Problem Set 6

Due date: 30 October

Please upload your completed assignment to the ELMs course site (under the assignments menu). Remember to include an annotated script file for all work with R and show your math for all other problems (if applicable, or necessary). Please also upload your completed assignment to the Github repository that you have shared with us. *We should be able to run your script with no errors.*

Total points: 40

Question 1

Points: 10

I hypothesize that the mean mathematics SAT score for University of Maryland students is different than 600. I take an SRS of 625 students and find that the mean mathematics score among those students is 610. The standard deviation of my sample is 75.

Part A

Based on my sample, what is a point estimate for the mean mathematics SAT score for University of Maryland students?

Ans 1A.

The point estimate is the best measure of the sample of the mean that is in this case 610 itself.

Part B

What is a 95% confidence interval around that estimate?

Ans1B:

We know that,

$$Confidence = \overline{x} + / - t(SD/\sqrt(n))$$

In this case,

$$\overline{x} = 610$$

$$SD = 75$$

$$n = 625$$

t for (625 - 1) degrees of freedom at a 95% confidence interval is approx 1.965

Hence confidence interval would be:

$$Confidence = 610 + / - 1.965(75/\sqrt(625))$$

$$Confidence = 610 + / - 5.895$$

Which means the confidence interval ranges from [604.105, 615.895]

Part C

What is the null hypothesis?

Ans 1C:

The null hypothesis, is a statement that there is no significant effect in a statistical hypothesis test. It serves as the default assumption to be tested. In your case, whether the mean mathematics SAT score for University of Maryland students is different from 600, the null hypothesis would be:

:The mean mathematics SAT score for University of Maryland students is equal to 600.

In other words, the null hypothesis assumes that there is no difference between the population mean SAT score and the hypothesized value of 600. Our goal is to evaluate whether there is enough evidence to reject this null hypothesis in favor of an alternative hypothesis that suggests a difference.

Part D

What is the two-tailed p-value and what does it mean?

Ans 1D:

We need to calculate the given Z score first in order to perform the two tailed test.

We know that ,

$$Z = \overline{x} - \mu / sigma / \sqrt{n}$$

Plugging in the values we get ,

$$Z = 610 - 600/75/\sqrt{625}$$

Which is about 10/3 or 3.33.

Hence for the two tailed test, we want to see the probability of getting a score that is greater than 3.33 or less than 3.33.

The p values corresponding to the given Z score is around 0.001 hence it is very small and the result is statistically significant.

What it means:

- A small p-value (typically less than your chosen significance level, in this case, 0.05) suggests strong evidence against the null hypothesis.

- In the context of this problem, a p-value of 0.001 is much smaller than 0.05, so you would reject the null hypothesis.
- This means that based on your sample data, you have strong evidence to conclude that the mean mathematics SAT score for University of Maryland students is different from 600.

In summary, a two-tailed p-value measures the strength of evidence against the null hypothesis. A very small p-value indicates strong evidence against the null hypothesis, while a larger p-value suggests weaker evidence against the null hypothesis.

Part E

Do the data support my hypothesis? Why, or why not?

Ans 1E:

The initial hypothesis was that the mean math SAT score was not equal to 600.

After conducting a hypothesis test, we have found that the two-tailed p-value was very small, approximately 0.001. This means that there is strong evidence against the null hypothesis that the mean SAT score is equal to 600. As a result, you can reject the null hypothesis.

In practical terms, this suggests that the sample mean SAT score of 610 that we have obtained from the 625 students is significantly different from the hypothesized population mean of 600. The data indicate that, based on our sample, the mean SAT score for University of Maryland students is not 600, but rather 610.

So, in conclusion, the data from our sample do support your hypothesis, and we have strong statistical evidence to suggest that the mean mathematics SAT score for University of Maryland students is different from 600.

Question 2

Points: 10

I take an SRS of 900 citizens in a public opinion survey and find that the mean feeling thermometer rating (on a 0 to 100 scale) of President Biden is 51.5 with a standard deviation of 25. Now, suppose I hypothesize that feelings toward Biden should be greater than 50.

Part A

What is the null hypothesis?

Ans 2A:

The null hypothesis : There is no significant difference or effect, or that feelings toward President Biden are not greater than 50.

SD = 25

Mean SRS = 51.5

In mathematical terms, you can represent this null hypothesis as:

H0:   50 H0: 50

Where:

- H0 is the null hypothesis.
- is the population mean feeling thermometer rating for President Biden.
- 50 is the hypothesized value (in this case, equal to or less than 50).

So, the null hypothesis is that there is no significant evidence to suggest that feelings toward President Biden are greater than 50.

Part B

What is a 95% confidence interval around the sample statistic? Does this interval indicate support for a two-tailed test of my hypothesis (`p < 0.05` as the threshold)? Why, or why not?

Ans 2B:

We know that the CI

$$Confidence = \overline{x} + / - t(SD/\sqrt(n))$$

The t value at (900 -1) and 95% confidence is around 1.645.

Plugging in the rest of the values,

$$Confidence = 51.5 + / - 1.645(25\sqrt(900))$$

Hence the final confidence will be:

$$Confidence = 51.5 + / - 1.37$$

Hence the confidence at 95% will be around [50.086, 52.914]

So, the 95% confidence interval for the mean feeling thermometer rating of President Biden is approximately (50.086, 52.914). This interval represents a range of values that we are 95% confident contains the true population mean.

For the p_value question,

The confidence interval does not directly indicate support for a two-tailed test with a significance level of p_value $<$ 0.05 because it represents a range of values around the sample mean (51.5) and is used for estimating the population mean. It is not the same as conducting a hypothesis test.

If we want to perform a two-tailed test with a significance level of p_value $<$ 0.05 , we would need to recalculate the test with a null hypothesis that the mean feeling thermometer rating is equal to 50, and then compare the test statistic to the appropriate critical value(s) for a two-tailed test at the desired significance level. The confidence interval alone does not determine the outcome of the hypothesis test.

Part C

What is a 90% confidence interval around the sample statistic? Does this interval indicate support for a one-tailed test of my hypothesis (`p < 0.05` as the threshold)? Why, or why not?

Ans 2C:

For a 90% interval, we can use the same formula:

$$Confidence = \overline{x} + / - t(SD/\sqrt(n))$$

The t value at (900-1) and a 90% confidence interval will be approx 1.645

Question 3

Points: 5

Suppose I conduct a difference of means test and obtain a t-statistic of 2.50. What does this indicate about the statistical significance of the sample mean compared to the null hypothesis?

Ans3:

A t-statistic of 2.50 indicates the difference between the sample mean and the null hypothesis mean in terms of the number of standard errors. The t-statistic is used to assess the statistical significance of the difference between the sample mean and the null hypothesis mean.

In the context of a t-test, if our t-statistic is 2.50, it means that the sample mean is 2.50 standard errors away from the null hypothesis mean. This difference is calculated using the formula for the t-statistic:

We can write the t score as,

$$t = \overline{x} - \mu / sigma / \sqrt{n}$$

The t-statistic quantifies how many standard errors the sample mean is away from the null hypothesis mean. In our case, a t-statistic of 2.50 suggests that the sample mean is 2.50 standard errors greater than what you would expect based on the null hypothesis.

To assess the statistical significance of this difference, we would compare the t-statistic to a critical value from the t-distribution for a given significance level and degrees of freedom. If the t-statistic is larger (in absolute value) than the critical value, it suggests that the observed difference is statistically significant, and you would likely reject the null hypothesis.

The specific critical value depends on the chosen significance level and degrees of freedom, and it's used to determine whether the observed difference is unlikely to occur by random chance. If the t-statistic is greater than the critical value, it indicates that the sample mean is significantly different from the null hypothesis mean. If it's smaller, it suggests that the difference is not statistically significant, and we would fail to reject the null hypothesis.

Question 4

Points: 5

I hypothesize that fewer than 40% of registered voters approve of President Biden. I conduct an SRS of 720 registered voters and find that 37.9% of them approve of President Biden.

Part A

Based on my sample, what is a point estimate for the proportion of the population that approves of President Biden?

Ans 4A:

N = 720.

Point estimate of a sample is 37.9 itself.

Part B

What is a 90% confidence interval around that estimate?

Ans 4B:

We can use the formula:

$$Confidence = p + / - z * \sqrt{p * (1 - p)/n}$$

The Z score for a 90% confidence interval is 1.645 approx

Plugging in the values , we can say that:

$$Confidence = 0.379 + / - 1.645 * \sqrt{0.379 * (1 - 0.379)/720}$$

Which means confidence is around Confidence Interval = 0.379±0.0429

Totally, CI = [0.3361, 0.4219]

Part C

What is the null hypothesis?

Ans 5C:

The Null hypothesis is that the proportion of registered voters who approve of President Biden is equal to 40% (p = 0.40).

Part D

What is the one-tailed p-value and what does it mean?

Ans 4D,

The test statistic (z) for a one-sample proportion test can be calculated as follows:

$$Confidence = p + / - z * \sqrt{p * (1 - p)/n}$$

Where:

- p is the sample proportion.
- pp is the hypothesized population proportion.

- nn is the sample size.

In our case,

- p (sample proportion) = 0.379
- pp (hypothesized population proportion) = 0.40
- nn (sample size) = 720

Now, plug these values into the formula:

z is approx: 0.371

For a one-tailed test, where you're testing whether the proportion is less than 40%, you need to find the area to the left of this test statistic on the standard normal distribution. Using a standard normal distribution table or calculator, you can find the one-tailed p-value.

The one-tailed p-value for a z-value of -1.03 is approximately 0.150. This p-value suggests that if the true proportion of registered voters who approve of President Biden is 40% (according to the null hypothesis), you would expect to obtain a sample proportion as extreme as 37.9% or less about 15% of the time by random chance alone. Since this p-value is greater than the typical significance level of 0.05 (5%), you would fail to reject the null hypothesis.

Now, you can find the one-tailed p-value using the standard normal distribution. Since you're testing whether the proportion is less than 40% ($p < 0.40$), you need to find the area to the left of this test statistic (z) value.

Using a standard normal distribution table or calculator, you can find that the one-tailed p-value for your test statistic (z) is approximately 0.1757.

This means that if the true proportion of registered voters who approve of President Biden is 40% (according to the null hypothesis), you would expect to obtain a sample proportion as extreme as 37.9% or less about 17.57% of the time by random chance alone. Since this p-value is greater than the typical significance level of 0.05 (5%), you would fail to reject the null hypothesis. In other words, your sample data does not provide enough evidence to conclude that the proportion of voters who approve of President Biden is less than 40%.

Part E

Do the data support my hypothesis? Why, or why not?

Ans5E:

Based on the information provided and the results of your hypothesis test, the data do not strongly support our hypothesis. Our hypothesis was that fewer than 40% of registered voters approve of President Biden, but the one-tailed p-value, which measures the evidence against the null hypothesis, is approximately 0.150. This p-value is greater than the typical significance level of 0.05 (5%).

In hypothesis testing, a p-value greater than the chosen significance level suggests that the observed sample data is not strong enough to reject the null hypothesis. In your case, this means that the data do not provide enough evidence to conclude that the proportion of registered voters who approve of President Biden is less than 40%.

It's important to note that while the data do not support your hypothesis, this does not necessarily mean that the null hypothesis is true (i.e., it doesn't confirm that exactly 40% of registered voters approve of President Biden). It simply means that, based on your sample data, there is not enough evidence to conclude that the proportion is less than 40%.

Hypothesis tests provide a measure of the strength of evidence against the null hypothesis, and in this case, the evidence is not strong enough to reject it.

Question 5

Points: 5

I hypothesize that in countries where Islam is the predominate religious group, women have a greater number of children on average (i.e., the total fertility rate is higher). Use the `world` dataset to test my hypothesis. State the null hypothesis and interpret what the results tell us about the null and alternative hypotheses. Be sure to show all work necessary to find the answer (i.e., you may use R to assist you, but you should show the necessary computations by hand).

> Note
>
> The `world` data set can be found in `poliscidata::world`.

Ans 5:

```
world_data <- poliscidata::world
```

```
Registered S3 method overwritten by 'gdata':
  method         from
  reorder.factor gplots
```

```
head(world_data)
```

```
      country gini10    dem_level4 dem_rank14 dem_score14 lifeex_f lifeex_m
1 Afghanistan   29.4 Authoritarian        151        2.77    45.25    44.79
2     Albania   33.0        Hybrid         88        5.67    80.30    74.82
3     Algeria   35.3 Authoritarian        117        3.83    76.31    72.78
4      Angola   58.6 Authoritarian        133        3.35    39.83    37.74
5   Argentina   48.8    Part Democ         52        6.84    80.36    73.71
6     Armenia   30.2        Hybrid        113        4.13    77.31    69.59
  literacy     oil pop_0_14 pop_15_64 pop_65_older fertility govregrel
1     28.1       0     42.3      55.3          2.4      5.39    10.000
2       NA    5400     21.4      68.1         10.5      1.48     0.000
3     69.9 2125000     24.2      70.6          5.2      1.75     8.611
4     67.4 1948000     43.2      54.1          2.7      5.97     0.556
5     97.2  796300     25.4      63.6         11.0      2.31     0.000
6     99.4       0     17.6      72.4         10.1      1.37     6.944
                 regionun           religoin spendeduc spendhealth spendmil
1                    Asia             Muslim        NA         1.8      1.9
2                  Europe             Muslim       2.9         2.9      2.0
3                  Africa             Muslim       4.3         3.6      3.0
4                  Africa           Catholic       2.6         2.0      3.0
5 Latin America/Caribbean           Catholic       4.9         5.1      0.8
6                    Asia Orthodox Christian       3.0         2.1      3.3
    hdi pop_age sexratio pop_total pop_urban gender_unequal gender_unequal_rank
1 0.349    16.9    106.0      29.1      22.6          0.797                 134
2 0.719    30.0    107.0       3.2      51.9          0.545                  61
3 0.677    26.2    104.6      35.4      66.5          0.594                  70
4 0.403    17.4     99.9      19.0      58.5             NA                  NA
5 0.775    30.4    103.6      40.7      92.4          0.534                  60
6 0.695    32.0    116.5       3.1      64.2          0.570                  66
  arda lifeex_total debt       colony confidence
1    1        45.02   NA           UK         NA
2    3        77.41 59.3 Soviet Union  49.335926
3    4        74.50 25.7       France  52.055735
4    7        38.76 20.3     Portugal         NA
5   11        76.95 50.3        Spain   7.299325
6   12        73.23   NA Soviet Union  27.132735
                                             decent08 dem_other dem_other5 democ
1                                  No local elections      10.5        10%    No
2     Legislature and executive are locally elected      63.0 Approx 60%   Yes
3 Legislature is elected but executive is appointed      40.8 Approx 40%    No
4                                  No local elections      40.8 Approx 40%    No
5                                                <NA>      87.5 Approx 90%   Yes
6     Legislature and executive are locally elected      63.0 Approx 60%   Yes
  democ11 democ_regime democ_regime08    district_size3 durable effectiveness
```

```
1      NA           No             No     single member       4      13.71158
2       9          Yes            Yes              <NA>       3      35.46099
3       3           No             No 6 or more members       5      32.62411
4       2           No             No              <NA>       3      19.14894
5       8          Yes            Yes 6 or more members      17      34.98818
6       5          Yes            Yes              <NA>       2      36.64303
   enpp3_democ enpp3_democ08 dnpp_3         eu fhrate04_rev fhrate08_rev
1        <NA>          <NA>     NA Not member          2.5            3
2 1-3 parties   1-3 parties      1 Not member          5.0            8
3        <NA>          <NA>      3 Not member          2.5            3
4        <NA>          <NA>      1 Not member          2.5            3
5 1-3 parties   1-3 parties      1 Not member          6.0           10
6 6-11 parties  6-11 parties      3 Not member          3.5            4
  frac_eth frac_eth2 frac_eth3 free_business free_corrupt free_finance
1   0.7693      High      High            NA           NA           NA
2   0.2204       Low       Low          68.0           34           70
3   0.3394       Low    Medium          71.2           32           30
4   0.7867      High      High          43.4           19           40
5   0.2550       Low       Low          62.1           29           30
6   0.1272       Low       Low          83.4           29           70
  free_fiscal free_govspend free_invest free_labor free_monetary free_property
1          NA            NA          NA         NA            NA            NA
2        92.6          74.2          70       52.1          78.7            35
3        83.5          73.4          45       56.4          77.2            30
4        85.1          62.8          35       45.2          62.6            20
5        69.5          75.6          45       50.1          61.2            20
6        89.3          90.9          75       70.6          72.9            30
  free_trade free_overall free_overall_4 gdp08 gdp_10_thou gdp_cap2 gdp_cap3
1         NA           NA           <NA>  30.6          NA     <NA>     <NA>
2       85.8         66.0          MidHi  24.3      0.1535      Low   Middle
3       70.7         56.9         MidLow 276.0      0.1785      Low   Middle
4       70.4         48.4            Low 106.3      0.0857      Low   Middle
5       69.5         51.2            Low 571.5      0.2797     High   Middle
6       80.5         69.2           High  18.7      0.0771      Low      Low
  gdpcap2_08 gdpcap3_08 gdpcap08_2 gdppcap08 gdppcap08_3 gender_equal3 gini04
1        Low        Low        Low        NA          NA          <NA>     NA
2        Low        Mid        Low      7715           2          <NA>   28.2
3       High        Mid       High      8033           2          <NA>   35.3
4       High        Mid       High      5899           2          <NA>     NA
5       High       High       High     14333           3          High   52.2
6        Low        Mid        Low      6070           2          <NA>   37.9
  gini08   hi_gdp indy muslim     natcode       oecd      pmat12_3 polity
1     NA     <NA> 1919    Yes afghanistan Not member          <NA>     NA
```

```
2   31.1  Low GDP 1991    Yes     albania Not member  Low post-mat      9
3   35.3  Low GDP 1962    Yes     algeria Not member          <NA>      2
4     NA  Low GDP 1975     No      angola Not member          <NA>     -2
5   51.3 High GDP 1816     No   argentina Not member High post-mat      8
6   33.8  Low GDP 1991     No     armenia Not member  Low post-mat      5
  pr_sys protact3        regime_type3 rich_democ unions unnetgro unnetuse
1     No     <NA>        Dictatorship         NA     NA       NA      1.7
2     No Moderate Parliamentary democ          0     NA    21329     23.9
3    Yes     <NA>        Dictatorship          0     NA     2633     11.9
4    Yes     <NA>        Dictatorship          0     NA     3567      3.1
5    Yes Moderate  Presidential democ          1   25.4      331     28.1
6     No     High                <NA>          0     NA      378      6.2
  unpovnpl unremitp unremitt vi_rel3 votevap00s votevap90s women05 women09
1     42.0       NA       NA    <NA>         NA         NA      NA      NA
2     18.5      476     12.2  20-50%      59.56   85.25755     6.4    16.4
3       NA       64      1.3    >50%         NA   71.43356      NA      NA
4       NA        5      0.1    <NA>         NA   88.28227      NA      NA
5       NA       17      0.2  20-50%      70.88   79.68567    33.7    41.6
6     50.9      345      8.9  20-50%         NA   53.32164     5.3     8.4
  women13 ipu_wom13_all womyear       womyear2 dem_economist democ.yes
1      NA          27.7      NA           <NA>             0         0
2    15.7          15.7    1920 1944 or before             0       100
3      NA          31.6    1962     After 1944             0         0
4      NA          34.1    1975     After 1944             0         0
5    37.4          37.4    1947     After 1944             1       100
6    10.7          10.7    1921 1944 or before             0       100
     country1
1 Afghanistan
2     Albania
3     Algeria
4      Angola
5   Argentina
6     Armenia
```

Hypothesis: Our hypothesis is that countries that have Islam as the predominant religion , the average fertility rate is higher.

Null hypothesis: The total fertility rate is the same in countries where Islam is the predominant religious group as in countries where Islam is not the predominant religious group.

Splitting the data and extracting what we need:

```r
library(tidyr)

#Cleaning data
cleaned_data <- cbind(as.numeric(world_data$fertility), as.character(world_data$muslim))

cleaned_data<- drop_na(as.data.frame(cleaned_data))
colnames(cleaned_data) <- c('Fertility_Rate', 'Is_Muslim')
cleaned_data$Fertility_Rate <- as.numeric(cleaned_data$Fertility_Rate)

#Splitting data
Countries_Islam <- cleaned_data[cleaned_data$Is_Muslim == 'Yes', ]
Countries_NonIslam <- cleaned_data[cleaned_data$Is_Muslim == 'No', ]

Mean_Islam = mean(Countries_Islam$Fertility_Rate)
Mean_NonIslam = mean(Countries_NonIslam$Fertility_Rate)

Mean_fertility_rate <- mean(cleaned_data$Fertility_Rate)

#Performing a t test
result <- t.test(Countries_Islam$Fertility_Rate, Countries_NonIslam$Fertility_Rate)
result
```

```
    Welch Two Sample t-test

data:  Countries_Islam$Fertility_Rate and Countries_NonIslam$Fertility_Rate
t = 3.4385, df = 71.372, p-value = 0.0009801
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3733271 1.4037285
sample estimates:
mean of x mean of y 
 3.408444  2.519917 
```

Note that I am using the Welchs t test as compared to Students , as Welch assumes that the variance between the two samples is different

As we can see from the results, the null hypothesis is rejected at a 95% confidence interval. Hence we reject the null hypothesis and assume that Our hypothesis that countries that have Islam as the predominant religion , the average fertility rate is higher is correct.