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

OUTSTANDING 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

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

(1)

(a) The scatterplot shows a positive, moderate, linear relationship between the number of pages and number of words in the 65 most popular English fiction books, where as the number of pages increases, the number of words also tends to increase. For every 500 ~~page~~ additional pages we see, on average, around a 150,000 word increase in the total number of words in this sample of popular fiction books.

The data ranges from around 10 to 2000 pages and from roughly 10,000 to 550,000 words, with very few books of pages greater than 1000 (only 2). The variation in number of words about the line of best fit is relatively constant between books of page length 0 to 1000 at around 200,000 words (50,000 below to 150,000 above). This is quite a range and shows the variability in word density for popular english fiction books, regardless of page number.

There are 3 unusual data values, one at ~~th~~ around (400, 400,000) (1900, 150,000) and at (2000, 500,000). The (2000, 500,000) point appears to ~~fit the~~ be close to the line of best fit and follows the average word per page density of just under 300 words per page so it is likely just a very long popular book. However, the books at (400, 400,000) and (1900, 150,000) have much different word densities at ~~about~~ roughly 1000 words per page and 80 words per page. These are unusual ~~in~~ but likely represent different print layouts than

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

a standard novel, possibly with large page sizes ((400, 400,000)) or lots of pictures ~~(page, 500,000)~~ (1900, 150,000)

(ii) One potential reason is that ~~the~~ the number of words is also impacted by the physical page size. Popular fiction books come in many different formats and some have particularly large/small page sizes. This impacts the variability in the number of ~~pg~~ words that can fit on each page and so will make the prediction using the number of pages less precise.

A second potential reason is that some fiction books contain pictures as well as words on their pages. Again, this impacts the number of words that can be printed on each page, increasing the variability of the number of words per page. This impacts the precision of the predictions made and so the number of pages may not precisely predict the number of words in the book.

(iii) Observing the coefficient of the explanatory variable on the equation for the linear model, we can adjust it to represent the average number of words per page:

$0.2202 \times 1000 = 220.2$

or 220 average words per page from the fitted linear model.

Although this value of 220 average words per page is 30 words less than the claim, we cannot disprove the claim

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

as the ~~[illegible]~~ fitted linear model was based ~~on~~ a specific sample of the 85 most popular fiction books in English. this sample may not be representative of all fiction books as it only includes popular English titles. Hence, our average words per page value of 220, although different to the claim, cannot be used to comment on the suitability of the statement 'for fiction books there are, on average, 250 words per page.'

(b)

(i) One claim made in the report is, for NZ adults who had read or started to read between may 2018 and march 2017, the difference of 2% (86% vs 88%) ~~was not~~ statistically significant.

For the March 2017 survey:
88% of NZ adults had read/started to read at least 1 book.

$n = 2082$

$MOE = \frac{1}{\sqrt{n}}$

$= \frac{1}{\sqrt{2082}}$

$= 0.021916$

$= 2.19\%$ (3SF)

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

For the ~~M~~May 2018 survey
86% of NZ adults had read/started to read at least 1 book.

$n = 2261$

$MOE = \frac{1}{\sqrt{n}}$

$= \frac{1}{\sqrt{2261}}$

$= 0.0210305$

$= 2.10\%$ (3SF)

Using the two independent surveys rule of thumb:

CI = difference $\pm$ 1.5 (Average MOE)

$= 2\% \pm 1.5\left(\frac{2.10 + 2.19}{2}\right)$

$= 2\% \pm 3.22\%$ (3SF)

$= [-1.22\%, 5.22\%]$

Since this confidence interval is between 1.22% in favour of May 2018 reading percentage and 5.22% ~~read~~ March 2017 reading percentage, we cannot claim the 2% difference between the two surveys was statistically significant. So the claim can be supported.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(c)(ii) potential

One non-sampling error is that the sample was collected through an online survey. This ignores people without access to the internet in the sampling population. Hence, the sample results may not be representative of all New Zealand adults reading tendencies.

potential non-

A second sampling error is that the sample was collected from people who responded to the survey. This means data was collected from people who self-selected themselves to participate. This could potentially create bias towards people who are more outgoing / want to share their reading tendencies. Hence, the sample's results may not be representative of all NZ adults.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(2)

(a)

(i) The bootstrap confidence interval between a median age of (108, 145) can be converted to the year of publish:

2020 - 108 = 1912

2020 - 145 = 1877

so the bootstrap confidence interval for the median publish date is (1877, 1912). Since this interval is entirely before 1924 (1912 < 1924), we can conclude that with 95% certainty most of the eBooks on the Project Gutenberg website are before 1924. Hence the claim 'most of their eBooks were published before 1924' can be supported.

(ii) The sample shows a mean difference in the age of the books of 4.58 years, in favour of Librivox over Project Gutenberg. The bootstrap distribution of this sample forms a 95% confidence interval of the mean age difference of the two website's books between 14.10 years in favour of Project Gutenberg and 22.76 years in favour of books from Librivox. Since this interval includes 0, we cannot conclude which which website has the greater mean age of books for their audio library.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(iii) We can use a binomial distribution to calculate the probability that chance alone could replicate the 19/24 english books from the sample with a realistic probability.

It is appropriate here as:

- We have a fixed number of trials (24 books)
- There are only two outcomes (either English or not English)
- The probability of success is constant (probability = 0.5 if exactly equal English and not-english)
- The probability of a book being English is independent of the previous book (random selection).

using: let x = number of english books.

$p = 0.5$

$n = 24$

$P(x \geq 19) = 1 - P(x \leq 18)$

$P(x \leq 18) \sim B(0.5, 24) = 0.996695$

$\therefore P(x \geq 19) = 1 - 0.996695$

$= 0.003305$

$= 0.331\%$ (3SF)

Since this is less than the 5% threshold, we can claim that the proportion of English books available from Project Gutenberg must be greater than half. So the claim can be supported.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(b) Firstly, a sample would need to be collected. This could be done by having everyone who visits the library record the time they enter and the time they leave when visiting the library. By doing this over a number of different days, and if possible, different times of the year, a representative sample mean of all people who visit the local library can be collected.

Secondly, the sample could be analysed. By processing the data into the two categories of morning and afternoon and the length of time they stay (numerical/discrete minutes), a multivariate box and whisker plot can be constructed and the mean stay time can be calculated.

Thirdly, a bootstrap distribution can be applied to form a 95% confidence interval of the difference in the mean stay times of the two groups for the total population. If this interval does not include 0, then the library can conclude that either the morning visitors stay longer than the afternoon visitors, or vice versa. They can use the mean difference from the sample to know how much longer, on average, they stay.

QUESTION NUMBER

Start each question on a new page.

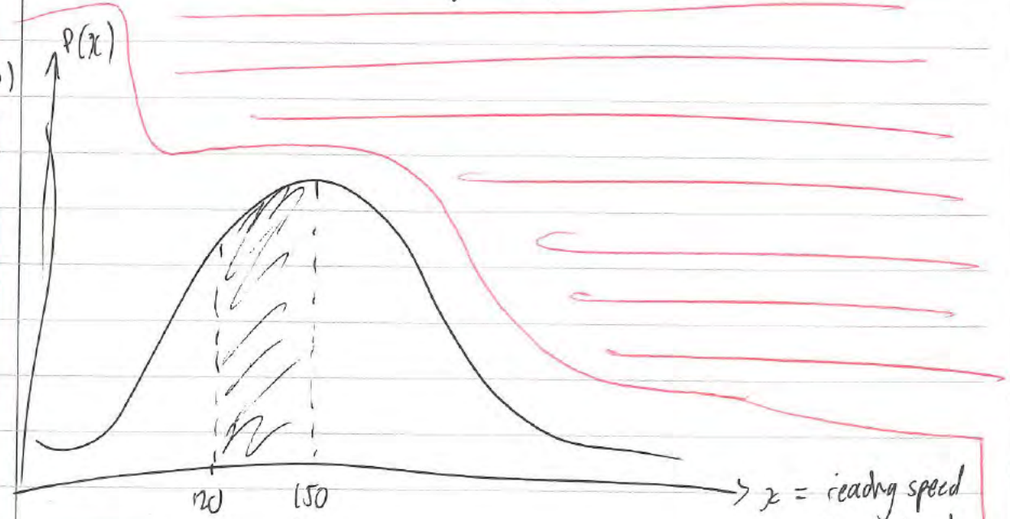
ASSESSOR'S USE ONLY

3

(a) Both readers of fiction and non-fiction books were least likely to read between 6 am and 8am (hours 6 & 7 of the day) with a probability of around 0.5%.

Non-fiction reading peaks at the 11th hour and 22nd hour of the day, with a probability of around 6%, whereas fiction reading peaks in the 23rd and 0th hours of the day with a probability of around 6.3%. This shows non-fiction readers are most likely to read in the late morning (11am-12pm) and late evening (10pm to 11pm) whereas fiction readers are most likely to read close to midnight (11pm to 1am).

(b)



let x = reading speed (words/min)

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

$P(120 \leq x \leq 150) = \frac{3310}{8000}$

$= 0.41375$

using: $z = \frac{x - \mu}{\sigma}$

using $N(0, 0.41375)$

$z = -0.217909$

$-0.217909 = \frac{120 - 150}{\sigma}$

$-0.217909\sigma = -30$

$\sigma = 137.672$

$= 138$ words/min

$z = -1.36$ (3SF)

$-1.36 = \frac{120 - 150}{\sigma}$

$-1.36\sigma = -30$

$\sigma = 22$ words/min (ODP)

113.8
106.9

186.2
193.1

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

using a normal distribution, $\mu = 150$

$\sigma = 22$

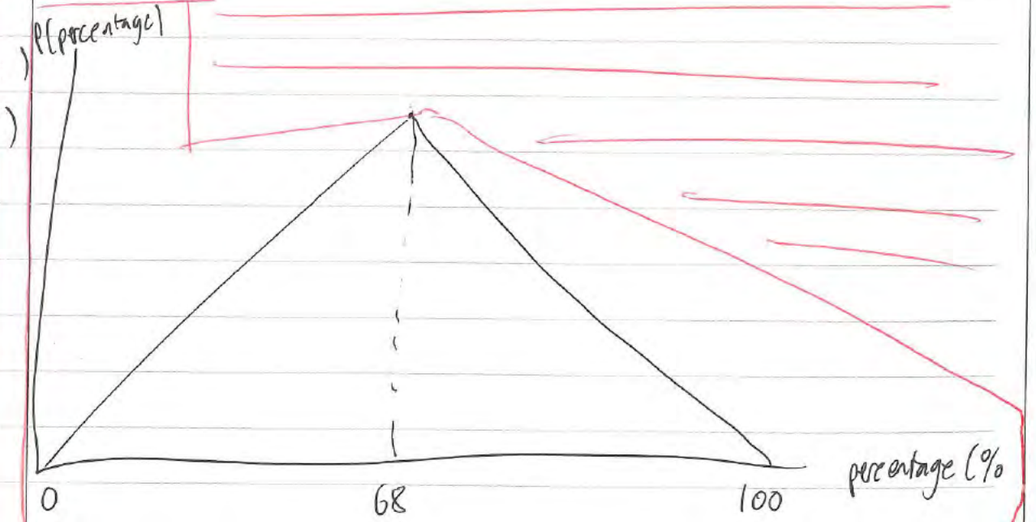
invN(150, 22, 0.025) for upper 2.5% value = 193.1192 = 193 words/min

for lower 2.5% value = 106.8807 = 107 words/min

So for 95% of the reading speeds, using the normal distribution, it be between

[107, 193] words a minute.

(c) (i)



QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

Looking at "about half of the books' mean completion value of 64% or less"

$P(\text{percentage} \leq 64) = \frac{1}{2} \times 64 \times \frac{2(64-0)}{(100-0)(68-0)}$

$= \frac{64^2}{100 \times 68}$

$= 0.60235$

$= 60\%$ (nearest percent)

Looking at "only about 5% of the books had a mean completion value higher than 90%"

$P(\text{percentage} > 90) = \frac{1}{2} \times 10 \times \frac{2(100-90)}{(100-0)(100-68)}$

$= \frac{10^2}{100 \times 32}$

$= 0.03125$

$= 3\%$ (nearest percent)

Although the second statement of about 5% being >90% completed" is close to our triangular model of 3%, the first statement of "about half being of 64% or less" has a probability of around 60% in our triangular model. This difference of 10% is too great for me to say the triangular model would be a good fit as it overrepresents the probability of lower completion books compared to the eBook subscription

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

companies' probabilities by too large an amount.

(ii) One factor that might explain the variation is the variation in the length of eBooks. Longer books require a greater time commitment to read, so will impact the completion variation of books from the eBook subscription company

A second factor is the variation in genres of books available. Not all books are of the same type and have varying content that is more/less likely for a reader to read completely. For instance a kids picture book will likely have a higher mean completion percentage than the Bible. Hence, this might explain the variation in mean percentage completion.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(4)

(a) the experiment of a sample of 238 university students used random allocation to form two testing groups to observe the reading ability of serif vs sans-serif fonts. Random allocation was used in an attempt to make the mean reading abilities of the two groups as similar as possible so that they would read the different font texts the same way, on average. The study used control variables such as keeping the text itself constant and the time allowed to read constant so only the explanatory variable: the font, would impact the response variable: the reading time. The experiment used blind groups so they do not know which treatment they are receiving

(b) the output shows a 1.75 words/min difference in mean reading speeds in favour of the Sans serif group.

The re-randomisation distribution shows a tail proportion of 0.4 ($\frac{400}{1000}$). This shows that chance alone was able to beat or equal the 1.75 words/min difference in the experiment with a probability of 40%. Since this is greater than the 5% threshold, we cannot reject chance as being a potential cause of the observed experiment results above (1.75 words/min mean difference towards sans-serif)

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

(c)

(i) The study collected information that shows:

- 80% of students were female, 20% were male
- [redacted] Paediatrics students, just over 80% were female, just under 20% male
- Of therapy students, just under 80% were female, just of 20% male.
- roughly 40% of students specialised in paediatrics, roughly 60% therapy.
- The median age for therapy students was 21 whereas for Paediatrics was between 22 & 23 (a roughly 1-2 year difference towards paediatrics)
- The spread of therapy students age was much greater than paediatrics: from 20 to 40 years (20 years) vs from 22 to 28 years (6 years) (14 year overall difference).

(ii) The experiment could have been modified so the sample was more representative of the population. They could have selected students so the gender distribution was 50/50. They could also have attempted to make the age distribution more representative, by likewise modeling the age distribution reflect the total population more. Centred at the mean with more [redacted] medical students of older ages (25+).

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

(d) Since this experiment was conducted with text in Russian (Cyrillic) the results can only be generalised for the reading speeds of the Russian language for serif and sans-serif fonts. Russian is a unique language with it's own alphabet (Cyrillic) and unique characters. Only other medical students who speak read in Russian can then be generalised to.

Because this experiment was conducted with well known text (history of Russian medicine for Russian medical students) the results from this study can only be concluded for the reading speeds of serif and sans-serif for common known texts. As reading speeds are quite different for reading new / unknown texts, we cannot generalise these results further than for reading well-known texts.

These results can also only be generalised to fluent speakers of / readers of Russian as the survey sample were all Russian. Non-fluent speakers will have much different reading speeds for the two fonts so cannot be generalised to.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(5)

(a)

(i) The overall long term trend shows an increasing trend in the number of new articles published in wikipedia from early 2001 to early 2007, from 0 to 50,000 new monthly works. This then begins to decline, showing a decreasing trend onwards to 2019, from 50000 monthly articles in early 2007 to around 20,000 new monthly articles at the end of 2018. This rate of decrease appears to be slowing from 2011 onwards,

There is an overall peak in July 2007, at around 65,000 new monthly wikipedia articles. This represents the highest point of volunteer contribution, which has since decreased. This is likely caused by either a reduction in the number of volunteers or the proportion of easily producable new topics for articles.

The seasonal trend shows relatively constant fluctuations each year besides a peak in the July-August months of around 3000 additional monthly new articles from the LTT. This is likely from the summer holiday season in the Northern Hemisphere.

There is a large peak in October 2002 at 38,000 new monthly articles, 30,000 above the LTT. This is likely when wikipedia opened it's site to volunteer contribution from the public.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(c)(i) the Holt-Winters model forecast predicts a central value of 15,195.98, or 15,196 new monthly articles for wikipedia in November 2020. This has a forecast interval of between: [-10,365, 40,757].

This interval is very wide, with a total range of 51,122 monthly articles. This is a large amount of variation and so does not produce a very precise/reliable prediction interval for monthly articles in November 2020.

Additionally, the lower bound of the forecast interval is negative, which is impossible for monthly wikipedia articles. Since this must be >0 monthly articles, it makes this prediction interval both unreliable and inaccurate.

these two reservations with the forecast interval also make me question the validity of the central forecast of 15,196 monthly articles for November 2020. Because whilst this central prediction produces an unreliable an inaccurate forecast interval, the Holt-Winters model may also produce a potentially unreliable central prediction from this dataset too. this model, therefore, is not appropriate to use for a forecast of New articles in November 2020.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

(b)

(i) One graphical technique is the use of colour to show the periods when a book is number in the Visualisation of New York Times best selling fiction books data (Figure 8)*. I believe this should be interpreted as the weeks in which the particular book was at #1 overall on the best sellers list.

A second graphical technique is the use of icons next to the book titles in figure 8. I believe this is to show if there is a visual format (either TV series or movie, 🎥 or 🎞) of the book available.

A third graphicle technique is the ordering of Book titles on the 'y' axis in order of first apperance on the NYT best sellers list. This is in figure 8. This shows the books in chronological order of their first apperance on the NYT best sellers list to help with clarity of interpretation.

* On the weekly lists data points

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

(ii) This visulisation reveals the recurring, almost exactly, annual presence of the "A Game of Thrones" book on the best sellers list. Every year, except 2018, the book has returned to the list during the middle weeks of the year (April to August). This coincides with the release of the Game of Thrones TV seasons and shows how interest in the book renews each time the TV show releases a new season.

This visulisation also reveals that it is uncommon for any of these popular fiction books to remain at number 1 on the best seller list for any longer than 6 months. The only two exceptions are "Fifty shades of Grey" and "How the mardads stay" (although it was on and off #1). Even then, no book remained longer than roughly 40 weeks. This shows the relatively short turnover of books at #1 on the best sellers list.

All of the 10 highest total books with the number of weeks on the list had at least 60 weeks total on the list the least was "A woman called Ove" with around 60, so at the 10th spot shows that to be within the top 10, your a book must remain popular longer than a year (at least 60 weeks) which, as seen in the other books, is achieved by

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

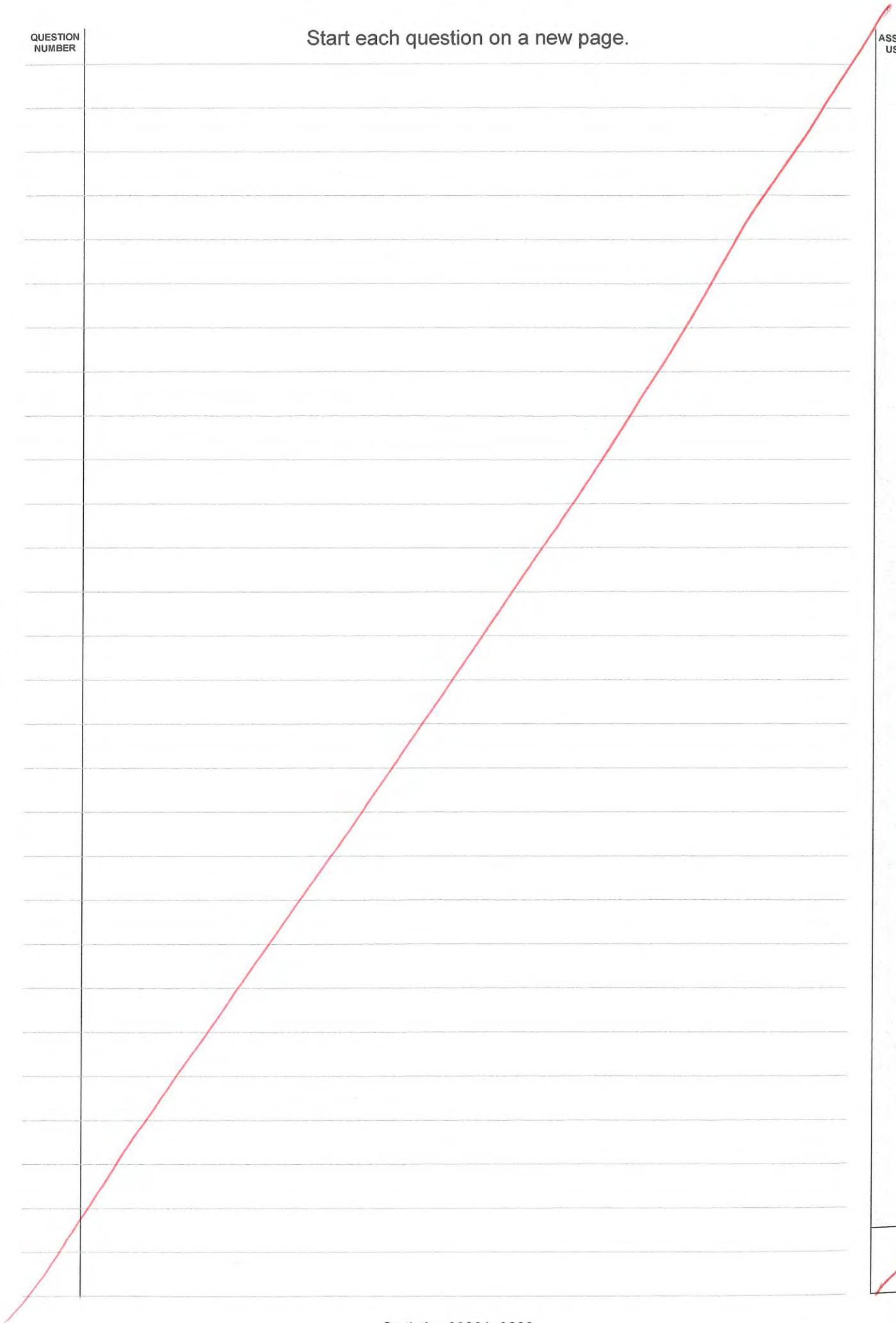
Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY



93201A

Subject	Statistics	***Standard***	93201	***Total score***	33

Q	***Grade score***	***Annotation***
1	7	For question 1 aiii) the candidate has provided a sufficient critique of the suggested linear model by linking the idea that the next occasion a sample for 65 fiction books is taken the linear model developed could be different due to sampling variation. Their suggested reasons for non-sampling errors (NSE) for 1bii) are too generic and do make the key link that the question was asking for, that the NSE are linked to how many books were read in the last year.
2	7	This question would have gained the full 8 marks had at least one more feature in the samples been discussed over and above the sample difference in means of 4.58 years. It was expected that candidates could also discuss the greater range in the sample for LibriVox book ages or the similar shape in observed sample data between 90 to 190 years for both LibriVox and Project Gutenberg books.
3	6	For question 3a) the candidate has not given sufficient detail in both the similarities and differences between fiction and non-fiction book reading rates at different times of the day and they have not provided any evidence of a “numeric comparison of likelihood” that the question asked for [this was one of the Outstanding marks for question 3].
4	7	The candidate would have gained the full 8 marks of this question but in 4b) provided no evidence for why the result was expected between the mean reading speed of serif and sans-serif fonts. The candidates bullet point approach to Qu 4ci) is one that helps encourage a succinct approach to summarising the information collected.
5	6	In question 5aii) the candidate has not sufficiently picked on the idea up the lack of regular seasonality in the time series is a major reason for the lack of reliability in future forecasts the wide prediction intervals. In Qu 5bi) the candidate has explained that a technique to do with colour exists but has not sufficiently explained what the technique is in terms of the infographic.

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