No part of the candidate evidence in this exemplar material may be presented in an external assessment for the New Zealand Scholarship award.

93201A

932011

S

SUPERVISOR'S USE ONLY

SCHOLARSHIP EXEMPLAR

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

Tick this box if there is no writing in this booklet □

Scholarship 2020
Statistics

2.00 p.m. Friday 20 November 2020
Time allowed: Three hours
Total score: 40

ANSWER BOOKLET

Check that the National Student Number (NSN) on your admission slip is the same as the number at the top of this page.

Write all your answers in this booklet.

Show ALL working. Start your answer to each question on a new page. Clearly number each question.

Check that this booklet has pages 2–24 in the correct order and that none of these pages is blank.

YOU MUST HAND THIS BOOKLET TO THE SUPERVISOR AT THE END OF THE EXAMINATION.

Question	Score
ONE	
TWO	
THREE	
FOUR	
FIVE	
TOTAL	/40

ASSESSOR'S USE ONLY

1) a) i) There is a moderate, positive, linear relationship between the number of pages and number of words for these books. There are 3 noticeable outliers. One outlier is located at approximately (400, 420,000), and has a significantly larger number of words than most books with a similar number of pages. [redacted] There is another outlier located at around (2900, 15000). This book has a very large number of pages, and has a significantly lower number of words than would be expected. Another outlier also has a very larger number of pages, and is located at around (2000, 520000). This data value is situated on approximately where it would be expected to given the trend shown in lower values, however, it also has a much larger number of pages than the majority of other data values.

ii) One possible reason is that books may have different font sizes. Books with smaller font sizes would have more words per pages, and thus the number of pages of a book may not precisely predict the number of words. [redacted] Another reason is that some books may contain pictures. Books with pictures would have less words per page, therefore meaning that the number of pages could not precisely predict the number of words in a book.

iii) If there were 250 words per page, then a book of 300 pages would have around:

$300 \times 250 = 75,000$ words

Looking at the graph, books with around 300 pages appear to have between around 30,000 and 220,000 words. This is a large amount of variation, thus potentially reducing the accuracy of the statement. However, most books of around close to 300 pages are clusted at around 50,000-100,000 words, which is close to what would be expected if books were to contain 250 words per page. Using the linear model equation for a book of 300 pages:

No. of words $= 57.84 + (0.2202 \times 300)$

$= 123.9$

So the linear model predicts that for a book of 300 pages, that it would have 123,000 words, which is significantly higher than the 75,000 that the statement would indicate. Therefore the statement is overall not suitable given the data.

b) i) The report makes the claim that "Maori made up most (69% of those adults who did not read a book in the last year". [redacted] $0.14 \times 2261 = 316.54$

$n \approx 317$

MoE : $1/\sqrt{n}$ = $1/\sqrt{317}$ = 5.6% [redacted]

95% CI : 63.4% – 69% – 74.6%

69% − 5.6% 69% + 5.6%

With 95% confidence I can infer that the percentage of people who didn't read a book in the past year who are Maori is likely to be somewhere between 63.4% and 74.6%. As the lower end of this confidence interval is above 50%, therefore the claim ~~of~~ that "Males made up most of those adults who did not read a book in the past year" is supported.

[redacted]

ii) Firstly, there may be some behavioural effects. Reading is typically seen as an intelligent hobby, and that ~~so~~ people may lie about how many books they read in order to ~~be~~ appear smarter. So the actual average number people read may be lower than 35 books per year. ~~Additionally~~

[redacted]

Secondly, a year ago is a long time for people to remember back to. People may have forgotten exactly how many books they read and thus are only providing an estimate of the number, which may differ from the actual amount they read.

[redacted]

2)a)i) A book published in 1924 would have an age of 96. As the bootstrap confidence interval for the median age is (108, 143), the lower limit of this interval ~~mode~~ is a book published in 1912. Therefore the claim that "most of these books were published before 1924" can be supported.

[redacted]

ii) Both distributions for Project Gutenberg and Librivox have modal clusters around 100-150 years old. The ~~[illegible]~~ sample of books from Librivox is significantly larger than that from Project Gutenberg, ~~[illegible]~~ with samples of 80 and 24 respectively. Based on the bootstrap distribution, the mean age of books from ~~Project~~ Librivox is likely to be somewhere between 14.10 years younger and 22.76 years older than the mean age of books from Project Gutenberg. Therefore a significant difference between the mean age of books from Librivox and Project Gutenberg cannot be claimed based on this data, as the bootstrap ~~[illegible]~~ distribution includes 0.

[redacted]

iii) I am using the Binomial distribution:

~~[illegible]~~ $P(\text{Not English}) = 5/24 = 0.2083$

~~$X \sim Bin(60{,}000, 0.7927)$~~ $\therefore P(\text{English}) = 0.7927$

~~$P(X \geq 30{,}000) = 1 - P(X \leq 29{,}999)$~~

~~[illegible]~~ Use sample of 1000 books from Project Gutenberg to make a estimate.

~~P~~ $X \sim Bin(1000, 0.7927)$

$P(X > 500) = 1 - P(X \leq 499)$

$= 1 - \text{~~[illegible]~~}(4.05 \times 10^{-97})$

$= 1$

With use of the binomial distribution there is sufficient evidence to claim that over half of the books available from Project Gutenberg are written in English. However, due to the small sample size, it should be considered that there will likely be significant sampling variability, which could reduce the accuracy of ~~[illegible]~~ my calculated probability * continued on page 7.

b) The amount of time a person who arrives in the morning spends at the library should be recorded for a sample of people who arrive in the morning, and the same should be done for a sample of people who arrive in the ~~after~~ afternoon. The mean of the ~~[illegible]~~ time spent by each group (morning and afternoon) should then be calculated. Then, a randomisation test could be carried out in order to determine whether the difference in mean ~~time~~ time spent at the library between the two groups is significant. ~~[illegible]~~ The times from each person would be randomly allocated into two groups by chance, and the mean of ~~difference~~ these random groups ~~between the means~~ would be calculated. Then the difference between the two means would be calculated. This process would be repeated 1000 times. ~~The~~ If the same or longer ~~[illegible]~~ observed difference between

[redacted]

the means of ~~the~~ morning sample and afternoon sample could be obtained by chance more than 5% of the time, then it could be concluded that there is no significant difference between the mean time a person spends at the library in the morning and the mean time a person spends at the ~~b~~ library in the afternoon. However, if the same or larger difference can only be obtained by chance, less than 5% of the time, then a significant difference could be claimed.

Seen

iii) *continued: I have used the binomial distribution as there is only two possible outcomes, a book is either written in English or it is not. Also there is a fixed number of books in the Project ~~[illegible]~~ Gutenberg database (60,000) so there is a fixed number of trials. The probability that one book in the database is written in English should also be independent of any others. Also the probability of a book being written in English should remain constant (I have used the probability of 0.7917 which is based on the sample of 24). Although the probability calculated would indicate that the claim can be made, the very small sample size on which the probability that a book was written in English was calculated from means that I believe there is not sufficient evidence to make the conclusion.

[redacted]

[redacted]

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

3) a) Both reading activity of fiction and non-fiction books decrease significantly from around ~~[illegible]~~ the 3rd - 8th hour of the day, ~~which~~ with both genres reaching a low of around 0.4% during the 6th hour of the day. The largest difference between reading activity of the two genres can be seen during the 11th hour of the day, where reading activity for non-fiction is significantly higher than for fiction. During the 11th hour, a person is approximately 1.2 times more likely to be reading ~~[illegible]~~ non-fiction than fiction.

[redacted]

[redacted]

b) $3310/8000 = 0.41375$

I am using the Normal distribution, with a mean of 150, as the distribution is approximately bell shaped

$P(120 < X < 150) = 0.41375$

$Z = \frac{120-150}{\sigma} = -1.364$

$\therefore -30 = -1.364\sigma$

$\sigma = 22.0$ ~~[illegible]~~ words per minute (wpm)

[redacted]

$P(a < X < b) = 0.95$

$a = 106.9 \approx 107 \quad b = 193.1 \approx 193$

Using continuity correction: $P(119.5 \leq X < 150.5) = 0.41375$

$Z = \frac{119.5 - 150.5}{\sigma} = -1.364$

$\therefore -31 = -1.364\sigma$

$\sigma = 22.7$ words per minute (wpm)

$\therefore P(a < X < b) = 0.95$

$a = 105.5 \approx 106$ ~~an~~ $b = 194.49 \approx 194$

[redacted]

As these values are very similar whether a continuity correction is used or not, it is likely to be

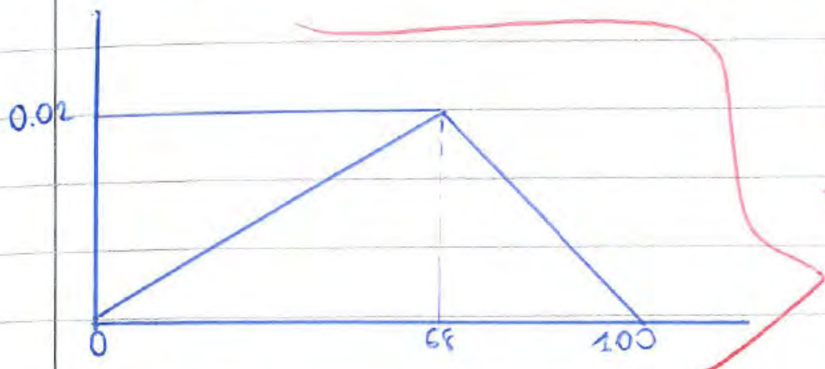
QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

unnecessary. Therefore the middle 95% of reading speeds for people in the study is between approximately 107 wpm (lower limit) and 193 wpm (upper limit).

c) i) For a triangular distribution:



$a = 0, \ b = 100$
$c = 68$

Actual: $P(X \leq 64) = 0.5$

Using Triangular distribution:

$f_{(64)} = \frac{2(64-0)}{(100-0)(68-0)} = 0.01882$

$P(X \leq 64) = 0.5 \times (64-0) \times 0.01882$
$= 0.6024$

Actual: $P(X > 90) = 0.05$

Using Triangular distribution:

$f_{(90)} = \frac{2(100-90)}{(100-0)(100-68)} = 0.00625$ ~~$P(X>90) = 0.5 \times (100-90)$~~

So, the ~~actual~~ percentage of books that had a mean completion value of 64% or less was significantly lower than what was calculated using the triangular distribution (0.5 as opposed to 0.6024). However, ~~Also~~ the ~~[illegible]~~ percentage of books that had a

[redacted]

$P(X>90) = 0.5 \times (100-90) \times 0.00625$
$= 0.03125$

mean completion value of ~~[illegible]~~ higher than 90% was slightly ~~[illegible]~~ closer ~~than~~ to what was calculated

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

using the triangular distribution (0.05 as opposed to 0.03325). Thus overall, I believe that the triangular distribution is not a good model for this data, as the predictions using this model are different from the observed probabilities, particularly on the left side of the distribution.

[redacted]

(i) The books used had been read by several thousand people in the study, so they were probably popular books. People may have heard about the books from lots of people and decided to read it, even if it was not their preferred genre. Thus people may enjoy the book to different extents, and the probability that they would continue to read it would be different.

[redacted]

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

4) a) The sample is 238 university medical students. They were randomly allocated into 2 groups, and then had different treatments applied to each of the two groups. One group was given the treatment of a text in sans-serif font, and the other was given the same text but in serif font. Thus the explanatory variable is the font type used. The response variable is number of words read from the text in one minute. The difference between the mean reading speeds of the two groups can then be calculated in order to investigate whether the type of font used affects the difficulty of reading the text.

[redacted]

b) The difference between the mean reading speed of the two groups is 2.75 words per minute. A difference of 1.75 words per minute was obtained 400/1000 (40%) of the time in the randomisation test. As this is above the 5% threshold, it therefore can be concluded that there is not a significant difference between mean reading speed for sans-serif and serif font. This result could have been expected in this context as since the participants were aware that their reading speed was being measured, they likely would have read faster than usual anyway, regardless of what font type they were using. Also, all of the people in the study are university medical students who would be used to reading lots for their study. Thus they would likely be particularly good readers and their reading speed would not be very easily affected by whether a font is ~~different~~ more difficult or not.

AS

c) i) There are significantly more females in the study than males, with around 80% of the participants being female and only around 20% being male. In terms of degree specialisation, around 40% of participants were studying paediatrics and 60% were studying therapy. By gender, slightly more females were studying paediatrics than therapy, and for males this was the opposite, slightly more males were studying therapy than paediatrics. The age of both those studying paediatrics and therapy is positively skewed, with most of the data points clustered around the lower ages. This is to be expected as most university students would be young people, with a few older people who have potentially done other degrees before this one or have come back to university after working for a while. The lower quartile of the age of participants studying paediatrics is higher than the median age of participants studying therapy, indicating that the median age of those studying paediatrics is significantly higher than those studying therapy.

[redacted]

ii) Ideally, the researchers should use stratified sampling in order to obtain a representative sample of the demographics in the population, such as

[redacted]

gender, ages and specialisation. For example, there would likely be closer to 50% males and 50% females in the target population. By using stratified sampling, a more proportional number of males and females could be obtained.

[redacted]

d) The Russian (Cyrillic) alphabet is likely to be very different in appearance from other alphabets in different languages. Therefore the results of the study may not be able to be applied to ~~the~~ texts using other alphabets, as the different fonts would cause words and letters to appear differently, thus affecting the ease and speed with which they could be read.

[redacted]

[redacted]

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

5) a) i) From 2000 to 2004, the trend in the number of new articles published in the English language per month is steadily increasing, from around 0 to 10,000. In the month of November 2002, there was a very large number of new English language articles, around 37,000. From around the start of 2004 until the start of 2007, there was a rapid increase in the number of new English language articles per month. ~~with the trend increasing~~ from around 10,000 to 50,000. This gives an average increase of 1,111 ~~approx~~ monthly new English language articles per month. From around mid 2007 until ~~[illegible]~~ 2011 the number of English language articles published per month fluctuated heavily, and showed an overall decreasing trend. The highest number of English language articles published per month was around ~~[illegible]~~ 65,000 in mid 2007. From the start of 2011 ~~[illegible]~~ onwards the trend continued to decrease, however at a slower rate. For example, from the beginning of ~~2010~~ 2007, the number of monthly new English articles published was 50,000, which decreased to around 30,000 by the start of 2011 ~~[illegible]~~, an average decrease of around 416.67 monthly new English language articles per month. However from the start of 2012 up until the start of 2020, this number only decreased from 30,000 to 20,000, an average monthly decrease of approximately 104.17 monthly new English language articles per month. There does appear to be some kind of a seasonal component, as there is regular peaks

[redacted]

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

around July to August each year.

ii) The Holt-Winters model gives a prediction of ~~[illegible]~~ ~~[illegible]~~ 15,195.98 for the total number of new articles published in November of 2020. It has a lower limit of -10,364.78 and an upper limit of 40,756.74. This is a large variation between the lower and upper limits, thus the prediction is likely to be incorrect. Also, the lower limit of the prediction is a negative value, which is not possible given the context, as a negative number of new articles cannot be published. A negative value has been given for the lower limit as an additive Holt-Winters model will ~~place~~ give more weight to recent values, and ~~[illegible]~~ the recent years have shown a decreasing trend in the total number of new articles published per month. Thus this model ~~would~~ is likely not be very appropriate to use to make a forecast for the total number of new articles published in November 2020, as a significant part of the prediction range ~~[illegible]~~ is in the negative values, which is not possible in the given context.

b) i) Firstly, ~~a~~ when a book was number one on the best-sellers list ~~for a particular week~~, it is coloured in yellow. For example, the Nightingale was 1st in the week of 20/12/15, and it has a ~~[illegible]~~ yellow bar for that week. Secondly, the darker the blue colour of a bar on the graph the closer it was to being number 1 on the best-sellers list ~~[illegible]~~ for that corresponding week. For example, the ~~[illegible]~~ blue bar for the week of 22/02/2015 for the Nightingale

[redacted]

[redacted]

was darker than that of the next week (the week of the 1/03/2015), as it was ranked 3rd and 11th respectively.

Thirdly, the height of the bar also indicates the ranking of a book in a particular week, as the higher the bar is, the higher the book is ranked for the corresponding week. For example, for the Nightingale, the height of the bar for the week of the 22/02/2015 was higher than that of the next week, as it was ranked 3rd and 11th respectively. * continued over on p 17

ii) A Game of Thrones appears to have appeared on the best-seller list over the longest stretch of time, as it first appeared in mid 2013 and its latest appearance was in mid 2019. There are sporadic appearances of A Game of Thrones in the best-seller list across the years. This is likely due to the fact that it is a TV show, and the sporadic surges in the book's popularity would likely correspond with when the seasons of the TV show were released. Fifty Shades of Grey held the no. 1 spot on the best-sellers ~~spot~~ list for the longest consecutive amount of weeks, for around all of mid 2012, likely around 30 weeks in a row. The other books in this series (Fifty Shades Darker and Fifty Shades Freed) also occupied very high spots on the best-sellers list throughout this time period, so this may have

L5

been when the movies version was released, causing a subsequent surge in the book's popularity. The Girl on the Train also held the number 1 spot for around 15 consecutive weeks in early 2015. The book then saw another surge in popularity when it rose to number 1 again for a few consecutive weeks in late 2016, which was likely due to the movie version being released. Looking at the books that were turned into movies, they often have 2 peaks, the first is likely when the book was first released, and the second is likely when the movie version was realised. Where the Crawdads Sing is the most recent book shown to have been no. 1 on the best-sellers list, appearing 1st throughout mid 2019.

lo

i) * continued: seen Additionally, the TV symbol next to the book title (as seen beside A Game of Thrones) indicates that the book was later developed into a TV series. The movie reel symbol next to the book title (as seen beside Gone Girl) indicates that the book was later developed into a movie.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

93201A

Subject	Statistics		*Standard*	93201	*Total score*	27
Q	***Score***	***Annotation***				
1	6	If the candidate for Qu 1aiii) had substituted more values into the equation of the linear model and worked out the average number of words per page, based on these values then at least one of the two marks available for this question could have been gained.				
2	4	For part b) the candidate has wrongly identified the randomisation test as the one to use to come up with a conclusion. The correct procedure to come up with a conclusion was a confidence interval based on bootstrapped resamples of the difference in mean (or median) times spent at the library between morning and afternoon.				
3	6	The candidate has gained one of the 2 Outstanding marks for part 3a) because they have provided evidence as the question asked for of "at least one numeric comparison of likelihood". In this instance the candidate has correctly calculated a relative likelihood that at 11am there is a 1.2 times as likely chance of reading non-fiction than fiction. For 3 cii) the candidate has not provided two factors (as asked for in the question) for sources of variation in the mean percentage completion values of reading e-books.				
4	5	In question 4b) the candidate has not been explicit enough about the insufficient evidence to make a causal claim that sans-serif fonts are easier to read. The statement made by the candidate – "there is not a significant difference between mean reading speed for sans-serif and serif font" – is more aligned with a concluding statement we would expect to see for a sample to population inference, rather than an experimental situation where an explanatory variable has been deliberately manipulated. Candidates were also expected to make a link to the distribution to gain the Outstanding mark for 4b) in response to why the results could have been expected, rather than just stating the participants were medical students and little differences exist between them.				
5	6	The candidate could have gained another mark on question 5 by giving a bit more focus, with numerical evidence on the lack of seasonality in the time series data. Linking this aspect to the wide prediction intervals and the difficulty in then providing a reliable forecast for November 2020 could also have ended in the first of the two Scholarship marks for this question. For Question 5 bii) the candidate has successfully compared several features of the books on the NYT Best Seller's list across different examples to gain the Outstanding mark.				

Confirmation of check	***Y / N***
This exemplar has been checked for similarities with current online exemplars.	Y/N