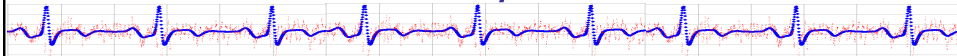


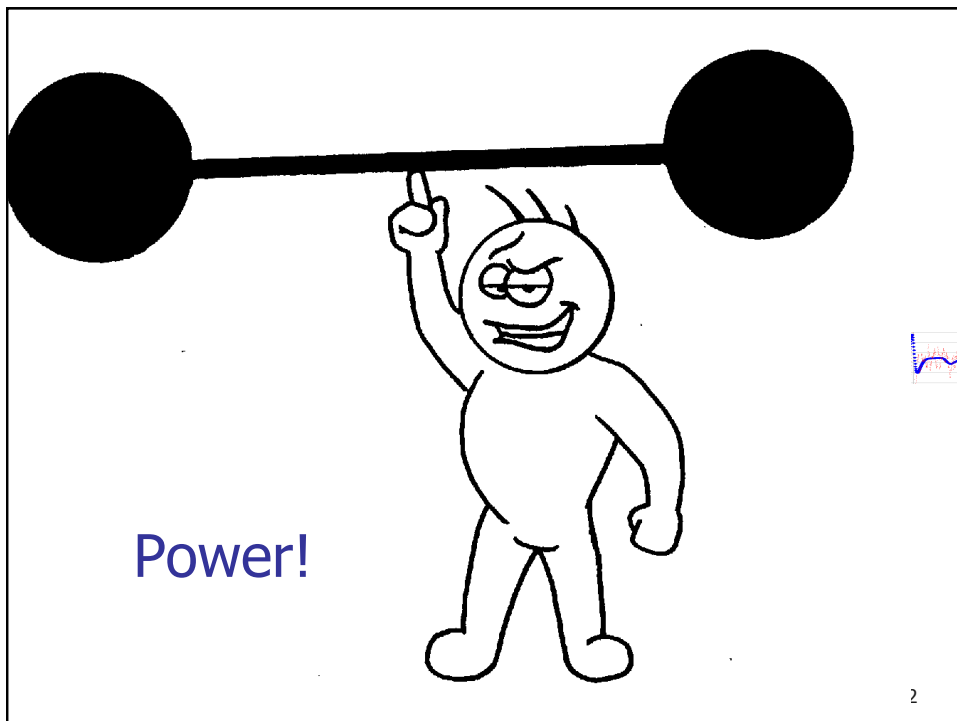
# Empirical Research Methods in Information Science

IS 4800/CS 6350



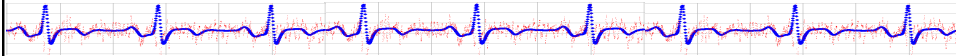
Lecture 15  
Power and Effect Size  
Midterm Prep

1



2

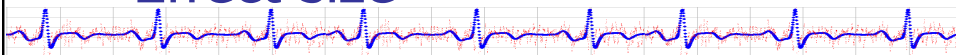
## Power



- The “power” of a statistical test is its ability to detect differences in data that are inconsistent with the null hypothesis.
  - $p(\text{rejecting } H_0|H_1)$
- aka – the ability to find a significant result, if your hypotheses are actually true.
- What is it called when this fails (ie, accepting  $H_0$  when  $H_1$  is true)?
- Why is this a bad situation?

3

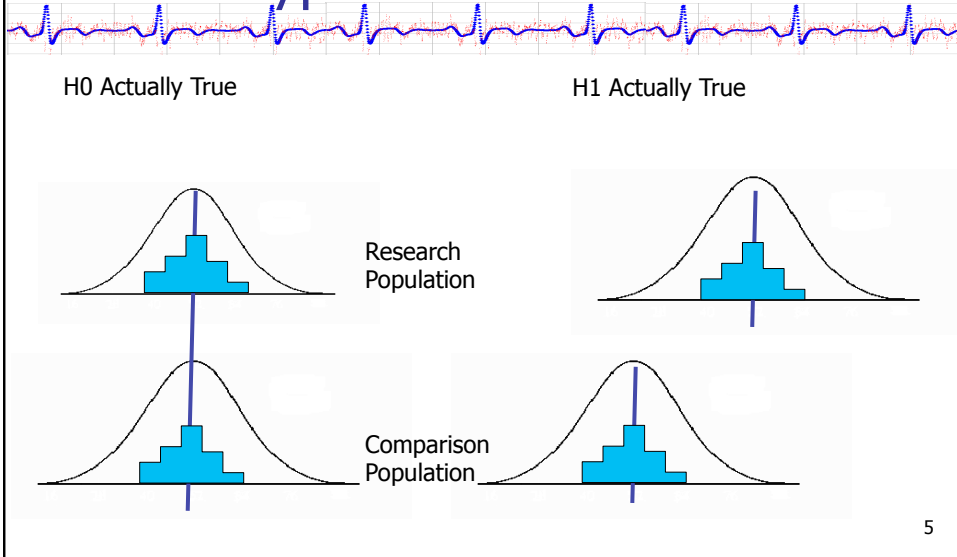
## Effect size



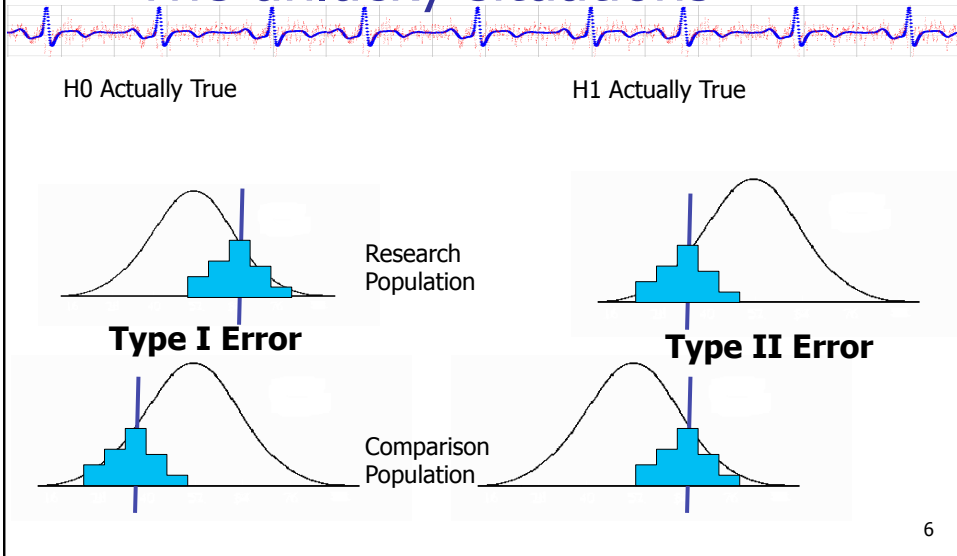
- The *amount* of measured difference between study conditions.
- The greater the effect size, the easier it is to show there is a significant difference in your study (ie, the greater the power).
- Effect size formula is different for each hypothesis test procedure.
- Tabulated standard values for “small”, “medium”, and “large” effect sizes.
- Only talk about effect size IF significance is established – but then DO present it in your results.

4

# The typical situation



# The unlucky situations



## Relationship between alpha, beta, and power.

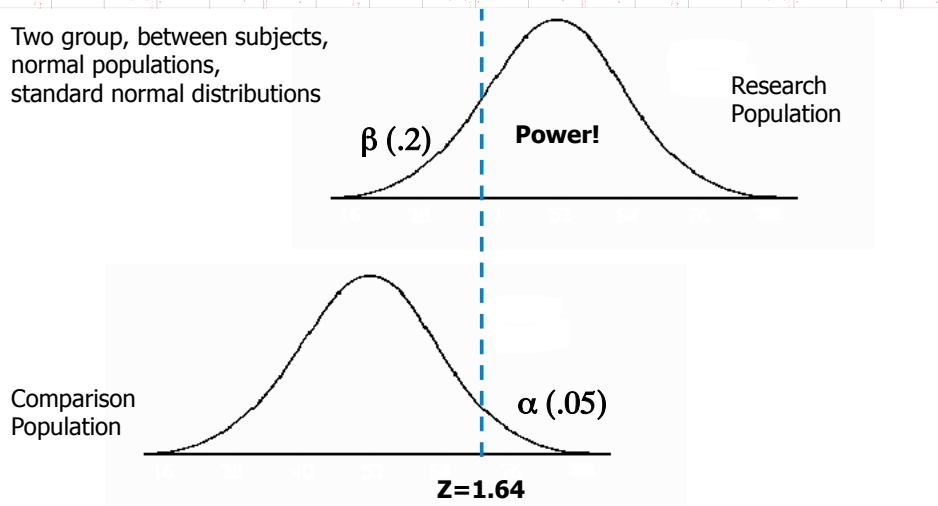
What is the probability of each of these situations occurring?

**"The Truth"**

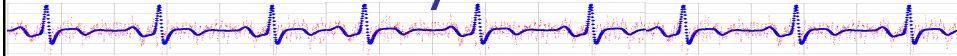
	H1 True	H1 False
Decide to Reject H0 & accept H1	Correct $p = \text{power}$	Type I err $p = \alpha$
Do not Reject H0 & do not accept H1	Type II err $p = \beta$	Correct $p = 1 - \alpha$

## Relationship between power and effect size

Two group, between subjects, normal populations, standard normal distributions



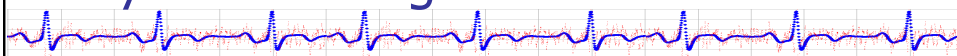
## Power Analysis



- Should determine number of subjects you need ahead of time by doing a 'power analysis'
- Standard procedure (part of your study plan):
  - Determine statistic you will use
  - Fix alpha and beta (1-power) (and number of tails if appropriate)
  - Estimate expected effect size from prior studies
  - Then: Determine number of subjects you need
- Note: Power
  - Increases with effect size
  - Increases with sample size
  - Decreases with decreasing alpha

10

Power analyses are different depending on the statistical test you are using...



First up: t-test for independent means.

11

## Effect Size

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

Parameters for population of individuals.  
(so, use SD-pooled for t-test of indep means)

Cohen:  
d~0.2 small  
d~0.5 medium  
d~0.8 large

12

## More Useful and Concise

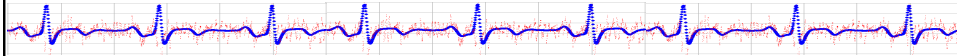
(for practical purposes use a power calculator)

**TABLE 8-5** Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

14

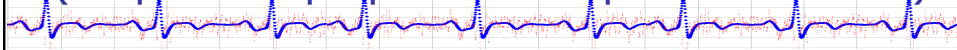
## Power Analysis Exercise



- Based on related research, we expect that there will be a medium effect size in our study of an LED sign in the Food Court affecting wait times.
- How many subjects do we need for a two-group, two-tailed test at  $\alpha=0.05$ , 80% power?

15

## More Useful and Concise (for practical purposes use a power calculator)



**TABLE 8-5** Approximate Number of Participants Needed in Each Group (Assuming Equal Sample Sizes) for 80% Power for the *t* Test for Independent Means, Testing Hypotheses at the .05 Significance Level

	Effect Size		
	Small (.20)	Medium (.50)	Large (.80)
One-tailed	310	50	20
Two-tailed	393	64	26

16

## But, I can't study 786 subjects!

- Increase effect size
  - Increase difference in population means (change manipulation)
  - Decrease population variance (better measures, control more extraneous vars)
  - Redesign study to collect many trials of measures per subject
- Relax criteria for Type I error
  - Increase  $\alpha$  threshold
  - Change from Two-tailed => one-tailed test
  - *Decreases credibility of your findings*
- Decrease power
  - *Decreases likelihood of getting a significant result*
- Use a different statistic
  - *If possible, maybe consult a statistician*
- Practically
  - usually, redesign experiment so that we have increased effect size or better measures for decreased variance
  - OR, call it a "pilot study"

17

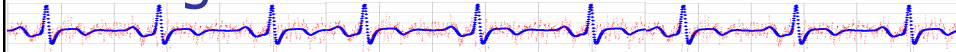
## Interpreting results: Significance & Effect Size

- Significance
  - Just indicates that it is likely there is a non-zero difference between populations.
  - Says nothing about how big the difference is.
- Effect Size
  - Only meaningful if result is significant.
  - Indicates how big the difference is (usually normalized to number of std-deviations)

18



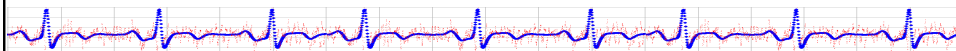
## Interpreting results: Significance & Effect Size



- Significant & small effect => ?
  - Real difference, but slight.
  - Probably not of practical importance.
- Significant & large effect => ?
  - Real difference, likely meaningful.
- Significant & small sample => ?
  - Significant & possibly important.
- Non-significant & small sample => ?
  - Inconclusive
- Non-significant & large sample => ?
  - Evidence there really is no difference

19

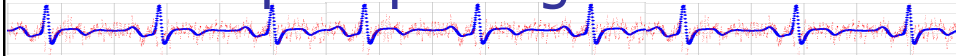
## Group Exercise



- Compute effect size for the study
- Characterize as small/medium/large
  
- You are now going to do a follow-up study using similar interventions and measures (ie assume same effect size)
- Do a power analysis to determine how many subjects you would need for a two-group between-subjects experiment with 80% power, alpha=0.05, two-tailed test.

20

## Power Analysis in R Use 'pwr' package



```
> require("pwr")      #every session  
> pwr.t.test(d=0.5,power=.8)
```

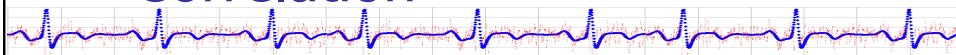
Two-sample t test power calculation

```
      n = 63.76561  
      d = 0.5  
sig.level = 0.05  
  power = 0.8  
alternative = two.sided
```

NOTE: n is number in \*each\* group

21

## Power & Effect Size for Correlation



- Effect size =  $|r|$
- Power, see table 11-7, pg 465 Aron
  - Usually, given
    - Expected effect size
    - Test criteria
      - Desired significance level (usually 0.05)
      - Desired power (usually 0.8)
      - Directionality of test

22

**Table 11-7** Approximate Power of Studies Using the Correlation Coefficient ( $r$ ) for Testing Hypotheses at the .05 Level of Significance

		Effect Size		
		Small ( $r = .10$ )	Medium ( $r = .30$ )	Large ( $r = .50$ )
<b>Two-tailed</b>				
Total $N$ :	10	.06	.13	.33
	20	.07	.25	.64
	30	.08	.37	.83
	40	.09	.48	.92
	50	.11	.57	.97
	100	.17	.86	*
<b>One-tailed</b>				
Total $N$ :	10	.08	.22	.46
	20	.11	.37	.75
	30	.13	.50	.90
	40	.15	.60	.96
	50	.17	.69	.98
	100	.26	.92	*

\*Power is nearly 1.

## Table 11-8, Aron

Approximate number of participants needed for 80% power for a study using the correlation coefficient ( $r$ ) for testing a hypothesis at the .05 significance level

Effect size		
Small ( $r=0.1$ )	Medium ( $r=0.3$ )	Large ( $r=0.5$ )
<b>783</b>	<b>85</b>	<b>28</b>

## Effect Size & Power for $\chi^2$ test for independence

- Completely different formulas than for Pearson  $r$  or  $t$ -test.
- Dependent on  $df$ .
- For  $2 \times 2$ , effect size = "phi"

$$\sqrt{\frac{\chi^2}{N}}$$

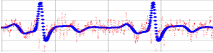
25

## Effect Size & Power for $\chi^2$

**Table 13-10** Approximate Total Number of Participants Needed for 80% Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

Total $df$	Effect Size		
	Small	Medium	Large
1	785	87	26
2	964	107	39
3	1,090	121	44
4	1,194	133	48

26

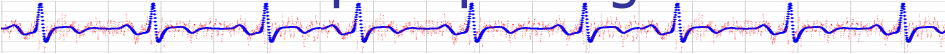


**Table 13-9** Approximate Power for the Chi-Square Test for Independence for Testing Hypotheses at the .05 Significance Level

Total <i>df</i>	Total <i>N</i>	Effect Size		
		Small	Medium	Large
1	25	.08	.32	.70
	50	.11	.56	.94
	100	.17	.85	*
	200	.29	.99	*
2	25	.07	.25	.60
	50	.09	.46	.90
	100	.13	.77	*
	200	.23	.97	*
3	25	.07	.21	.54
	50	.08	.40	.86
	100	.12	.71	.99
	200	.19	.96	*
4	25	.06	.19	.50
	50	.08	.36	.82
	100	.11	.66	.99
	200	.17	.94	*

\*Nearly 1.

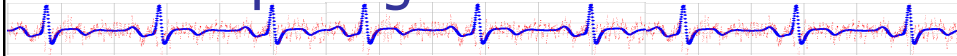
## Also in 'pwr' package



- `pwr.t.test`      t-tests (one sample, 2 sample, paired)
- `pwr.chisq.test`      chi-square test
- `pwr.r.test`      correlation

28

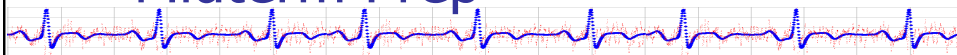
## Computing effect size



- Some thoughtless authors do not include means & stddevs (per group) in their article...
- Package `compute.es` contains a variety of methods for computing effect size given other info (e.g., t score, N1, N2)
- Morale: Always include means & stddevs
- Better: Report effect sizes yourself!

29

## Midterm Prep



- Every test has a long question of the following form:

**Study Proposal (25%).** Sketch a study proposal to prove which search engine is best (Google or Bing) for individuals who have never used a computer before. Your primary outcome measure is learnability (from Nielsen).

30