Module 7: Additions to Significance Testing

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

UNT UNIVERSITY OF NORTH TEXAS
Discover the power of ideas.

Introduction to Statistics for the Social Sciences

RSS
Research and Statistical Support

The RSS short courses

The Research and Statistical Support (RSS) office at the University of North Texas hosts a number of "Short Courses". A list of them is available at:

`http://www.unt.edu/rss/Instructional.htm`

Outline

Outline

1. Effect Size
2. Statistical Power

Outline

Outline

1. Effect Size
2. Statistical Power
3. Practical Significance
4. Summary of Module 7

From a score to a distribution of scores

Effect Size

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**

RSS
Research and Statistical Support

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**
     - Note sometimes there is no need to standardize (units of the scale have inherent meaning).

RSS Research and Statistical Support

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**
     - Note sometimes there is no need to standardize (units of the scale have inherent meaning).
  2. Measures of Variance Accounted for.

RSS
Research and Statistical Support

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**
     - Note sometimes there is no need to standardize (units of the scale have inherent meaning).
  2. Measures of Variance Accounted for.
     - Amount of explained variance vs. total variance.

RSS
Research and Statistical Support

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**
     - Note sometimes there is no need to standardize (units of the scale have inherent meaning).
  2. Measures of Variance Accounted for.
     - Amount of explained variance vs. total variance.
     - Such as R^2 and R^2_{adj}

From a score to a distribution of scores

Effect Size

- Keep in mind, there are two types of effect sizes:
  1. Measures of Difference
     - Allows comparison across samples and variables with differing variance.
     - **Equivalent to Z-scores**
     - Note sometimes there is no need to standardize (units of the scale have inherent meaning).
  2. Measures of Variance Accounted for.
     - Amount of explained variance vs. total variance.
     - Such as R^2 and R^2_{adj}
- For now, we will deal with *Measures of Difference*.

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.
  - Effect size is the **magnitude of experimental effect**.

RSS
Research and Statistical Support

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.
  - Effect size is the **magnitude of experimental effect**.
- Effect size:

RSS
Research and Statistical Support

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.
  - Effect size is the **magnitude of experimental effect**.
- Effect size:
  - Increases with greater differences between means,

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.
  - Effect size is the **magnitude of experimental effect**.
- Effect size:
  - Increases with greater differences between means,
  - Decreases with greater standard deviations in the population but,

RSS
Research and Statistical Support

Effect Size

- Effect size is a standardized measure of difference (lack of overlap) between populations.
  - Effect size is the **magnitude of experimental effect**.
- Effect size:
  - Increases with greater differences between means,
  - Decreases with greater standard deviations in the population but,
  - Is not affected by sample size.

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

RSS
Research and Statistical Support

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Notice within this formula, we are *removing* the influence of population standard deviation.

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Notice within this formula, we are *removing* the influence of population standard deviation.
  - This produces the *standardized* effect size.

RSS
Research and Statistical Support

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Notice within this formula, we are *removing* the influence of population standard deviation.
    - This produces the *standardized* effect size.
    - Raw score effect size (i.e., without dividing by σ) is virtually useless.

RSS
Research and Statistical Support

Calculating Effect Size

- There are many measures of effect size, for now we will be using Cohen's *d*.

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

- Notice within this formula, we are *removing* the influence of population standard deviation.
  - This produces the *standardized* effect size.
  - Raw score effect size (i.e., without dividing by σ) is virtually useless.
- The standardization allows us to compare effect sizes obtained from different research studies.

RSS
Research and Statistical Support

Remember Scooby...?

Remember Scooby...?

- Population 1: Dogs on cartoons.

Remember Scooby...?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Pluto, and Goofy ($\overline{X} = 133.67$).

Remember Scooby...?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Pluto, and Goofy ($\overline{X} = 133.67$).
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

RSS Research and Statistical Support

Remember Scooby...?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Pluto, and Goofy ($\overline{X} = 133.67$).
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

$$d = \frac{133.67 - 100}{15} = \frac{33.67}{15} = 2.24$$

RSS
Research and Statistical Support

Remember Scooby...?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Pluto, and Goofy ($\overline{X} = 133.67$).
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

$$d = \frac{133.67 - 100}{15} = \frac{33.67}{15} = 2.24$$

- Please note the effect size is greater than 1. This may not always be the case, but the value of Cohen's *d* can be greater than 1.

RSS
Research and Statistical Support

Remember Scooby...part 2?

[1] From the Module 6 handout

Remember Scooby...part 2?

- Population 1: Dogs on cartoons.

[1] From the Module 6 handout

Remember Scooby...part 2?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Underdog, and Scrappy ($\overline{X} = 107.67$)[1].

[1] From the Module 6 handout

RSS
Research and Statistical Support

Remember Scooby...part 2?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Underdog, and Scrappy ($\overline{X} = 107.67$)[1].
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

[1] From the Module 6 handout

RSS
Research and Statistical Support

Remember Scooby...part 2?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Underdog, and Scrappy ($\overline{X} = 107.67$)[1].
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

$$d = \frac{107.67 - 100}{15} = \frac{7.67}{15} = 0.51$$

[1] From the Module 6 handout

Remember Scooby...part 2?

- Population 1: Dogs on cartoons.
  - Sample: Scooby, Underdog, and Scrappy ($\overline{X} = 107.67$)[1].
- Population 2: Dogs not on cartoons ($\mu = 100, \sigma = 15$)

$$d = \frac{107.67 - 100}{15} = \frac{7.67}{15} = 0.51$$

- The effect size is not greater than 1, but this may still be considered a large effect size.

[1] From the Module 6 handout

RSS
Research and Statistical Support

Interpreting Cohen's *d*

- One way: Effect size conventions suggested from Cohen.

Interpreting Cohen's *d*

- One way: Effect size conventions suggested from Cohen.
  - Small = 0.20
  - Medium = 0.50
  - Large = 0.80 and greater

Interpreting Cohen's *d*

- One way: Effect size conventions suggested from Cohen.
  - Small = 0.20
  - Medium = 0.50
  - Large = 0.80 and greater
- A better way: Rational judgment based on a thorough understanding of the phenomena and the previous literature.

RSS
Research and Statistical Support

Interpreting Cohen's *d*

- One way: Effect size conventions suggested from Cohen.
  - Small = 0.20
  - Medium = 0.50
  - Large = 0.80 and greater
- A better way: Rational judgment based on a thorough understanding of the phenomena and the previous literature.
  - It may be that an effect size of 0.90 is small based on previous findings where $d = 1.20$ to 1.90.

RSS Research and Statistical Support

Statistical Power

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.
  - The ability to detect a significant effect is one is present.

RSS
Research and Statistical Support

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.
  - The ability to detect a significant effect is one is present.
- Important to note: '*if the null hypothesis is false*'

RSS
Research and Statistical Support

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.
  - The ability to detect a significant effect is one is present.
- Important to note: '*if the null hypothesis is false*'
  - If you get a significant result when the null is true, then you have committed a Type I error.

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.
  - The ability to detect a significant effect is one is present.
- Important to note: '*if the null hypothesis is false*'
  - If you get a significant result when the null is true, then you have committed a Type I error.
- General equation for power:

Statistical Power

- Definition: The probability that the study will produce a statistically significant result if the null hypothesis is false.
  - The ability to detect a significant effect is one is present.
- Important to note: '*if the null hypothesis is false*'
  - If you get a significant result when the null is true, then you have committed a Type I error.
- General equation for power:
  - Power = 1 - beta
  - Power = 1 - β

RSS Research and Statistical Support

Two kinds of Power analysis

- A priori Power

Two kinds of Power analysis

- A priori Power
    - Used when planning a study

RSS
Research and Statistical Support

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.
- Post hoc Power

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.
- Post hoc Power
  - Used when evaluating a study.

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.
- Post hoc Power
  - Used when evaluating a study.
  - What chance did a study have of finding significant results?

RSS
Research and Statistical Support

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.
- Post hoc Power
  - Used when evaluating a study.
  - What chance did a study have of finding significant results?
  - Not really useful. If you do the power analysis and conduct your study accordingly, then you did what you could.

RSS
Research and Statistical Support

Two kinds of Power analysis

- A priori Power
  - Used when planning a study
  - Used to determine the sample size necessary to achieve a specified power level.
- Post hoc Power
  - Used when evaluating a study.
  - What chance did a study have of finding significant results?
  - Not really useful. If you do the power analysis and conduct your study accordingly, then you did what you could.
    - To say afterward: “I would have found significance but did not have enough power or enough participants is not going to impress anyone”.

RSS
Research and Statistical Support

A priori Power

Can use all the following to calculate how many subjects / participants we need for our study.

A priori Power

Can use all the following to calculate how many subjects / participants we need for our study.

- Decide an acceptable level of power.

A priori Power

Can use all the following to calculate how many subjects / participants we need for our study.

- Decide an acceptable level of power.
- Set the significance level (usually .05).

A priori Power

Can use all the following to calculate how many subjects / participants we need for our study.

- Decide an acceptable level of power.
- Set the significance level (usually .05).
- Figure out the desirable or expected effect size.

A priori Power

Can use all the following to calculate how many subjects / participants we need for our study.

- Decide an acceptable level of power.
- Set the significance level (usually .05).
- Figure out the desirable or expected effect size.
- Calculate *n* needed to achieve significance with those levels of power and effect size.

A priori Effect Size?

- Figure out an effect size before I conduct my study?

A priori Effect Size?

- Figure out an effect size before I conduct my study?
- Several ways to do this:

A priori Effect Size?

- Figure out an effect size before I conduct my study?
- Several ways to do this:
    - Base it on substantive knowledge.

RSS
Research and Statistical Support

A priori Effect Size?

- Figure out an effect size before I conduct my study?
- Several ways to do this:
  - Base it on substantive knowledge.
    - What you know about the situation and scale of measurement.

A priori Effect Size?

- Figure out an effect size before I conduct my study?
- Several ways to do this:
  - Base it on substantive knowledge.
    - What you know about the situation and scale of measurement.
  - Base it on previous literature / research.

RSS
Research and Statistical Support

A priori Effect Size?

- Figure out an effect size before I conduct my study?
- Several ways to do this:
  - Base it on substantive knowledge.
    - What you know about the situation and scale of measurement.
  - Base it on previous literature / research.
  - Use Cohen's conventions (not recommended).

RSS
Research and Statistical Support

An acceptable level of power?

Why not set power at .99?

An acceptable level of power?

Why not set power at .99?

- Practicalities.

An acceptable level of power?

Why not set power at .99?

- Practicalities.
  - Cost of increasing power (usually done by increasing sample size) can be high.

An acceptable level of power?

Why not set power at .99?

- Practicalities.
  - Cost of increasing power (usually done by increasing sample size) can be high.
- Increasing power decreases the Type II error rate (good), but also increases Type I error rate (bad).

RSS
Research and Statistical Support

An acceptable level of power?

Why not set power at .99?

- Practicalities.
  - Cost of increasing power (usually done by increasing sample size) can be high.
- Increasing power decreases the Type II error rate (good), but also increases Type I error rate (bad).
- Power has a range of 0 to 1 (it is a probability); with a higher number indicating greater power.

RSS Research and Statistical Support

Influences on Power

Table 1: Influences on Power

Feature of Study	High Power	Low Power
Effect Size	larger	smaller
Sample Size	larger	smaller
Sig. Level	high (.10)	low (.001)
Tailed Test	1-tailed	2-tailed
Type of analysis	varies	varies

Carrying out the calculation of Power

The easiest way.

Carrying out the calculation of Power

The easiest way.

- When you have to implement power calculations, you can use specialist programs.

Carrying out the calculation of Power

The easiest way.

- When you have to implement power calculations, you can use specialist programs.
  - Many websites offer free applications to conduct power analysis.

Carrying out the calculation of Power

The easiest way.

- When you have to implement power calculations, you can use specialist programs.
  - Many websites offer free applications to conduct power analysis.
- G-power:

  `http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3`

RSS
Research and Statistical Support

Calculating Power

The more difficult way.

Calculating Power

The more difficult way.

- First, convert your critical value ($Z_{crit} = 1.64$) into a raw score.

Calculating Power

The more difficult way.

- First, convert your critical value ($Z_{crit} = 1.64$) into a raw score.

$$(Z_{crit}) * (\sigma_M) + \mu = (1.64) * (8.67) + 100 = 114.22$$

- This defines the point on your **Null Distribution** where the rejection region begins.

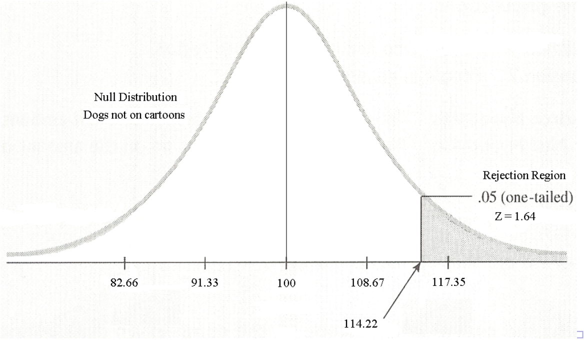
Calculating Power

The more difficult way.

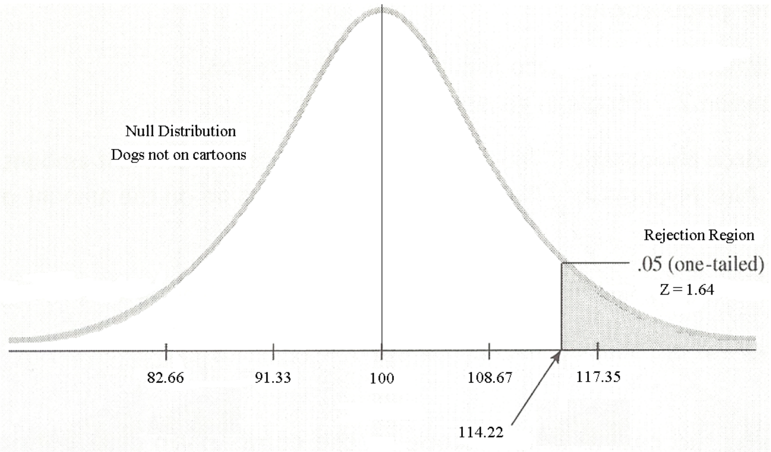
- First, convert your critical value ($Z_{crit} = 1.64$) into a raw score.

$$(Z_{crit}) * (\sigma_M) + \mu = (1.64) * (8.67) + 100 = 114.22$$

- This defines the point on your **Null Distribution** where the rejection region begins.



Null Distribution



Calculating Power continued

- Next, calculate the Z-score for a raw score of 114.22 **on the Alternative Distribution.**

Calculating Power continued

- Next, calculate the Z-score for a raw score of 114.22 **on the Alternative Distribution.**

$$\frac{X-\overline{X}}{\sigma_M} = \frac{114.22-133.67}{8.67} = \frac{-19.45}{8.67} = -2.243$$

RSS
Research and Statistical Support

Calculating Power continued

- Next, calculate the Z-score for a raw score of 114.22 **on the Alternative Distribution.**

$$\frac{X-\overline{X}}{\sigma_M} = \frac{114.22-133.67}{8.67} = \frac{-19.45}{8.67} = -2.243$$

- Finally, look in the Z-score table to identify beta and power.

  http://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm

RSS
Research and Statistical Support

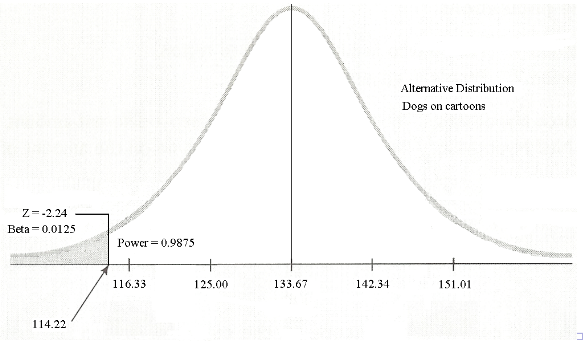
Calculating Power continued

- Next, calculate the Z-score for a raw score of 114.22 **on the Alternative Distribution.**

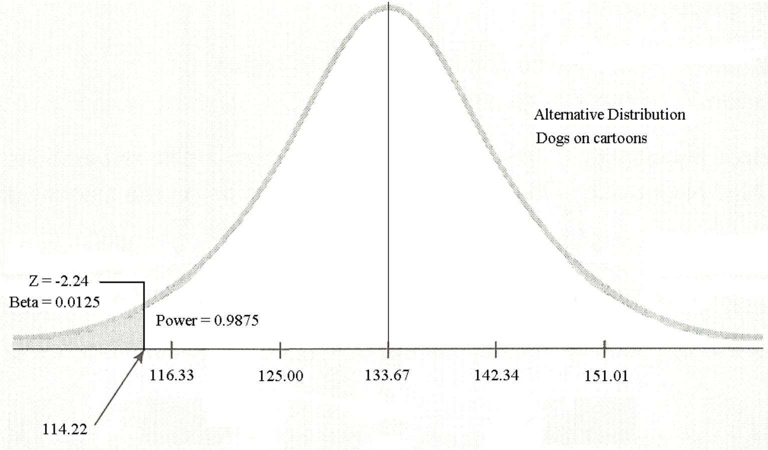
$$\frac{X - \overline{X}}{\sigma_M} = \frac{114.22 - 133.67}{8.67} = \frac{-19.45}{8.67} = -2.243$$

- Finally, look in the Z-score table to identify beta and power.

http://www.sjsu.edu/faculty/gerstman/EpiInfo/z-table.htm



Alternative Distribution



KSS
Research and Statistical Support

Statistical vs. Practical Significance

- Statistical significance is determined by a dichotomous decision based on the *p* value.

Research and Statistical Support

Statistical vs. Practical Significance

- Statistical significance is determined by a dichotomous decision based on the *p* value.
  - If $p < .05$; then reject the null hypothesis.

RSS
Research and Statistical Support

Statistical vs. Practical Significance

- Statistical significance is determined by a dichotomous decision based on the *p* value.
  - If $p < .05$; then reject the null hypothesis.
  - If $p < .05$; then fail to reject the null hypothesis.

RSS
Research and Statistical Support

Statistical vs. Practical Significance

- Statistical significance is determined by a dichotomous decision based on the *p* value.
  - If $p < .05$; then reject the null hypothesis.
  - If $p < .05$; then fail to reject the null hypothesis.
- Practical significance has more to do with the effect size and meaningfulness of the results in practical terms.

RSS
Research and Statistical Support

Statistical vs. Practical Significance

- Statistical significance is determined by a dichotomous decision based on the *p* value.
  - If $p < .05$; then reject the null hypothesis.
  - If $p < .05$; then fail to reject the null hypothesis.
- Practical significance has more to do with the effect size and meaningfulness of the results in practical terms.
  - If $p = .001$, reject the null, but if $d = .12$; then your results are not likely to be influential or useful.

RSS
Research and Statistical Support

More on Practical Significance

- Keep in mind, anything will be significant with a large enough sample!!!

More on Practical Significance

- Keep in mind, anything will be significant with a large enough sample!!!
- However, the results may not be meaningful or useful.

More on Practical Significance

- Keep in mind, anything will be significant with a large enough sample!!!
- However, the results may not be meaningful or useful.
- Remember Scooby and Friends...

More on Practical Significance

- Keep in mind, anything will be significant with a large enough sample!!!
- However, the results may not be meaningful or useful.
- Remember Scooby and Friends...
  - Example 1: $n = 3, p < .00007, d = 2.24$; reject the null because $p < .05$
  - Example 2 (from Module 6 handout): $n = 3, p = .1894, d = .051$; fail to reject the null because $p > .05$

More on Practical Significance

- Keep in mind, anything will be significant with a large enough sample!!!
- However, the results may not be meaningful or useful.
- Remember Scooby and Friends...
  - Example 1: $n = 3, p < .00007, d = 2.24$; reject the null because $p < .05$
  - Example 2 (from Module 6 handout): $n = 3, p = .1894, d = .051$; fail to reject the null because $p > .05$
- Hypothetically, you could get a result like this: $n = 25000, p = .000001, d = 0.000001$

RSS
Research and Statistical Support

Concluding Thoughts

- Always report as much information as you can; meaning:

Concluding Thoughts

- Always report as much information as you can; meaning:
  - The calculated sample statistic
  - The sample size
  - The critical level (.05)
  - The obtained p value ($p < .00007$
  - The effect size ($d = 2.24$)
  - The power
    - If it was used a-priori to calculate sample size and the appropriate sample size was obtained (G-power application).

RSS
Research and Statistical Support

Concluding Thoughts

- Always report as much information as you can; meaning:
  - The calculated sample statistic
  - The sample size
  - The critical level (.05)
  - The obtained p value ($p < .00007$
  - The effect size ($d = 2.24$)
  - The power
    - If it was used a-priori to calculate sample size and the appropriate sample size was obtained (G-power application).
- Remember, p values are not related to effect sizes.

RSS
Research and Statistical Support

Concluding Thoughts

- Always report as much information as you can; meaning:
  - The calculated sample statistic
  - The sample size
  - The critical level (.05)
  - The obtained *p* value ($p < .00007$
  - The effect size ($d = 2.24$)
  - The power
    - If it was used a-priori to calculate sample size and the appropriate sample size was obtained (G-power application).
- Remember, *p* values are not related to effect sizes.
- Use a-priori power and effect size to determine the minimum sample size (and gather that amount of data) prior to collecting the data.

RSS
Research and Statistical Support

Concluding Thoughts

- Always report as much information as you can; meaning:
  - The calculated sample statistic
  - The sample size
  - The critical level (.05)
  - The obtained *p* value ($p < .00007$
  - The effect size ($d = 2.24$)
  - The power
    - If it was used a-priori to calculate sample size and the appropriate sample size was obtained (G-power application).
- Remember, *p* values are not related to effect sizes.
- Use a-priori power and effect size to determine the minimum sample size (and gather that amount of data) prior to collecting the data.
  - Post hoc power is virtually meaningless.

Summary of Module 7

Module 7 covered the following topics:

- Cohen's *d* Effect Size

Summary of Module 7

Module 7 covered the following topics:

- Cohen's *d* Effect Size
- Statistical Power

Summary of Module 7

Module 7 covered the following topics:

- Cohen's *d* Effect Size
- Statistical Power
- Practical Significance

RSS
Research and Statistical Support

Summary of Module 7

Module 7 covered the following topics:

- Cohen's *d* Effect Size
- Statistical Power
- Practical Significance

Many of these topics will be revisited consistently in future modules.

RSS
Research and Statistical Support

This concludes Module 7

Next time Module 8.

- Next time we'll begin covering Introduction to *t* tests.
- Until next time; have a nice day.

These slides initially created on: October 8, 2010
These slides last updated on: October 12, 2010

- The bottom date shown is the date this Adobe.pdf file was created; LaTeX[2] has a command for automatically inserting the date of a document's creation.

[2] This document was created in LaTeX using the Beamer package