ADM 2304 -- ASSIGNMENT 1

Note that Minitab output does not substitute for the different elements of a hypothesis test—it should be viewed strictly as a calculator. The hypotheses, decision and conclusion must be written or typed and show separately from any computer output for marks. No marks are given for a solution that only provides Minitab output.

1. [9 marks]

The file **toronto.mtw** contains data on the median incomes for neighbourhoods in Toronto.

a. Treating the 849 incomes as the population, use Minitab to calculate the population mean. Now set aside all population information until part d.

```
Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3
medinc    849   0  25614       327   9529     8332  18924   23429  30127
```

1 mark for the population mean of $25614.

b. Using Minitab (Calc Menu – Random Data – Sample from Columns), draw twenty samples of size n = 31 from the population. This procedure must be replicated twenty times (note that if you open up the same sampling dialog box each time from the menu, then you only have to replace the last destination column with the next one). For each sample, use Minitab to calculate a 90% confidence interval estimate for the population mean, assuming you do not know the population standard deviation (this interval estimation can be done in one operation on all twenty columns).

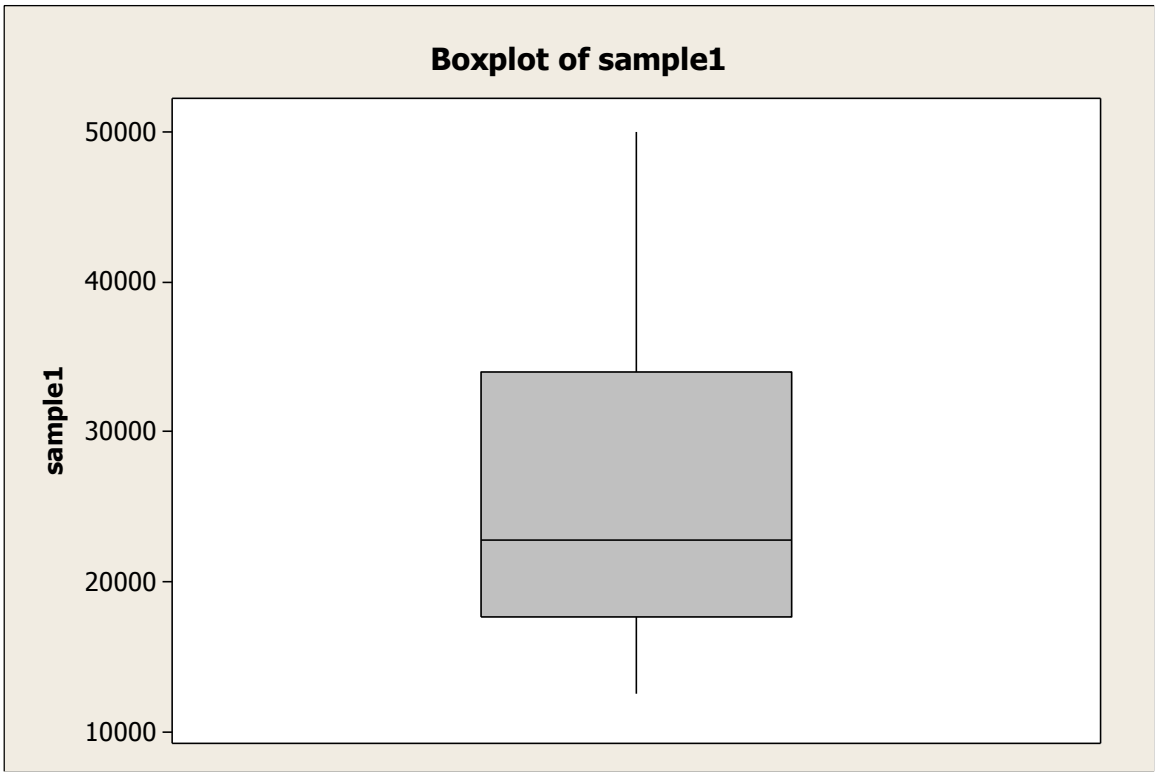
One-Sample T: sample1

```
Variable   N     Mean   StDev  SE Mean         95% CI
sample1   31  25562.8  9689.5   1740.3  (22008.7, 29117.0)
```

Every student should have twenty of these intervals and they should all be different.

3 marks for showing the twenty CIs.

c. For your first sample, confirm the Minitab generated interval by calculating the interval manually. Display the sample data graphically and comment on whether the relevant assumption regarding the population distribution is warranted (state clearly the assumption needed to justify the interval estimation. Note that the population values do not have to be normally distributed).



The boxplot shows that the population income data can be assumed to be ***not extremely skewed***; but in general, the sample will be somewhat skewed.

Since the sample is considered large (n > 30), we can safely assume that the sample mean has a sampling distribution which is normally distributed.

3 marks
-1 mark for any graph
-1 mark for manual calculations
-1 mark for commenting on commenting on large sample size and reasonableness of assumption that the distribution is not extremely skewed (no marks for comment that requires the population to be approximately normal since sample is large).

d. Of your twenty intervals, how many contain the value of the population mean from part a?

1 mark for the count

e. What is the (binomial) probability that a student counts eighteen of his/her twenty intervals cover the population mean?

```
Binomial with n = 20 and p = 0.9

 x  P( X = x )
18    0.285180
```

1 mark for the binomial probability of 0.285, if the Minitab or manual calculation is shown.

2 **[10 marks]**

Ohio is an important "swing state" in every presidential election—it has traditionally voted for the winning candidate. In 2012, Barack Obama won 50.67% of the vote in Ohio, with 47.69% going to Mitt Romney. A recent poll of 1000 respondents found that Hillary Clinton was favoured by 46% with Donald Trump at 39%.

a. Test at the 0.01 level of significance whether this is sufficient evidence to show that Clinton's support is lower than Obama's popular vote was in 2012. Show your manual calculations.

```
Test of p = 0.5067 vs p < 0.5067


Sample    X     N  Sample p  99% Upper Bound  Z-Value  P-Value
1       460  1000  0.460000         0.496665    -2.95    0.002

Using the normal approximation.

Ho: p = 0.5067; Ha: p < 0.5067

Z = (0.46 - 0.5067) / sqrt(0.5067*0.4933/1000) = -0.0467/0.0158 = -2.96

Reject Ho since z < -2.326, conclude support has dropped.
```

*4 mar*ks:
1 for hypotheses (stated separately from Minitab output)
1 for showing how the z-statistic is calculated
1 for showing p-value (from Minitab or as P(z < -2.95)) or rejection region of < -2.326.
1 for showing decision to reject Ho and conclusion support has dropped (0.5 each)

M = 0.01, z = 2.575 for 99% CI, p-hat =.044, q-hat =.956

$n = pq\,(z / M)^2 = .044*.956 * (2.575/0.01)^2 = 2789$ (accept range from 2778 to 2799)

2 marks:

b. What sample size would be required to obtain a 99% 2-sided confidence interval for the true level of Clinton's support with a margin of error of ±1%?

M = 0.01, z = 2.575 for 99% CI, p-hat =.46, q-hat =.54

$n = pq\,(z / M)^2 = 46*.54 * (2.575/0.01)^2 = 16471$ (accept range from 16406 to 16534)
or
$n = 0.5*0.5*(2.575/0.01)^2 = 16577$ (accept range from 16512 to 16641)

2 marks:
-1 mark for z-value and M value (accept z=2.57 or 2.58)
-1 mark for use of proper formula and calculation, using either p=0.46 or p=0.50

c. Suppose that, in a random sample of 17 Ohio State students, only 3 indicated a preference for Donald Trump. Test whether this is sufficient evidence to indicate that the level of support for Trump among Ohio State students is lower than the 47.69% share of the popular vote captured

by Romney in 2012. Use the .05 level of significance and explain how you would calculate the p-value for this test. (Hint: explain first whether the normal approximation should be used to answer this question.)

```
Test of p = 0.4769 vs p < 0.4769

                                                Exact
Sample  X   N  Sample p  95% Upper Bound  P-Value
1       3  17  0.176471         0.395641    0.011



Test of p = 0.4769 vs p < 0.4769


Sample  X   N  Sample p  95% Upper Bound  Z-Value  P-Value
1       3  17  0.176471         0.328553    -2.48    0.007

Using the normal approximation.
The normal approximation may be inaccurate for small samples.
```

The second test uses the normal approximation which is inappropriate, and the first is the exact calculation using the binomial distribution.

4 marks:

1 for hypotheses of Ho: p=.4769, Ha: p<.4769
1 for recognizing normal approximation not appropriate since we only observe 3 out of 17 or np = 17.4769 < 10.*
1 for calculating p-value of 0.011 (whether using Minitab or binomial calculation)
*1 for **decision to reject** Ho and conclusion that there is sufficient evidence to conclude support lower than .4769.*

Note that if the solution uses the z-statistic and the normal approximation to find the p-value, then give the 2 marks only for hypotheses and decision/conclusion.

Some students may make the adjustment from section 11.6, using p-hat = (3 + 2)/(17+4)=0.238

The z-statistic would be (0.238 - .4769)/sqrt(.4769*.5231/21) = -.2389/.1090 = -2.19, with a p-value of P(z < -2.19) = 0.0143, which is close to the binomial calculation of 0.011.

A 95% CI would be an upper bound of 0.238 + 1.645 * sqrt(.238 * .762/ 21) or 0.238 + 1.645*0.093 or 0.238 + 0.153 = 0.391, which is closer to the binomial calculation above.

This solution can get one more mark for a total of 3 out of 4, if the confidence interval estimate is given without a p-value.

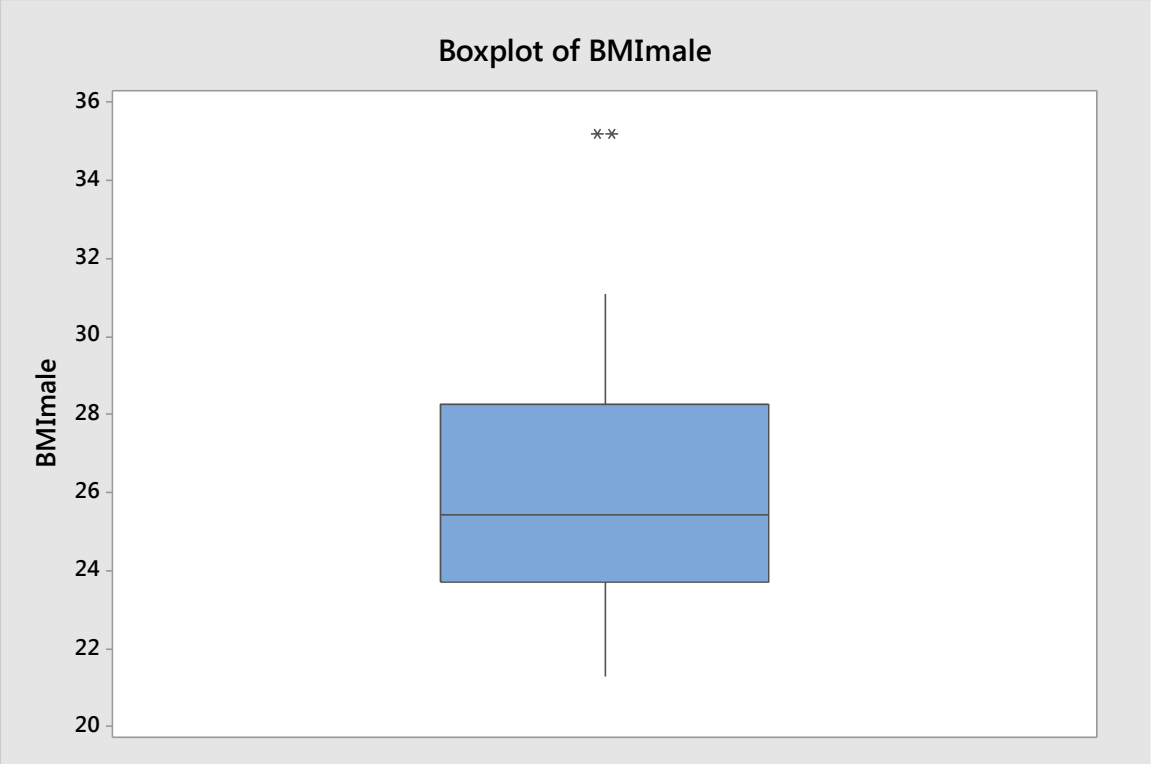
3 **[6 marks]**

The file **BMIsamples.mtw** contains two samples of BMI values from the male and female populations. Test at the 0.05 level of significance whether there is sufficient evidence here to show that the average ***male*** BMI (in the population) exceeds 25. Explain whether your test satisfies the underlying assumptions, with reference to graphical evidence, and show your manual calculations.

One-Sample T: BMImale

```
Test of μ = 25 vs > 25

Variable   N    Mean  StDev  SE Mean  95% Lower Bound     T      P
BMImale   46  26.030  3.232    0.476           25.230  2.16  0.018
```



The sample is skewed but not extremely so. We can assume the sample mean is normally distributed since the sample size is > 30. Do not accept the reason that the sample is not symmetric or that the population is not normally distributed since this is not required.

6 marks
-1 mark for graph
-1 mark for comment on underlying distribution, given large sample size
-1 mark for hypotheses, separately from Minitab
-1 mark for showing calculation of t-statistic
-1 mark for decision to reject null H
-1 mark for conclusion that average male BMI exceeds 25

4 **[10 marks]**

Two of the columns, **OWmale** and **OWfemale**, in the same dataset code the BMI values as:

0 - if BMI ≤ 25.4 (these are considered "not overweight");
1 - if BMI ≥ 25.5 (these are considered "overweight").

a. Test whether there is sufficient evidence to show that the proportion of overweight males (proportion of males who are overweight) is different than the proportion of overweight

females in the population. Use the critical value approach and the 0.05 level of significance. Perform the test manually after using Minitab to summarize the data (note that the mean coded value in each sample is the sample proportion).

b. Now find the p-value for your sample result and explain how you would find the p-value if you did not have statistical software to perform the test for you.
c. Finally calculate manually the 95% 2-sided confidence interval for the true difference between the proportions of overweight males and overweight females.
d. Explain how the results in parts b and c are consistent with your conclusion in part a.
e. Explain why the normal approximation is warranted in this question.

Test and CI for Two Proportions: OWfemale, OWmale

```
Event = 1


Variable   X   N  Sample p
OWfemale  10  35  0.285714
OWmale    23  46  0.500000


Difference = p (OWfemale) - p (OWmale)
Estimate for difference:  -0.214286
95% CI for difference:  (-0.422316, -0.00625526)
Test for difference = 0 (vs ≠ 0):  Z = -1.94  P-Value = 0.052

Fisher's exact test: P-Value = 0.069

Ho: p1-p2=0 vs Ha: p1-p2 ≠ 0
p-bar = (10+23)/(35+46)=33/81=0.4074

z-statistic = (0.214)/sqrt((33*48/(81*81)*(1/35+1/46))
                                = 0.214 / 0.1102 = ± 1.94

Do not reject the null hypothesis since 1.94 is not > 1.96
Conclude insufficient evidence to show a difference.


-1 mark for decision not to reject null H
-1 mark for conclusion that there is insufficient evidence to show a difference.
```

5 marks:
-1 mark for statement of hypotheses, separately from Minitab output
-1 mark for pooling proportions
-1 mark for z-statistic (must show manual calculation)
-1 mark for decision not to reject null H
-1 mark for conclusion that there is insufficient evidence to show a difference.

Some may do this test as a 2-sample test with similar calculated values:

```
Two-sample T for OWfemale vs OWmale

           N   Mean  StDev  SE Mean
OWfemale  35  0.286  0.458    0.077
OWmale    46  0.500  0.506    0.075


Difference = mu (OWfemale) - mu (OWmale)
Estimate for difference:  -0.214286
95% CI for difference:  (-0.431168, 0.002596)
T-Test of difference = 0 (vs not =): T-Value = -1.97  P-Value = 0.053  DF = 79
Both use Pooled StDev = 0.4858
```

Using the pooled stdev of 0.4858, the standard error for the t-statistic is 0.4858 sqrt(1/35 +1/46) = 0.1090 with t = 0.2143/0.109 = 1.97

```
Two-sample T for OWfemale vs OWmale

           N   Mean  StDev  SE Mean
OWfemale  35  0.286  0.458    0.077
OWmale    46  0.500  0.506    0.075


Difference = mu (OWfemale) - mu (OWmale)
Estimate for difference:  -0.214286
95% CI for difference:  (-0.428406, -0.000165)
T-Test of difference = 0 (vs not =): T-Value = -1.99  P-Value = 0.050  DF = 76
```

Here the standard error is sqrt(0.458^2/35 + 0.506^2/46) = 0.1075,
and t = 0.2143/0.1075 = ± 1.99

If the 2-sample t-test is done, then the hypotheses
Ho: μ1 - μ2 = 0 vs Ha: μ1 - μ2 ≠ 0 does not get a mark,
and the manual calculation of t does not get a mark.
However, if they get the decision and conclusion consistent with their t-statistic, then they can get 2 marks for these.

(For the 2-sample t-test, the exact t-critical values are closer to ±1.99 whereas the normal approximation has ± 1.96 and both are acceptable).

(b)

For the 2-proportions test, the p-value is
Prob(|Z| > 1.94) = 2 * P(Z < -1.94) = 2 * P(Z > 1.94) = 2*0.0262 = 0.052

For the 2-sample test, however,
this could be 2P(Z > 1.97) = 2*0.0244 = 0.0488 or 2 P(Z > 1.99) = 2*0.0233 = 0.0466,
Note these are both < 0.05, or
2 P(t > 1.97) > 0.05 or 2 P(t > 1.99) = 0.05, both not < 0.05.

1 mark for any calculation close to any of the above.

(c)

For the 2-proportions test, the CI is
0.50 – 0.2857 ± 1.96 * sqrt(0.5 *.5 / 46 + .2857 * .7143 / 35)
= 0.214 ± 1.96 * 0.1061 = 0.214 ± 0.208 = (0.006, 0.422) or (-0.422, -0.006,)

For the 2-sample t-test, the CI can be:

0.2143 ± 1.99 * 0.1090 = 0.2143 ± 0.2169 or 0.2143 ± 1.96 * 0.1090 = 0.214 ± 0.2136
or 0.2143 ± 1.99 * 0.1075 or 0.214 ± 0.2139 or 0.214 ± 1.96 * 0.1075 = 0.214 ± 0.2107

Only the first CI above covers zero.

2 marks for any CI like the above.

(d) Depending on the p-value calculated, and the CI calculated, the decisions may or may not be consistent with (a).

0.5 mark for any reasonable comparison of p-value to alpha;
0.5 mark for any reasonable check of whether the CI covers zero.

(e) The normal approximation is warranted since the sample sizes are large enough (because there are 10, 23, 25 and 23 female and male overweight individuals and female and male non-overweight individuals in the two samples, respectively (that is, all counts are at least equal to 10.

1 mark