Cheat Sheet: Confidence Intervals for Population Proportions

Essential Concepts

- A point estimate is a single value based on representative sample data that is a plausible estimate of the population parameter. The sample proportion is used as a point estimate of the population proportion.
- When taking many, many random samples of size n from a population distribution with proportion p:
    - The mean of the distribution of sample proportions is p.
    - The standard deviation of the distribution of sample proportions is $\sqrt{\frac{p(1-p)}{n}}$.
    - If $np \geq 10$ and $n(1-p) \geq 10$, then the Central Limit Theorem (CLT) states that the distribution of the sample proportions follows an approximate normal distribution with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$.
    - When the sample size is large enough, we can use $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ in place of $\sqrt{\frac{p(1-p)}{n}}$. This is called the standard error, which is the estimated standard deviation of sample proportions.
- A confidence interval for a population proportion is a reasonable range of values where we expect the population proportion to fall within, with a chosen degree of confidence.
    - A confidence interval is calculated using the point estimate and the margin of error.

- The margin of error (*ME*) is what determines the width of the interval. A confidence interval will have a width of twice the margin of error.
- $ME = z^* \cdot (\text{standard error})$, where: standard error = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, and z^* is the point on the standard normal distribution such that the proportion of area under the curve between $-z^*$ and $+z^*$ is C, the confidence level.
- The confidence level, C, tells us how much confidence we have in the method used to construct the interval. It corresponds to the percentage of all intervals we would expect to contain the true population parameter.
- For proportions, each confidence level has a corresponding z critical value (z*). This is the point on the standard normal distribution such that the proportion of area under the curve between $-z^*$ and +z* is C, the confidence level.

Confidence Level	z *
90%	1.645
95%	1.96
99%	2.576

- Confidence level is a measure of our confidence in the method. The interpretation of a confidence interval depends on the confidence level. For example, using a 95% confidence level, the interpretation would be: We can be 95% confident that the interval between ___ and ___ captures the true population proportion of ______.
    - Formula to determine the minimum sample size needed to produce a given margin of error: $n = \hat{p}(1-\hat{p})(\frac{z^*}{ME})^2$
    - Using the conservative $\hat{p} = 0.5$ approach always yields a larger than necessary sample size.
- There are two different methods for calculating confidence intervals for the difference in proportions.
    - If the two groups are independent, the sample for one group is drawn independently of the other group. Knowing the observations of one group does not provide useful information about the other sample. Additionally, the groups can be different sizes.
    - If the two groups are dependent (also known as paired or matched pairs), the samples for the two groups are not drawn independently of one another. Knowing the observations of one group does provide useful information about the other sample. Additionally, both groups must be the same size.

- When our goal is to estimate a difference between two population proportions (or the size of a treatment effect), we select two independent random samples and use the difference in sample proportions as an estimate.
- The three conditions when calculating confidence intervals for the difference between two proportions are:
    - Random samples: The observations represent a random sample of the population.
    - Independence: The samples are independently selected.
    - Sample size: $n_1\hat{p}_1 \geq 10$ and $n_2\hat{p}_2 \geq 10$
- When certain conditions apply (more on those later), the sampling distribution tells us three things about the distribution of $\hat{p}_1 - \hat{p}_2$:
    - For large samples, the distribution is normal.
    - The distribution has a mean of $\hat{p}_1 - \hat{p}_2$, the true population difference.
    - The distribution has a standard deviation of $\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.
- The formula for the confidence interval is:

$$\text{Estimate} \pm \text{Margin of Error}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Key Equations

confidence interval for a population proportion

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

confidence interval for the difference between two population proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

margin of error (*ME*)

$$ME = z^* \cdot (\text{standard error})$$

standard error

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

minimum sample size needed to produce a given margin of error

$$n = \hat{p}(1-\hat{p})(\frac{z^*}{ME})^2$$

standard deviation for the difference in proportions

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Glossary

confidence interval for a population proportion

reasonable range of values where we expect the population proportion to fall within, with a chosen degree of confidence

confidence level

how much confidence we have in the method used to construct the interval

margin of error (ME)

determines width of a confidence interval

point estimate

a single value based on representative sample data that is a plausible estimate of the population parameter