Faiz Surani, Rahul Araza, Victor Bai
Period 5
May 3, 2019

Spring Project Inference Test

We set out to test whether there is a statistically significant difference in mean sentiment of commenters on liberal and conservative Reddit communities. Below are the four steps of the two-sided t-test for mean difference that we conducted.

1 State

- Population: Comments on liberal and conservative partisan communities
- Parameter: Sentiment scores for comments on liberal and conservative communities, μ_L and μ_C, respectively.
- Hypotheses:

$$H_0 : \mu_L - \mu_C = 0 \qquad H_a : \mu_L - \mu_C \neq 0$$

  where μ_L is the mean sentiment score of liberal community users and u_C is the mean sentiment score of conservative community users.

  In words, our null hypothesis is that there is no difference in mean sentiment between users on liberal and conservative political communities, while our alternative hypothesis states that there is a difference in the true mean sentiments.

- Significance Level: $\alpha = 0.10$ (90% confidence level)

2 Plan

- Test: Two-sample t-test for a difference of means
- Conditions for Inference:
  - Randomness: Both data sets were collected using a Simple Random Sample of $n = 500$ after a collection period of three days in which every single comment on the selected partisan communities was scraped.[1]
  - Independence: To satisfy the independence condition when sampling without replacement, we must meet the 10% condition: $N \geq 10n$ for both samples. N is defined in this case as the number of comments posted over the course of our three day collection period which we treat as representative of the whole population of comments on partisan Reddit communities due to the impracticality of sampling

[1]The random samples were taken from the full data sets using the Python pandas library. The sample data used as well as the code written to collect, randomly sample, and get sentiments for the comments can be found here: `https://github.com/ProbablyFaiz/RedditPoliticalSentiment`

from all comments ever posted in the communities' existences. This condition is met as follows:

$$N_L = 127876,\ n_L = 500 \qquad N_C = 56601,\ n_C = 500$$
$$N_L \geq 10n_L \qquad N_C \geq 10n_C$$
$$127876 \geq 10(500) \qquad 56601 \geq 10(500)$$
$$127876 \geq 5000\ \checkmark \qquad 56601 \geq 5000\ \checkmark$$

– Normality/Large Sample: This condition can be met by either normally distributed values or $n \geq 30$ for both samples. This condition is met as follows:

$$n_L = 500 \qquad n_C = 500$$
$$n_L \geq 30 \qquad n_C \geq 30$$
$$500 \geq 30\ \checkmark \qquad 500 \geq 30\ \checkmark$$

3 Do

Below are the observed sample means and standard deviations:[2]

$$\bar{x}_L = -0.23481,\ \bar{x}_C = -0.19917,\ s_L = 0.452468,\ s_C = 0.512785,\ n_L = 500,\ n_C = 500$$

3.1 Significance Test

$$t = \frac{\bar{x}_L - \bar{x}_C}{\sqrt{\frac{s_L^2}{n_L} + \frac{s_R^2}{n_R}}}$$

Use test statistic formula for two-sample t-test for a difference of means. Denominator is the SE for two samples with unequal variances.

$$t = \frac{-0.23481 - (-0.19917)}{\sqrt{\frac{0.452468^2}{500} + \frac{0.512785^2}{500}}}$$

Substitute in observed mean and sample standard deviation values.

$$t = -\frac{0.03564}{0.03058} = -1.165$$

Simplify. Note that $SE = 0.03058$.

$$\nu = \frac{\left(\frac{s_L^2}{n_L} + \frac{s_C^2}{n_C}\right)^2}{\frac{s_L^4}{n_L^2\,(n_L - 1)} + \frac{s_C^4}{n_C^2\,(n_C - 1)}}$$

Use Welch-Satterthwaite equation to approximate degrees of freedom for our two populations with (assumed) unequal variances.

[2] All standard deviations given are sample standard deviations (i.e. with Bessel's correction).

$$\nu = \frac{\left(\frac{0.452468^2}{500} + \frac{0.512785^2}{500}\right)^2}{\frac{0.452468^4}{500^2\,(500-1)} + \frac{0.512785^4}{500^2\,(500-1)}}$$ Substitute standard deviations and sample sizes.

$$\nu = \frac{8.5600 \times 10^{-7}}{8.7101 \times 10^{-10}} = 982.523$$ Simplify.

$P\left(t_{982.523} > |-1.165|\right) = 0.244$ Calculate two-sided p-value for the test statistic $t = -1.165$ in a t-distribution with 982.523 degrees of freedom.

3.2 Confidence Interval

We will calculate a 90% confidence interval for the value $\mu_L - \mu_C$. Below are the variables we bring in:

$SE = 0.03058,\ \nu = 982.523$ (from Section 3.1), $\alpha = 0.10,\ t^*\left(\alpha = 0.10, \nu = 982.523\right) = 1.646$

$CI : (\bar{x}_L - \bar{x}_C) \pm t^* SE$	Formula for our confidence interval with a t critical value.
$CI : (-0.23481 - (-0.19917)) \pm (1.646)(0.03058)$	Substitute.
$CI : (-0.0860,\ 0.0147)$	Simplify.

4 Conclude

4.1 Significance Test

At $\alpha = 0.10$, we fail to reject $H_0 : \mu_L - \mu_C = 0$ (P-value = 0.244) and find insufficient evidence for a difference in mean sentiment of comments on liberal and conservative Reddit communities.

4.2 Confidence Interval

We are 90% confident that the true difference in mean sentiment between comments on liberal and conservative Reddit communities ($\mu_L - \mu_C$) is captured by the interval $(-0.0860,\ 0.0147)$.