transport providers were contacted to explain the research and generously provided free passes for the two researchers and a covering letter of support. The two researchers worked together for safety reasons and avoided late night trips.

Researchers recorded passenger characteristics and behaviour over a 4-minute period, on a range of routes and times, using 12 pre-set codes. During the four-minute observation period, a passenger might be recorded as carrying out only one or more than one activity at a time (multitasking), for example, reading a book while wearing headphones or texting while eating. To accommodate this diversity, the data analysis refers to the numbers of passengers who were "ever observed" doing the activity.

Most passengers (65.3%) were "looking ahead/out of the window" at some point in the observation period, more on buses than on trains. About one-fifth of all passengers observed were seen reading, more on trains. Other activities included listening on headphones, talking, texting, and sleeping/eyes closed (see Table 2 for all activities and additional results).

Table 2: Ever-observed activities on bus and train (*N* = 812)

Activities	Bus		Train		Total	
	Number	*% of bus sample*	*Number*	*% of train sample*	*Number*	*% of total sample*
Looking ahead/out of window	270	76.5	260	56.6	530	65.3
Reading	44	12.5	132	28.8	176	21.7
Headphones in	60	17.0	96	20.9	156	19.2
Talking	48	13.6	77	16.8	125	15.4
Texting	29	8.2	46	10.0	75	9.2
Sleeping/eyes closed	15	4.2	57	12.4	72	8.9
Handling wallet, etc.	16	4.5	42	9.2	58	7.1
Other	15	4.2	28	6.1	43	5.3
Eating/drinking	13	3.7	25	5.4	38	4.7
Using computer	1	0.3	34	7.4	35	4.3
Writing	4	1.1	22	4.8	26	3.2
On phone	6	1.7	6	1.3	12	1.5

Gender and broad age group were recorded (young = about 18 to 30–34; middle age = 35 to 60; older = over 60). There were 402 women observed. Activities were compared using odds ratios on the basis of gender, age group, mode, and time of day. An odds ratio compares whether the probability of an event is the same for two groups; an odds ratio of 1 means that the event is equally likely for each group. It was found that females were 2.1 times as likely to be observed talking than males and 0.2 times as likely to be observed using a computer than males.

Adapted from: Russell, M., Price, R., Signal, L., Stanley, J., Gerring, Z., & Cumming, J. (2011). What do passengers do during travel time? Structured observations on buses and trains. *Journal of Public Transportation*, 14(3), 7.

(a) Evaluate the design of the study described in the report, including discussion of the following points:

- two strengths of the study design
- two challenges with how the data was collected.

(b) Suppose the data from this study is used to construct confidence intervals for:

- the difference between the percentage of all Wellington bus passengers who look ahead or out the window and the percentage of all Wellington train passengers who look ahead or out the window
- the difference between the percentage of all Wellington bus passengers who look ahead or out the window and the percentage of all Wellington bus passengers who read.

(i) Identify the relevant information from the report that would be needed to construct each confidence interval. **Do not calculate or construct the confidence intervals.**

(ii) Explain why the calculation of the margin of error for each confidence interval will be different.

(iii) Discuss TWO reservations you would have with using the confidence intervals to make inferences about all current bus or train passengers in New Zealand.

(c) The report states that "females were 2.1 times as likely to be observed talking than males".

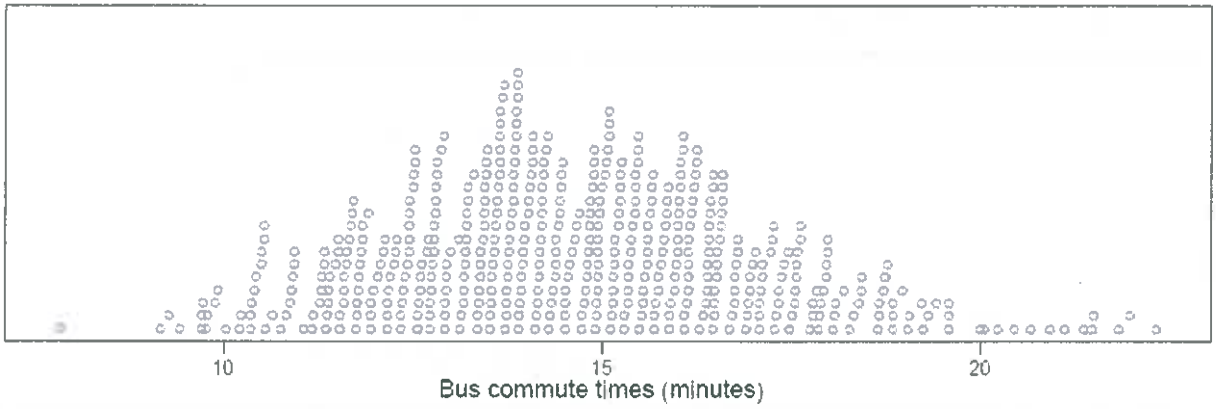
Using the data and results reported from the study, calculate an estimate for the proportion of passengers observed talking in the study who were male.

2019

QUESTION FIVE

(a) Amelia regularly commutes to work by bus. There are 11 bus stops and five intersections with traffic lights along the 3.8 km long journey. Most of her bus commutes are during the morning, but some occur later in the day. The length of each of Amelia's bus commutes (in minutes) was recorded over several years and is displayed in Figure 13.

Figure 13: Amelia's bus commute times



(i) Give TWO potential reasons for the variation in Amelia's bus commute times, and discuss how each could affect the length of her bus commute.

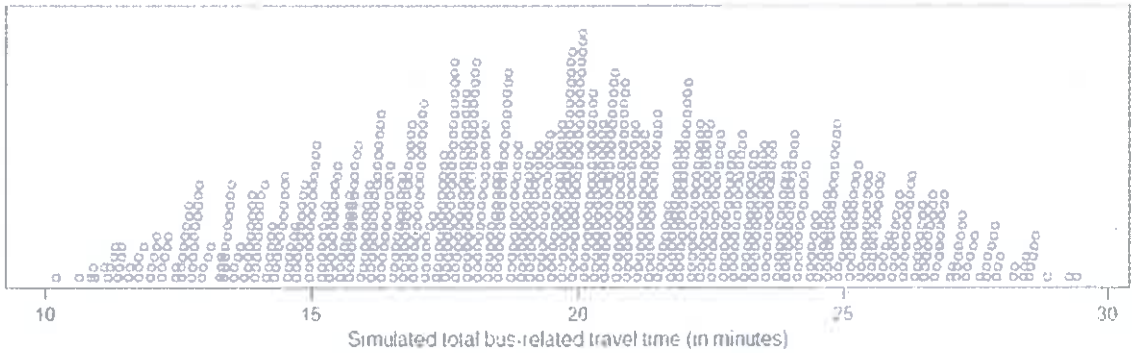
(ii) Discuss the appropriateness of using a normal distribution as a probability model for Amelia's future bus commute times.

(b) Jacob also commutes to work by bus at roughly the same time each day. He developed a model for his total bus-related travel time (the time spent waiting for the bus plus the time spent commuting to work by bus), using one triangular distribution and the following information:

- the time it will take for the next bus to arrive at the bus stop is between 0 and 6 minutes, with all times in between equally likely
- the commute to his work by bus takes between 10 and 24 minutes, with the most likely commute time around 17 minutes.

Jacob then simulated 1000 total bus-related travel times using his triangular distribution model. These times are displayed in Figure 14.

Figure 14: Jacob's simulated travel times

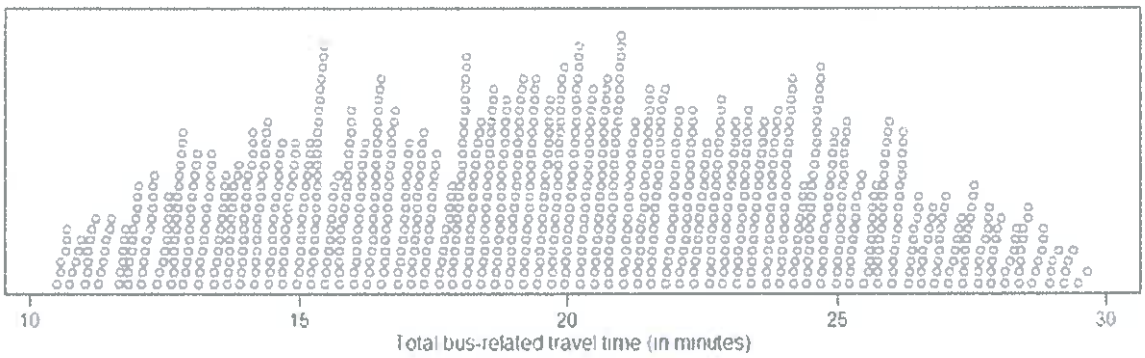


(i) Give the parameters for Jacob's model of his total bus-related travel times, and explain how Jacob appears to have determined these parameters.

(ii) Use Jacob's model to calculate an estimate for the probability that his total bus-related travel time is longer than 28 minutes, given that it is longer than 25 minutes.

(iii) Jacob then recorded his total bus-related travel times over an extended period. These times are displayed in Figure 15.

Figure 15: Jacob's total bus-related travel times



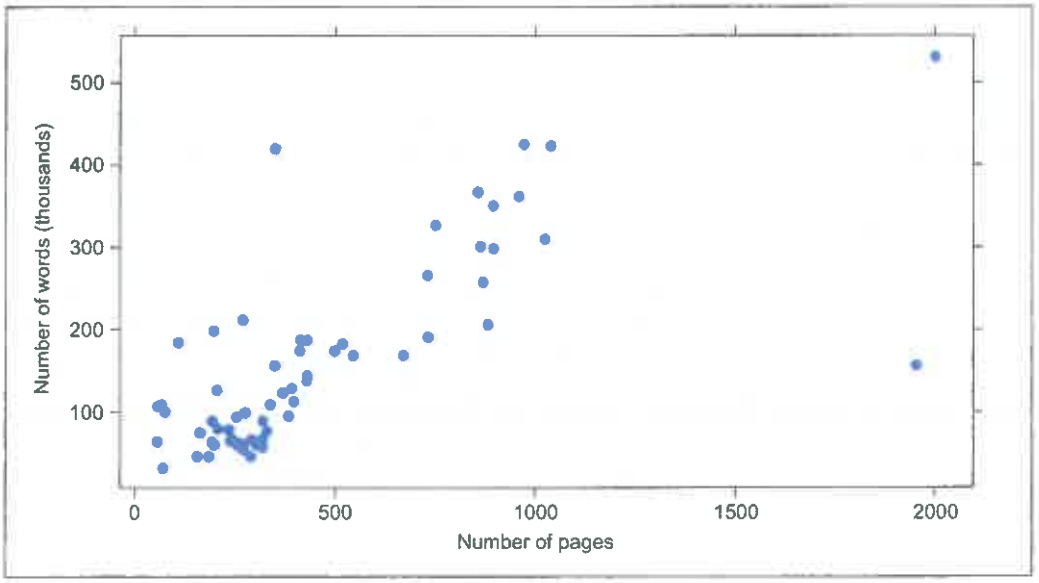
Discuss how the distribution of the data shown above challenges the model Jacob developed, and why assumption(s) he may have made when developing his model may not be valid.

2020

QUESTION ONE

(a) A website for authors states that for fiction books there are, on average, 250 words per page. Data was obtained on the number of words (in thousands) and the number of pages for 65 of the most popular fiction books written in English. Figure 1 shows the scatterplot produced from these books and variables.

Figure 1: Scatterplot of popular fiction books written in English



(i) Describe the relationship between the number of words and the number of pages for these books, identifying any notable features of the data.

(ii) Give TWO potential reasons why the number of pages of a book might not precisely predict the number of words in the book.

(iii) A linear model was fitted to the data shown in Figure 1.

The equation of this model is given below:

$$\text{Number of words (thousands)} = 57.84 + 0.2202 \times \text{Number of pages}$$

With reference to the data, the features of the scatterplot, and this linear model, discuss the suitability of the statement "for fiction books there are, on average, 250 words per page".

(b) Read the following report.

> This report summarises the results of the second survey of book reading in New Zealand.
>
> Between 2 and 25 May 2018, 2261 adult New Zealanders responded to the online survey conducted by Horizon Research Limited for the New Zealand Book Council. Participants were recruited to represent the New Zealand population. The sample was weighted to match national demographics for age, gender, personal income, education level, employment status, and ethnicity.
>
> This research into the reading habits of New Zealanders confirms that we are a nation that loves to read. 86% of New Zealand adults had read or started to read at least one book in the past year, with on average 35 books per reader. While this is a lower percentage than in the March 2017 survey (88%, $n = 2082$), the difference is not statistically significant.
>
> It is wonderful that New Zealanders love to read, and to see that books remain an important touchstone in our society. But it's worrying to see how many of us didn't pick up a book in the past year. 14% of Kiwis didn't read a book in the past year. Males made up most (69%) of those adults who did not read a book in the past year.
>
> Adapted from: Book Reading in New Zealand, August 2018, New Zealand Book Council. https://www.read-nz.org/Downloads/Assets/Download/56854/1/2018%20Book%20Reading%20in%20NZ%20August%2027%20high-res%20final.pdf

(i) Identify ONE claim made in the report that is based on at least one survey percentage.

Evaluate the claim using a point estimate and a "rule of thumb"-based margin of error.

(ii) The report states that "86% of New Zealand adults had read or started to read at least one book in the past year, with on average 35 books per reader".

Discuss TWO potential non-sampling errors associated with people responding to being asked how many books they read in the past year.

2020

QUESTION TWO

(a) Project Gutenberg is an online library with over 60 000 free eBooks. The eBooks have been digitised by volunteers, with a focus on older works that are in the public domain.

A random sample of 24 books was taken from all the eBooks available from Project Gutenberg. For each book in the sample, the language it was written in and year that it was first published was determined by examining the digitised book.

(i) The age of each book was calculated using the difference, in years, between 2020 and the year the book was first published. The sample was then used to construct a bootstrap confidence interval for the median age, and this interval had limits (108, 143).

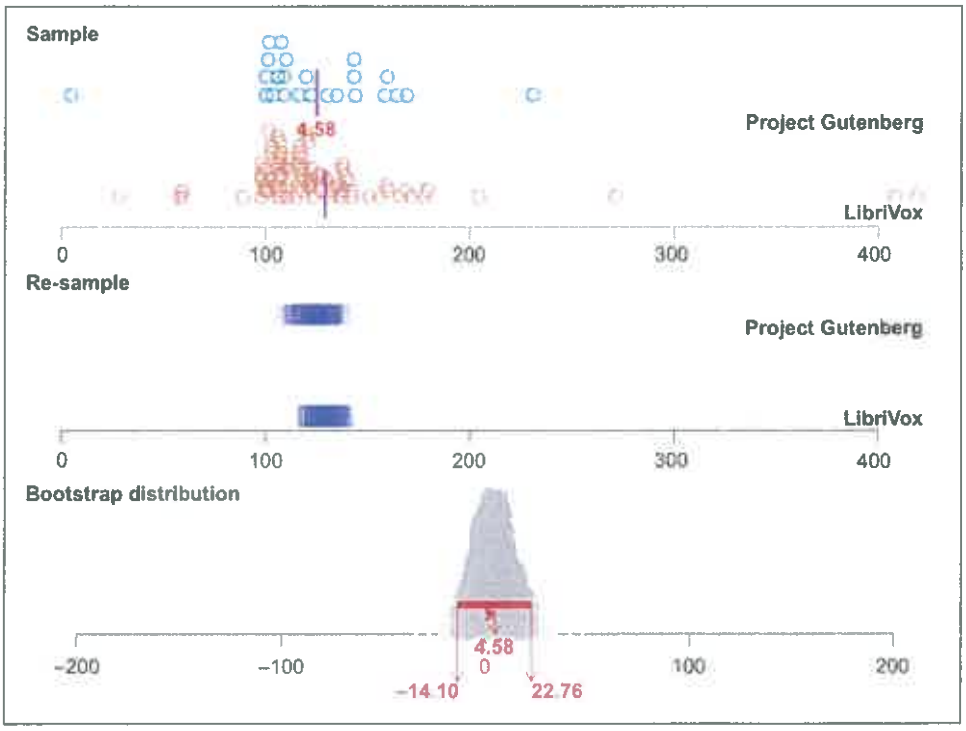
Project Gutenberg states on their website that most of their eBooks were published before 1924.

Discuss whether this claim can be supported by interpreting the confidence interval given above.

(ii) LibriVox is a website that provides over 15 000 free public domain audiobooks read by volunteers from around the world. A random sample of 80 books was taken from all the audiobooks available from the website, and for each book in the sample, the year that it was first published as a printed book was recorded.

The random sample of books from LibriVox was compared to the random sample of 24 books from Project Gutenberg. The samples were used to construct a bootstrap confidence interval for the difference between the mean age of books from LibriVox and the mean age of books from Project Gutenberg. The output from this analysis is shown in Figure 2.

Figure 2: Bootstrap confidence interval output



Discuss what can be concluded from both the features of the sample data distributions and the confidence interval constructed using the sample data.

(iii) Five of the 24 books in the sample from Project Gutenberg were not written in English.

Use a probability distribution model to evaluate whether there is sufficient evidence to conclude that more than half of the books available from Project Gutenberg are written in English.

In your answer justify the selection of the probability distribution model that you used.

(b) A local library wants to find out how long people spend at the library when they visit in person. They suspect that people who arrive in the morning stay for longer than people who arrive in the afternoon, and want to know how much longer, on average, they stay.

Apply the steps of the statistical enquiry cycle to this situation, giving a short description of what each step would involve.

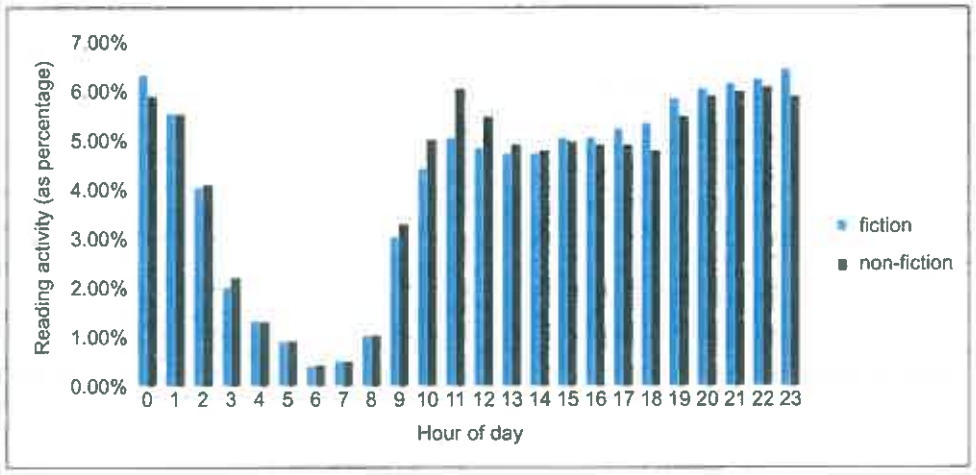
2020

QUESTION THREE

A study was carried out using 10 months of reading data from an eBook subscription company. 8000 people were studied over three million reading sessions.

(a) Figure 3 shows the distribution of reading activity for fiction and non-fiction genres throughout a day.

Figure 3: Distribution of reading activity throughout a day



Write TWO comments comparing the similarities and differences for the reading activity of fiction and non-fiction books throughout a day. Include at least one numeric comparison of likelihood.

(b) The distribution of reading speeds (number of words read per minute) for people in the study was described in a report as "bell-shaped, with a mean around 150 words per minute".

3310 of the people in the study had a reading speed between 120 and 150 words per minute.

Using an appropriate probability distribution model, estimate the lower and upper limits for the middle 95% of reading speeds for people in the study.

(c) The study explored what percentage of a book, on average, people read before they stopped reading the book. Several thousands of books that had each been read by at least 40 people during the study were used to calculate a mean percentage completion value for each book. The mean percentage completion values for these books ranged from 0% to 100%, with 68% being the most likely value.

About half of the books had a mean completion value of 64% or less. Only about 5% of the books had a mean completion value higher than 90%.

(i) Investigate whether a triangular distribution would be a good model for the mean percentage completion values for books available from the eBook subscription company.

Support your answer with statistical reasoning and calculations.

(ii) Discuss TWO factors that might explain the variation in the mean percentage completion values for books available from the eBook subscription company.