**LAB 5: HYPOTHESIS TESTING, T-TEST**

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

**Having a representative probability sample matters when running a t-test because the goal is to utilize the sample data to reflect the distribution and draw conclusions of the entire population. In this survey the sampling was framed to mimic the target population of the adult population in the U.S. According to the methodology document, probability-based sampling methods were used stating “Ipsos migrated to an ABS recruitment methodology via the U.S. Postal Service’s Delivery Sequence File (DSF)” which allows for the survey to improve population coverage by increasing the inclusion of hard-to-reach individuals like young adults and minorities as well as households without internet connection who were “provided with a web-enabled device and free internet service.” This sample can be considered representative of the target population because of the methods used in recruitment which ensured the “geodemographic composition of our panel members mimic those of the adult population in the U.S.” through random sampling and the latest version of DSF which proved successful with its relatively high response rate at 60%.**

1. 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.
   1. 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.
   2. 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”.
   3. 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): 1106**

**N-size (Repubs): 685**

1. 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.

| **t-Test: Two-Sample Assuming Equal Variances** |  |  |
| --- | --- | --- |
|  |  |  |
|  | ***Dems – Feelings Toward Journalists*** | ***Repubs – Feelings Toward Journalists*** |
| **Mean** | **72.30** | **36.66** |
| **Variance** | **509.94** | **715.57** |
| **Observations** | **1106.00** | **685.00** |
| **Pooled Variance** | **588.56** |  |
| **Hypothesized Mean Difference** | **0.00** |  |
| **df** | **1789.00** |  |
| **t Stat** | **30.21** |  |
| **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** |  |

1. 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.

**Twitter users were asked about their average feelings towards journalists on a scale from 0 (as cold and negative as possible) to 100 (as warm and positive as possible) based on their political party. The average feelings toward journalists from Democrats was 72.30, showing that, on average, Democrats have significantly more positive feelings towards journalists than Republicans who reported almost half that average at 36.66.**

1. 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: Democrat Twitter Users have mean feelings towards journalists that are not different from Republican Twitter Users**

**H1: Democrat and Republican Twitter users have different feelings towards journalists.**

1. 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.

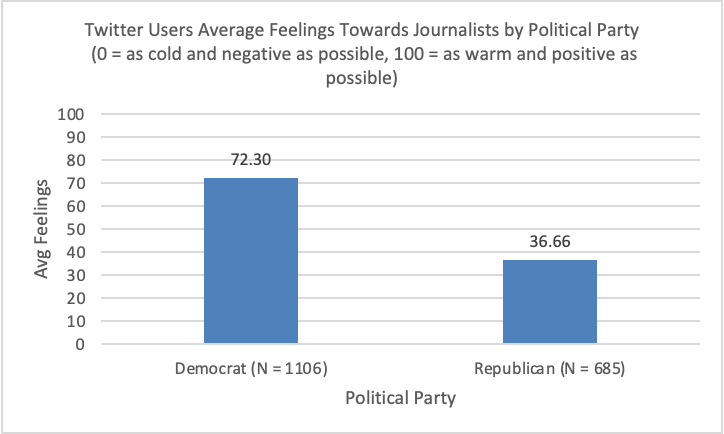
| t-Test: Two-Sample Assuming Unequal Variances | |  |
| --- | --- | --- |
|  |  |  |
|  | *Dems – Feelings Toward Journalists* | *Repubs – Feelings Toward Journalists* |
| Mean | 72.30 | 36.66 |
| Variance | 509.94 | 715.57 |
| Observations | 1106.00 | 685.00 |
| Hypothesized Mean Difference | 0.00 |  |
| 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** |  |

1. 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)*
   * *confidence level*
   * *t-stat*
   * *degrees of freedom*
   * *p-value*
   * *statistical significance*
   * *what you can conclude about your hypotheses*

**Using a two-sample t-test assuming unequal variances and required confidence level of 95% (alpha = 0.05), I find that the mean feelings about journalists for Democrats (mean = 72.30, sd = 22.58, N = 1106), on a scale from 0 (as cold and negative as possible) to 100 (as warm and positive as possible), was significantly higher than for Republicans (mean = 36.66, sd = 26.75, N = 685). With a p-value of 0.000 (<alpha = 0.05) and a t-value of 29.04 (> critical t = 1.96), and degrees of freedom of 1268, the difference is significant at the 99.9% level of confidence. We can reject the null hypothesis that the mean feelings are equal among Twitter Users of both Republican and Democrat political parties, and accept the research hypothesis that Democrats' average feelings about journalists differ from Republicans' average feelings towards journalists.**

1. 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.
   1. In 2-3 sentences, offer a modified reporting of the results compared to what you said in #7 above.
   2. 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).

**Democrats and Republicans were both asked about their feelings towards journalists on a scale from 0 (as cold and negative as possible) to 100 (as warm and positive as possible). The average feelings toward journalists for Democrats was 72.30, showing that, on average, Democrats feel significantly more positive about journalists than Republicans, who reported a lower average of 36.66 (t = 29.04, df = 1268, p = 0.000).**

****

***(t = 29.04, df = 1268, p = 0.000)***

1. 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: Democrat Twitter Users have mean feelings towards journalists that are not different from Republican Twitter Users**

**H1: Democrat Twitter Users have higher mean feelings towards journalists than Republican Twitter Users**

1. 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?
   1. Describe your variables, how they are measured, and what your two hypotheses would be. It can be a one-tailed or two-tailed test.

**A social science problem I’d be interested in running a two-sample t-test on would be examining whether there is a significant difference in the average income between college graduates from a two parent household and those in a single parent household in a specific region. The variables would be household type, a categorical variable representing if a college graduate comes from a two-parent household or single-parent household, and income, a continuous variable measuring annual income in dollars. The null hypothesis for a two-tailed test would be there is no significant difference in the average income between college graduates from two-parent households and those from single-parent households in a specific region while the research hypothesis would be that there is a significant difference in the average income between college graduates from two-parent households and those from single-parent households in a specific region.**

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