Chapter 10: Inference on Proportions

DSCC 462

Computational Introduction to Statistics

Anson Kahng

Fall 2022

Inference on Proportions

Inference on Proportions

- So far, we have considered inference for when we have continuous data

Inference on Proportions

- So far, we have considered inference for when we have continuous data
- We can also extend inferential methods to cover count data

Inference on Proportions

- So far, we have considered inference for when we have continuous data
- We can also extend inferential methods to cover count data
- In particular, we are often interested in the proportion of times a dichotomous (i.e., yes/no) event occurs

Sampling Distribution of a Proportion

Sampling Distribution of a Proportion

- Recall that the sample mean is distributed like $\hat{p} \sim N\left(p, \sqrt{\frac{(p(1-p)}{n}}\right)$, given that $np \geq 5$ and $n(1-p) \geq 5$

$$p = \frac{\text{Binom}(n, p)}{n} \xrightarrow{\text{CLT}} \frac{N\left(np, \sqrt{np(1-p)}\right)}{n}$$

$$\rightarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Sampling Distribution of a Proportion

$\bar{x}$

- Recall that the sample mean is distributed like $\hat{p} \sim N\left(p, \sqrt{\frac{(p(1-p)}{n}}\right)$,

  given that $np \geq 5$ and $n(1-p) \geq 5$

- Thus, $Z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$ has a standard normal distribution

Confidence Intervals for Proportions

Confidence Intervals for Proportions

- Confidence intervals for population proportions follows the same procedure as what we used for population means

Confidence Intervals for Proportions

- Confidence intervals for population proportions follows the same procedure as what we used for population means
- Draw a sample of size n and compute $\hat{p} = \dfrac{x}{n}$

Confidence Intervals for Proportions

- Confidence intervals for population proportions follows the same procedure as what we used for population means
- Draw a sample of size n and compute $\hat{p} = \dfrac{x}{n}$
- $\hat{p}$ is a point estimate of population proportion p

$\hat{p} \sim N(p,$

Confidence Intervals for Proportions

- Confidence intervals for population proportions follows the same procedure as what we used for population means
- Draw a sample of size n and compute $\hat{p} = \dfrac{x}{n}$
- $\hat{p}$ is a point estimate of population proportion p
- We know from above that $Z = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ is a standard normal random variable, given that n is sufficiently large (i.e., $np \geq 5$ and $n(1-p) \geq 5$)

Confidence Intervals for Proportions

Confidence Intervals for Proportions

- For a standard normal distribution, 95% of possible outcomes lie between `qnorm(0.025) = -1.96` and `qnorm(0.975) = 1.96`

Confidence Intervals for Proportions

- For a standard normal distribution, 95% of possible outcomes lie between `qnorm(0.025) = -1.96` and `qnorm(0.975) = 1.96`
- Thus, $\Pr\left(-1.96 \leq \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \leq 1.96\right) = 0.95$

$$\Pr\left(-1.96\sqrt{\frac{p(1-p)}{n}} + \hat{p} \leq p \leq 1.96\sqrt{\frac{p(1-p)}{n}} + \hat{p}\right) = 0.95$$

Confidence Intervals for Proportions

- For a standard normal distribution, 95% of possible outcomes lie between `qnorm(0.025) = -1.96` and `qnorm(0.975) = 1.96`

- Thus, $\Pr\left(-1.96 \leq \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \leq 1.96\right) = 0.95$

- This can be rearranged to give

$$\Pr\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$

Confidence Intervals for Proportions

- For a standard normal distribution, 95% of possible outcomes lie between `qnorm(0.025) = -1.96` and `qnorm(0.975) = 1.96`
- Thus, $\Pr\left(-1.96 \leq \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} \leq 1.96\right) = 0.95$
- This can be rearranged to give
$$\Pr\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) = 0.95$$
- Note that this confidence interval depends on the (unknown) value of p!

Confidence Intervals for Proportions

Confidence Intervals for Proportions

- So how do we estimate p? Use $\hat{p}$, our sample estimate (Wald)

Confidence Intervals for Proportions

- So how do we estimate p? Use $\hat{p}$, our sample estimate (Wald)
- Therefore, our confidence interval calculation becomes

$$\Pr\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

Confidence Intervals for Proportions

- So how do we estimate p? Use $\hat{p}$, our sample estimate (Wald)

- Therefore, our confidence interval calculation becomes

$$\Pr\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 0.95$$

- In other words, we are 95% confident that the interval $\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$ contains the true population proportion p

given Wald estimate $p = \hat{p}$.

Confidence Intervals for Proportions

Confidence Intervals for Proportions

- In general, an approximate two-sided $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)$

Confidence Intervals for Proportions

- In general, an approximate two-sided $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)$

- A one-sided lower $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(\hat{p} - z_{\alpha}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ 1\right)$

Confidence Intervals for Proportions

$p = \hat{p}$

- In general, an approximate two-sided $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(\hat{p} - z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)$
- A one-sided lower $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(\hat{p} - z_{\alpha}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\ 1\right)$
- A one-sided upper $(1-\alpha)\cdot 100\,\%$ confidence interval for p is given by $\left(0,\ \hat{p} + z_{\alpha}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right)$

Wald vs. Wilson Intervals

Wald vs. Wilson Intervals

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)

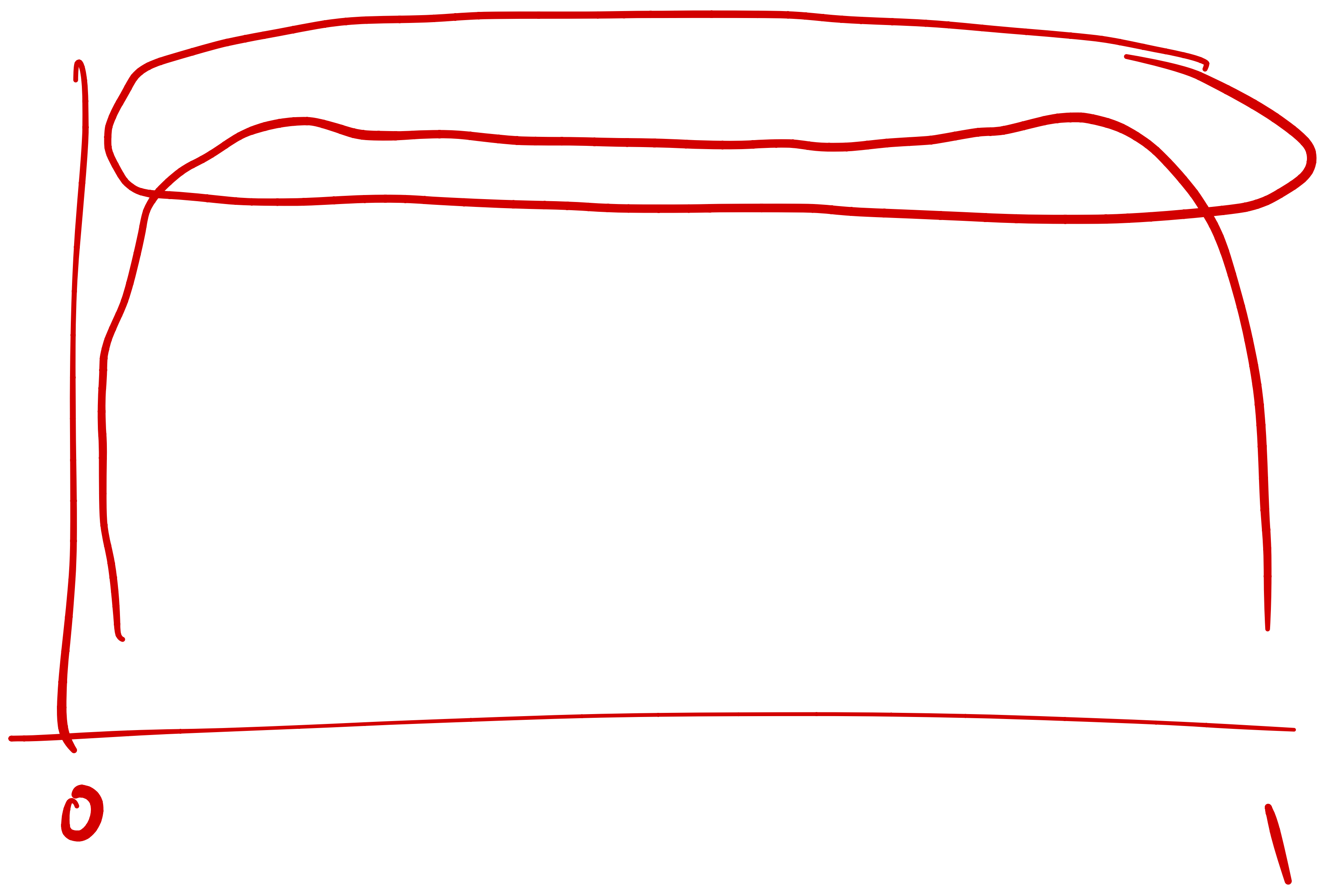
Wald vs. Wilson Intervals

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)

Wald vs. Wilson Intervals

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

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)
    - Less than $(1-\alpha)\cdot 100\,\%$ confidence interval



Wald vs. Wilson Intervals

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)
  - Less than $(1 - \alpha) \cdot 100\,\%$ confidence interval
- (One) alternative method: Wilson (what `prop.test()` uses in R)

Wald vs. Wilson Intervals

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)
  - Less than $(1-\alpha)\cdot 100\,\%$ confidence interval
- (One) alternative method: Wilson (what `prop.test()` uses in R)
  - Solve for p in terms of $Z, \hat{p}$ from the approximation $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$

Wald vs. Wilson Intervals

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)
  - Less than $(1-\alpha)\cdot 100\,\%$ confidence interval
- (One) alternative method: Wilson (what `prop.test()` uses in R)
  - Solve for p in terms of $Z, \hat{p}$ from the approximation $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$
  - Get an estimate of p that is a weighted average of $\hat{p}$ and $\dfrac{1}{2}$, where the weight on $\hat{p}$ increases with n

~~Wald~~ vs. Wilson Intervals R

- Recall that we estimate p using $\hat{p}$, our sample estimate (Wald)
- In general, this provides poor coverage when $\hat{p}$ is close to extremes (0 or 1)
  - Less than $(1-\alpha)\cdot 100\,\%$ confidence interval
- (One) alternative method: Wilson (what `prop.test()` uses in R) MN
  - Solve for p in terms of $Z, \hat{p}$ from the approximation $Z = \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}}$
  - Get an estimate of p that is a weighted average of $\hat{p}$ and $\dfrac{1}{2}$, where the weight on $\hat{p}$ increases with n
  - Better coverage!

Confidence Intervals for Proportions: Example

Confidence Intervals for Proportions: Example

- Setup: We are interested in determining what proportion of a population is right-handed. Suppose we have a sample of $n = 62$ subjects and 53 of these subjects are right-handed. Find a 95% confidence interval for the population proportion of right-handed people in the population

Confidence Intervals for Proportions: Example

- Setup: We are interested in determining what proportion of a population is right-handed. Suppose we have a sample of $n = 62$ subjects and 53 of these subjects are right-handed. Find a 95% confidence interval for the population proportion of right-handed people in the population

- Find $\hat{p}$ (our estimate of p):

Confidence Intervals for Proportions: Example

- Setup: We are interested in determining what proportion of a population is right-handed. Suppose we have a sample of $n = 62$ subjects and 53 of these subjects are right-handed. Find a 95% confidence interval for the population proportion of right-handed people in the population

- Find $\hat{p}$ (our estimate of p):

- Check normality assumptions:

Confidence Intervals for Proportions: Example

- Setup: We are interested in determining what proportion of a population is right-handed. Suppose we have a sample of $n = 62$ subjects and 53 of these subjects are right-handed. Find a 95% confidence interval for the population proportion of right-handed people in the population

- Find $\hat{p}$ (our estimate of p): $\hat{p} = \frac{53}{62} = 0.855$

- Check normality assumptions: $n\hat{p} = 53$ $\quad n(1-\hat{p}) = 9$ $\quad \alpha = 0.05$

- Apply a two-sided 95% confidence interval: $\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

"by hand" $(0.767, \; 0.943)$

prop.test $\rightarrow (0.75, \; 0.95)$

Normal Approximations of Binomial Distributions

Normal Approximations of Binomial Distributions

- Note that the true distribution for proportions is a binomial distribution (number of “successes” out of a certain number of trials)

Normal Approximations of Binomial Distributions

- Note that the true distribution for proportions is a binomial distribution (number of “successes” out of a certain number of trials)

- However, confidence intervals are based on the normal distribution

Normal Approximations of Binomial Distributions

- Note that the true distribution for proportions is a binomial distribution (number of “successes” out of a certain number of trials)
- However, confidence intervals are based on the normal distribution
- We are using the normal distribution as an approximation for a binomial distribution

Normal Approximations of Binomial Distributions

- Note that the true distribution for proportions is a binomial distribution (number of “successes” out of a certain number of trials)
- However, confidence intervals are based on the normal distribution
- We are using the normal distribution as an approximation for a binomial distribution
- Normal approximation (Wilson) confidence intervals can be calculated in R using `prop.test(x,n)`

Normal Approximations of Binomial Distributions

- Note that the true distribution for proportions is a binomial distribution (number of "successes" out of a certain number of trials)
- However, confidence intervals are based on the normal distribution
- We are using the normal distribution as an approximation for a binomial distribution
- Normal approximation (Wilson) confidence intervals can be calculated in R using `prop.test(x,n)`
- Exact binomial (Clopper-Pearson) confidence intervals can be calculated in R using `binom.test(x,n)`

Sample Size Estimation

Sample Size Estimation

- For confidence intervals on proportions, we have that the margin of error is $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (half the length of the confidence interval)

Sample Size Estimation

- For confidence intervals on proportions, we have that the margin of error is $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (half the length of the confidence interval)
- Just like before, if we want a certain margin of error at the same confidence level, we can determine the number of subjects (n) needed to get the desired results

Sample Size Estimation

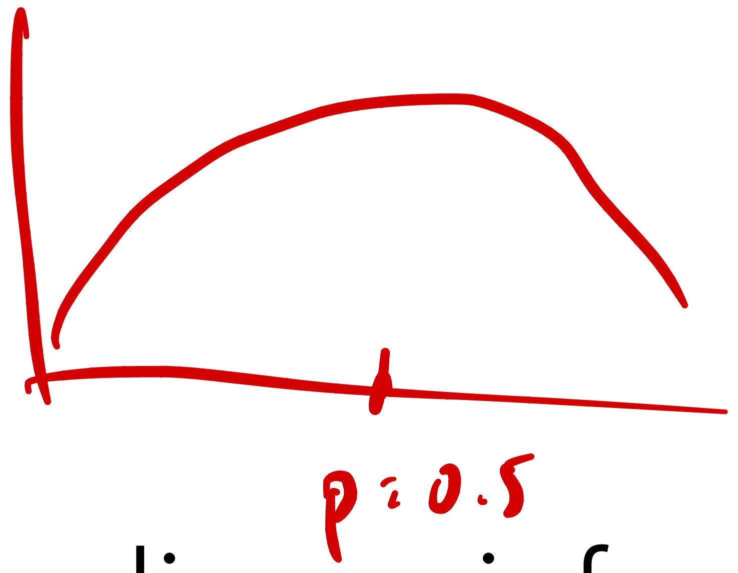
- For confidence intervals on proportions, we have that the margin of error is $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (half the length of the confidence interval)
- Just like before, if we want a certain margin of error at the same confidence level, we can determine the number of subjects (n) needed to get the desired results
- Thus, $n = \left\lceil \dfrac{z_{\alpha/2}^2 p(1-p)}{m^2} \right\rceil$

Sample Size Estimation

- For confidence intervals on proportions, we have that the margin of error is $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (half the length of the confidence interval)
- Just like before, if we want a certain margin of error at the same confidence level, we can determine the number of subjects (n) needed to get the desired results
- Thus, $n = \left\lceil \dfrac{z_{\alpha/2}^2 p(1-p)}{m^2} \right\rceil$
- If we can estimate p based on previous studies or information, use that

Sample Size Estimation

- For confidence intervals on proportions, we have that the margin of error is $m = z_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$ (half the length of the confidence interval)
- Just like before, if we want a certain margin of error at the same confidence level, we can determine the number of subjects (n) needed to get the desired results
- Thus, $n = \left\lceil \dfrac{z_{\alpha/2}^2 p(1-p)}{m^2} \right\rceil$



- If we can estimate p based on previous studies or information, use that
- Otherwise, use $p = 0.5$ to get the most conservative estimate of the standard error (overestimate of the number of subjects needed)

Sample Size Estimation: Example

Sample Size Estimation: Example

- Setup: We want to determine what proportion of college students have an iPhone within a $\underbrace{\text{margin of error of 8}}_{m = 0.08}$ percentage points with 95% confidence

Sample Size Estimation: Example

- Setup: We want to determine what proportion of college students have an iPhone within a margin of error of 8 percentage points with 95% confidence

- Q1: How large of a sample should you take?

qnorm(0.975)

1.96

0.5   0.5

$m = 0.08$

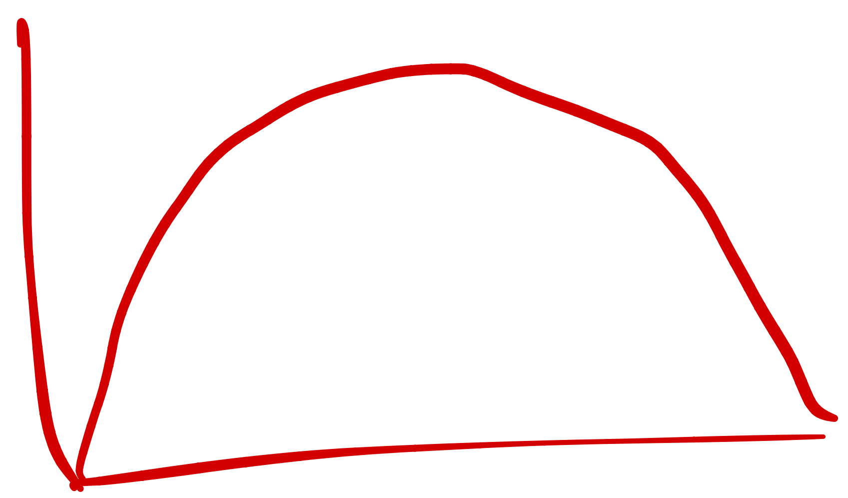
$p = 0.5$

$$n = \left\lceil \frac{z_{\alpha/2}^2 \cdot p(1-p)}{m^2} \right\rceil = \underline{151}$$

Sample Size Estimation: Example

- Setup: We want to determine what proportion of college students have an iPhone within a margin of error of 8 percentage points with 95% confidence
- Q1: How large of a sample should you take?

151.



- Q2: A national study determined that 38% of all Americans own iPhones. Now, how large of a sample should you take?

$m = 0.08$

$p = 0.38$

$1-p = 0.62$

$$n = \left\lceil \frac{1.96^2 \cdot 0.38 \cdot 0.62}{0.08^2} \right\rceil = \cancel{142}$$

40

Hypothesis Testing for Proportions

Hypothesis Testing for Proportions

- Just as we used hypothesis tests to see if a population mean was equal to some hypothesized value, we can also test whether a population proportion is equal to some value

Hypothesis Testing for Proportions

- Just as we used hypothesis tests to see if a population mean was equal to some hypothesized value, we can also test whether a population proportion is equal to some value

- Consider a two-tailed test at the $\alpha = 0.05$ significance level

Hypothesis Testing for Proportions

- Just as we used hypothesis tests to see if a population mean was equal to some hypothesized value, we can also test whether a population proportion is equal to some value

- Consider a two-tailed test at the $\alpha = 0.05$ significance level

- $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$

Hypothesis Testing for Proportions

- Just as we used hypothesis tests to see if a population mean was equal to some hypothesized value, we can also test whether a population proportion is equal to some value

- Consider a two-tailed test at the $\alpha = 0.05$ significance level

- $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$

- Draw a random sample of size n observations from the underlying population (each observation is a dichotomous yes/no)

$\hat{p} = \frac{x}{n}$

Hypothesis Testing for Proportions

- Just as we used hypothesis tests to see if a population mean was equal to some hypothesized value, we can also test whether a population proportion is equal to some value

- Consider a two-tailed test at the $\alpha = 0.05$ significance level

- $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$

- Draw a random sample of size n observations from the underlying population (each observation is a dichotomous yes/no)

- Calculate a z-statistic: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1 - p_0)}{n}}}$

CI: $z = \dfrac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}}$

Hypothesis Testing for Proportions

Hypothesis Testing for Proportions

- For sufficiently large n and when H_0 is true, we can compare z to a standard normal distribution to calculate the probability of obtaining a proportion as extreme or more extreme than $\hat{p}$

Hypothesis Testing for Proportions

- For sufficiently large n and when H_0 is true, we can compare z to a standard normal distribution to calculate the probability of obtaining a proportion as extreme or more extreme than $\hat{p}$
- Calculate p-value in R by `p=2*pnorm(-abs(z))`

Hypothesis Testing for Proportions

- For sufficiently large n and when H_0 is true, we can compare z to a standard normal distribution to calculate the probability of obtaining a proportion as extreme or more extreme than $\hat{p}$
- Calculate p-value in R by `p=2*pnorm(-abs(z))`
- If $p \leq 0.05$, we reject the null hypothesis and conclude that $p \neq p_0$

Hypothesis Testing for Proportions

- For sufficiently large n and when H_0 is true, we can compare z to a standard normal distribution to calculate the probability of obtaining a proportion as extreme or more extreme than $\hat{p}$
- Calculate p-value in R by `p=2*pnorm(-abs(z))`
- If $p \leq 0.05$, we reject the null hypothesis and conclude that $p \neq p_0$
- If $p > 0.05$, we fail to reject the null hypothesis and conclude that there is not significant evidence to say that $p \neq p_0$

Hypothesis Testing for Proportions: Example

Hypothesis Testing for Proportions: Example

- Consider the dominant hand example ($n = 62$, $\hat{p} = 53/62 = 0.855$)

Hypothesis Testing for Proportions: Example

- Consider the dominant hand example ($n = 62$, $\hat{p} = 53/62 = 0.855$)
- Test at the $\alpha = 0.05$ level whether the true population proportion of right-handed people is equal to 0.9

Hypothesis Testing for Proportions: Example

- Consider the dominant hand example ($n = 62$, $\hat{p} = 53/62 = 0.855$)
- Test at the $\alpha = 0.05$ level whether the true population proportion of right-handed people is equal to 0.9
- $H_0 : p = 0.9$ vs. $H_1 : p \neq 0.9$

Hypothesis Testing for Proportions: Example

- Consider the dominant hand example ($n = 62,\ \hat{p} = 53/62 = 0.855$)
- Test at the $\alpha = 0.05$ level whether the true population proportion of right-handed people is equal to 0.9
- $H_0 : p = 0.9$ vs. $H_1 : p \neq 0.9$
- Check normality assumptions based on p:

Hypothesis Testing for Proportions: Example

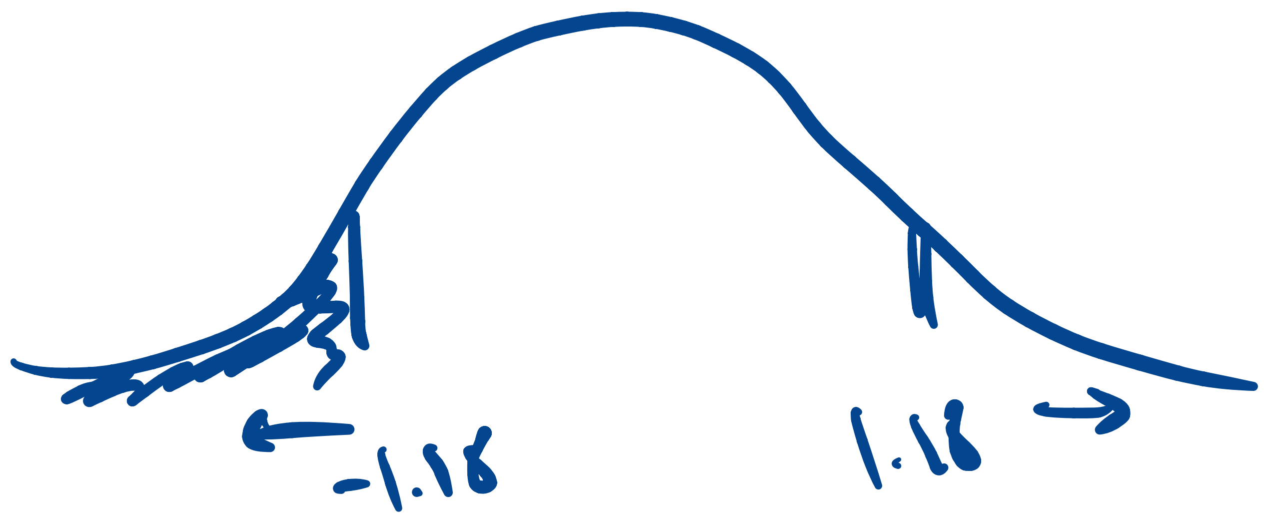
- Consider the dominant hand example ($n = 62,\ \hat{p} = 53/62 = 0.855$)
- Test at the $\alpha = 0.05$ level whether the true population proportion of right-handed people is equal to 0.9
- $H_0 : p = 0.9$ vs. $H_1 : p \neq 0.9$
- Check normality assumptions based on p:
- Calculate z-score:

Hypothesis Testing for Proportions: Example

- Consider the dominant hand example ($n = 62$, $\hat{p} = 53/62 = 0.855$)
- Test at the $\alpha = 0.05$ level whether the true population proportion of right-handed people is equal to 0.9
- $H_0 : p = 0.9$ vs. $H_1 : p \neq 0.9$

$$\boxed{\mu = p_0 , \quad \sigma = \frac{p_0(1-p_0)}{n}}$$

- Check normality assumptions based on p: $np = 62 \cdot 0.9 > 5 \qquad n(1-p) = 62 \cdot 0.1 = 6.2$
- Calculate z-score: $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}} = \dfrac{0.855 - 0.9}{\sqrt{\dfrac{0.9 \times 0.1}{62}}} = -1.18$
- Calculate p-value:



$2 \cdot \text{pnorm}(-1.18) = 0.238$

Confidence Intervals vs. Hypothesis Tests for Proportions

Confidence Intervals vs. Hypothesis Tests for Proportions

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent

Confidence Intervals vs. Hypothesis Tests for Proportions

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent
- This is no longer the case for proportions!

Confidence Intervals vs. Hypothesis Tests for Proportions

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent
- This is no longer the case for proportions!
    - Intuition: For sample means, there are two parameters of interest (μ, σ), whereas for proportions, p determines both the mean and variance

Confidence Intervals vs. Hypothesis Tests for Proportions

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent
- This is no longer the case for proportions!
    - Intuition: For sample means, there are two parameters of interest (μ, σ), whereas for proportions, p determines both the mean and variance
- For proportion hypothesis tests, we calculate the standard error based on p_0 as $\sqrt{\frac{p_0(1-p_0)}{n}}$ (i.e., our frame of reference is centered at the null hypothesis)

Confidence Intervals vs. Hypothesis Tests for Proportions

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent
- This is no longer the case for proportions!
  - Intuition: For sample means, there are two parameters of interest (μ, σ), whereas for proportions, p determines both the mean and variance
- For proportion hypothesis tests, we calculate the standard error based on p_0 as $\sqrt{\frac{p_0(1-p_0)}{n}}$ (i.e., our frame of reference is centered at the null hypothesis)
- For proportion (Wald) confidence intervals, we calculate the standard error based on $\hat{p}$ as $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (i.e., our frame of reference is centered at our observed sample proportion)

$H_0: p = 0.3 \qquad H_1: p \neq 0.3$

Confidence Intervals vs. Hypothesis Tests for Proportions

$\hat{p}$

- When looking at sample means, confidence intervals and hypothesis tests are essentially equivalent
- This is no longer the case for proportions!
  - Intuition: For sample means, there are two parameters of interest (μ, σ), whereas for proportions, p determines both the mean and variance
- For proportion hypothesis tests, we calculate the standard error based on p_0 as $\sqrt{\frac{p_0(1-p_0)}{n}}$ (i.e., our frame of reference is centered at the null hypothesis)
- For proportion (Wald) confidence intervals, we calculate the standard error based on $\hat{p}$ as $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (i.e., our frame of reference is centered at our observed sample proportion)
- As with confidence intervals, we can perform an exact test based on the binomial distribution instead of using the normal approximation: `binom.test(x,n,p=p₀)`

$\Pr(\hat{x} \mid p_0, n)$

One-Sided Hypothesis Tests

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
    - $H_1 : p \neq p_0$

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
    - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
    - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction
    - $H_1 : p > p_0$

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
  - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction
  - $H_1 : p > p_0$
  - $H_1 : p < p_0$

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
  - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction
  - $H_1 : p > p_0$
  - $H_1 : p < p_0$
- For two-sided tests, we are concerned with the area in both tails of the distribution

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
  - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction
  - $H_1 : p > p_0$
  - $H_1 : p < p_0$
- For two-sided tests, we are concerned with the area in both tails of the distribution
- For one-sided tests, we are concerned with the area in only one tail of the distribution

One-Sided Hypothesis Tests

- With two-sided hypothesis tests, we were only concerned with whether or not there was a difference from the postulated population proportion
  - $H_1 : p \neq p_0$
- However, we are sometimes interested in deviations only in one direction
  - $H_1 : p > p_0$
  - $H_1 : p < p_0$
- For two-sided tests, we are concerned with the area in both tails of the distribution
- For one-sided tests, we are concerned with the area in only one tail of the distribution
- Analyses follow directly as they did for one-sided tests for sample means

Comparison of Two Proportions

Comparison of Two Proportions

- We can extend hypothesis tests to situations where we compare proportions for two groups

Comparison of Two Proportions

- We can extend hypothesis tests to situations where we compare proportions for two groups
- Interested in testing whether the proportions from two independent populations are the same

Comparison of Two Proportions

- We can extend hypothesis tests to situations where we compare proportions for two groups
- Interested in testing whether the proportions from two independent populations are the same
  - $H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$

Comparison of Two Proportions

- We can extend hypothesis tests to situations where we compare proportions for two groups
- Interested in testing whether the proportions from two independent populations are the same
    - $H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$
- Our alternative hypothesis is that there is a difference between these groups

Comparison of Two Proportions

- We can extend hypothesis tests to situations where we compare proportions for two groups
- Interested in testing whether the proportions from two independent populations are the same
    - $H_0 : p_1 = p_2$ or $H_0 : p_1 - p_2 = 0$
- Our alternative hypothesis is that there is a difference between these groups
    - $H_1 : p_1 \neq p_2$ or $H_1 : p_1 - p_2 \neq 0$

Comparison of Two Proportions

Comparison of Two Proportions

- We draw a sample of size n_1 from the first population and a sample of size n_2 from the second population

Comparison of Two Proportions

- We draw a sample of size n_1 from the first population and a sample of size n_2 from the second population
- There are x_1 successes in the first sample and x_2 successes in the second sample

Comparison of Two Proportions

- We draw a sample of size n_1 from the first population and a sample of size n_2 from the second population

- There are x_1 successes in the first sample and x_2 successes in the second sample

- Sample proportion for each group:

Comparison of Two Proportions

- We draw a sample of size n_1 from the first population and a sample of size n_2 from the second population
- There are x_1 successes in the first sample and x_2 successes in the second sample
- Sample proportion for each group:
  - $\hat{p}_1 = \dfrac{x_1}{n_1}$

Comparison of Two Proportions

- We draw a sample of size n_1 from the first population and a sample of size n_2 from the second population

- There are x_1 successes in the first sample and x_2 successes in the second sample

- Sample proportion for each group:

    - $\hat{p}_1 = \dfrac{x_1}{n_1}$

    - $\hat{p}_2 = \dfrac{x_2}{n_2}$

Comparison of Two Proportions

$H_0: \quad p_1 = p_2$

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter
  - $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

$\hat{p}_1 = \dfrac{x_1}{n_1}$

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter
  - $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
  - $\hat{p}$ is the weighted average of the two sample proportions (or total successes over total trials)

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter
  - $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
  - $\hat{p}$ is the weighted average of the two sample proportions (or total successes over total trials)
- The estimator of the standard error of $\hat{p}_1 - \hat{p}_2$ can now be based on this common $\hat{p}$

Comparison of Two Proportions

$\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right)$

$\hat{p}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter

  - $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

    $\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \right.$ var: $\left(\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$

  - $\hat{p}$ is the weighted average of the two sample proportions (or total successes over total trials)

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

- The estimator of the standard error of $\hat{p}_1 - \hat{p}_2$ can now be based on this common $\hat{p}$

  $p_1 = p_2 = p$

  - Standard error: $\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$

Comparison of Two Proportions

- Under the null hypothesis, $p_1 = p_2 = p$
- Thus, the data from both samples can be combined to estimate this common parameter
  - $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$
  - $\hat{p}$ is the weighted average of the two sample proportions (or total successes over total trials)
- The estimator of the standard error of $\hat{p}_1 - \hat{p}_2$ can now be based on this common $\hat{p}$
  - Standard error: $\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$
  - Similar to the "pooled" estimate for sample means

Comparison of Two Proportions

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2,\ \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right)$$

Comparison of Two Proportions

- Putting these pieces together, we get our z-statistic:

Comparison of Two Proportions

- Putting these pieces together, we get our z-statistic:

$H_0: p_1 = p_2$

- $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Wald

$p_1 \overset{?}{=} p_2$

$p(1-p)$

Comparison of Two Proportions

- Putting these pieces together, we get our z-statistic:
- $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
- If n_1 and n_2 are sufficiently large, this z statistic is approximately a standard normal (mean 0, standard deviation 1)

Comparison of Two Proportions

- Putting these pieces together, we get our z-statistic:
- $$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$
- If n_1 and n_2 are sufficiently large, this z statistic is approximately a standard normal (mean 0, standard deviation 1)
- Typically, we want $n_1\hat{p}_1,\ n_1(1 - \hat{p}_1),\ n_2\hat{p}_2,$ and $n_2(1 - \hat{p}_2)$ to all be greater than 5 (this is a conservative standard)

Comparison of Two Proportions

Comparison of Two Proportions

- With these conditions satisfied, we compare the value of the z statistic with the critical value to find a p-value, p

Comparison of Two Proportions

- With these conditions satisfied, we compare the value of the z statistic with the critical value to find a p-value, p
- If $p \leq \alpha$, we reject the null hypothesis

Comparison of Two Proportions

- With these conditions satisfied, we compare the value of the z statistic with the critical value to find a p-value, p
- If $p \leq \alpha$, we reject the null hypothesis
- If $p > \alpha$, we fail to reject the null hypothesis

Comparison of Two Proportions

Comparison of Two Proportions

- Consider our dominant hand example. Suppose we are interested in knowing whether the right-handedness rate is different for Group A and Group B

Comparison of Two Proportions

- Consider our dominant hand example. Suppose we are interested in knowing whether the right-handedness rate is different for Group A and Group B

- At the $\alpha = 0.01$ significance level, we will test the following hypotheses:

Comparison of Two Proportions

- Consider our dominant hand example. Suppose we are interested in knowing whether the right-handedness rate is different for Group A and Group B

- At the $\alpha = 0.01$ significance level, we will test the following hypotheses:

    - $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$

Comparison of Two Proportions

- Consider our dominant hand example. Suppose we are interested in knowing whether the right-handedness rate is different for Group A and Group B

- At the $\alpha = 0.01$ significance level, we will test the following hypotheses:

  - $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$

- We take samples of $n_A = 54$ and $n_B = 62$

Comparison of Two Proportions

- Consider our dominant hand example. Suppose we are interested in knowing whether the right-handedness rate is different for Group A and Group B

- At the $\alpha = 0.01$ significance level, we will test the following hypotheses:

    - $H_0 : p_A = p_B$ vs. $H_1 : p_A \neq p_B$

- We take samples of $n_A = 54$ and $n_B = 62$

- We observe $x_A = 48$ and $x_B = 60$ subjects being right handed

Comparison of Two Proportions

Comparison of Two Proportions

- Calculate proportions:

$$\hat{p}_A = \frac{48}{54} = 0.889$$

$$\hat{p}_B = \frac{60}{62} = 0.968$$

Comparison of Two Proportions

- Calculate proportions:

0.889

0.968

- Is this difference too large to be attributed to chance?

Comparison of Two Proportions

- Calculate proportions:

$\hat{p}_A = .889$

$\hat{p}_B = 0.968$

- Is this difference too large to be attributed to chance?

- Under H_0, $p_A = p_B$, so we can estimate their common value p

$$p = \frac{X_A + X_B}{n_A + n_B} = \frac{48+60}{54+62} = \frac{108}{116} = 0.931$$

Comparison of Two Proportions

Comparison of Two Proportions

- Checking normality assumptions:

Comparison of Two Proportions

- Checking normality assumptions:
  - $n_A p_A = 48 > 5$

Comparison of Two Proportions

- Checking normality assumptions:
  - $n_A p_A = 48 > 5$
  - $n_A(1 - p_A) = 6 > 5$

Comparison of Two Proportions

- Checking normality assumptions:
  - $n_A p_A = 48 > 5$
  - $n_A(1 - p_A) = 6 > 5$
  - $n_B p_B = 60 > 5$

Comparison of Two Proportions

- Checking normality assumptions:
  - $n_A p_A = 48 > 5$
  - $n_A(1 - p_A) = 6 > 5$
  - $n_B p_B = 60 > 5$
  - $n_B(1 - p_B) = 2 < 5 \implies$ proceed with caution

Comparison of Two Proportions

Comparison of Two Proportions

- Calculate z-statistic:

Comparison of Two Proportions

$H_0: p_A - p_B = 0$

- Calculate z-statistic:

$$z = \frac{(\hat{p}_A - \hat{p}_B) - \cancel{(p_A - p_B)}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} = \frac{0.889 - 0.963}{\sqrt{0.931(0.069)\left(\frac{1}{54} + \frac{1}{62}\right)}}$$

$\hat{p}$: 0.931

$= -1.675$

$p = 2 \cdot Pr(Z < -1.675) = \underline{0.09}$

- Conclusion:

$\alpha = 0.01$

Comparison of Two Proportions

Comparison of Two Proportions

- We can also calculate a confidence interval for the difference of two proportions

Comparison of Two Proportions

- We can also calculate a confidence interval for the difference of two proportions
- As in the one-sample case, the standard error is not the same for the confidence interval and hypothesis test

Comparison of Two Proportions

- We can also calculate a confidence interval for the difference of two proportions

- As in the one-sample case, the standard error is not the same for the confidence interval and hypothesis test

- For a two-sided confidence interval, we are $(1-\alpha)\cdot 100\,\%$ confident that the interval $\left(\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$ contains the true population difference, $p_1 - p_2$

$\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

(Wald)

Miettinen - Nurminen (MN)

Comparison of Two Proportions

Comparison of Two Proportions

- Continuing with our dominant hand by group example, we can construct a two-sided 95% confidence interval for $p_A - p_B$ as follows:

Comparison of Two Proportions

- Continuing with our dominant hand by group example, we can construct a two-sided 95% confidence interval for $p_A - p_B$ as follows:

$$\left(\underset{\hat{p}_A}{0.889} - \underset{\hat{p}_B}{0.968}\right) \pm 1.96\sqrt{\frac{0.889 \cdot 0.111}{54} + \frac{0.968 \cdot 0.032}{62}}$$

$$= (-\underline{0.173}, \underline{0.016})$$

$$\sqrt{\frac{p_A(1-p_A)}{n_1} + \frac{p_B(1-p_B)}{n_2}}$$

Comparison of Two Proportions

- Continuing with our dominant hand by group example, we can construct a two-sided 95% confidence interval for $p_A - p_B$ as follows:

$$(0.889 - 0.968) \pm 1.96\sqrt{\frac{0.889 \cdot 0.111}{54} + \frac{0.968 \cdot 0.032}{62}}$$

$$= (-0.173, 0.016)$$

- We are 95% confident that the interval (-0.173, 0.016) contains the true difference in the proportion of members of Group A and Group B who are right-handed

Comparison of Two Proportions

Comparison of Two Proportions

- One tail, lower bound:

Comparison of Two Proportions

- One tail, lower bound:

$$\left(\hat{p}_1 - \hat{p}_2 - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ 1\right)$$

Comparison of Two Proportions

- One tail, lower bound:

$$\left(\hat{p}_1 - \hat{p}_2 - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ 1\right)$$

- One tail, upper bound:

Comparison of Two Proportions

- One tail, lower bound:

$$\left(\hat{p}_1 - \hat{p}_2 - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},\ 1\right)$$

- One tail, upper bound:

$$\left(-1,\ \hat{p}_1 - \hat{p}_2 + z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

DSCC 462 Midpoint Survey

$$z_{1-\alpha} = \frac{\bar{x} - \mu}{\sqrt{\sigma/n}}$$

$$z_{1-\alpha} \cdot \sqrt{\sigma/n} + \mu_0 = \bar{x}$$

$$z_\beta = \frac{\bar{x} - \mu_1}{\sqrt{\sigma/n}}$$

$$z_\beta \cdot \sqrt{\frac{\sigma}{n}} = \mu_0 + z_{1-\alpha} \cdot \sqrt{\sigma/n} - \mu_1$$

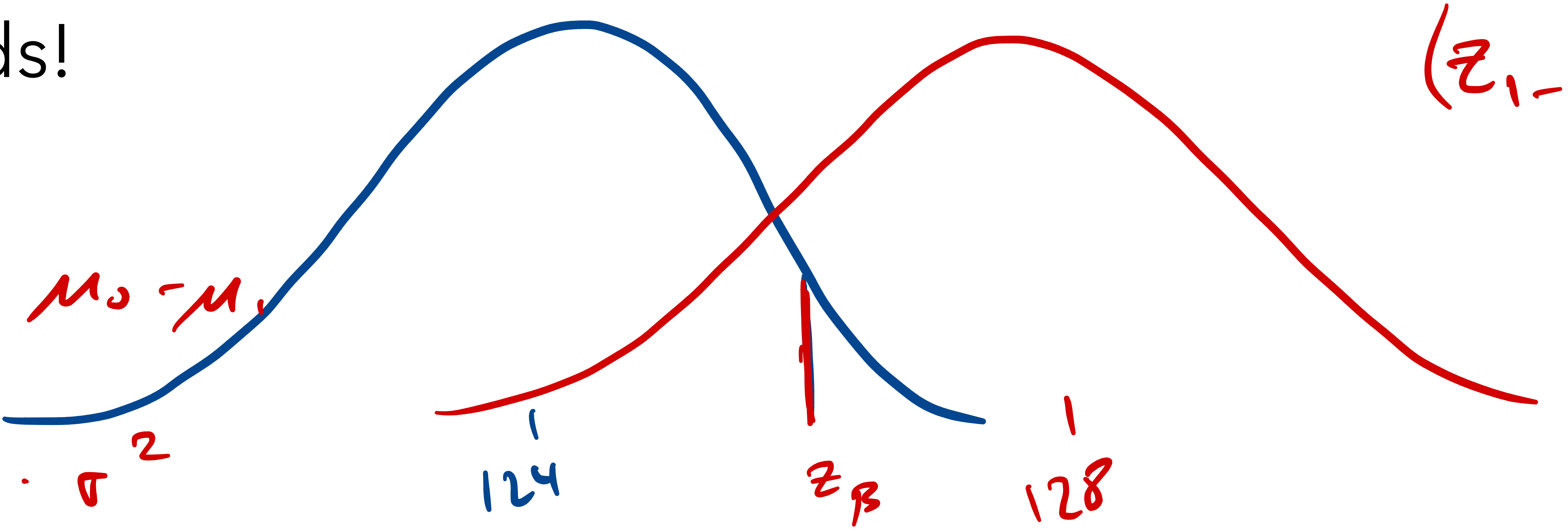
$$(z_\beta - z_{1-\alpha}) \sqrt{\frac{\sigma}{n}} = \mu_0 - \mu_1$$

- https://forms.gle/Zt3Qzrb7S7UXFXY28
- If you fill it out: +2.5% on midterm
- If at least 90% of the class fills it out: +2.5% on each person's midterm
  - Tell your friends!

$$(z_{1-\alpha} + z_{1-\beta})$$

$$(z_\beta + z_\alpha) \sqrt{\frac{\sigma}{n}} = \mu_0 - \mu_1$$

$$n = \left(\frac{z_\alpha + z_\beta}{\mu_0 - \mu_1}\right)^2 \cdot \sigma^2$$

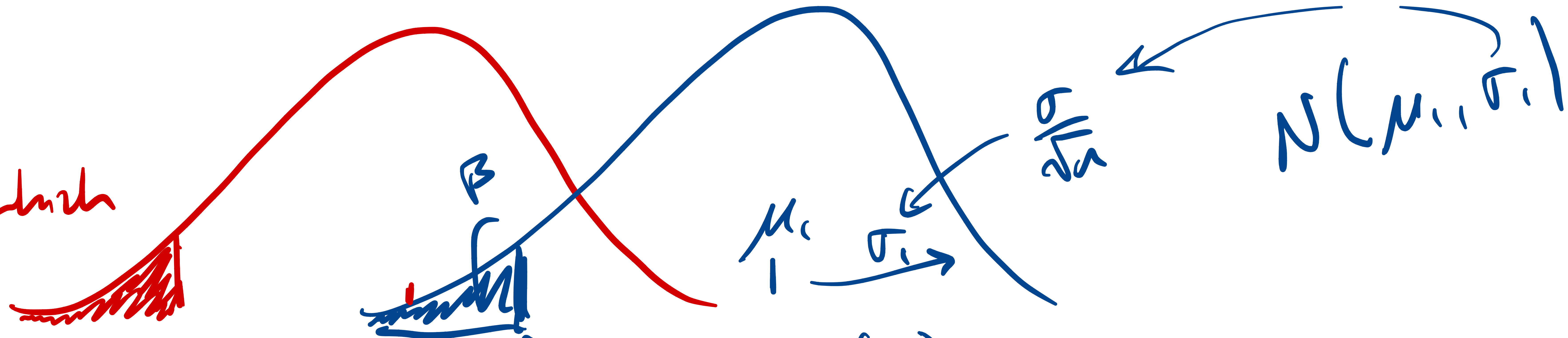


DSCC 462 Midpoint Survey

$\bar{X} \leftarrow \alpha$

plot(mu1, some function of mu1, ...)

- https://forms.gle/Zt3Qzrb7S7UXFXY28
- If you fill it out: +2.5% on midterm
- If at least 90% of the class fills it out: +2.5% on each person's midterm
  - Tell your friends!



z_α is value for cutoff

$Pr(Z < z_\alpha) = \alpha$

z_α = qnorm(α)

z_β = qnorm(β)

DSCC 462 Midpoint Survey

- https://forms.gle/Zt3Qzrb7S7UXFXY28
- If you fill it out: +2.5% on midterm
- If at least 90% of the class fills it out: +2.5% on each person's midterm
  - Tell your friends!