MATH 254: Introduction to Statistics **Mintaek Lee**

Lesson 5: Inference for a Difference in Two Proportions

Corresponding Workbook Module: 5

Example 5.1 *In September, Mintaek randomly sampled 100 BSU students and asked if they love statistics or not, where 70 students responded they do. Mintaek is curious to see if the students' love for statistics has gone cold over time. In November, Mintaek randomly sampled 90 BSU students and asked them the same question. Much to Mintaek's dismay, only 45 of them responded they do this time. Use the appropriate statistical methods and set up the hypotheses that best match Mintaek's interests.*

We use the two-sample z-test for proportions for this problem. It is because we have collected two different sets of categorical data. First one is responses from a set of 100 students whether they love statistics or not. Second one is responses from a different set of 90 students whether they love statistics or not

Before we do any testing, we first find the following statistics:

- $\hat{p}_1 = \dfrac{70}{100} = 0.7$: sample proportion of BSU students who love statistics in September.
- $\hat{p}_2 = \dfrac{45}{90} = 0.5$: sample proportion of BSU students who love statistics in November.
- $\hat{p} = \dfrac{70 + 45}{100 + 90} = 0.605263$: pooled sample proportion of BSU students who love statistics (both in September and November). We will only use this to calculate the standard error, and nowhere else.

Step 1: Mintaek is interested in finding out whether the students' love for statistics has gone cold or not. So, he is interested in seeing if population proportion of BSU students who love statistics have decreased over time.

- H_0: $p_1 = p_2$, there is **no difference** between the population proportion of BSU students who love statistics in September and the population proportion of BSU students who love statistics in November.
- H_A: $p_1 > p_2$, the population proportion of BSU students who love statistics in September is **greater than** the population proportion of BSU students who love statistics in November.

Step 2: We are using the two-sample z-test for proportions here. We check all the underlying conditions (See workbook page 134 for description of conditions and workbook page 98 for the conditions for binomial settings). For the significance level, I will use $\alpha = 0.05$.

For **both** groups (September group and November group), we can check:

- **B**inary: we only have two possible outcomes here: love statistics or not.
- **I**ndependence: we are given that students in both groups were randomly selected.
- **N**umber: $n_1 = 100$ (number of students sampled in September) and $n_2 = 90$ (number of students sampled in November) were both set in advance
- **S**uccess: since there are certainly more than 1900 (10 times the combined sample size, $n_1 + n_2 = 190$) students in BSU, population size is at least 10 times the sample size. It would therefore be reasonable to assume that the probability of a student responding that he/she loves statistics (success probability) is stable (i.e. same for every students asked).

By checking BINS for both groups, we verified that our setting indeed is the binomial setting.

Now, we need to check if we can use the z-test by checking whether we have a large enough sample or not. This condition can be satisfied if the number of successes and number of failures in both groups is at least 5.

For the September group, 70 responded they love statistics (number of success) and 30 responded that they do not (number of failure). For the November group, 45 responded they love statistics (number of success) and 45 responded that they do not (number of failure). All of them are at least 5.

Since all conditions are met, we can use the two-sample z-test for proportions to make an inference on differences between p_1 and p_2 (population difference between proportions of BSU students who love statistics in September and November).

Step 3: We can now find the z-test statistic value and corresponding P-values. Using $\hat{p}_1 = 0.7$, $\hat{p}_2 = 0.5$, $\hat{p} = 0.605263$ from earlier, we find

$$\begin{aligned} z &= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \\ &= \frac{0.7 - 0.5}{\sqrt{0.605263(1-0.605263)\left(\frac{1}{100} + \frac{1}{90}\right)}} \\ &= \frac{0.2}{0.071020} \\ &= 2.82 \end{aligned}$$

From the standard normal distribution (z) table, we can see that our z-test statistic $z = 2.82$ corresponds to the area of 0.9976, which is the area left from $z = 2.82$. But, we are interested in the *upper* tail area, the area right from $z = 2.82$. We can find this area by subtracting 0.9976 from 1. We find the *upper* tail area corresponding to our z-test statistic $z = 2.82$ as $1 - 0.9976 = 0.0024$. Since this is a one-sided test and the sign in the alternative hypothesis is '>', we are only interested in the upper tail area. Therefore, our P-value is 0.0024.

Step 4: We reject the null hypothesis since P-value = $0.0024 < \alpha = 0.05$. There is a significant evidence suggesting that the population proportion of BSU students who love statistics in September is **greater than** the population proportion of BSU students who love statistics in November. In other words, we can say that BSU students' love for statistics has gone cold, or proportions of BSU students who love statistics have decreased over time.

Interpretation of the P-value: The probability of observing a difference in proportions of 0.2 ($\hat{p_1} - \hat{p_2}$) or more if the population proportions are the same (hence $p_1 = p_2$) is 0.0024. That is, there is 0.24% chance of obtaining a sample proportion difference of larger than 0.2 if the true mean difference was 0 (= H_0 was true). Since the chance is less than $\alpha = 0.05$, this would be a surprising enough result to reject H_0.

IF THIS WAS A TWO-SIDED TEST WHERE $\mathbf{H_A : p_1 \neq p_2}$

Step 1: H_0 is the same, just H_A needs to be changed.

- H_0: $p_1 = p_2$, there is **no difference** between the population proportion of BSU students who love statistics in September and the population proportion of BSU students who love statistics in November.
- H_A: $p_1 \neq p_2$, the population proportion of BSU students who love statistics in September is **different from** the population proportion of BSU students who love statistics in November.

Step 2: Same

Step 3: For a two-sided test, we are interested in both lower and upper tails. We found 0.0024 as our upper tail area. We need to find the lower tail area. Here, the lower tail area is the area left from $z = -2.82$ (notice that it is negative here). Since the standard normal distribution is symmetric, this lower tail area would be the same as the upper tail area. Therefore, we can obtain area of both tails by doubling one tail area. Since the lower tail area was 0.0024, our P-value would be $2 \times 0.0024 = 0.0048$.

Step 4: We reject the null hypothesis under the significance level $\alpha = 0.05$. It is because our P-value was smaller than our pre-determined significance level. We do have a significant evidence to suggest that the population proportion of BSU students who love statistics in September is **different from** the population proportion of BSU students who love statistics

in November. We won't necessarily know whether BSU students' love for statistics has gone cold or hotter, since we only tested to see if two proportions were not the same.

Interpretation of the P-value: The probability of observing a difference in proportions this far (or farther) away from 0 if the population proportions are the same is 0.0048. That is, there is 0.48% chance of obtaining a sample proportion difference of larger than 0.2 OR smaller than -0.2 if the true mean difference was 0 (= H_0 was true). Since the chance is less than $\alpha = 0.05$, this would be a surprising enough result to reject H_0.

Example 5.2 *Mintaek now wants to estimate the difference in population proportions of BSU students who love statistics in September and population proportions of BSU students who love statistics in November with the 95% certainty.*

Technically, you still would need to check all conditions for the two-sample z-confidence intervals for proportions. However, you may notice that they are identical to ones for the two-sample z-test for proportions. It will be omitted here for the sake of space.

We first find the standard error of the difference in proportions

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.7(1-0.7)}{100} + \frac{0.5(1-0.5)}{90}} = 0.069841$$

Note that this is **different** from the pooled standard error used to calculate the z-test statistic in the Question 1.

We want to construct 95% confidence interval. At the **t-distribution table**, we look at the row corresponding to z^*. We then find a column corresponding to the confidence level 95% from the last row. You should find $z^* = 1.960$ as the appropriate critical value.

Margin of error is defined as: $z^* \times SE = 1.960 \times 0.069841 = 0.137$. Then we can finally calculate the 95% confidence interval for the difference in population proportions of BSU students who love statistics in September and in November.

$$\begin{aligned}
&(\hat{p_1} - \hat{p_2}) \pm z^* \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}} \\
&= (0.7 - 0.5) \pm 1.960 \times 0.069841 \\
&= 0.2 \pm 0.137 \\
&= (0.2 - 0.137, 0.2 + 0.137) \\
&= (0.06, 0.34)
\end{aligned}$$

We are 95% certain/sure/confident that the interval (0.06, 0.34) contains the difference in population proportions of BSU students who love statistics in September and population proportions of BSU students who love statistics in November, $(p_1 - p_2)$.