

# Lecture 8: Statistical Power

Dewayne E Perry

ENS 623

Perry@ece.utexas.edu

# Errors Revisited

## → One reality

↪ H0 (Null Hypothesis) is True

↪ H1 (Alternative Hypothesis) is False

↪ There is *no* relationship, *no* difference, theory is *wrong*

## → We accept H0, reject H1

↪ Match reality

↪ Confidence level:  $1-\alpha$  (eg, .95)

- The odds of saying there is no relationship or difference when in fact there is none
- The odds of correctly not confirming our theory
- Ie, 95 time out of 100 when there is no effect, we will say there is none.

## → Type I Error: we reject H0, accept H1

↪ Contradict reality - say there is a relationship when there is none

↪ Significance level:  $\alpha$  (eg, .05)

- The odds of saying there is a relationship or difference when there is none
- The odds of confirming our theory incorrectly
- 5 times out of 100, when there is no effect, we will say there is
- We should keep this small when we can't afford/risk wrongly concluding our treatment works

# Errors Revisited

## ➔ The other reality

↪  $H_0$  (Null) is False

↪  $H_1$  (Alternative) is True

↪ There *is* a relationship, *is* a difference, and our theory *is supported*

## ➔ Type II Error: we *accept* $H_0$ , *reject* $H_1$

↪ Contradict reality - say there is no relationship when there is one

↪  $\beta$  (eg, .20)

➤ The odds of saying there is no relationship or difference when in fact there is one

➤ The odds of not confirming out theory when it is true

➤ 20 times out 100, when there is an effect, we will say there isn't

## ➔ We *accept* $H_1$ , *reject* $H_0$

↪ Match reality

↪ Power:  $1-\beta$  (eg, .80)

➤ The odds of saying there is a relationship or difference when there is one

➤ The odds of confirming our theory correctly

➤ 80 times out 100 when there is an effect we will say there is

➤ We generally want this to be as large as possible

# Decreasing Errors

- ➔ Decrease Type I Error by setting a *more* stringent  $\alpha$ 
  - ↳ Eg, .01 instead of .05
  - ↳ Decreasing Type I increases the likelihood of Type II Error
- ➔ Decrease Type II Error by setting *less* stringent  $\alpha$ 
  - ↳ Eg, .10 instead of .05
- ➔ Seek a balance between the two
  - ↳ As Type I goes up, Type II goes down and vice versa

# Purpose of Power Analysis

## ➔ Planning of research

- ↳ Determine size of sample needed
- ↳ To reach a given  $\alpha$  level
- ↳ For any particular size of effect expected

## ➔ Evaluation of research completed

- ↳ Determine if failure to detect an effect at a given  $\alpha$  is primarily due to too small a sample

# Power

## ➔ Level of Power determined by

- ↪ Statistic used to determine the level of significance
- ↪ The level of  $\alpha$  selected
- ↪ The size of the sample
- ↪ The size of the effect

## ➔ Increasing Power can be achieved by

- ↪ Raising the level of significance required,
- ↪ Reducing the standard deviation,
- ↪ Increasing the magnitude of the effect by using strong treatments, and
- ↪ Increasing the size of the sample

# Example

- ➔ X compares OO programming against standard programming randomly assigning 40 programmers to use OO and 40 as the control group
  - ↳ The OO treatment programs have significantly fewer bugs
  - ↳ Using  $t$  test (comparing means),  $t(78) = 2.21, p < .05$
- ➔ Y is skeptical and replicates X's work
  - ↳ Assigns 10 programmers to each
  - ↳ Results:  $t(18) = 1.06, p > .30$
  - ↳ Y claims X results unrepeatable
- ➔ Misleading conclusions
  - ↳ Y's results in the same direction as X's
  - ↳ Y's effect size same as X's ( $1/2\sigma = 2t / \sqrt{df}$ )
  - ↳ Y's sample size too small: X's power = .6, Y's power = .2

## Effect Size (ES)

- ➔ Effect Size: *standardized measure of the change in the dependent variable as a result of the independent variable*
- ➔ Standardization of effect size is done in the simplest case by dividing the change in the dependent measure by the standard deviation of the control group
- ➔ If  $ES=1$ , the experimental and control results differ by 1 standard deviation

# Effect Size

- ➔ Effect Sizes are usually less than 1
- ➔ Cohen 1988 argues
  - ↳ Small effect size = 0.2
  - ↳ Medium effect size = 0.5
  - ↳ Large effect size = 0.8
- ➔ Enables us to compare the effects in different studies of the same phenomena
- ➔ Enables us to combine results from different studies in meta-analyses

# Example

## ➔ Comparison:

↳ Treatment: 8 designers, design method X

↳ Control: 8 designers, std design method Y

## ➔ Results in terms of errors:

↳ Treatment: 5 6 9 4 8 3 7 6

↳ Control: 10 11 10 9 9 8 9 14

## ➔ Means:

↳ Treatment: 6

↳ Control: 10

## ➔ Standard deviations

↳ Calculate sum of squared deviations from the mean via shortcut formula:

$$\sum x^2 - (\sum x)^2 / n$$

# Example

## ➔ Treatment:

↪ Squares: 25, 36, 81, 16, 64, 9, 49, 36

↪ Sum = 48, sum of squares = 316

↪  $316 - 2304/8 = 316 - 288 = 28$

↪ Std dev is  $\sigma = \sqrt{(28/7)} = \sqrt{4} = 2$

## ➔ Control:

↪ Squares: 100, 121, 100, 81, 81, 64, 81, 196

↪ Sum = 80, sum of squares = 824

↪  $824 - 6400/8 = 824 - 800 = 24$

↪ Std dev is  $\sigma = \sqrt{(24/7)} = \sqrt{3.53} = 1.85$

## ➔ Effect size $d = \text{mean 1} - \text{mean 2} / \sigma$

↪  $(6 - 10) / 1.85 = 2.16$

↪ A very large effect (Cohen: 0.8 is a large effect)

# Power Tables

## ➔ Cohen 1969, 1977, 1988

- ↪ Comprehensive, elegant and useful discussion of power analysis in behavioral research
- ↪ Defines small, medium and large effects for 7 statistics from  $t$  to  $F$
- ↪ Tables provide sample sizes vs power and significance

# Neglect of Power

➔ Behavioral researcher faces a high risk of committing Type II errors

- ↪ For medium effect sizes and  $\alpha = .05$  the odds are better than 50:50 that the null hypothesis would not be rejected when its false
- ↪ Since Cohen's work, situation has gotten worse apparently
- ↪ Continue to work at low power
- ↪ Continue to rate Type I errors as more significant than Type II errors
- ↪ Almost completely lacking in SE empirical studies

# Neglect of Power

- ➔ Assessing relationship of Type I vs Type II errors
  - ↳ Use ratio  $\beta/\alpha$ 
    - Remember  $\beta$  is the likelihood we will make a Type II error,  $\alpha$  the likelihood of making a Type I error
  - ↳ Eg,  $\alpha = .05$  and power =  $.40$ ,
    - $\beta/\alpha = .6/.05 = 12$ , ie Type I errors are considered to be 12 times more serious than Type II
  - ↳ What would we need to do if we wanted  $\alpha = .05$  and power =  $.95$ ,  $\beta/\alpha = .05/.05 = 1$ 
    - ie, consider I & II equally serious