93201A

932011

S

12.30

SUPERVISOR'S USE ONLY

SCHOLARSHIP EXEMPLAR

NZQA

NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

Scholarship 2014
Statistics

9.30 am Wednesday 12 November 2014
Time allowed: Three hours
Total marks: 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 NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

1) a) The Diabetic rate vs Obesity rate graph shows a weak positive correlation between the two variables. As the percentage of the adult population who are obese increases, the diabetes rate also increases. For every 1% rise in the obesity rate, the Diabetic rate tends to rise by 0.0344%. The regression line intercepts the y axis at 2.3718%. I expect that this is due to the proportion of the population who have diabetes as a result of their genetic makeup. This will be approximately constant across the countries measured, and is not affected by the relationship between a correlation relationship with the Obesity rate of a country. The diabetic rate vs GDP per capita exhibits a moderate positive relationship, where as the GDP increases, the Diabetic rate also tends to increase. For every $10 000 increase in GDP per capita, the Diabetic rate tends to increase by 0.6445%. It is interesting to note that there is far more variation towards the lower end of the GDP in the Diabetic rate for countries with low a GDP. This may just be a fluke of our sample however, as the distribution of GDPs in our sample is significantly in the favour of low GDP values, which may provide a perception in the graph that there is more variation around the low GDPs. The Life expectancy vs Diabetic rate shows the strongest relationship of all the three graphs. It shows us that as Diabetic rate increases, the Life expectancy also increases. Where a 1% increase in the Diabetic rate would tend to mean a 2.0677 year increase in the Life

5

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

Expectancy. On the graph, it is interesting to note that the countries exhibit similar variation to and around the regression line throughout the graph, meaning that if we plotted the residuals of the data would exhibit no definite pattern. This means that the regression line calculated will be an accurate representation of the relationship a good representation of the relationship between the two variables of Obesity rate and Life expectancy.

5)c) Using GDP as an indicator:

$y = 0.6445(0.5) + 2.1002$

$= 2.42245\%$ $= 2.4\%$ (2sf)

Using obesity rate as an indicator:

$y = 0.0344(71) + 2.3718$

$= 2.4424 + 2.3718$

$= 4.81429\%$ $= 4.8\%$ (2sf)

- I would be much more confident using the ~~GDP~~ via regression estimated as using the GDP, as the relationship between the GDP and diabetes rate is stronger than the relationship between Obesity rate and diabetes rate shown in ~~Exhibit 2 maps~~ Fig 1 & 2. The stronger relationship means I can be ~~more~~ confident that the calculated value will be close to the estimate when I use the GDP to estimate than when I use the Obesity rate. I am reasonably confident ~~of this estimate~~, as I assume this assumption, as the two y intercepts on the two diabetes graphs are within an acceptable distance of each other (2.3718 in Fig 1 compared to 2.1002 in Fig 2). If I am correct in this assumption then these would be close to identical as neither the obesity rate or GDP would have any effect on the ~~regression~~ of the prediction with diabetes

It may be possible to generate a more accurate model using multivariate analysis to find a relationship between both Obesity rate and GDP, and how they affect the Diabetes rate.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

b) i) Clearly, my ~~[illegible]~~ estimates of 2.4% and 4.8% were drastically off from the actual value of 7.2%. Therefore there must be significant confounding variables that must have a major exception from the trends identified in the graphs. Looking at the information presented, I expect the reason to be that poor state of Nauru's healthcare. The poor condition of a country's healthcare system would mean that they lack the ~~[illegible]~~ prevention and treatment of diabetes patients and ~~[illegible]~~ the education of those who are at risk of developing diabetes due to their lifestyle habits would suffer. This could be the reason for the ~~[illegible]~~ significantly higher diabetes rate than the model's predictions. S

ii) With 95% confidence, I can say that in 2006/7 the Obesity rate ~~[illegible]~~ for New Zealanders could have been as low as 25.5% or as high as 27.5%, and that in 2012/13 the ~~[illegible]~~ obesity rate could have been as low as 27.4% or as high as ~~[illegible]~~ 29.8%. As there is overlap between these two ranges, (i.e. ~~[illegible]~~ it is possible that the obesity rate dropped from 27.5% in 2006/7 to 27.4% in 2012/13), the claim that the obesity rate is declining is unsupported. // P

Q1. SSP

6

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

2) a) ii) Random assignment was used when forming the groups to ~~[illegible]~~ minimise the effect of non-sampling errors such as self selection bias. For example If the students were able to choose what group they were placed in, then those who were wanting to lose weight may have chosen to go into the group that did activities. ~~[illegible]~~ ~~[illegible] affected by this, as [illegible] students [illegible] a lot more variables uncontrolled, which may~~ ~~[illegible]~~ Not randomly assigning students leaves a lot more variables uncontrolled (desire to lose weight for example), which may have a confounding effect on the results obtained.

iii) From this output, I can conclude that there is a 8.4% probability that the difference visible at the end of the experiment was due to chance acting alone. The rerandomisation process ~~[illegible]~~ randomly assigns ~~[illegible]~~ the ~~[illegible]~~ results of each participant into one participant ~~[illegible]~~ many times, the purpose of this is to emulate the variability within the population, and to find out how likely it would have been to obtain ~~[illegible]~~ the same, or greater difference if we had divided our two groups identically. The 8.4% result generated ~~[illegible]~~ is ~~[illegible]~~ large enough to make me question whether the difference visible is due to chance acting alone, ~~however~~ or due to the different treatment of the groups. I am not confident that the claim that the activities provided ~~[illegible]~~ help children stop being obese is supported by this study.

P

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

4) i) The two surveys found that with the sample of people who took the survey, there was a drop in the mean number of sugary drinks consumed of 0.6 (4.9-4.3). Interestingly, the second sample results exhibited less variation than the first. This is reflected in the drop in the standard deviation from 2.6 to 2.1, and the decrease in the ~~interquartile~~ range covered by the middle 50% of the sample from 4 in the first survey (7-3), to 2.5 in the second survey (5.5-3). Both graphs show the distribution to be approximately normal, ~~although the data~~ although the second survey has a normal distribution more than the first, as it ~~was~~ is skewed slightly towards higher values.

ii) The purpose of the bootstrap distribution is to emulate what would happen if we took a very large number (usually 1000 using ~~weights~~) of samples from the population. It does this by randomly resampling ~~the sample groups from~~ ~~to participate from~~ each the two survey response groups from the existing results obtained with replacement, ie ~~the~~ ~~person's response would~~ may be present in the resampled sample multiple times, while others may not be present at all). The ~~data~~ ~~is~~ is then plotted on the ~~to~~ bootstrap distribution. ~~The~~ process is then repeated many times to produce the distribution shown. ~~Finally the top and bottom 2.5% of the data on the bootstrap distribution are removed and~~

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

~~enough and we note~~ Finally, we take the range covered between the 2.5th percentile and the 97.5th percentile and use this for the range in which we can be 95% confident that the ~~difference~~ population difference ~~in the number of sugary drinks consumed as a result of the~~ in the number of sugary drinks consumed prior to, and following the education campaign lies between. Using the bootstrap distribution, I can be 95% confident that on average NZ secondary school students would drink between 0.1d and 1.24 sugary drinks less following the ed education and advertising campaign. Therefore, as 0 is not within this range, I can ~~can be confident in saying~~ the claim that the campaign made NZ secondary school students drink less sugary drinks is supported.

Q1) P

Q.2. PSP

5

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

3)a) Figure 8 shows a clear upward trend in the Obesity rate of both Men and women in New Zealand. The percentage of men who were obese grew from 25.5% in 2004 to 29.0% in 2013, at an average rate of 0.39% per year

av rate of growth = $\frac{\text{final} - \text{initial}}{\#\text{ time periods}}$

$= \frac{29 - 26.5}{4}$

$= 0.388889$

$= 0.39\%$ (2sf)

The male line shows a far greater consistency in the rate of increase of the Obesity rate than the female, as evidenced by the much straighter line. The female line also shows a clear upward trend, although it is much more erratic and varies widely to the point where it actually drops in the Obesity rate in the 2006-07, 2008-09 and 2011-12 periods. The female obesity rate climbed from 26.4% to 29.5% over the period from 2004 to 2013, at an average rate of 0.34%

av rate of growth = $\frac{\text{final} - \text{initial}}{\#\text{ time periods}}$

$= \frac{29.5 - 26.4}{9}$

$= 0.3444$ ✓

$= 0.34\%$ (2sf)

The female obesity rate is consistently higher than the male obesity rate across this period //

QUESTION NUMBER

Across the period Con 2004-2013 Start each question on a new page.

ASSESSOR'S USE ONLY

Figure 9 shows an increasing trend in the Obesity rates of both 15-24 year olds from 12.6 - 22.3% at an average rate of 1.1% per year and 35-44 year olds from 27.6% to 32.8% at an average rate of 0.58% per year, It also shows a decreasing trend in the Obesity rate of 55-64 year olds from 37.1% in 2004 to 31.9% in 2013, at an average rate of -0.58% per year. The 15-24 and 55-64 year old age groups show similar or fairly consistent rate of change across the period, while the trend lines for the 35-44 age group does not follow an approximately linear path. From 2004-2008, the Obesity rate for 35-44 year olds rapidly climbs at an increasing rate till it reaches its maximum at 33.9% in 2008. In the next period the an Obesity rate dropped to 32.5% and remained relatively constant through till the end of the time period //

Also answer for 3d

5

4) Using my graphics calculator I plotted the data points for the ~~across~~ male Obesity rates across the period and used the calculator to generate a linear regression line that I ~~will~~ will use to model the trend. The model generated was //

$$y = 0.3861x + 25.5$$

where y is the obesity rate and x is years since 2004. This line had a correlation coefficient of 0.9449 (3sf) so I can be confident that it is a good tool for

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

predicting the Obesity rate in 2015.

Using this model to predict the male Obesity rate in 2014

$y = 0.3561(10) + 24.5$

$= 29.4\%$ (3 sf).

I used a similar method to generate a regression line for the female obesity rate. ~~[illegible]~~ This regression line had the following equation:

$y = 0.3061x + 26.54$

This line had a correlation coefficient with the time series factor of 0.924. While this still means that the model generated will be reasonably accurate, the less consistent nature of the female trend means that I will be less confident with this estimate than I was with the one for the males.

Using the equation to predict the ~~Obesity~~ female Obesity rate in 2015:

$y = 0.3061(10) + 26.54$

$= 29.6$ (3sf)

c)) I notice that the female Obesity rate tends to decrease every 3 years (i.e. in 2006, 2009, 2012). The cyclic nature of this may mean that there is some confounding variable ~~[illegible]~~ affecting female obesity rates which acts on a 3 year cycle.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

P

f) The difference in the two samples (assuming ideal experimental design) is due to sampling variation. When ever we take a survey, we are looking only at a small proportion of the total population. These two surveys would have taken different samples, ~~from the same population,~~ and what they would be from the same population, it is highly unlikely that they would have produced the same results. The 0.2% difference in the two conclusions is due to the two samples each containing different members of the same population. This difference would be well within the range for sampling error, and therefore it is not a concern. This is assuming the experiments were ideally designed however, as there may be some non-sampling error in one of the surveys that is partially to blame for the difference of 0.2%.

g) See Barber information on this //

Q3 SNP

5

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

41

65-74

people in 65-74 age group Obese = 6.14 × 111000
= 196100

people in 65-74 age grp = $\frac{156100}{386} \times 100$
= 404404
≈ 404,000 (3 sf)

N

	Obese	Not
65-74	156100	404,000

~~don't~~ P(Obese | 65-74) = $\frac{156100}{}$

a) X ~ N Binomial (1000, 0.386)
I am using Binomial as:
- we can treat each of the 1000 people as a trial
- 1 person is either Obese or not
- The probability of one person in our random chosen sample being obese is not affected by whether any other person is obese
- There is a constant probability of someone being obese of 0.386

S

P(X ≤ 360) = 0.04836 (4 sf) //

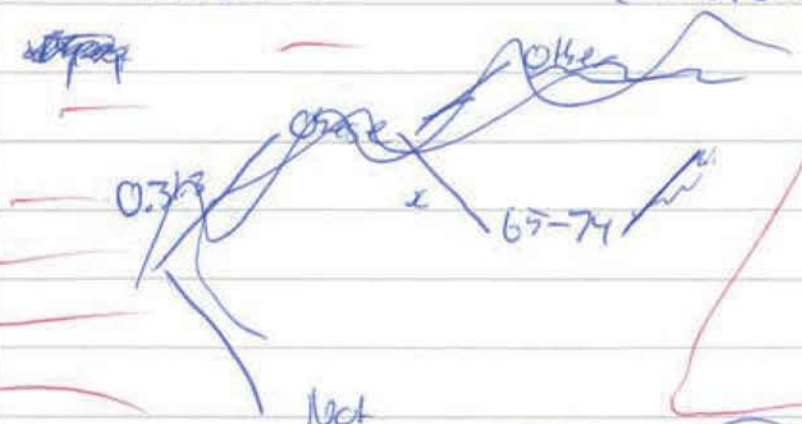
QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

4) # people in 65–74 age group: 0.14 × 1,114,000

& obese = 156100



$P(\text{Obese} \cap 65\text{–}74) = 0.313 \times x$

$P(\text{Obese} \mid 65\text{–}74)$

$P(65\text{–}74 \mid \text{Obese}) = \frac{P(65\text{–}74 \cap \text{Obese})}{P(\text{Obese})}$

people in 65–74 age group

$= \frac{156100}{386} \times 100$

$= 404{,}000$ (3sf)

people over 15 $= \frac{1{,}114{,}000}{31.3} \times 100$

$= 3{,}562{,}300$

$= 3{,}560{,}000$ (3sf)

$\therefore P(65\text{–}74 \cap \text{Obese}) = \frac{156100}{3{,}562{,}000}$

$= 0.04382$

$P(65\text{–}74 \mid \text{Obese}) = \frac{P(65\text{–}74 \cap \text{Obese})}{P(\text{Obese})}$

$= \frac{0.04382}{0.313} = 0.14$

P

QUESTION NUMBER

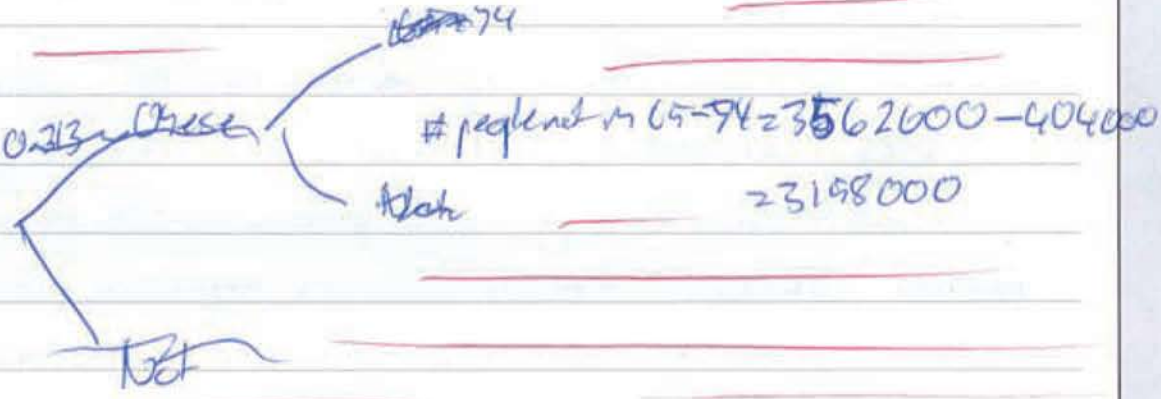
Start each question on a new page.

ASSESSOR'S USE ONLY

c) # obese people in 65-74 = 156100

obese not 65-74 = 1,115,000 - 156100

= 958900



P(Obese given not in 65-74) = $\frac{958900}{3158000}$

= 0.3036 (4 sf)

$\frac{P(\text{Obese given 65-74})}{P(\text{Obese given not 65-74})} = \frac{0.386}{0.3036}$ ✓

= 1.271 (4 sf)

A person in the 65-74 age group is 1.3 times as likely to be obese as someone not within this age range. //

Q.4.

NPO

7

across the 23 years from 1974 to 1997

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

5) c) i) In Brazil, the prevalence of overweight children (6-9) increased from 4.9 to 17.4 in our samples, an increase of 12.5%, although the increase could be as much as 14.03 or as low as 10.97% in the total population of Brazilians aged between 6-9. This ~~is~~ ~~was a certain confidence interval~~ ~~of~~ sample growth was 0.54% per year, although this could be as low as 0.4% or as high as 0.61% in the population. The trend shown in Brazil is totally different to the trend in overweight prevalence in the same age group in Russia, where the ~~gap~~ decreased by 16.2 percentage points across the period from 1992 - 1998, although this change could be as low as 13.84 or as high as 18.56. Whilst we can make a comparison across these two countries, there is a large potential for non-sampling error. This is because the time span the information was taken across is far longer in Brazil than it is in Russia, whilst the end dates are similar, the starting date of 1974 for Brazil and 1992 for Russia are far to far apart for any meaningful insight to be gained. The Brazil values may have experienced significant change due to the same global trend that would not have affected the data from Russia as they were yet to start recording their data.

In Russia, the prevalence of thinness amongst adolescents ~~from~~ dropped from 11.5 to 8.5 across this period, although this ~~drop~~ ~~is~~ whilst

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

The prevalence of obesity amongst adolescents in China rose from [illegible] - [illegible]. The Russian data show a clear drop, as there is no overlap in the confidence intervals. However, the Chinese ~~data~~ ~~show only a small increase~~, which could be as ~~there~~ low as 0.68 or as high as 7.72 percentage points. The change in Russia ~~is a result of a~~ suggests ~~either~~ the change of one or more factors which are strongly influencing the prevalence of obesity amongst adolescents. The author suggests this to be the period of tremendous economic stress the country went through

It is interesting to note that in America, it is the lowest income households that ~~have~~ have the highest prevalence of obesity. This ~~is~~ contrasts with Brazil, where highest income households have a higher prevalence of obesity, ~~as was~~ in 1997, (although their 1974 results show the lower income households to have a higher prevalence of obesity), and China and ~~Russia~~ Russia, who have relatively even ~~distributions of~~ proportions of obesity across the three income levels measured.

iii) The prevalence of underweight children in Brazil dropped significantly from 13.3 to 6.1% across the period from 1974 to 1997, at an average rate of 0.27% per year. ~~Russia~~ Whilst this

/

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

is the largest drop //

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

4) The visually recognisable significant variation in the time periods covered, across the four countries mean that it is difficult to make any comparisons between them. My suggestion would be to adapt the table to include an average rate of change per year column, as this would be very useful to gauge longer trends across the countries. This would minimise the influence of the confounding variable of the time span between the two surveys in each country, by allowing readers and analysts to clearly compare the rate at which obesity prevalence is changing across these three countries.

2P

My suggestion for the Figure 1 would be to include the information on the prevalence of underweightness. Whilst the figure is a good tool to see how income affects overweightness across the three countries, it provides no information on the underweight, which is needed to show changes. //

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

c) The question is asking me to look at the changes of over and under weight producers. I would expect that an analysis of the changes in these variables would include an analysis of the rate at which they are changing, especially since the absolute values are difficult to compare across the 4 countries as they are measured across sequentially different time periods. N

The accompanying text has suggested to help the source identify changes in the prevalence of underweight, which is not required to support the argument. The table currently provides no information that can be used to argue any thing about changes in underweight. N

Q.4. 2P 2

Scholarship

1. Three points were correct from the first two graphs so an S was scored.
2. The spurious correlation was not picked up due to the fact that the obesity rate was correlated with both life expectancy and the diabetic rate.
3. Three points were picked up namely the best prediction value, an appropriate justification as to why it was best along with why Nauru could be regarded as an outlier with its very high obesity rate.
4. There was no mention of the difference in the point estimates which would have scored an S.
5. There was no mention of the actual difference along with the tail probability so answer was incomplete so a P was scored.
6. All three distinct comparisons were made; the drop in mean, the reduction in variation and a comparison of distributions.
7. There was no mention that we were looking at a range pertaining to the mean number of sugary drinks so a P was scored.
8. The prediction for the male obesity rate was incorrect and no account was taken of the fluctuations in the women's obesity rate in the other forecast calculation.
9. This question part was asking for the percentage change in the overall obesity rates and wasn't a repetition of part (a)'s requirements.
10. Only two points earning credit were made; a correct trend comparison between Brazil and Russia for overweight and a comment on how the data presentation in the table could be improved.