93201A

932011

S

SUPERVISOR'S USE ONLY

TOP SCHOLAR

NZQA

NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

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

Scholarship 2019
Statistics

2.00 p.m. Wednesday 13 November 2019
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	/8
TWO	/8
THREE	/8
FOUR	/8
FIVE	/8
TOTAL	/40

ASSESSOR'S USE ONLY

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

Q1

(a) From 1A:

A greater proportional of the new buses ran on diesel compared to the proportional of used ~~cars~~ buses that ran on diesel, with about 90% of the new buses being diesel but only about 70% of the used buses being diesel. On the other hand, while only about 10% of the new ~~cars~~ buses were petrol ~~about~~ about 3 times more of the used buses were petrol, 30% being so. While a very low (less than 5%) of new buses used an 'other' motor type this is not seen at all in the used buses. Overall, both new and used buses show the same trend in relative amounts of motor types, with diesel being the most frequent, followed by petrol ~~and~~ while other is rare or non existant. However, while of the new ~~cars~~ buses there were 9 times more diesel than petrol motors, there was only just ~~all~~ over 2 times the amount of diesel compared to petrol motors in the used buses.

From ~~2A~~ 1B:

The diesel buses in this sample have a much larger range of ages of 40 years and a much larger inter-quartile range (IQR) ~~than~~ of ages of about 18 years, than the petrol buses, whose

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

range is ~~about~~ only 16 years and IQR a mere 7 years. On the other hand, the range ~~of~~ of the 'other' motors is comparable to that of the diesel buses while its IQR is slightly larger at 20 years. ~~Of the medians~~ In terms of shift, all 3 IQR's overlap, with the petrol IQR lying within those of the diesel and 'other' buses. However, the median age for the petrol buses in this sample is about 4 years higher than the median age of the diesel buses, which is in turn about 7 years higher than the median age of the 'other' buses. This may be due to the ages of the diesel and 'other' buses being positively skewed (skewed to the right) while the petrol buses is slightly skewed to the right, but much more symmetrical in shape.

From 1C:

The 'other' buses have the greatest proportion of new buses*, with 100% new buses and 0% used. This is followed by the diesel cars ~~of which~~ which has the next highest proportion of new to used ~~cars~~ buses, with almost 80% of the diesel buses being new and just over 20% used. The petrol buses have the most similar proportions of new to used buses and is the only motor type * and lowest proportion of used buses,

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

whereby there is a greater proportion of used buses, at about 60%, than new buses, at about 40%.

From 20:
The new buses show a positively skewed distribution of ages, unlike the used buses which are almost symmetrical. The IQR of ages of the new buses is shifted to the left (to lower ages) compared to the used buses; there is very little overlap. The median age for new buses at about 5 years is significantly lower than the median age of the used buses at 20 years. Finally, as the lower quartile age for the used buses at 12 years is higher than the median age for the new buses at 5 years, the older 75% of used buses are older than the newer 50% of new buses.

(b)
(i) There is a positive, linear relationship between the bus mass and engine capacity of service coaches, where as the bus mass increases, the engine capacity also tends to increase. This relationship is of medium to strong strength for lower bus masses, but weak in strength for higher bus masses due to

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

the amount of scatter increasing as the bus mass increases – thus the data appears to be fanning.* Two outliers exist with master around 21000 kg but with unusually low engine capacities less than 1000 cc.

* The correlation coefficient at r=0.73 indicates that overall there is a moderately strong relationship between bus mass and eg engine capacity.

(ii) The correlation coefficient is linked to how close a fit the data is to a linear regression trend line. The outlier causes any regression line applied to the relationship between bus mass and engine capacity for minibuses to be much steeper compared to if the outlier was removed, where the trend line would be able to go directly through the centre of the main cluster of data. As a result, removal of the outlier (the minibus with a 5800 kg mass and a 5400 cc engine capacity) would improve the fit of the model to the data, therefore causing the value of the correlation coefficient to increase.

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

(c)(i) Based on the general linear model for all used buses:

Years registered with NZTA = $-0.655 + 0.6706 \times 40$
$= 26.169$
$= 26$ years (0 d.p.)

Based on the linear model for minibuses only:

Years registered with NZTA = $-2.796 + 0.8441 \times 40$
$= 30.968$
$= 31$ years (0 d.p.)

(ii) The second prediction of 31 years should be used as it is calculated from the trend line applied to the minibuses only. This is more accurate as it can be seen from figure 4 that the minibuses and service coaches show significantly different trends, thus using the general prediction would be less reliable as the inclusion of the service coaches which have a lower gradient causes the model to underpredict the potential number of years that a 40-year-old minibus has been registered with NZTA.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(iii) Reservations exist firstly because the buses used ~~in~~ in this data analysis are only a sample of all the minibuses registered with the NZTA. It cannot be certain whether these 500 buses are representative of the overall population*, and thus whether any predictions based on the data are truly valid. Furthermore, the prediction made in (i) is an extrapolation as the 40-year-old minibus is outside of the domain of the ages of the minibuses in the sample; one cannot know whether the trend in years registered with NZTA will continue in the same manner with respect to current age outside of the domain of the data provided from the sample.

* and their age and number of years registered with NZTA

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

Q2

(a) Both the total wage of Auckland transport and the population of the Auckland region show an increasing trend from 2006 to 2018, and in both cases the rated annual increase becomes greater (the gradien increases) from 2013 onwards.

Overall, there is an increase in transport wage from about ~~27500~~ 52500 thousand in 2006 to about ~~43000~~ 97000 thousand in 2018, overall an increase of about ~~55500~~ 44 thousand trips, or

average annual increase = ~~44500~~ / (2018 − 2006) = ~~7425~~ 3708

so there is an annual increase of ~~7425~~ 3708 about 308 thousand trips per year.

For the population, there is an increase from about ~~1875~~ 1375 thousand people in 2006 to about 1695 thousand people in 2018, an increase overall of 320 thousand people.

average annual increase = 320 / (2018 − 2006) = 26.7

So there is an increase of ~~26.7~~ 27 thousand people each year.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

Therefore, this claim is not justified as the number of trips using Auckland Transport increased by 3708 thousand each year while the population only increased by ~~26.9 people~~ 27 thousand people each year.

(b) Both trains and ferries show an increasing trend in usage from 2006 to 2018. However, ~~while~~ almost the train usage is significantly higher for all years, beginning at about 750 thousand trips in 2006 to about 5750 thousand trips in 2018, while the ferries have ~~only~~ about 1000 thousand trips in 2006, slightly higher than the trains, but quickly become much proportionally less used reaching only about 1550 thousand trips in 2018. Thus the trains show a much greater increase in usage, being used 7.7 times more in 2018 compared to 2006 ~~and~~ with on average a yearly increase of about 417 trips per year, while for the ferries they are used only 1.6 times more in 2018 than in 2006 with an average yearly increase of ~~46~~ 45.8 trips per year.

Interestingly, while both trains and ferries show distinct seasonality in their usage, this seasonality is inverse to one another. ~~While~~ The trains begin below the

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

trend by about 100 trips in January – March and slowly increase in usage to about 200 trips more than the trend in July – September where usage peaks, before falling to a minimum usage of about 200 trips below the trend line in October – December. On the other hand, ferries show the opposite trend and with slightly more swing - beginning at a peak usage of almost 250 trips above the trendline in January – March, then falling below the trendline to a minimum of about 200 trips below the trendline in July - September, then rising once more in October – December. Thus it seems that while trains became more popular as a use of transport over the winter, the ferries became more frequently used over the summer months.

(c)

(i) Model 1 predicts a bus usage of 19549.90 thousand for the second quarter of 2020 with a confidence interval of 17850.78 – 21249.01 (in thousands), while Model 2 predicts a bus usage of 20543.03 thousand for the same quarter with a confidence interval of 19680.01 – 21406.06 (in thousands).

Therefore, Model 2 predicts a higher forecast,

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

993.13 thousand trips higher, & and also a much smaller confidence interval of 1726.05 thousand than the confidence interval made by model 1 of 3398.23, almost 2 times larger. # Thus Model 2 ~~predi~~ makes a much more ~~or~~ confident forecast than Model 1.

(ii) The forecasts differ, for while Model 1 takes into account the bus usage from 2006 to 2018, Model 2 only considers more recent data from 2016 to 2018. ~~where~~ Therefore, the confidence intervals in which the ~~forecasts~~ ~~of~~ for the second quarter of 2020 are much larger for Model 1 as the model is fitted to data more widely varied, where slightly different trends are seen in periods between 2006 and 2018. On the other hand, the trend is ~~the~~ ~~data~~ comparatively more constant when simply considered over the last 3 years, hence ~~the~~ the confidence interval for model 2 is much smaller; it is more likely that the bus usage in 2019 and 2020 will continue

(both models give a forecast for quarter 2 of 2020 with a similar upper limit.)

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

in a similar manner as 2016-2018 rather than all of 2006-2018 as a whole. Furthermore, the trend from 2016-2018 has a slightly steeper gradient than other periods within 2006-2018, hence when only 2016-2018 is considered the forecast for quarter 2 of 2020 is higher than when Model 2 is used which relies on the average gradient of the trend from 2006-2018, which is much lower.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

Q3

(a) For the New Zealand sample, the mean busing travel time is 13.87 minutes higher than the mean walking travel time. Likewise, for the United Kingdom sample there is a similar difference between mean with the mean busing travel time being 11.49 minutes longer than the mean walking travel time.

From analysis of the bootstrapping distribution ~~of travel times~~ for the New Zealand sample, the population parameter is expected with 95% confidence to lie between the mean busing travel time being between 10.35 minutes and 17.26 minutes longer than the mean walking travel time. Since this interval does not include zero and the difference between means is positive for all re-samples, a claim can be made that in the New Zealand student population the mean busing travel time is higher than the mean walking travel time.

In a similar manner from analysis of the bootstrap distribution for the United Kingdom sample, it is expected with

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

95% confidence that the population mean busing travel time is between 7.90 and 15.20 minutes larger than the mean walking travel time. Again, this confidence interval excludes zero, so the claim can also be made for the United Kingdom student population that the mean busing travel time is higher than the mean walking travel time.

It can also be seen from the sample data distributions that students in the United Kingdom sample have a much greater range of travel times for both walking and busing than students in the New Zealand sample. For the United Kingdom sample, the range of walking travel times is 60 minutes while the range of busing travel times is 80 minutes, whereas in New Zealand this is only 40 minutes for both busing and walking. ☆ [see over page] Furthermore, while the busing and walking travel times for the sample of United Kingdom students are positively skewed for both transportation methods, for the sample of New Zealand students only the walking travel times are positively skewed; the busing travel times

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

appear almost negatively skewed.

[from ~~over~~ previous page]
In addition, for both walking and bus, in the sample of New Zealand students there are only 4 discrete travel times, whereas for the United Kingdom students there are far more. Coupled with the earlier distribution observation it can be concluded that there is much greater variation in travel times in the UK students sample than in the NZ students sample.

(b)

(i) Students, through random sampling where each student has an equal probability of being sorted into each group, have been placed in ~~each~~ either a control group or a treatment group of equal size (68 students in each). The control group received no daily motivational messages to walk or cycle to the university campus from the app whereas the treatment group did. Thus the explanatory variable in this experiment is whether or not a student received daily messages while the response variable is whether a student used

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

walking or cycling to the university each day or not. The experiment is essentially blinded as students would not know that some had been receiving daily messages and others not, meaning that they would not know whether ~~or not~~ they were part of the treatment or control group.

(ii) The mean number of days where a student who was sent daily messages walked or cycled is 1.72 days higher than the number of days where a student who was not sent daily messages walked or cycled. From the re-randomisation test, whereby the students were randomly resampled into two groups 1000 times and the difference between means plotted, it can be seen that the difference between means was greater than the difference between the means of number of days walked or cycled into the two groups of no daily messages and daily messages 26.6% of the time. Since this is above the statistical threshold of 10%, it cannot be claimed that ~~in the overall student population that~~ receiving daily messages increased the numbe

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

of days walked or cyded as opposed to being sent no daily messages by the app, as the difference between means seen in the experiment could easily have been due to chance based on the randomisation test output.

However, the result could have been expected in this output, since all of the students included in the experiment were part of ~~the trial~~ an environment club. Therefore, all of the students, regardless of the app, would be eager to help protect the environment by using sustainable and non-polluting transport methods. Furthermore, being asked to install an app on their phones and record on it the number of days which they walked or cycled could have introduced bias, as students may act differently to normal (that is, cycle or walk more often) knowing that their actions were being recorded, or just as easily they might lie to lift their self-respect or image of being fit and environmentally friendly, which would reduce the effects of the app.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

(iii) It would be helpful to know how far away each student lived from the university campus. This is because the further away a student resided, the less likely they would be to use cycling or walking due to time and physical constraints. This would ~~affect it~~ act as a confounding study in the experiment as so could therefore reduce the reliability of the result in (ii).

In addition, it would be necessary to know whether the students had been honest about their input of data. For example, those who received daily messages could be guilt-tripped into claiming that they had cycled or walked more days than they actually had, which could affect the difference between the means in each group and consequently the randomisation output.

Finally, whether or not the students knew they were part of a study* could have similar consequences ~~to~~ such as causing them to act differently or lie, introducing bias to the conclusions obtained from the study.

*and whether they had volunteered to do so

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

Q4.

(a) One strength of the study design is that it considered a range of trips in order to gain an overall unbiased representation of passenger activities across bus and train rides. This was first ensured by selecting a range of different lengths of both long and short distances, as for passengers on longer trips may be more inclined to engage in activities such as sleeping, using devices, ~~reading~~ reading or listening to music to fill up their time, whereas passengers on shorter trips might not commit to such activities, instead looking ahead or out of the window, or talking. However, the study did not consider extreme distances such as very short (less than 20 minutes) or very long (more than 2 hours) during which passenger behaviour and activities could be very different again.
Likewise, ~~also~~ a range of times of day were included in the study to account for passenger behaviour at various times of day, for example, passengers may be more likely to engage in more relaxed activities such as sleeping during night commutes. However, a challenge in data collection existed for safety reasons in that late night trips were not

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

included which could create bias, furthermore, while working as a pair the researchers could be more likely to be noticed by passengers, who then could adjust their activities ~~knowingly~~ knowingly produced inaccurate results. However, as an advantage two researchers would be able to compare classifications and observe more passengers, improving the quality of the data.

Another strength to the study is the allowance for multi-tasking, as is the well-defined codes for each activity, allowing data to be collected accurately and ~~efficiently~~ efficiently. However, another challenge could be the poorly-defined timing of the four-minute observation periods, as it does not account for passenger activities across the entire journey; a passenger could change their activity part way through; the four-minute period might not be representative of the entire journey.

(b)

(i) For the first confidence interval, we would need to know the sample size ~~for~~ for each mode of transport (bus and train) and the percentage of passengers who looked out the window for each, as well as the difference between these. For the second confidence interval, we

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

would need to know the total number of bus passengers, # and the percentage of bus passengers who look ahead or out the window and the percentage of bus passengers who read, as well as the percentage difference between these.

(ii) The first confidence interval is calculated to compare ~~compare~~ across two groups, Wellington bus passengers and Wellington train passengers. Therefore, the @ confidence interval will equal 1.5 times the average of the margin of error for each group.
In contrast, the second confidence interval is calculated to compare two proportions between a single group, Wellington bus passengers, so the confidence interval will this time be equal to twice the margin of error for the Wellington bus passengers.

(iii) To begin with, it is not certain whether the activities of passengers on Wellington trains and buses is the same # for commutes throughout New Zealand. For example, certain activities might be favoured in Wellington that are not favoured elsewhere in the country and

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

vice versa, meaning that the confidence intervals created might not accurately depict passenger activities throughout New Zealand as a whole. Furthermore, this sample only represents certain times of day, routes, distances and so on so may not be representative of all commuters.

(c) $RR = \dfrac{P(\text{taking} \mid \text{female})}{P(\text{talking} \mid \text{male})} = 2.1$

There are 402 females, therefore, number of males = 812 − 402 = ~~40~~ 410

The number of people taking was 125.

Let f = the number of females who were talking, and m = the number of males who were talking.

$\therefore 2.1 = \dfrac{\left(\frac{f}{402}\right)}{\left(\frac{m}{410}\right)}$

$= \dfrac{f}{402} \times \dfrac{410}{m}$

$\therefore 2.1 = \dfrac{410f}{402m}$

$\therefore 844.2m = 410f$

$\therefore f = \dfrac{844.2}{410}m = 2.06m$ (3 s.f.)

$m + f = 125$

$\therefore m + 2.06m = 125$

$3.06m = 125$

$m = 40.8$ (4 s.f.)

~~Therefore, there it can be estimated that there~~

$\therefore P(\text{male} \mid \text{talking}) = \dfrac{40.8}{125} = 0.326$ (3 s.f.)

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

So it can be estimated that 32.7% of the passengers observed talking in the study were male.

Q5.

(a)

(i) Amelia passes 11 bus stops, therefore, whether or not the bus has to stop to pick up passengers at each, and the number of passengers it has to pick up at each stop, will cause variation in her commute times. ~~with~~ ~~the~~ As the number of passengers which the bus has to stop for increases, so will the length of her bus commute.
Amelia also passes five intersections, therefore ~~she~~ the bus may have to stop for more or longer amounts of time at each intersection depending on traffic. The more traffic, the longer her commute.
Finally, Amelia sometimes takes her bus commute at different times of the day, during which the number of passengers needing to be picked up along the way and the amount of traffic may differ, causing variance within her commute lengths.

(ii) As necessary for a normal distribution model, Amelia's commute lengths show a unimodal, ~~appro~~ almost symmetrical shape, which would make a normal distribution model appropriate. However, a normal distribution model has

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

no upper or lower limit, and while Amelia's travel times have no upper limit there is a clear lower limit of 0 minutes. Therefore, a normal distribution model may be suitable, but not entirely accurate. Finally, for the future commutes to be modelled using a normal distribution it must be assumed that her bus commutes lengths will continue in the same manner and distribution into the future, and that the length of each future bus commute is independent of all other bus commute lengths.

(b)

(i) *of 17 minutes

The mode parameter, $c = 20$ minutes as seen in Jacob's simulation, appears to have been determined by adding the most likely bus commute time* to the average waiting time of 3 minutes. The lower limit of $a = 10$ would have been calculated by adding the minimum wait time of 0 minutes to the minimum travel time of 10 minutes. Similarly, the upper limit of $b = 30$ minutes appears to have been found by adding the maximum waiting time of 6 minutes to the maximum commute time of 24 minutes.

(ii) [continued on spare paper]

Achievement Standard Number 9 3 2 0 1

Question Number 5bii

Write your NSN here → 1 3 0 4 8 1 8 2 4

Supervisor **must** print name & sign here : S. FREEMANTLE

Assessor's Use Only

5b ii) The height of the triangular distribution at time $x=25$ will be given by $f(25) = \frac{2(30-25)}{(30-10)(30-20)} = 0.05$

The height of the triangular distribution at time $x=28$.

$f(28) = \frac{2(30-28)}{(30-10)(30-20)} = 0.02$

$P(X>25) = \frac{1}{2} \times (30-25) \times 0.05 = 0.125$

$P(X>28) = \frac{1}{2} \times (30-28) \times 0.02 = 0.02$

Since $P(A|B) = \frac{P(A \cap B)}{P(B)}$ and $P(X>28 \cap X>25) = P(X>28)$

$P(X>28 \mid X>25) = \frac{0.02}{0.125} = 0.16$

∴ The probability that Jacob's total bus-related travel time is longer than 28 minutes given that it is longer than 28 minutes is 0.16 or 16%.

This extra answer page MUST be placed inside the plastic bag with your answer booklet and handed in.

[redacted]

Assessor's Use Only

5biii) Jacob assumed that the probability of his bus commute taking ~~each length of time~~ any particular length of time between 10 and 24 minutes each day will remain the same each day, however, unidentified potential confounding variables such as roadworks, bus breakdowns or unusual traffic could increase this or otherwise change the time of his commute on any day. Furthermore, he assumed that the triangular model would be appropiate for his commute travel times, whereby the frequency of each successive time changed at the same rate of the travel time increased for each of $x < 20$ and $x > 20$, however, this assumption is unlikely to be valid - while the simulated travel times are approximately triangular in distribution, his recorded total bus-related travel times are anything but triangular, instead seeming more similar to either a normal or rectangular distribution.

Furthermore, the assumption that has to be made to use a triangular distribution that the length of each commute time is independent of the length of any other commute time may also not be valid, as the obstacles which could cause his commute to be longer, such as roadworks, might be long term and therefore affect multiple commutes causing his commutes to no longer be independent of one another.

This extra answer page MUST be placed inside the plastic bag with your answer booklet and handed in.