Review of Module 9

Vocabulary Review

**Point estimate:** this is another word for the observed statistic or summary statistic

- The point estimate is the center of the confidence interval
- This week we are looking at two categorical variables, so the point estimate is the sample difference in proportion
- Since this value is calculated from the sample the notation used is $\hat{p}_1 - \hat{p}_2$
    - Make sure to use informative subscripts – what is group 1 and group 2

**Example Activity 16:**

- $\hat{p}_{helmet} - \hat{p}_{no\ helmet} = \frac{96}{752} - \frac{480}{2810} = 0.128 - 0.171 = -0.043$

**Interpreting the sample difference in proportion:**

***Include:***

- Summary measure (difference in proportion)
    - What numerical value are we calculating?
    - This is dependent on the type of variable(s) in our study
    - Give the value of the statistic and the order of subtraction
- Context
    - Observational units/cases – what or whom are we collecting data on
    - Variable of interest – what success are we focusing on in the research question.
    - ***For comparison studies we also need the explanatory variable groups.***

**Example Activity 16:**

- **The proportion of skiers and snowboarders involved in accidents with a head injury for those who wore a helmet is 0.043 lower than those who did not wear helmets.**

**Parameter of Interest:** What information do we want to know about the population?:

- The parameter of interest is used in the hypotheses statements, in conclusions, and in many interpretations!

***Include:***

- Reference of the population (true, long-run, population, all)
    - Clearly refer to the population
- Summary measure (difference in proportion)
    - What numerical value are we calculating?
    - This is dependent on the type of variable(s) in our study
- Context
    - Observational units/cases – what or whom are we collecting data on
    - Variable of interest – what success are we focusing on in the research question.
    - ***For comparison studies we also need the explanatory variable groups.***

Review of Module 9

**Example Activity 16:**

- ***The difference in population proportion of skiers and snowboarders involved in an accidents with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet).***

**Hypothesis Test (test of significance/inference):** test to show evidence based on the sample statistic against the null hypothesis

**Hypotheses:**

- Null Hypothesis: This is the known claim that we are trying to disprove; may be based on random chance
    - For comparison studies we assume there is no difference between groups
    - $H_0$: $\pi_1 = \pi_2$ or $H_0$: $\pi_1 - \pi_2 = 0$
- Alternative: this is the claim we are testing that is based on the research question
    - $H_a$: $\pi_1 \begin{Bmatrix} > \\ \neq \\ < \end{Bmatrix} \pi_2$ or $H_a$: $\pi_1 - \pi_2 \begin{Bmatrix} > \\ \neq \\ < \end{Bmatrix} 0$
        - The direction of the alternative (the sign) is determined by the research question
- **Example Activity 16:** Is there evidence that safety helmet use is associated with a reduced risk of head injury for skiers and snowboarders?
    - *$H_0$:* ***There is no difference in population proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet).***
        - *Note: we are assuming that the proportion of skiers and snowboarders with a head injury for those who wore helmets is equal to the proportion that did not wear helmets*
    - $H_o: \pi_{helmet} - \pi_{no\ helmet} = 0\ or\ H_o: \pi_{helmet} = \pi_{no\ helmet}$
    - *$H_A$:* ***The population proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet is less than for those who did not wear a helmet***
        - *The direction of the alternative is less than because the research question asks for evidence that wearing helmets is associated with a reduction in the proportion of head injuries!*
    - $H_A: \pi_{helmet} - \pi_{no\ helmet} < 0\ or\ H_A: \pi_{helmet} < \pi_{no\ helmet}$

Review of Module 9

**Theory-based Hypothesis Test for a difference in proportions:**
Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to follow an approximate normal distribution:

- Independence: The data are independent within and between the two groups. **(*Remember*: This also must be true to use simulation methods!)**

*Note: Theory-based methods will give similar p-values to simulation methods IF there is a large enough sample size. To check that we have a large enough sample size we need to check the success-failure condition.

- **Large enough sample size**: The success-failure condition holds for each group. Need to have at least 10 successes and 10 failures in each group. Equivalently, we check that all cells in the table have at least 10 observations.

**Example Skiing Helmet Study:**

- There are 96 head injuries (successes) in the helmet group – there are more than 10 successes (head injuries) in Group 1 (helmet group)
- There are 636 no head injuries (failures) in the helmet group – there are more than 10 failures (no head injuries) in Group 1 (helmet group)
- There are 480 head injuries (successes) in the no helmet group – there are more than 10 successes (head injuries) in Group 2 (no helmet group)
- There are 2330 no head injuries (failures) in the no helmet group - there are more than 10 failures (no head injuries) in Group 2 (no helmet group)

- Since there are more than 10 successes and 10 failures in each group, the theory-based p-value will be approximately the same as the simulation p-value and the theory-based confidence interval will be similar to the simulation confidence interval.

Review of Module 9

**Calculation of the Standardized Statistic:**

- For a theory-based hypothesis test first we will calculate the standardized statistic. The standardized statistic is compared to the standard normal distribution to find the p-value.

$$z = \frac{\hat{p}_1 - \hat{p}_2 - 0}{SE_0(\hat{p}_1 - \hat{p}_2)}$$

- For a hypothesis test, we always assume the null hypothesis is true. The best estimate for the standard error (the sample variability) is calculated using $\hat{p}_{pool}$ (the overall proportion of success)

$$\hat{p}_{pool} = \frac{total\ successes}{n_1 + n_2}$$

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{\hat{p}_{pool} \times (1 - \hat{p}_{pool}) \times \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

**Example Activity 16:**

total # successes

$$\hat{p}_{pool} = \frac{576}{3562} = 0.162$$

$\to n_1 + n_2$

assuming Ho is true

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{752} + \frac{1}{2810}\right)} = 0.015$$

$$z = \frac{0.128 - 0.171 - 0}{0.015} = \frac{-0.043}{0.015} = -2.867$$

**Interpretation of the standardized statistic:** *measures the number of standard errors the sample statistic is from the null value*

*Include:*

- Statistic (summary measure and value in context)
- Units of measure are standard errors *(value of z)*
- Above/below *the null value (give the value)*

**Example Skiing Activity 16:**

*

- *The sample difference in proportion of skiers/snowboarders with a head injury who wore a helmet and did not wear a helmet (helmet – no helmet) of -0.043 is 2.867 standard errors below the null value of zero.*

Review of Module 9

**To find the p-value:**

**Example Activity 16:**

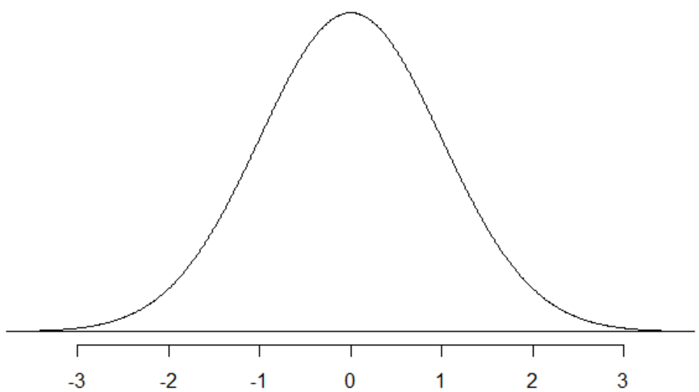
The pnorm function in Rstudio is used to find the p-value:

pnorm(-2.867, # Enter value of standardized statistic

m=0, s=1, # Using the standard normal mean = 0, sd = 1

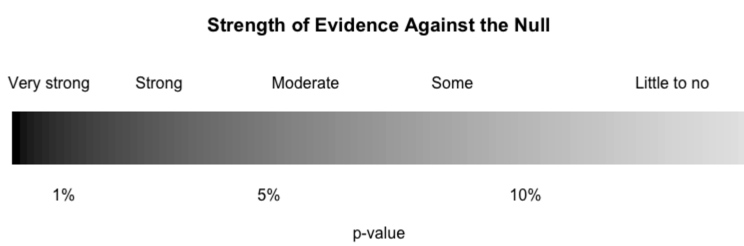
lower.tail=TRUE)

**Standard Normal Distribution:**




**Strength of Evidence:** How much evidence does the p-value provide **against** the null?

- Use the guidelines for the strength of evidence




- The smaller the p-value the MORE evidence there is against the null hypothesis

There are FOUR things we ask about the p-value

**Evaluation of a p-value:**

- How much evidence does the p-value provide AGAINST the null hypothesis?
- **Example Activity 16:** The theoretical p-value for this study was found to be 0.002.
    - ***There is very strong evidence against the null hypothesis that there is no difference in true proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet).***

Review of Module 9

**Interpretation of a p-value:**

- What the p-value measures: the probability of observing the sample statistic or more extreme if the null hypothesis is true (Don't forget the context!)
- Include in the interpretation:
    - Statement about probability (in x% of simulated samples, in x out of 1000 simulated samples, with a probability of x%)
    - Statistic in context (give the value and in words what the statistic represents)
    - more extreme (direction of the alternative)
    - If the null hypothesis is true in context (give the null value and in words what the null represents)
        - Note: context only needs to be included in either the statistic OR the null

- **Example Activity 16:** The theoretical p-value for this study was found to be 0.002
    - ***We would observe a sample difference in proportion of -0.043 or less with a probability of 0.002 if we assume there is no difference in true proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet).***
    ***OR***
    - ***If there is no difference in true proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet), we would observe a standardized difference in sample proportion of -2.867 or less with a probability of 0.2%.***
    **OR**
    - ***There is a 0.2% chance, we would observe a sample difference in proportion of -0.45 or less if we assume there is no difference in true proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet and those who did not wear a helmet (helmet – no helmet).***

**Conclusion:** Answers the research question. Write a conclusion as the amount of evidence in support of the alternative.

- **Example Activity 16:** Write a conclusion in context of the study.
    - ***There is very strong evidence*** *that the* ***population proportion of skiers and snowboarders involved in an accident with a head injury for those that wore a helmet is less than for those who did not wear a helmet***

**Decision:** compare the p-value to the set significance level

- If the p-value is less than the significance level (α), the decision will be to reject the null hypothesis
- If the p-value is greater than the significance level (α), the decision will be to fail to reject the null hypothesis
- **Example Activity 16**:
    - Since we have a very small p-value less than the significance level of 0.1 (considering a confidence level of 90%) we will reject the null hypothesis.
        - This means that the null value of zero will NOT be in the 90% confidence interval

Review of Module 9

**Theory-based Confidence interval:** Interval notation: (lower value, upper value)

**Calculation of the Confidence Interval:**

- To calculate the confidence interval we add and subtract the margin of error to the point estimate.

$$\boldsymbol{point\ estimate \pm margin\ of\ error}$$

- The margin of error is the z* multiplier times the standard error

$$ME = z^* \times SE(\hat{p}_1 - \hat{p}_2)$$

- Since we do NOT assume the null hypothesis is true when finding the confidence interval, we use the values for $\hat{p}_1$ and $\hat{p}_2$ to calculate the standard error.
- Remember that the notation $SE(\hat{p}_1 - \hat{p}_2)$ is function notation that reads the standard error of the difference in sample proportions. We are NOT multiplying the standard error times the difference in proportions.

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \times (1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (1-\hat{p}_2)}{n_2}}$$

not assuming $H_0$ is true

- The z* multiplier – determines the width of the confidence interval and is NOT the same value as Z (the standardized statistic).

qnorm(percentile)

  - *Dependent on the level of confidence*
    - 90% - 1.645
    - 95% - 1.96
    - 99% - 2.576

**Example Activity 16:** $\hat{p}_1 = \frac{96}{752} = 0.128, \hat{p}_2 = \frac{480}{2810} = 0.171, n_1 = 752, n_2 = 2810$

- To find a 90% confidence interval we will use a multiplier of z* = 1.645

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.128\times(1-0.128)}{752} + \frac{0.171\times(1-0.171)}{2810}} = 0.014$$

$$ME = 1.645 \times 0.014 = 0.023$$

$$0.128 - 0.171 \pm 0.023$$

$$(-0.043 - 0.023, -0.043 + 0.023)$$

$$(-0.066,\ -0.02)$$

Review of Module 9

**Interpretation of a confidence interval:**

- Include in the interpretation:
    - How confident you are (90%, 95%, 99%)
    - Parameter of interest *in context*
        - *Population word (true, long-run, population)*
        - *Summary measure (difference in proportion)*
        - *Observational units*
        - *Variable of interest*
    - Calculated Interval
    - Order of subtraction

**Example Activity 16:** ***The true proportion of skiers/snowboarders involved in an accident with a head injury for those who wear helmets is between 0.02 and 0.066 lower than for those who do not wear helmets, with 90% confidence.***

Or

***We are 90% confident that the difference in true proportion of skiers/snowboarders involved in an accident with a head injury for those who wear helmets and those who do not (helmet – no helmet) is between -0.066 and -0.02.***

OR

***We are 90% confident that the true proportion of skiers/snowboarders involved in an accident with a head injury for those who wear helmets minus the true proportion who do not wear helmets is between -0.066 and -0.02.***

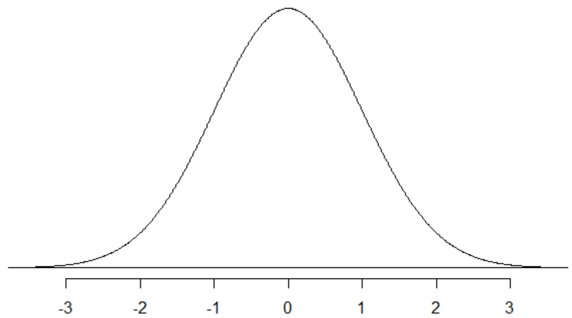
**Impacts on the p-value:**

- Increasing the sample size will decrease the standard error (sample to sample variability), thereby increasing the value of the standardized statistic (Z). A larger standardized statistic will result in a smaller p-value (more evidence against the null hypothesis).

From **Skiing Helmet Study**:

$$SE_0(\hat{p}_1 - \hat{p}_2) = \sqrt{0.162 \times (1 - 0.162) \times \left(\frac{1}{1056} + \frac{1}{3944}\right)} = 0.013$$

$$z = \frac{0.128 - 0.171 - 0}{0.015} = \frac{-0.043}{0.013} = -3.31$$




- Using a two-sided test rather than a one-sided test will double the size of the p-value (less evidence against the null hypothesis).

**Impacts of sample size on the width of the confidence interval**

- Increasing sample size will decrease the standard error (sample to sample variability). This decreases the margin of error so the width of the confidence interval will be narrower