1 Inference for Categorical Data: Proportions

1.1 Constructing a One Proportion z-Interval

Definition: A confidence interval for a population parameter is an interval of plausible values for that unknown parameter.

It is constructed in such a way so that, with a chosen degree of confidence, the value of the parameter will be captured inside the interval.

The chosen degree of confidence is called the confidence level. The confidence level gives information about how much "confidence" we will have in the method used to construct the interval.

To create an interval of plausible values for a parameter, we need two components:

- A point estimate is a single value used to estimate the population parameter such as a sample population.
- A margin of error represents the maximum expected difference between the true population parameter and the sample estimate.

To calculate the interval you use the formula

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p}$ is the point estimate, z^* is the critical value and $z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the margin of error.

To calculate the critical value, you can use invNorm. Generally for a 90% confidence interval, $z^* = 1.64$, for a 95% confidence interval, $z^* = 1.96$ and for a 99% confidence interval, $z^* = 2.58$.

1.2 Constructing a One Proportion z-Test

A significance test is another inference method that assesses evidence provided by data about a claim. Significance tests tell us if sample data gives us convincing evidence against a null hypothesis.

- A null hypothesis (H_0) is the claim being assess in a significance test. Usually, the null hypothesis is a statement of "no change from the expected value."
- An alternative hypothesis (H_A) proposes what we should conclude if we find the null hypothesis to be unlikely.

Hypotheses always refer to the population not the sample.

A p-value is the probability of getting results as extreme or more extreme in the direction of the null hypothesis by random chance alone assuming the claim of the null hypothesis is true.

- Small p-values give convincing evidence against the null hypothesis since the result we got is unlikely to occur.
- Large p-values fail to give convincing evidence against the null hypothesis since the result we got is likely to occur.

The significance level (α) is a fixed value that we will regard as the decisive value that determines if the p-value is small or large.

- Typically we choose $\alpha = 0.05$ which says we need data so strong that it would happen by chance less than 5% of the time.

To construct a test follow the steps:

- Define the parameter: $p =$ true proportion of {parameter in context}
- State the hypotheses. (If you are not given a claimed proportion, we use a conservative estimate which is 0.50)
- Check the Assumptions and Conditions
- Name the Inference method
- Calculate the test statistic
- Obtain the p-value
- Make a decision
- Write your conclusion in context

1.3 Relating Confidence Intervals and Significance Tests

Margin of error is point estimate $\pm$ margin of error.

Increasing confidence level increases critical value and margin of error and gives a wider interval.

Decreasing confidence level decreases critical value and margin of error and gives a narrower interval.

Increasing sample size decreases standard error and decreases margin of error and gives a narrower interval.

Decreasing sample size increased standard error, margin of error and gives a wider interval.

Keep in mind that the margin of error in a confidence covers only accounts for sampling variability and does not account for bias in the sampling methods.

1.4 Inference for Comparing Two Population Proportions

To construct a two proportion z-interval,

- Define the parameter. p_1 and p_2 are the true proportion of parameters in context for a population
- Check the assumptions and conditions (randomness, independence, and normality)
- Name the inference method
- Calculate the interval: $(\hat{p}_1 - \hat{p}_2 \pm z * \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_1(1-\hat{p}_2)}{n_2}})$
- Write the conclusion in context

To construct a two proportion z-test,

- Define the parameter
- State the hypothesis - The null is $H_0 : p_1 = p_2$ and the altternate is $p_1 < p_2, p_1 > p_2$ or $p_1 \neq p_2$
- Check the assumptions and conditions
- Name the Inference Method
- Calculate the Test Statistic (z-score). The null hypothesis states that there is no difference between the two population proportions. If this is true, the observations really come from a single population.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

  where $\hat{p}_c = \frac{x_1+x_2}{n_1+n_2}$, where x is the number of successes and n is sample size.
- Obtain the p-value (using normalcdf)
- Make a decision, if the p-value is less than α you reject the null otherwise you fail to reject

- State the conclusion in context

1.5 Errors & Power

A Type I error occurs when

- Reject H_0 incorrectly
- The probability of a Type I error is equal to the significance level
- P(Type I error) $= \alpha$
- Our significance level tells us what p-value is "low enough"

A type II error occurs when

- Fail to reject H_0 incorrectly
- P(Type II Error) $= \beta$

Power

- Reject H_0 correctly
- The probability we reject the null correctly is 1 minus the probability we reject the null incorrectly
- Power $= 1 - \beta$

Errors and their relationships

- Type I and Type II errors have an indirect relationship: as the probability of one increases, the probability of the other decreases
- Our Type I error is set with our significance level
- The higher our significance level is, the lower our probability of failing to reject becomes, which is why α and β have an indirect relationship
- The higher our significance level, the more likely we will reject the null, which increases the likelihood we do that incorrectly (as well as correctly)
- Therefore, Type I error and Power have a direct relationship: as the probability of Type I Error increases, the higher the Power of the test