LAB 5: HYPOTHESIS TESTING, T-TEST

Two tailed t test in excel: https://www.statology.org/two-sample-t-test-excel/

Background Info

- This week, you'll continue working with the same dataset from last week's assignment, the political data from Twitter users. The methodology document / codebook is the same.

- Type your answers directly into this Word document, and submit this Word document along with your Excel workbook on Canvas.

Assignment

1. As a whole, can this survey data be considered a representative probability sample? Look at the survey methodology document, and write a 4-5 sentence paragraph that answers the following questions. Be specific in your answers. Include 1 or 2 quotes from the methodology document to back up your thoughts.

On the whole, the survey data is a representative probability sample. If your sample isn't representative, you won't be able to confidently generalize your t-test findings to the whole population. The target population of this survey is "non-institutionalized adults age 18 and older residing in the United States who use Twitter" (codebook 3). Ipsos gathered sample data by inviting adults who have a twitter account to take the survey at their earliest convenience (codebook 3). This sample can be considered representative of the target population because the samples are based on an equal probability selection method (EPSEM) (codebook 9). EPSEM ensures that each member of a population has an equal chance of being included in the sample.

a. Why does having a representative probability sample matter when running a t-test?
b. What is the target population of this survey?
c. How was this sample data gathered?
d. Can this sample be considered representative of the target population? Why or why not?

2. Your t-test will build on Q2 from Lab 3, about average feeling toward journalists by political party. Clean the data to prepare for the analysis.
    a. On a new sheet, paste *PARTY* and *THERMOe* variables and delete the rows where *THERMOe = Refused.* Then only keep the rows where *PARTY = Republican* or *Democrat* and delete the rest.
    b. Reshape the data so that Dems' Thermo scores are in one column, and Repubs' Thermo scores are in another. Label the columns "Dems – Feelings Toward Journalists" and "Repubs – Feelings Toward Journalists".
    c. Highlight the observations in each column, and look at the bottom right of the excel screen to determine the n-size for Dems and Repubs (where it says "Count"), and report the n-size below. Careful not to include column headers in your count!

N-size (Dems): 11106

N-size (Repubs): 685

3. Using the "Data Analysis" button in excel, run summary statistics on Dems and Repubs feelings toward journalists (the two columns in #1). Reformat your output to 2 decimals. Paste your table below.

Democrat vs. Republican Feelings Towards Journalists

Democrats – feeling towards journalists		*Republicans – Feeling towards journalists*	
Mean	72.28	Mean	36.72
Standard Error	0.68	Standard Error	1.02
Median	76.00	Median	35.00
Mode	50.00	Mode	50.00
Standard Deviation	22.58	Standard Deviation	26.73
Sample Variance	509.71	Sample Variance	714.65
Kurtosis	0.97	Kurtosis	-0.87
Skewness	-1.05	Skewness	0.32
Range	100.00	Range	100.00
Minimum	0.00	Minimum	0.00
Maximum	100.00	Maximum	100.00
Sum	79865.00	Sum	25114.00
Count	1105.00	Count	684.00

4. Write 2-3 sentences: compare the averages we see in this sample data. Make sure to mention the variable's scale so it's clear what the averages mean.

The summary statistics show a higher mean for Democrats' feelings towards journalists than Republicans on a scale of 0-100, 0 being cold and negative feelings towards journalists and 100 being warm and positive feelings towards journalists. On average, Democrats report feelings towards journalists at 72.30/100 , which is more than twice the mean of Republicans at 36.66/100.

5. Now we want to test whether what we've seen in the sample data is true in the population. Write a Null Hypothesis (H0) and Research Hypothesis (H1) that could be tested with a **two-tailed** t-test, comparing Democrats' and Republicans' feelings toward journalists.

   H0: Democrats' and Republicans' feeling towards journalists are the same.

   H1: Democrats' and Republicans' feeling towards journalists are not the same.

6. Run a two-sample t-test assuming unequal variances on the data. Highlight the p-value in the table that is relevant to your hypotheses above. **Bold** the cells with the t-statistic and relevant t critical value. Reformat your output to 2 decimals. Paste your table below.

Democrat vs. Republican Feelings Towards Journalists

	Avg Feeling for Dems	*Avg Feeling for Repubs*
Mean	72.30	36.66
Variance	509.94	715.57
Observations	1106.00	685.00
Hypothesized Mean Difference	0.00	
df	1268.00	
t Stat	**29.04**	
P(T<=t) one-tail	0.00	
t Critical one-tail	1.65	
P(T<=t) two-tail	0.00	
t Critical two-tail	**1.96**	

7. What does the table output tell you about your hypotheses in #5? Were your results statistically significant? Assume you are writing for a technical audience of statisticians. Write 3-4 sentences summarizing the results and include the following concepts somewhere in your response:
   - *average feeling for Dems, average feeling for Repubs*
   - *N of each group (sample size)*
     The N
   - *confidence level*
   - *t-stat*
   - *degrees of freedom*

- *p-value*
  - *reject the null*
- *statistical significance*
- *what you can conclude about your hypotheses*

The table output displays the results of a two-sample t-test assuming unequal variances and a 95% confidence level (alpha = 0.05). Out of sample sizes N = 1106 for Democrats and N = 685 for Republicans, the mean average feeling towards journalists on a scale of 0-100 is higher for Democrats (72.30/100) than Republicans (33.66/100). The p-value of 0.00 (<alpha = 0.05), t value of 29.04 (>critical t = 1.96) and degrees of freedom (1268) indicate that the difference is significant at the 99.9% level of confidence, meaning we can reject the null hypothesis that average feelings towards journalists is the same for Democrats and Republicans, and accept the research hypothesis that average feelings towards journalists are not the same across Democrats and Republicans.

8. Now assume you are writing for a general, but professional audience. Think of a group that needs to act on the information but is not trained in statistics. They trust you as the statistical analyst to run the correct analyses and give the correct interpretation.
   a. In 2-3 sentences, offer a modified reporting of the results compared to what you said in #7 above.
   b. Include a basic bar chart comparing means that has an appropriate axis, title, and a parenthetical footnote of the statistical detail under the chart (t=, df=, p=; this is for professional readers who may want to see this technical information).

Republicans and Democrats were both asked their feelings towards journalists on a scale from 0-100, 0 being cold and negative feeling towards journalists and 100 being warm and positive feeling towards journalists. The average feeling for Democrats was 72.30 and the average feeling for Republicans was 36.66, demonstrating that, on average, Democrats feel significantly warmer and more positively towards journalists than Republicans (t=29.04, df = 1268, p = 0.000).

9. Now pretend we had instead wanted to run a **one-tailed** t-test on the same data. Write the H0 and H1 for a one-tailed test below.

   H0: Democrats' and Republicans' feeling towards journalists are the same.

   H1: Democrats' towards journalists are more warm and positive than Republicans' feelings towards journalists.

10. What's one example of a social science problem/question you'd be interested in running a two-sample t-test on, assuming you have access to any/all data necessary?
    a. Describe your variables, how they are measured, and what your two hypotheses would be. It can be a one-tailed or two-tailed test.

Variables: I would investigate with a two-tailed test whether or not male and female highschoolers who play school sports (independent variable) rank school athletics differently on their list of after school priorities (dependent).

Measurement: I would ask study participants to rank “school athletics” on a scale of 1-10, 1 being their least important after school priority and 10 being their most important after school priority.

H0: Male and female highschoolers who play school sports rank athletics the same on their list of their after school priorities.

H1: Male and female highschoolers who play school sports rank athletics differently on their list of after school priorities.

Submit this Word document along with your Excel workbook on Canvas.