# Unit 6: Inference for Categorical Data - Proportions

## 6.1 - Why Be Normal?

- the next four units will be focused on analyzing if there is convincing evidence for a claim
- can use sample proportions to calculate how often something is likely to occur, with z-scores

> 📢 - This unit will be focused on using proportions to prove if the categorical data of a study is consistent with a claim

## 6.2 - Constructing a Confidence Interval for a Population Proportion

- when estimating the proportion of successes in a single population, use a one-sample z interval for a population proportion
    - check the following conditions
        - the data was collected using a random sample from the population AND either the data was sampled with replacement OR that the sample size is less than 10% of the population
        - the shape of the sampling distribution is approximately normal, meaning that $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$
- calculating margin of error
    - describes how much a sample statistic is likely to vary in the population parameter
    - determined by how much the statistic typically varies from the parameter AND how confident we want to be in the estimate
        - **margin of error:** critical value multiplied by the standard error of statistic
            - the standard error is an estimate of the standard deviation of the sampling distribution
            - $SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $\hat{p}$ is the sample proportion (because we can't use the actual proportion)
        - **critical value:** multiplier that makes the margin of error large enough to give a specific amount of confidence
            - should represent the boundaries encompassing $C$% of the normal sampling distribution, where $C$% is approximately the confidence level
            - find the two z-scores representing the boundaries, for example, if I wanted to calculate a 95% confidence interval, I would find the z-score values of -1.96 and 1.96 (which encompass the middle 95%) and my critical value $z^*$ would be 1.96
- **confidence interval:** equal to a point estimate $\pm$ the margin of error
    - $CI = \hat{p} \pm z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where $\hat{p}$ is the point estimate, $z^*$ is the critical value, and $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is the margin of error
    - used to calculate a range of how likely a proportion is going to occur if it was compared to the rest of the sampling distribution
        - usually used when only one point estimate is available and not the population parameters
- when working backwards to find sample size (starting with a known confidence level and a margin of error), use 0.5 for $\hat{p}$ because it'll give you the upper bound for the size

> - **Margin of error:** critical value * standard error of statistic
> - **Critical value:** the multiplier that makes the margin of error large enough to satisfy a specific confidence level
> - **Confidence interval:** $CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where the two values represent the upper and lower limit of likeliness that the point statistic would occur if compared to the sampling distribution
> - The confidence interval estimates likeliness because we don't have the actual sampling distribution data
> - When estimating the proportion of successes in a single population, use a *one-sample z interval for a population proportion*

# 6.3 - Justifying a Claim Based on a Confidence Interval for a Population Proportion

- interpret a confidence interval with the sentence starter "we are $C$% confident that the interval from [lower bound] to [upper bound] captures the [population parameter]
    - when given only the confidence interval and the margin of error (and not the sample size), first find the interval bounds and then evaluate if the bounds can prove the claim
    - $C$% confidence means that if we took the proportion $p$ and sample size $n$ and calculated the $C$% confidence interval for each sample, approximately $C$% of the samples would capture $p$ in its interval
        - this is only true when there is repeated sampling - the confidence level does not describe the likelihood that ONE sample captures the $p$ value
- the margin of error is small when the sample size is large OR the confidence level is small
- steps to construct and interpret a confidence interval for a population proportion
    1. define the parameter you are trying to estimate - $C$% confidence interval for $p$ = [proportion parameter]
    2. check that the sampling is random, the sample size is less than 10% of the population size, and that the number of successes and failures are greater than 10
    3. calculate the value of $\hat{p}$ by getting the chance of success from the given sample, the margin of error, and the confidence interval
    4. interpret the confidence interval - "we are $C$% confident that the interval from [lower] to [upper] captures the [proportion parameter]"
- if the sample size is less than 10% of the population when you are sampling without replacement, the data will overestimate the data if you sampled *with* replacement, but the error is so small that it is negligible

> - The confidence interval means that $C$% of the taken samples will include the proportion within their interval bounds
> - The confidence *level* does NOT mean how likely one sample is to include the proportion $p$ because the sample will either include it or not include it - the confidence level represents the percentage of all possible samples that can be expected to be included in the interval

# 6.4 - Setting Up a Test for a Population Proportion

- **null hypothesis:** a claim that there is no difference or change
    - $H_0$: where $p = n$ so that the successes have the same chance of happening as the failures, and $p$ is the proportion of the population parameter
    - until evidence proves otherwise, the null hypothesis is true
    - always contains an equality reference
- **alternative hypothesis:** the claim that we are trying to prove with the evidence collected

        - $H_a$: where $p \neq n$, $p > 0.5$, or $p < 0.5$ to show that there was something that affected the data to sway a certain way
        - always contains an inequality
            - $\neq$ means that the alternative is two-sided, where both sides of data supports the claim
            - $<$ or $>$ means that the alternative is one-sided, where only one side of the data is considered
- when testing the claim about the proportion of successes in a single population, use a one-sample z test for a population proportion
    - check the following conditions
        - the data was collected using a random sample from the population AND either the data was sampled with replacement OR that the sample size is less than 10% of the population
        - the shape of the sampling distribution is approximately normal, meaning that $np_0 \geq 10$ and $n(1-p_0) \geq 10$ where $p_0$ is the proportion specified by the null hypothesis

> - **Null hypothesis:** the standard hypothesis that is assumed true and claims that nothing in the sample differs from the proportion successes
> - **Alternative hypothesis:** the claim that the data is supposed to prove
> - When testing a claim about the proportion of successes in a single population, use a *one sample z test for a population proportion*

## 6.5 - Interpreting P-Values

- the evidence $\hat{p}$ for $H_a$ should be greater than the predetermined threshold and consistent with the direction
- to figure out whether the evidence $\hat{p}$ is convincing, the likelihood of getting a $\hat{p}$ of this value or stronger needs to be determined
    - $z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ to calculate the standardized test statistic, or the z-score, then find the likeliness to get a score equal to or greater than that z-score
    - if the $H_a$ is a $\neq$ inequality, then the $p$ value has to be doubled
- interpret a $p$ value with the sentence starter "assuming [null hypothesis is true], there is a $p$ probability of getting a sample proportion of $\hat{p}$ or greater/less by chance alone in a random sample of $n$ [context about the sample size]"

> - To determine how strong the data is, the likelihood of getting that value or better needs to be calculated with a z-score probability chart
> - If the alternate hypothesis is an inequality, then both sides of the data need to be taken into account for the final probability

## 6.6 - Concluding a Test for a Population Proportion

- the smaller the $p$ value, the more convincing the statistical evidence for $H_a$
- significance level $\alpha$ is a predetermined boundary for whether the $p$ value is small or not, and usually include $\alpha = 0.01$, $\alpha = 0.05$, and $\alpha = 0.10$
    - if $p$ value $\leq \alpha$, we reject $H_0$ because there is convincing statistical evidence for $H_a$
- interpret a confidence level with the sentence starter "because the $p$ value of [p-value] is greater/less than [confidence level of $\alpha$], we (fail to) reject $H_0$. There is is/not convincing statistical evidence that [context about alternative hypothesis]"
    - if a confidence level isn't specified, use $\alpha = 0.05$
    - if you fail to reject $H_0$, don't conclude that $H_0$ is true

  - if $H_0$ is rejected, don't say that $H_a$ is "proven" true

> - The null hypothesis $H_0$ can only be rejected if the $p$ value is lower than the significance level $\alpha$
> - When the null hypothesis is not rejected, don't claim that the null hypothesis is true
> - When the null hypothesis is rejected, don't claim that $H_a$ is proven true

## 6.7 - Potential Errors When Performing Tests

- **type I error:** when there is convincing evidence for $H_a$ but $H_0$ is actually true
  - occurs when the null hypothesis is true but is rejected (false positive)
  - when the null hypothesis is true, the probability of a type I error is equal to the significance level
- **type II error:** when there is no convincing evidence for $H_a$ but $H_0$ is actually false
  - occurs when the null hypothesis is false but it is not rejected (false negative)
  - in a lot of cases, a type II error is worse
  - **power:** the probability that a test will correctly reject a false null hypothesis for a given $p$ value
    - probability of a type II error is $1 -$ power
    - power is greater (probability of a type II error is smaller) when the sample size increases, the significance level $\alpha$ increases, the standard error decreases, the true parameter is farther from the null

> - **Type I error:** when the null hypothesis is true but it gets rejected
> - **Type II error:** when the null hypothesis is false but it isn't rejected
> - **Power:** the probability that a test will correctly reject a false null hypothesis

## 6.8 - Confidence Interval for the Difference of Two Proportions

- when estimating the difference in proportions of two populations, use a two-sample z interval for a difference in proportions
  - this procedure is used when
    - two random samples have been selected from two populations
    - subjects are randomly assigned to two groups in an experiment
  - check the following conditions
    - the data was collected using a random samples from both populations AND either the data was sampled with replacement for each sample OR that the sample sizes is less than 10% of the population
    - the shape of the sampling distribution is approximately normal, meaning that $n_1\hat{p}_1 \geq 10$, $n_1(1-\hat{p}_1) \geq 10$, $n_2\hat{p}_2 \geq 10$, and $n_2(1-\hat{p}_2) \geq 10$
- calculate the standard error of statistic with $\hat{p}$ values because we don't have $p$ values
  - $SE_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ to calculate standard error
- $CI = \hat{p} \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ where $\hat{p}$ is the given statistic and $z^*$ is the critical value

> - When estimating the confidence interval for a difference in proportions, use a *two-sample z interval for a difference in proportions*

## 6.9 - Justifying a Claim Based on a Confidence Interval for a Difference Between Population Proportions

- interpret a confidence interval with the sentence starter "we are $C$% confident that the interval from [lower bound] to [upper bound] captures the difference (1 - 2) [value to be estimated]"
    - if 0 is included in the confidence interval, then there is a possibility that there is no difference between two estimates
- steps to construct and interpret a confidence interval for a difference in proportions
    1. define the confidence interval for the difference in proportions - $C$% confidence for (parameter 1 - parameter 2)
    2. check that the sampling is random, the sample size is less than 10% of the population size, and that the number of successes and failures are greater than 10
    3. calculate $\hat{p}_1$ and $\hat{p}_2$, margin of error, and confidence interval
    4. interpret the confidence interval - "we are $C$% confident that the interval from [lower] to [upper] captures the difference between (parameter 1 - parameter 2) [context]"
- the confidence level is the proportion out of all possible random samples that would capture the true difference

> 📢 - Because we're calculating a difference between two proportions, if the confidence interval contains 0, that means there is a chance that there is no difference between proportions

## 6.10 - Setting Up a Test for the Difference of Two Population Proportions

- the null hypothesis $H_0$ claims that there is no difference between the two proportions, where $p_1 = p_2$
- the alternative hypothesis $H_a$ claims that there is a difference in proportions, where $p_1 > p_2$ or $p_2 > p_1$
- when testing the claim about a difference in proportions between two populations, use a two-sample z test for a difference in proportions
    - this procedure is used when
        - two random samples have been selected from two populations
        - subjects are randomly assigned to two groups in an experiment
    - calculate the combined proportion of successes $\hat{p}_c$ in both groups and then check the following conditions
        - the data was collected using a random samples from both populations AND either the data was sampled with replacement for each sample OR that the sample sizes is less than 10% of the population
        - the shape of the combined sampling distribution is approximately normal, meaning that $n_1\hat{p}_c \geq 10$, $n_1(1-\hat{p}_c) \geq 10$, $n_2\hat{p}_c \geq 10$, and $n_2(1-\hat{p}_c) \geq 10$

> 📢 - The null hypothesis for a test claims that there is no difference between two populations, while the alternative hypothesis claims that there is a difference
> - When testing a claim about the proportion of successes between two populations, use a *two-sample z test for a difference in proportions*
> - Use the combined proportion of successes $\hat{p}_c$ when testing the claim instead of $\hat{p}_1$ and $\hat{p}_2$

## 6.11 - Carrying Out a Test for the Difference of Two Population Proportions

- calculate the standardized test statistic of the difference

  - $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)(\frac{1}{n_1}+\frac{1}{n_2})}}$ where $\hat{p}_c$ is the combined proportion of successes
- $p$ value is the probability of getting a statistic as strong or stronger than the observed test statistic
- interpret a $p$ value with the sentence starter "assuming [null hypothesis is true], there is a $p$ probability of getting a difference in proportions of $\hat{p}$ or greater/less by chance alone in the random assignment"
  - small $p$ values means that the test statistic is unlikely to occur while large $p$ values means that the test statistic is likely to occur by random chance
  - state the conclusion with the sentence starter "because the $p$ value of [p-value] is greater/less than [confidence level of $\alpha$], we (fail to) reject $H_0$. There is/not convincing statistical evidence that [context about alternative hypothesis]"

> 📢 - If $H_a$ is an inequality, double the $p$ value to count both sides of the normal distribution