No part of the candidate's evidence in this exemplar material may be presented in an external assessment for the purpose of gaining an NZQA qualification or award.

SUPERVISOR'S USE ONLY

S

93201A

932011

Draw a cross through the box (☒) if you have NOT written in this booklet ☐

TOP SCHOLAR

Mana Tohu Mātauranga o Aotearoa
New Zealand Qualifications Authority

Scholarship 2023
Statistics

Time allowed: Three hours
Total score: 32

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 your answers in this booklet.

Make sure that you have Formulae Booklet S–STATF.

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

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

Do not write in any cross-hatched area (AREA DO NOT WRITE). This area may be cut off when the booklet is marked.

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

© New Zealand Qualifications Authority, 2023. All rights reserved.
No part of this publication may be reproduced by any means without the prior permission of the New Zealand Qualifications Authority.

1a) ① The proportion of energy used from electricity is far greater in the residential sectors and commercial sectors as opposed to the industrial and agricultural sector. Electric energy makes up over 50% of energy used in residential and commercial sectors, while electric energy only makes up roughly 25% of industrial energy use and 30% of agricultural use. In these two sectors, electric energy is not the highest used energy source (2nd highest in both).

② Across all four sectors, electronics and lighting are always only powered by electric energy. However, in the residential and commercial sectors, electronics and lighting are responsible for a much greater proportion of energy use as opposed to the industrial and ~~agy~~ agricultural sectors (20% for Res., 25% for Comm., 5% roughly for Ind. and Agri.)

③ Natural gas is mostly only used for heating + cooling across all four sectors, but makes up a much greater proportion of energy production source in the industrial sectors (25% roughly) as opposed to the other sectors (10% for Res., 15% for Comm., 5% for Agri.) Very small proportions of natural gas may go towards other uses.

④ Diesel use is most common / widespread in the agricultural sector, making up just under 50% of all energy production sources. In the other 3 sectors, diesel use is much less common, making up 5% of energy production in residential sector, 15% in commercial sector, and 15% in industrial sector.

1b)

i) difference in proportions = 60 − 39 = 21%.
since sample is large, use MOE = $\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{3034}} = 1.8139\%$.
since only one sample tested, CI = difference ± 2 MOE (rule of thumb)
so confidence interval = 21% ± 3.6278%
= (17.37%, 24.63%)

since 0% is not in this confidence interval, we can say that the proportion of NZ households using heat pumps is significantly greater than the proportion using electric plug-in heaters. At a 95% significance level we can say this difference lies between 17.37% and 24.63% higher.

ii) ① The total cost of electricity in the summer months (namely January) was much greater after installing the heat pump as opposed to before installing the heat pump. In Jan 2022, total electricity cost was $178 while the total cost in Jan 2021 was $122. As such, the peak observed during summer 2022 is much taller than the peak observed during summer 2021.

② During all seasons, the mean electricity costs per hour before and after installing the heat pump both steadily rise from 5am to around 8pm, with each season experiencing a significant increase in electricity costs from 6pm to 10pm before and after the pump was installed. The greatest costs were observed during winter with mean costs at 8pm being $0.32 an hour after installation and $0.275 before installation, compared to the costs at the same time during summer being $0.19 an hour after installation and $0.125 before installation. This is expected as winter is generally colder so the heat pump or other warming devices are likely to be used more, using more electricity.

③ Figure 3 shows that the total cost of electricity per month during winter dropped slightly after installation of heat pump, but still showed a considerable seasonal peak. In [illegible] Aug 2021, total electricity cost was $[illegible] $210, while in Aug[illegible] 2020 total cost was $240. Although this shows a minor decrease, the seasonal winter peak still remains after installation and has not affected total costs that much during winter.

iii) The two [illegible] time series data points are plotted on two separate axes. Whilst this does clearly distinguish data before and after installation, it makes it hard to compare visually. Table 1 could instead be graphed on four separate pairs of axes, one for each season. On each graph, a separate coloured line could be plotted to indicate mean electricity costs per hour before and after installation of the heat pump. This would make it easier to visualise the difference before and after installation for each individual season, due to the data pairings in question being plotted on the same graph.

2 a)

i). Calculate the ~~sum of the~~ differences between predicted + actual production each month (0, 400, 0, 300, etc...) + convert to percentages. By summing these values and using unbiased estimates, the deviation can be calculated to 0.57%.

$\bar{x} = \frac{\Sigma x}{n}$ deviation^2 $= \frac{1}{n-1}\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)$

(% difference in January $= \frac{\text{predicted} - \text{actual}}{\text{actual}} = \frac{4850-5000}{5000} = -3\%$)

ii) Use a binomial distribution, as population size is low (12 ~~mont~~ months) and result can be one of two outcomes (higher or lower).

Assume probability of predicted production being higher than actual production is 50%.

Model data by $X \sim B(12, 0.5)$ where X is months where predicted production is higher than actual production.

$P(X \geq 7) = \binom{12}{7} 0.5^7 0.5^5 + \dots + \binom{12}{12} 0.5^{12} 0.5^0$

$= 0.3872 = 38.72\%$

At at 5% significance level, the chance of $P(X \geq 7)$ is statistically probable so there is not enough evidence to claim that predicted production is higher than actual production more than half the time.

2b)

i) The raw data for both alertness scores with LEDs and incandescent lighting show a large amount of variation. Both conditions have outliers at considerably lower alertness scores (10 for LED, 8 for incandescent). The data for incandescent lighting also appears to be positively skewed. The raw data shows that the mean alertness score with incandescent light in 0.85 points higher than the mean score with LED lighting. However when rerandomised this difference disappears and becomes negligible. The re-randomisation distribution shows that the chance of the difference between the mean scores being 0.85 or greater (weirder) is 17.8%, which is much greater than 5% (using a 95% significance level). As such, it is likely the difference in means is simply due to chance and is not significant.

ii) The study would be repeated, but instead, each student would complete the task twice - once in LED lighting, and once in incandescent lighting. This removes the variable of some students having a greater ability to process written information, as the exact same group of students are being tested at each light conditions. Additionally, 50% of the students should be randomly selected to be tested under LED lighting first, with the other 50% being tested with incandescent lighting first. Since after completing the test once, a student is likely to perform better the second time around, using this method helps to remove this added variable from the overall data analysis.

3a)

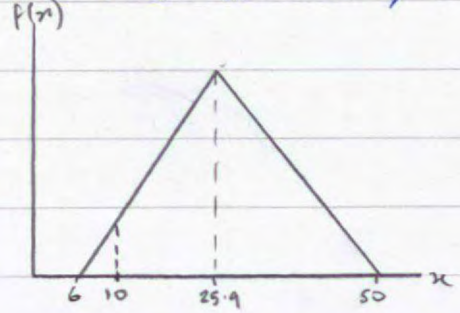
i) Using a median reduces the influence of outliers on the average as the median is simply the 'middle' value in a population, as opposed to the mean which is influenced by particularly large or small outlier values. Using the median means that the value of any outliers is irrelevant. When dealing with large quantities of data, it is often more useful to use a median as it removes the need for long calculations involving adding up every 'greenness' score and dividing by the total number of scores, which would likely be very tedious even with a computer software and would produce an answer with many decimal places in need of rounding. Calculating a median simply involves ordering all the data and taking the value in the middle position, with no rounding needed, which would influence accuracy.

ii). Vancouver likely has the higher standard deviation over Sydney for 'greenness' scores. The distribution for Sydney reflects a normal bell-shaped curve with a considerable peak at the median and steep slopes eitherside, showing minor variation about the mean. Vancouver's distribution is less normal with less steep slopes eitherside of the ~~mean~~. peak / median / mean. The tails of the distribution for Vancouver are much larger than for Sydney, showing more variation from the mean and more values further spread out, especially more lower values less than 15%. Hence, Vancouver's distribution shows greater spread and less normal-characters so likely has a ~~lower~~ higher standard deviation (a measure of spread from the mean).

3a iii)

Sydney: ~~normal distribution~~ triangular distribution

~~X ~ N~~ $a = 6\%$, $b = 50\%$, $c = 25.9\%$



when $6 \leq x \leq 25.9$, $f(x) = \frac{2(x-6)}{(50-6)(25.9-6)}$

so $f(10) = \frac{2 \times 4}{875.6} = 9.13659 \times 10^{-3}$

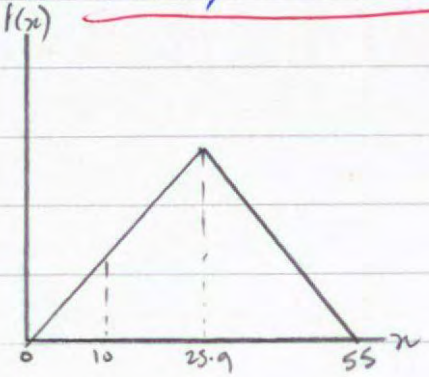
hence $P(x \leq 10) = (10-6) \times \frac{1}{2} \times f(10)$

$= 0.018273$

$=$ ~~1.827~~ 1.83% (3sf)

Vancouver: triangular distribution

$a = 0\%$, $b = 55\%$, $c = 25.9\%$



when $0 \leq x \leq 25.9$, $f(x) = \frac{2(x-0)}{(55-0)(25.9-0)}$

so $f(10) = \frac{2 \times 10}{55 \times 25.9} = 0.014040014$

hence $P(x \leq 10) = \frac{1}{2} \times (10-0) \times f(10)$

≈ 0.07020007

$= 7.02\%$ (3sf)

The use of triangular distribution is justified as we can observe the roughly highest + lowest values and know the highest value (median), but are not aware of deviation, variation, sample or population sizes, or means; hence, normal or binomial distributions are not suitable models.

3b) An observational study would involve randomly sampling a selection of light-coloured and dark-coloured roofed houses in Wellington, measuring the indoor temperature, and calculating mean values for houses with dark roofs and houses with light roofs. The sample data distributions could be bootstrapped to determine a confidence interval of differences of temperatures to then make a conclusion. An experiment would instead involve finding or setting up two identical houses in the same location, standardising all variables except roof colour. One roof is light, one is dark. This standardisation of variables is not present in an observational study, so any difference in temperatures observed may not be directly a cause of roof colour - instead, you can only determine a correlation if any. Since in an experiment, the only difference is roof colour, any difference in indoor temp. can be attributed to roof colour. An experiment requires continued measuring of the two houses (experimental units), whereas the observational study requires measuring of a much lager sample of households.

4a) Suburbs from each city were randomly selected as part of the sample, likely using a random number generator and a numbered list of suburbs in each city. While this sampling method is generally useful and random, there is a chance that the city with more suburbs (likely Auckland) will have more urban suburbs or more rural suburbs selected purely by chance as compared to the other city. As such, it could be useful to remove this influence by ensuring a specific amount of suburbs are selected from each smaller region in each city to ensure the sample provides an accurate representation of the city (20 from North Shore, 20 from South Auckland, 20 from West, 20 from East, 20 from central for instance in Auckland). Alternatively, select suburbs randomly from each region based on the proportion of total land area that region occupies (if South Auckland is larger than Central, select more suburbs from SA.) The coordinates + square used for each suburb were standardised by using the same method of sampling for each suburb: using the coordinates from Google Maps. However, this could be misleading, as Google Maps is likely to provide the coordinates of the suburb's town center which is more likely to be less green than other areas in the suburbs, simply due to urban sprawl. ~~and~~ A random sample of pixels were selected from the image then analysed to see how many were a shade of green. This sampling method, while certainly random, doesn't take into account any houses with green roofs, as not all green pixels will be a result of trees. This would hence distort the 'greenness' rating by likely increasing it above the actual result.

✗aA

4b) The mean greennes score for Auckland is 43·31 points higher than Christchurch. The sample data for Auckland seems to be positively skewed, with a few outliers at greeness scores much higher than the mean at over 900. The same can be said for Christchurch, with a few outliers at 620 points, but generally the data seems less varied about the mean. However, for both sample data distributions, there is no considerable increase in proportion at the means, indicating a large amount of variation for both data sets, even if Christchurch may be less varied. Christchurch also has a smaller range / spread of values (650) compared to Auckland (950). Even ignoring outliers, Auckland's spread is much greater (600 vs 400). By analysing the bootstrap distribution, we can see that the 95% confidence interval with mean difference being 43.31 is (-51.71, 135.14). Since there are negative values and zero included in this confidence interval, a difference in mean sample scores of 43.31 is not enough to say Auckland has a higher 'greennes' level, as the zero value being included indicates there may be no difference, and the inclusion of negative values in the CI indicates ~~that~~ Christchurch still could be more 'green' than Auckland. Therefore, the sample data provided is not significant enough to make any claims.

4c) The maximum distance to the city centre from a suburb sampled in Christchurch is 35km, although this seems to be an outlier. Distances to the city centre from Christchurch suburbs are mostly between 3km and 12km. This is to be expected as Christchurch is smaller in area so suburbs sampled are likely to be closer to the city centre as compared to Auckland. Among there Christchurch samples, there seems to be limited correlation between distance from CC (city centre) and greenness score. Two suburbs both with a score of 620 (roughly) differ in their distance from CC by around 30km. The same can be said for two suburbs both 5km from CC which differ in score by over 600 points. Alternatively, Auckland suburbs show a much greater spread of distances from CC (greatest being 66km) and also a much greater spread of greenness scores (greatest being 950). Generally, there seems to be a positive weak correlation between distance from CC and greenness score, with more variation experienced as both values increase. Both cities have outliers, with one Christchurch suburb significantly distenced from the majority of data points at coordinates (35, 620). Auckland has outliers below where the expected line of best fit would be, at very far distances from CC with rather low greenness scores.

4d) When selecting suburbs, ensure that no suburbs are selected with distance from CC greater than 40km, and the suburbs selected have distances from CC across the entire range of values from 0km to 40km. This removes the presence of outliers, especially in Auckland, and focuses the investigation more on suburbs closer to the ~~cen~~ centre of the city, making the greenness levels more directly comparable + fairer. 40km seems to be a good limit as this is where the majority of Auckland and Christchurch suburbs seem to fall under with only a few outliers having distances above 40km.