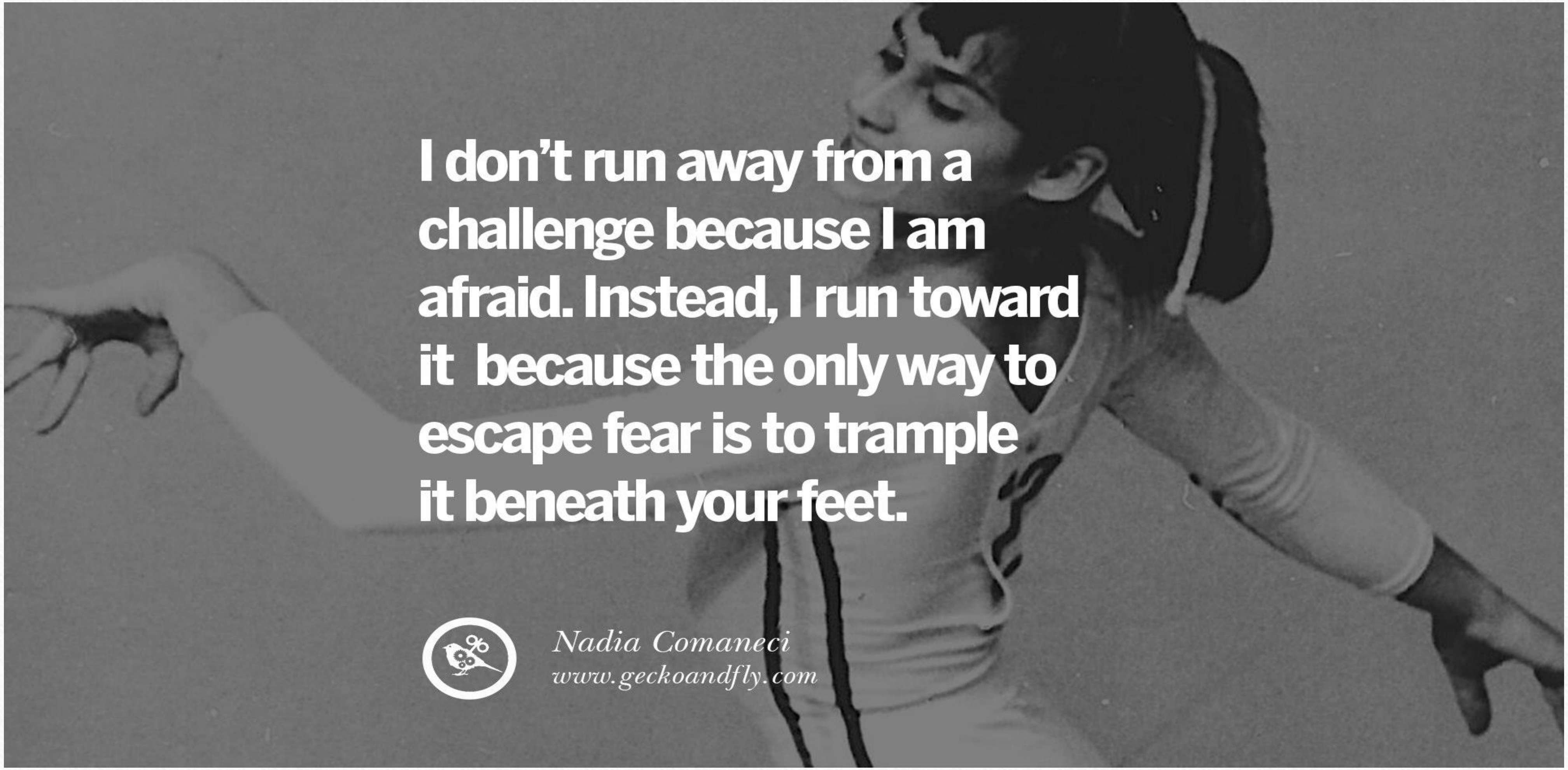
# Introduction to Biostatistics II



I don’t run away from a challenge because I am afraid. Instead, I run toward it  because the only way to escape fear is to trample it beneath your feet.

*Nadia Comaneci*
*www.geckoandfly.com*

# Why study statistics?

Because you’re 3x more likely to meet someone on Match.com than not using Match.com

# Why study statistics?

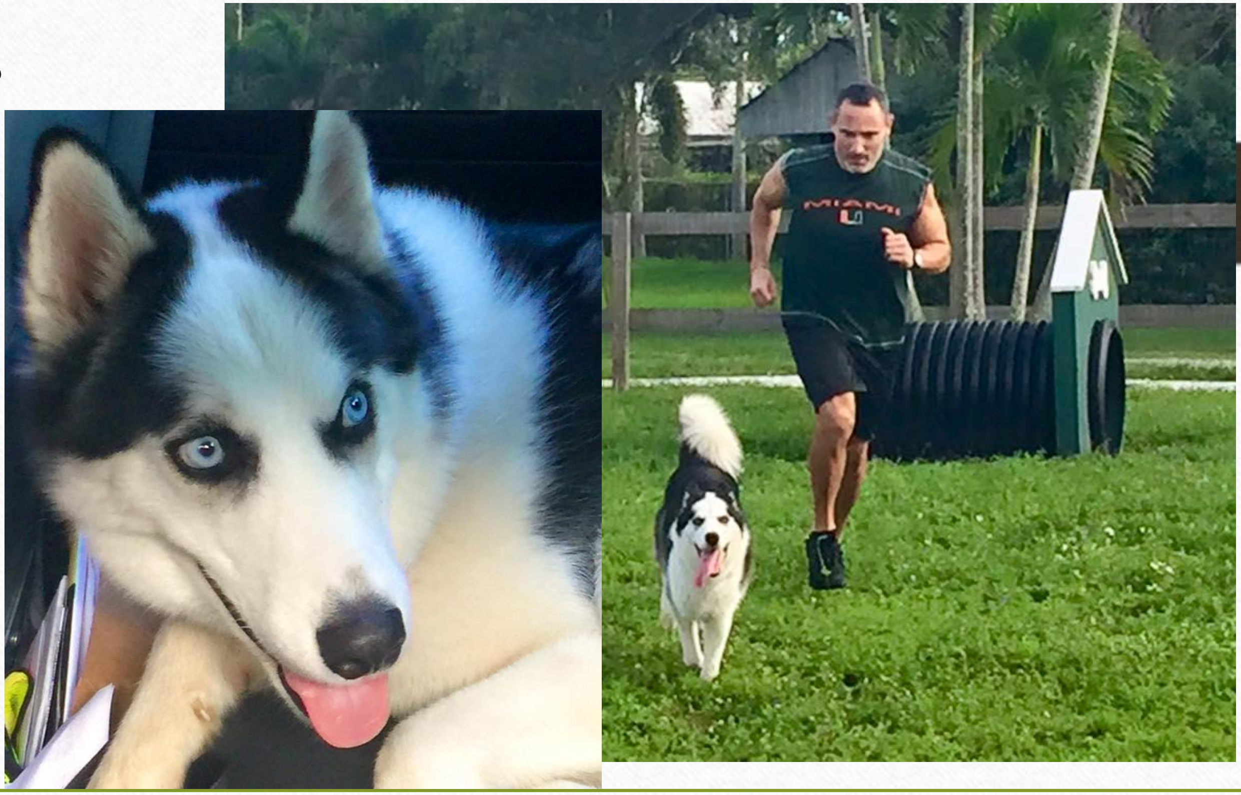
Because 4 out of 5 dentists recommend Colgate.

# Agenda

- Random "chance" factors in research
- *p* value
- Type I and Type II errors
- Beyond the *p* value ~ effect size
- ANOVA

# The Influence of Random / Chance Effects

- Random factors influence our lives
  - Overhearing a job opportunity
  - Meeting someone at a coffee shop
  - Finding a stray dog on the streets



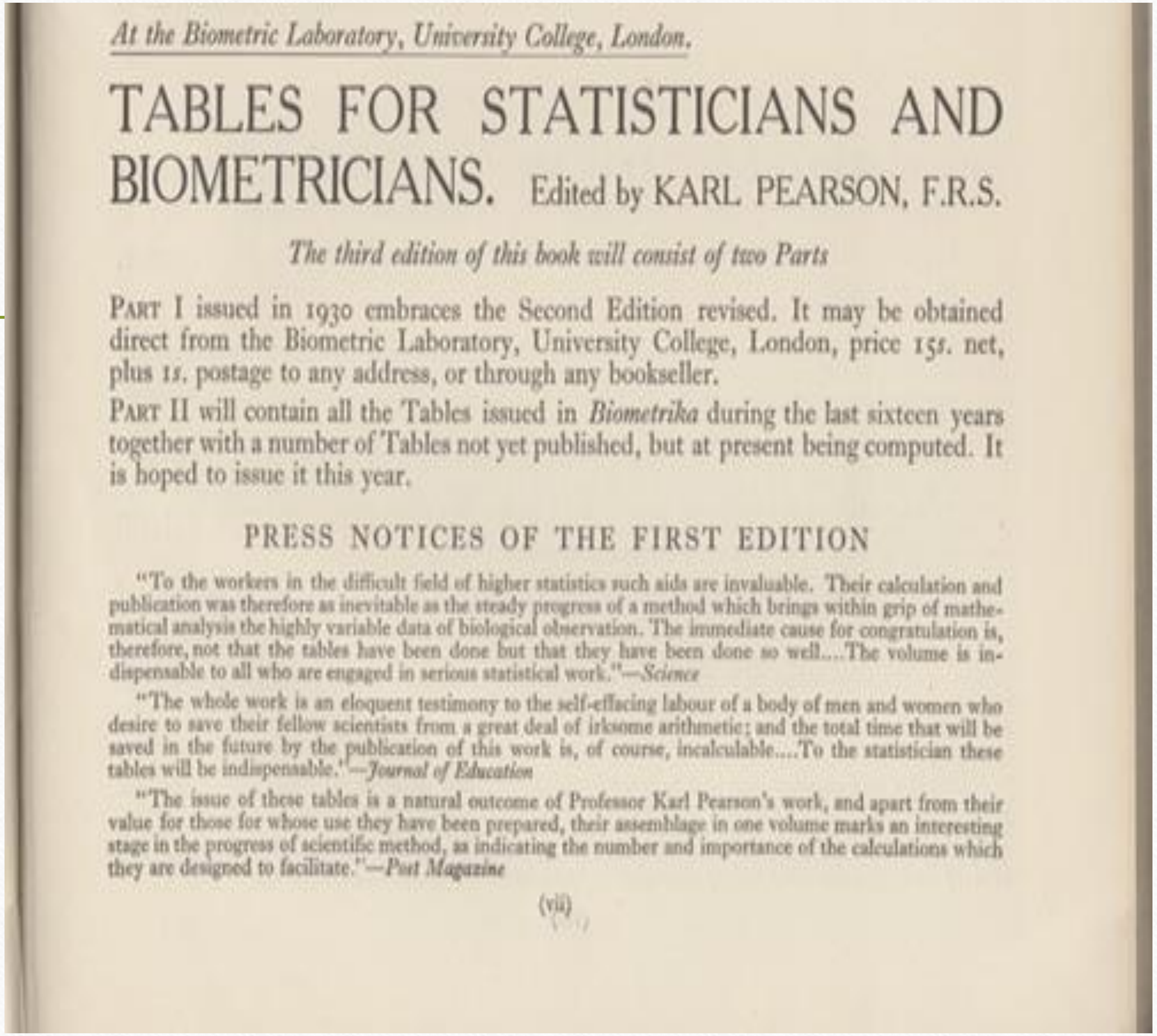
# The Influence of Random / Chance Effects

- Similar to life, randomness has an influence on the outcome of research
- Unlike life where randomness can have a positive impact, it's typically an unwelcome factor in research
    - Randomness may cause the researcher to believe the Intervention worked, when in reality it did not

# The P VALUE IN PLAIN ENGLISH

# *p*-value

- We need to verify randomness or chance factors are not overly influencing our results



*At the Biometric Laboratory, University College, London.*

## TABLES FOR STATISTICIANS AND BIOMETRICIANS. Edited by KARL PEARSON, F.R.S.

*The third edition of this book will consist of two Parts*

PART I issued in 1930 embraces the Second Edition revised. It may be obtained direct from the Biometric Laboratory, University College, London, price 15*s.* net, plus 1*s.* postage to any address, or through any bookseller.

PART II will contain all the Tables issued in *Biometrika* during the last sixteen years together with a number of Tables not yet published, but at present being computed. It is hoped to issue it this year.

### PRESS NOTICES OF THE FIRST EDITION

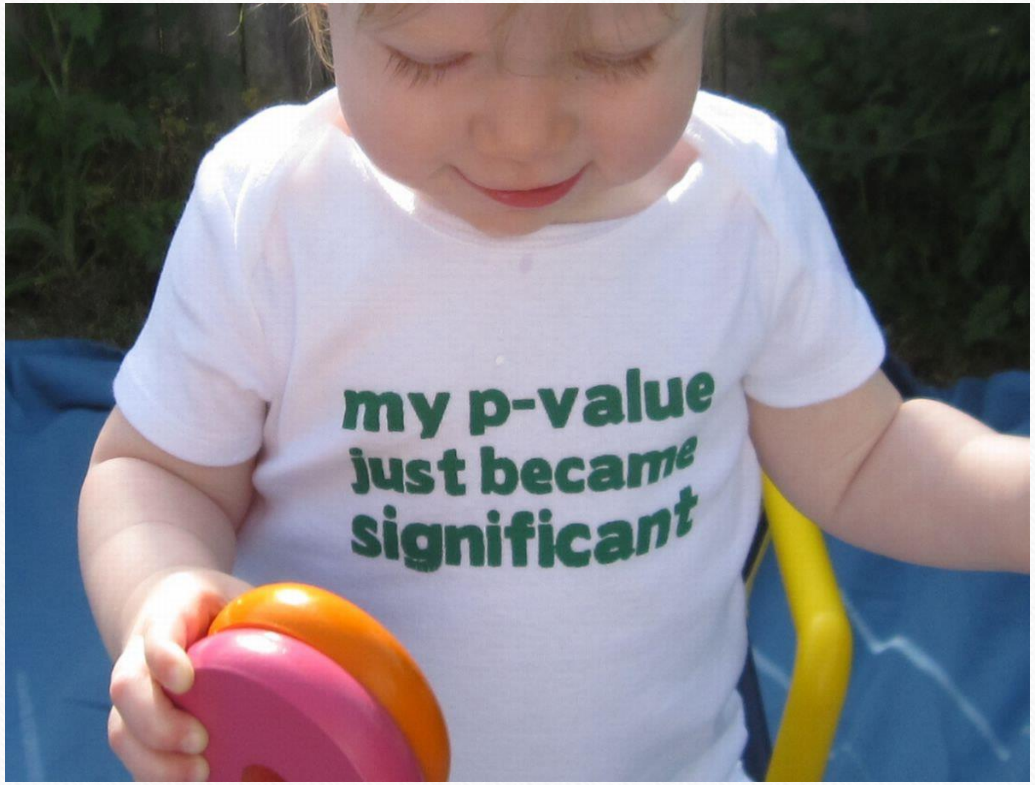
"To the workers in the difficult field of higher statistics such aids are invaluable. Their calculation and publication was therefore as inevitable as the steady progress of a method which brings within grip of mathematical analysis the highly variable data of biological observation. The immediate cause for congratulation is, therefore, not that the tables have been done but that they have been done so well....The volume is indispensable to all who are engaged in serious statistical work."—*Science*

"The whole work is an eloquent testimony to the self-effacing labour of a body of men and women who desire to save their fellow scientists from a great deal of irksome arithmetic; and the total time that will be saved in the future by the publication of this work is, of course, incalculable....To the statistician these tables will be indispensable."—*Journal of Education*

"The issue of these tables is a natural outcome of Professor Karl Pearson's work, and apart from their value for those for whose use they have been prepared, their assemblage in one volume marks an interesting stage in the progress of scientific method, as indicating the number and importance of the calculations which they are designed to facilitate."—*Post Magazine*

# What is a *p* value?

- What does a *p* value of less than .05 mean?



# $p < .05$ = There is less than a 5% probability, the findings occurred by chance

$p = \alpha$ (alpha)

# What is *p*-value

- Statistical significance:
  - *p* = .80 = There is a 80% probability the findings occurred by chance.
    - Evidence that chance factors (i.e. randomness) influenced the results
  - *p* < .05 = There is less than a 5% probability the findings occurred by chance

- A researcher conducted a study comparing the effect of an intervention vs placebo on reducing body weight, and found 5 lb reduction among the intervention group with *p*=0.01.

- Another researcher conducted a similar study comparing the effect of the same intervention vs the same placebo on reducing body weight, and found the same 5 lb reduction with the intervention group but could not claim that the intervention was effective because *p*=0.35.

*Why the different results?*

# What impacts a p-value?

- Variation in data
  - Larger variation can result in larger p-value
  - Sample bias
  - Measurement error
  - *and what else?*

# What impacts a *p-value*?

- Sample size!
    - Larger sample size can make p-value smaller!
        - Even a small, clinically meaningless effect can become significant if you keep enrolling patients indefinitely

# Why do we need a *p-value*?
# Validates a hypothesis

- Two possible outcomes:
  - There is a significant difference between Phenelzine (Nardil) vs. a placebo on treatment of depression. *Reject the null ($p$<.05)*
  - There is no significant difference between Phenelzine (Nardil) vs. a placebo on treatment of depression. . *Fail to reject the null ($p$>.05)*

*When making this inferential judgement, two possible (types) of errors can occur*

# Why do we need a *p-value*?

- There is a significant difference between Phenelzine (Nardil) vs. a placebo on treatment of depression. *Reject the null ($p$<.05)*
  - **This could be an accurate finding or an inaccurate finding – *false positive***

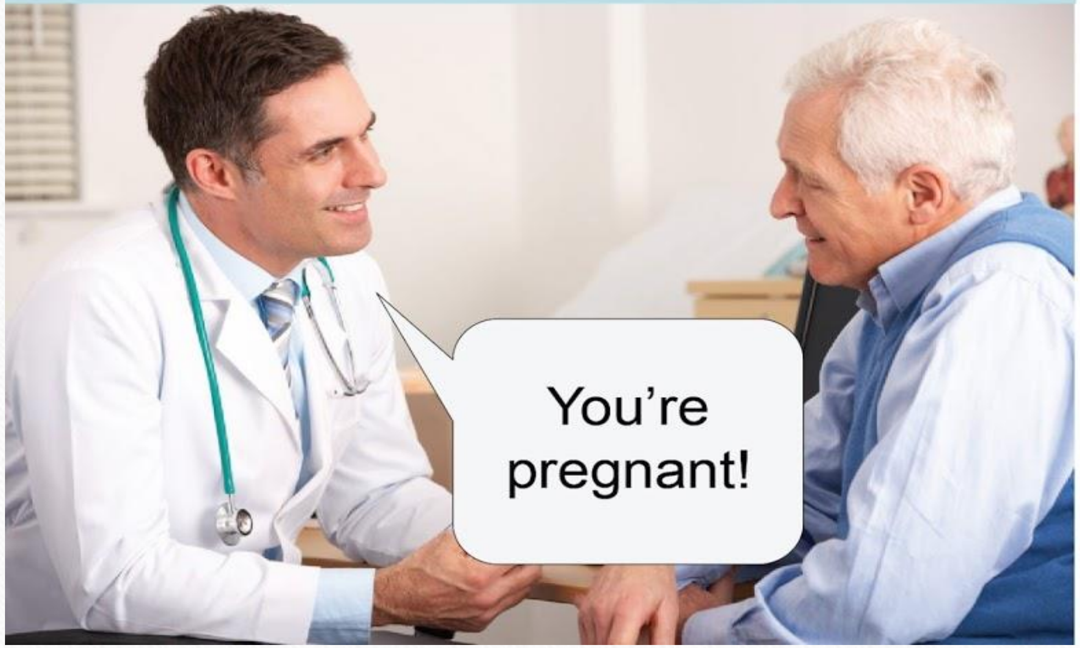
- There is no significant difference between Phenelzine (Nardil) vs. a placebo on treatment of depression.  *Fail to reject the null ($p$>.05):*
  - **This could be an accurate finding or an inaccurate finding – *false negative***

# Types of Error

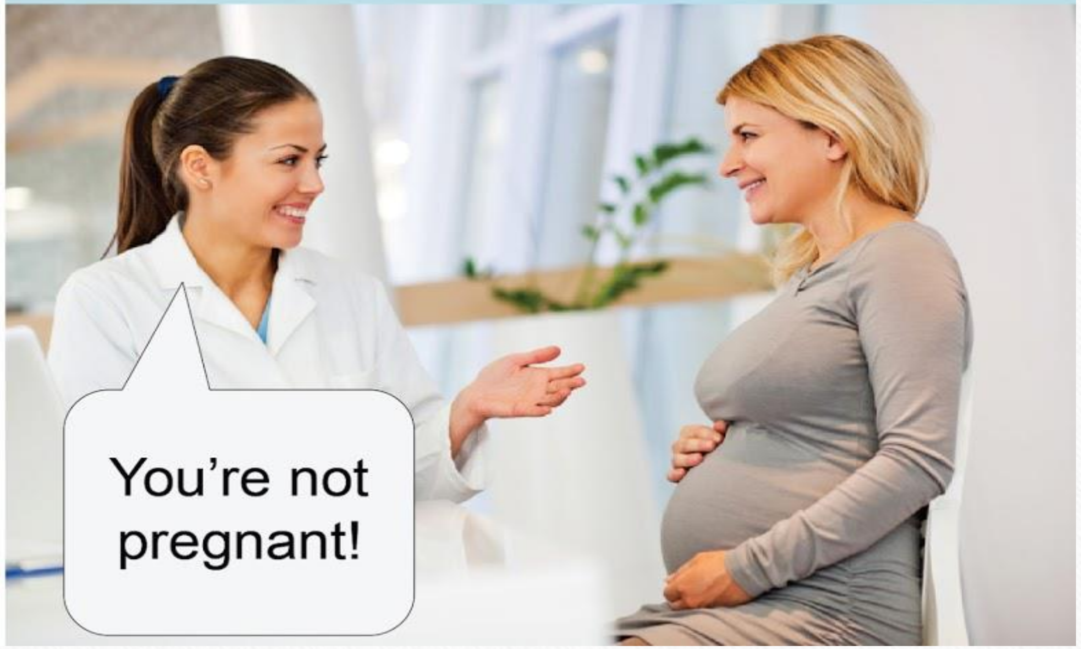
- Type 1 error (α): *False Positive*: Falsely concluding drug is effective when actually has no effect.
  - Pregnancy test shows positive, but in realty not pregnant
  - Fire alarm sounds, but no fire
  - Guilty verdict, but actually innocent
- Type II error (β): *False Negative*: Falsely concluding drug has no effect when actually effective.
  - Pregnancy test shows negative, but in realty pregnant
  - Fire alarm does not sound, but there is a fire
  - Innocent verdict, but actually guilty

# Type I and Type II Error

Type I Error



Type II Error



# Types of Error

We can adjust alpha and beta to avoid Type I or II errors

| | | Reality | |
|---|---|---|---|
| | | $H_o$ is true | $H_a$ is true |
| Research Decision | $H_o$ is true | Accurate ( $1 - \alpha$ ) | Type II Error<br>False Negative ( $\beta$ ) |
| | $H_a$ is true | Type I Error<br>False Positive ( $\alpha$ ) | Accurate ( $1 - \beta$ ) |

# Adjusting α and β ?

- Reduce α (0.01 – 0.05) when research question makes it particularly important to avoid Type I (false-positive) error
  - Important to ensure a intervention actually works
  - Judicial system: “Better that ten guilty persons escape than that one innocent suffer” Sir William Blackstone, 1765
- Reduce β (0.05 – 0.20) when important to avoid Type II (false-negative) error
  - Prefer an initial cancer diagnosis show a false positive than a false negative.

# Power

- **Power of a test** (1 – β): **the probability of correctly concluding the drug is effective when it is actually effective.** (β = 0.05 to 0.20)
  - 1 - β = .10 ~ Researcher willing to accept a 10% chance of missing an association of a given effect size between Phenelzine and depression.
  - (1 – β) = (1 - .10) = .90 ~ a 90% chance of finding an association of the given effect size

# How much power do we need?

## *Depends on the research question*

- Reducing the risk of a Type I error – Reduce the significance level
  - $p = .05$ → $p = .025$
  - Lowers the chance of a false positive (more stringent requirement), but increases the chance of a false negative (Type II error) – missing something that is actually occurring
    - Medical treatment – Important to verify treatment actually works. A false positive could be problematic in treating patients (thinking it helps when it does nothing.) $p = .025$

# How much power do we need?
# *Depends on the research question!*

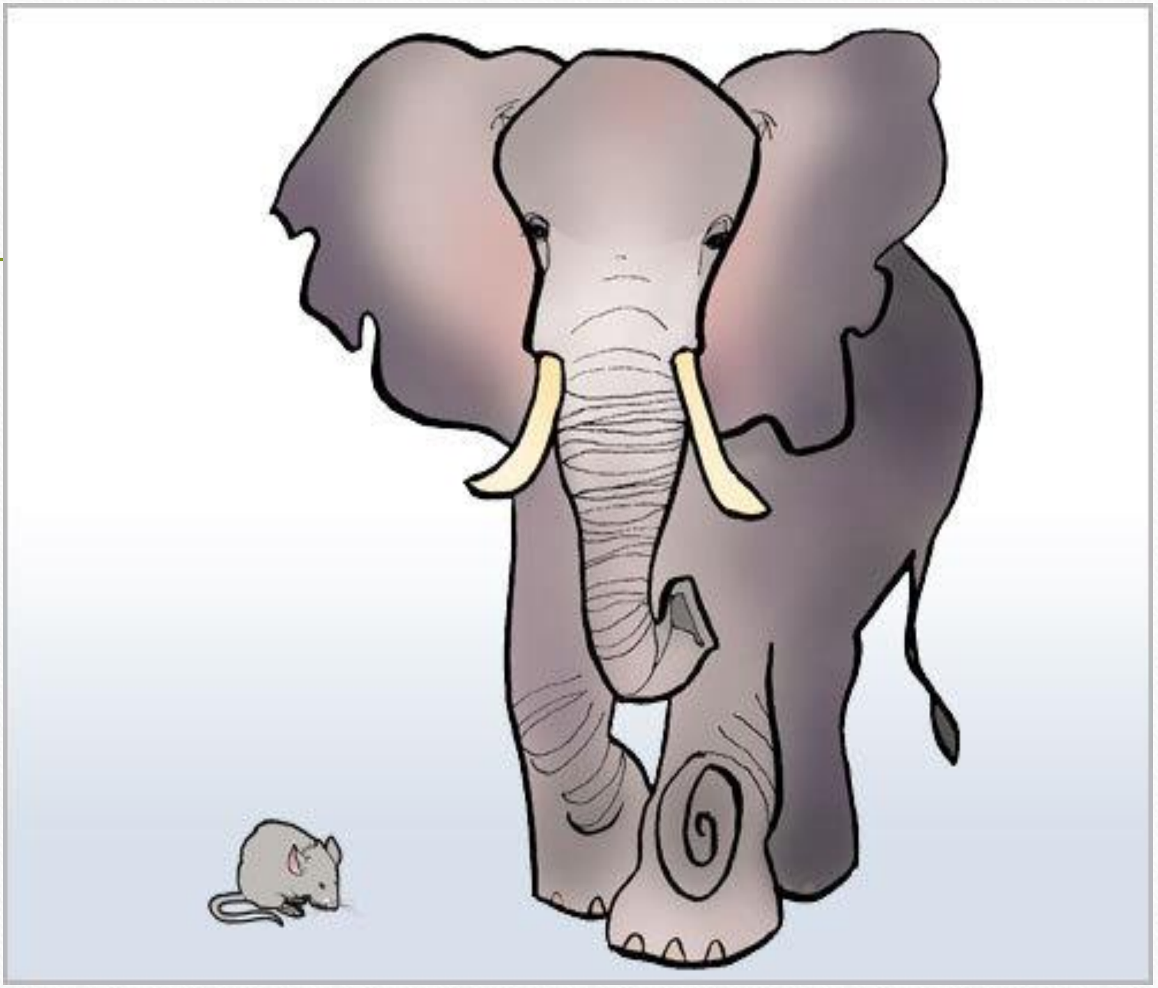
- Reducing the risk of a Type II error – Decrease $\beta$ increases power $(1 - \beta)$
    - $(1 - \beta) = (1 - .20) = .80 \rightarrow (1 - \beta) = (1 - .10) = .90$
    - Lowers the chance of a false negative
    - Increases chance of a false positive
        - E.g., Most would prefer a false positive in an initial cancer diagnosis than a false negative

**What if it cost $200.00**

# Beyond *p*-value

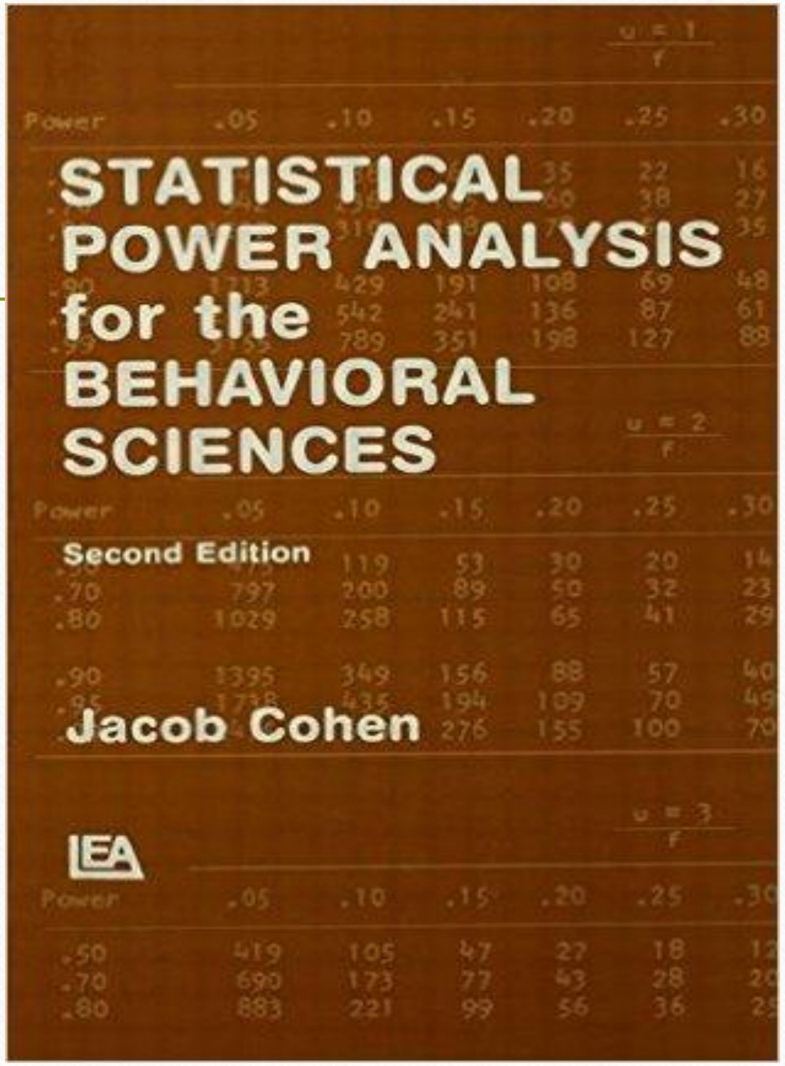
- FAU researchers recently found a statistically significant difference between a "safe and legal" Supplement and a Placebo on test performance
- Would you pay $ 1.00 for this safe and legal supplement that research has shown to result in a statistically significant difference (improvement) in test performance?

# Effect Size



# Why should we be concerned with Effect Size?

- Because Psychologist / Statistics author, Jacob Cohen (1923-1998) said so:
  - *“The primary product of a research inquiry is one or more measures of effect size, not p value.”*



# Why should we be concerned with Effect Size?

ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND P-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

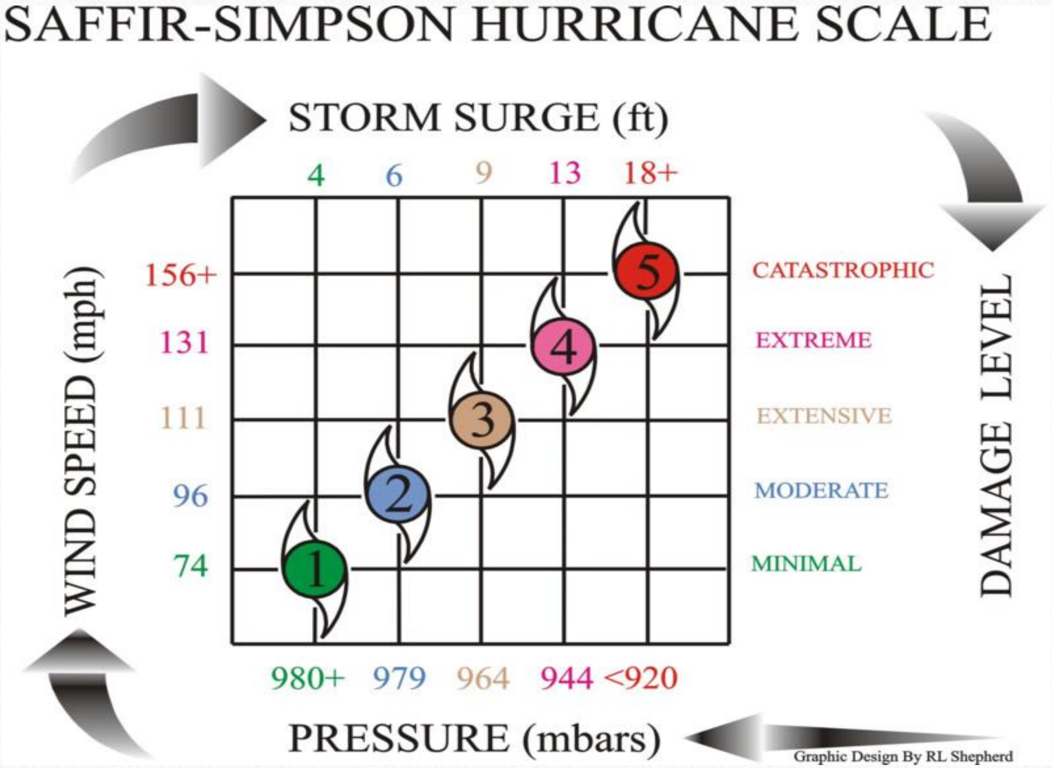
# Why should we be concerned with Effect Size?

- More and more journals are requiring the reporting of effect size in research

- "it is almost always necessary to include some measure of **effect size** in the Results section." (**APA** Publication Manual 2010, p. 34).

# Beyond *p-value*, Effect size!

- **Effect Size** – A name given to indices that measure the relative magnitude of treatment effect.



# Large sample size example

- Physicians Health Study (1989) of aspirin to prevent myocardial infarction (MI)
  - In more than 22,000 participants over an average of 5 years, aspirin was associated with a reduction of MI that was highly statistically significant: <u>***p* < .00001**</u>.
  - Due to the conclusive evidence, aspirin was recommended for general prevention.
  - <u>**However, the effect size was extremely small (*r* = .034, $r^2$ = .001)**</u>
  - As a result, many people were advised to take aspirin who would not experience the benefit, yet were also at risk for adverse effects (e.g., bleeding in the stomach or brain).
  - FDA (2014), revised position on the use of daily aspirin.

# Common Indices of Effect Size

- Comparison Studies
  - Cohen's *d*
  - Odds ratio (OR)
  - Relative risk or risk ratio (RR)
  - Number Needed to Treat (NNT)
- Relational studies (all correlations are effect sizes)
  - Pearson's r correlation
  - $r^2$ coefficient of determination

# Cohen’s *d*

- The difference between two means (e.g., treatment mean minus control mean) divided by the standard deviation of the two conditions

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

- What precisely the standard deviation (s) is, was not originally made explicit by Cohen
  - Defined as, the standard deviation of either population (since they are assumed to be equal)

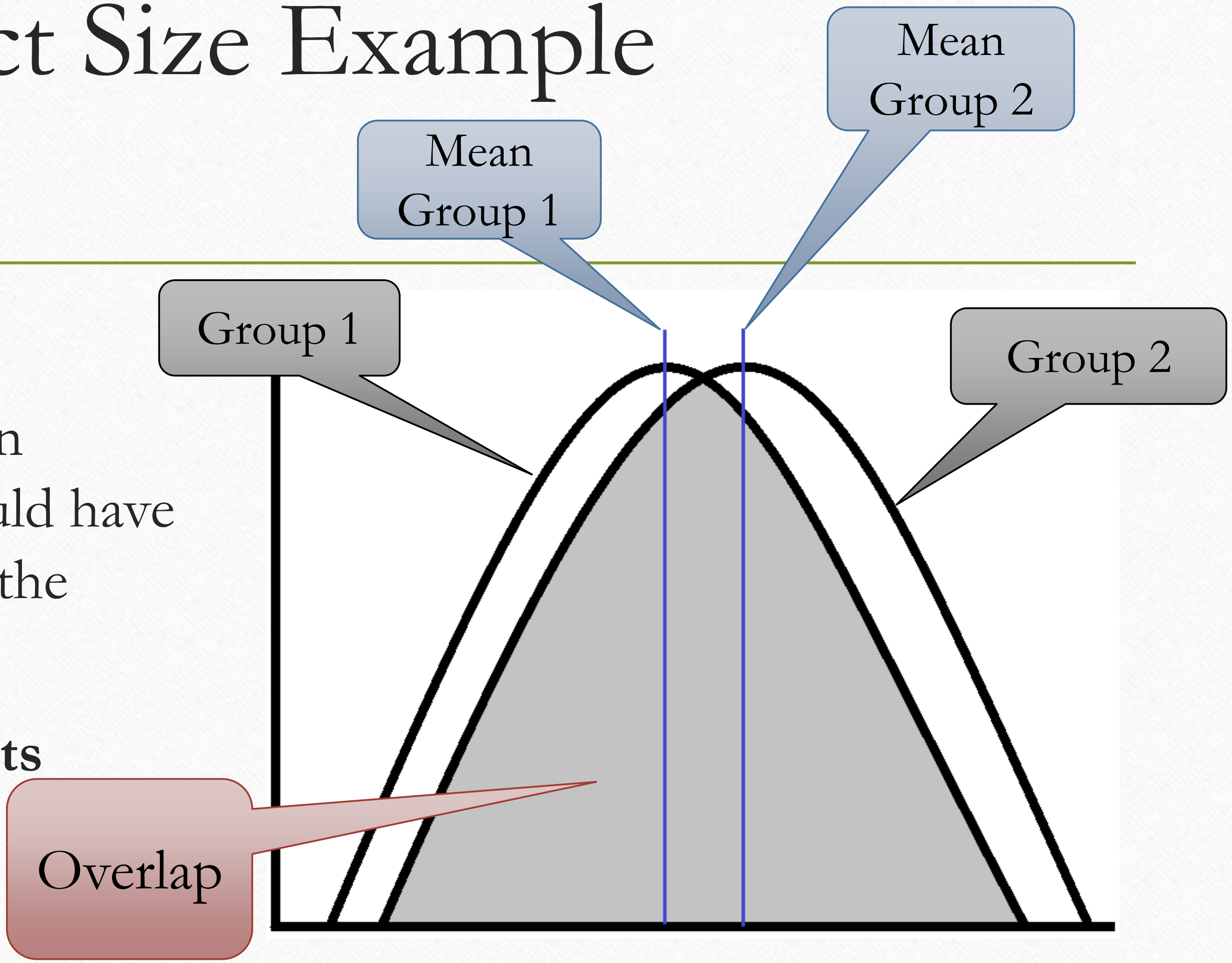
# Cohen's *d*

- Identified specific effect size values:
  - **.2 = small effect      .5 = medium effect      .8 = large effect**

*NOTE: Ideally, interpretation of results should be grounded in a meaningful context or by quantifying their contribution to knowledge. Where this is problematic, Cohen's effect size criteria may serve as a backup.*

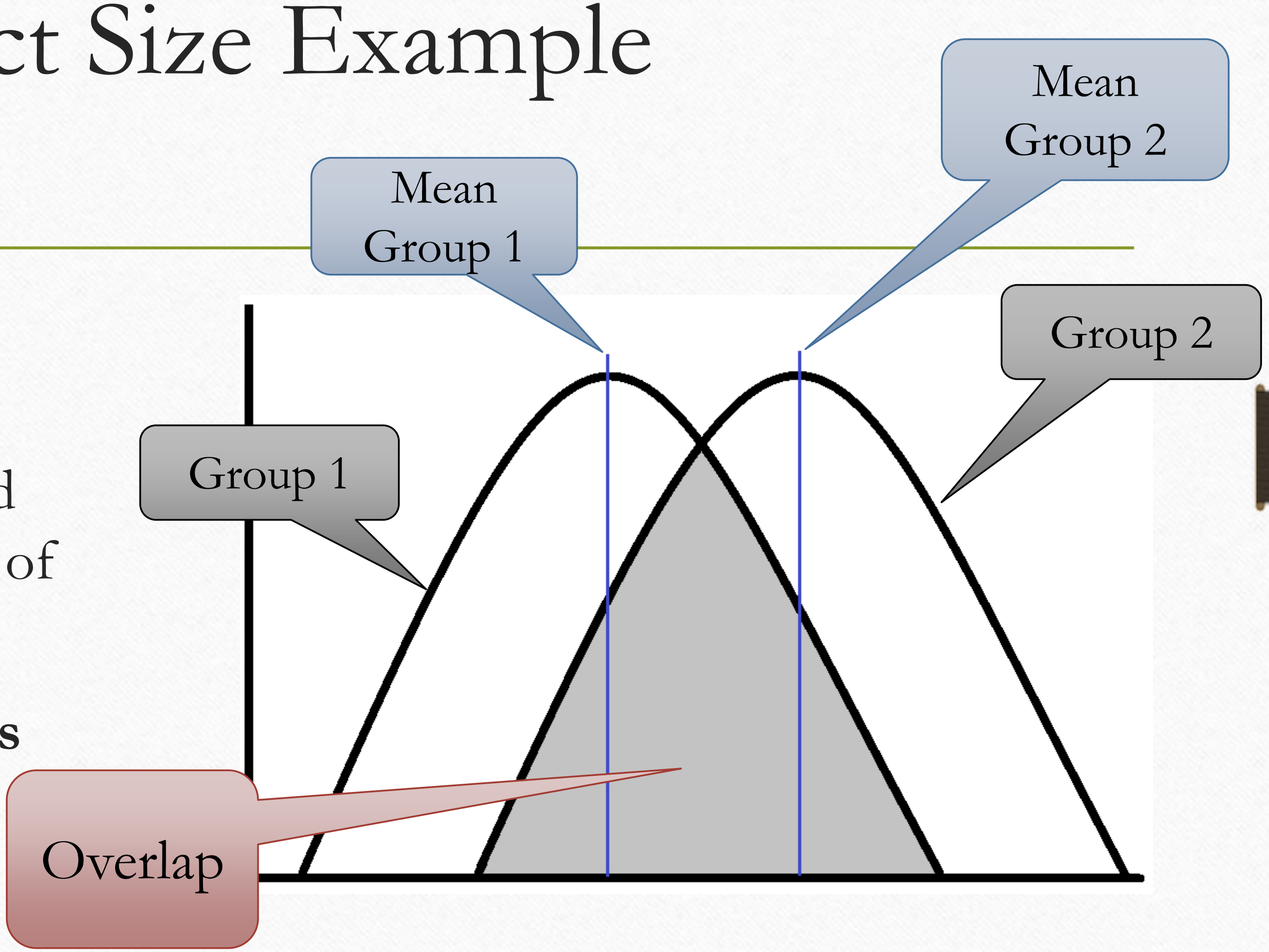
# Effect Size Example

- Effect size = 0.2
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 58% of the people in Group 1
- **85% overlap of participants**



# Effect Size Example

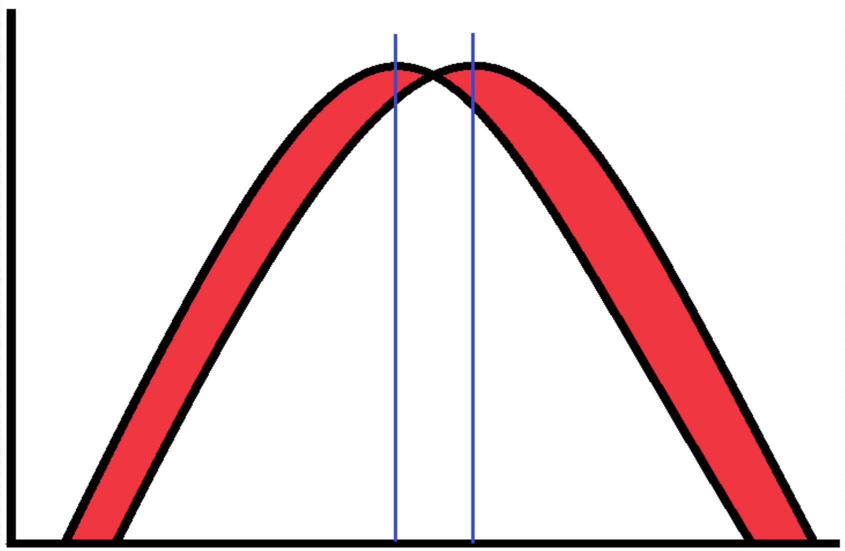
- Effect size = 0.8
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 79% of the people in Group 1
- **53% Overlap of participants**



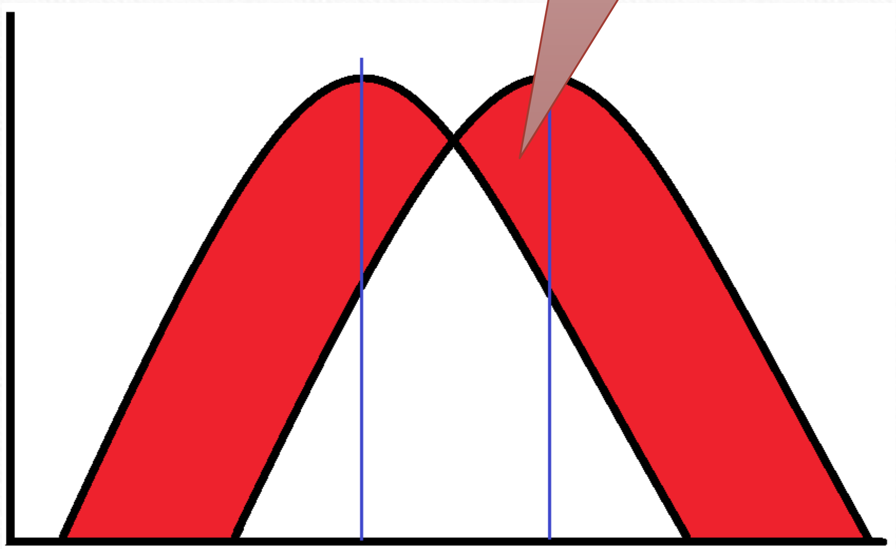
# Effect Size Example

As Effect Size increases, magnitude of difference increases

Effect Size = 0.2



Effect Size = 0.8



# Cohen's *d* clinical example

- The Research Unit on Pediatric Psychopharmacology Anxiety Study Group (RUPP; 2001). The influence of Fluvoxamine on Pediatric Anxiety
  - DV: Score on the Pediatric Anxiety Rating Scale
  - IV: Medication. Two levels (Fluvoxamine, Placebo) (N=128)
    - Mean SD: Treatment = 9.0 +/- 7.0, Control = 15.9 +/- 5.3, ***p* < .001**
    - Difference: -6.9 (95% CI: -4.6, -9.2), ***d* = 1.1**
    - 84% of placebo group had worse scores than the average score of fluvoxamine group.

# Be Mindful of Clinical Consequences!

- A small effect (e.g., *d* = 0.2) from  study comparing treatments related to mortality rates for two chemotherapies for breast cancer would have greater clinical consequence than a large effect (e.g., *d* = 0.80) from a study of treatment related to ADHD symptom reduction.

# Eta-squared ($\eta^2$)

- Eta-squared is a measure of effect size typically for use in ANOVA
- Interpret $\eta^2$ (Cohen):
  - .02 ~ Small
  - .13 ~ Medium
  - .26 ~ Large
- ***Remember! Interpretation of results should be grounded in a meaningful context, or by quantifying their contribution to knowledge.***

# Relative Risk (RR) aka risk ratio

- For categorical measures (e.g., Improved vs. Not Improved) consider RR and OR
- Example - Influence of Cognitive Behavioral Therapy (CBT) on children with Asperger's ability to pass a social awareness test
  - Control: 2 students pass for every 1 that fails
    - Probability of passing is 2/3 (or 0.67)
  - Treatment: 6 students pass for every 1 that fails
    - Probability of passing is 6/7 (or 0.86)
  - RR = 0.86 / .067 = 1.28 (Note: not comparable to Cohen's d)

# Relative Risk (RR) aka risk ratio

- In the RUPP example
  - 76% of the subjects receiving fluvoxamine were treatment responders according to Clinical Global Impressions (CGI) improvement ratings
  - 29% of the placebo group were treatment responders according to Clinical Global Impressions (CGI) improvement ratings
  - **RR = 2.6** (0.76 / 0.29)
  - **Results suggest the patients treated with fluvoxamine for anxiety disorders had almost a threefold greater probability of responding than those on placebo**

# Odds Ratio (OR)

- Research example - Influence of Cognitive Behavioral Therapy (CBT) on children with Asperger's ability to pass a social awareness test
  - Control: 2 students pass for every 1 that fails
    - Odds of passing are two to one (or 2/1 = 2)
  - Treatment: 6 students pass for every 1 that fails
    - Odds of passing are six to one (or 6/1 = 6)
  - OR = 6 / 2 = 3 (Note: not comparable to Cohen's d)
  - Odds of passing of Treatment group are three times higher than the Control group

# Number Needed to Treat (NNT)

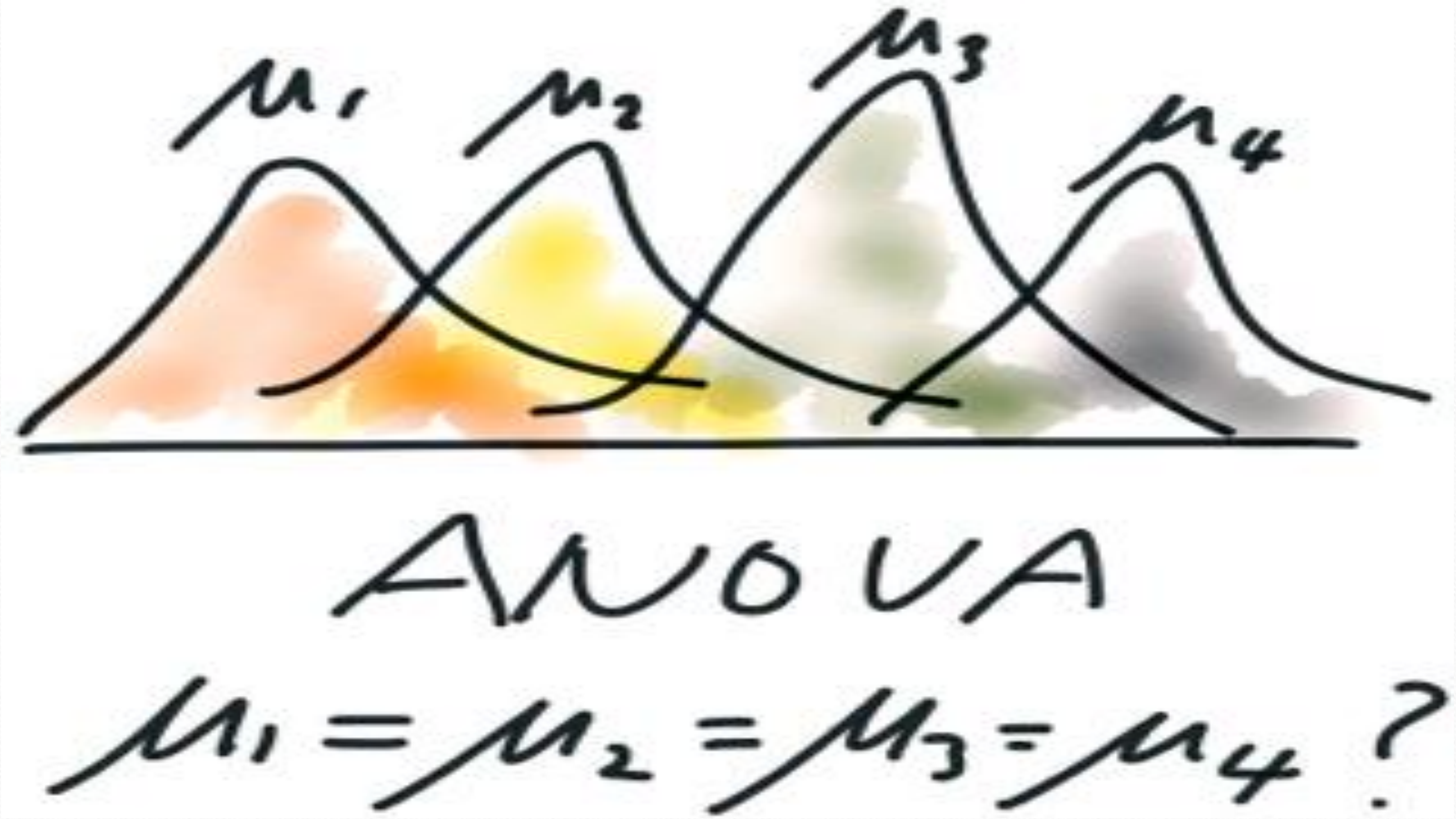
- Number of subjects one would expect to treat with agent A to have one or more successes (or one less failure) than if the same number were treated with agent B
- Well suited for binary (success/failure) outcomes
  - NNT = 100 / (% improved on Treatment - % improved on Placebo)
- RUPP study: 76% improved on fluvoxamine, 29% improved on placebo
  - NNT = 100 / (76-29) = 2.1
  - **For every two patients treated with fluvoxamine, at least one will have a better outcome than if treated with placebo**

# Coefficient of determination ($R^2$ or $r^2$)

- An output of regression analysis
- Interpreted as the proportion of the variance in the dependent variable (criterion) that is predictable from the independent variable (predictor).
- Range from 0 to 1

# Coefficient of determination ($R^2$ or $r^2$)

- $R^2 = 0$ ~ Dependent variable cannot be predicted from the independent variable
- $R^2 = 1$ ~ Dependent variable can be predicted without error from independent variable
- $R^2 = .20$ ~ 20% of the variance in dependent variable is predictable from the independent variable.
- Trait mindfulness explained 28% ($R^2 = .28$) of the variance in test anxiety (Altairi, 2014)
  - Note: 72% of the variance unexplained.


$\mu_1$
$\mu_2$
$\mu_3$
$\mu_4$
ANOVA
$\mu_1 = \mu_2 = \mu_3 = \mu_4$ ?

# ANOVA

- Testing variation among the means of three or more groups
  - Remember, t-tests are used to explore the variation between two groups
- Basic setup:
  - IV = Categorical (3 or more groups), 4th grade vs. 5th grade vs. 6th grade
  - DV = Continuous – test performance, score on self-efficacy task

# Basic Logic of ANOVA

- Estimating population variance from variation from **<u>within</u>** each sample
  - Within group variance affected by
    - Individual difference
    - Chance
    - Experimental error

# Basic Logic of ANOVA

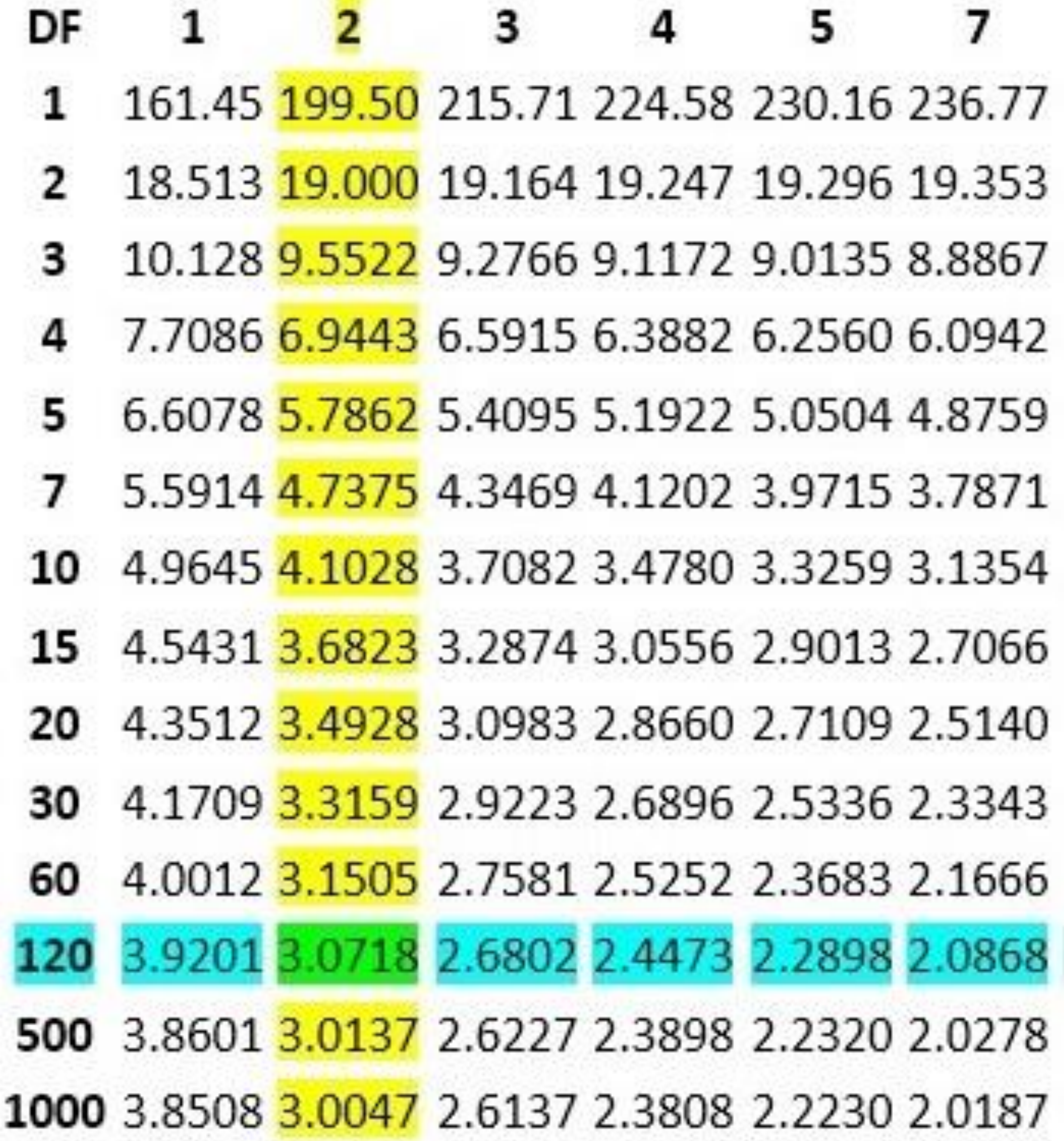
- Estimating population variance from variation between the means of the samples
  - Variation due to treatment, chance factors

# Basic Logic of ANOVA

- The *F* ratio
  - Ratio of the between-groups population variance estimate to the within-groups population variance estimate

$$F = \frac{\text{Between-group variability}}{\text{Within-group variability}}$$

F table to determine statistical significance



| DF | 1 | 2 | 3 | 4 | 5 | 7 |
|---|---|---|---|---|---|---|
| 1 | 161.45 | 199.50 | 215.71 | 224.58 | 230.16 | 236.77 |
| 2 | 18.513 | 19.000 | 19.164 | 19.247 | 19.296 | 19.353 |
| 3 | 10.128 | 9.5522 | 9.2766 | 9.1172 | 9.0135 | 8.8867 |
| 4 | 7.7086 | 6.9443 | 6.5915 | 6.3882 | 6.2560 | 6.0942 |
| 5 | 6.6078 | 5.7862 | 5.4095 | 5.1922 | 5.0504 | 4.8759 |
| 7 | 5.5914 | 4.7375 | 4.3469 | 4.1202 | 3.9715 | 3.7871 |
| 10 | 4.9645 | 4.1028 | 3.7082 | 3.4780 | 3.3259 | 3.1354 |
| 15 | 4.5431 | 3.6823 | 3.2874 | 3.0556 | 2.9013 | 2.7066 |
| 20 | 4.3512 | 3.4928 | 3.0983 | 2.8660 | 2.7109 | 2.5140 |
| 30 | 4.1709 | 3.3159 | 2.9223 | 2.6896 | 2.5336 | 2.3343 |
| 60 | 4.0012 | 3.1505 | 2.7581 | 2.5252 | 2.3683 | 2.1666 |
| 120 | 3.9201 | 3.0718 | 2.6802 | 2.4473 | 2.2898 | 2.0868 |
| 500 | 3.8601 | 3.0137 | 2.6227 | 2.3898 | 2.2320 | 2.0278 |
| 1000 | 3.8508 | 3.0047 | 2.6137 | 2.3808 | 2.2230 | 2.0187 |

# One-way ANOVA

- Independent Variable (IV) – Three or more levels
  - E.g., Treatment for depression: Sertraleen (Zoloft), Phenelzine (nardil), CBT (Cognitive Behavioral Therapy)
- Dependent Variable (DV) – One continuous variable
  - Score on a Pediatric Generalized Anxiety Disorder (GAD) scale

# One-way ANOVA

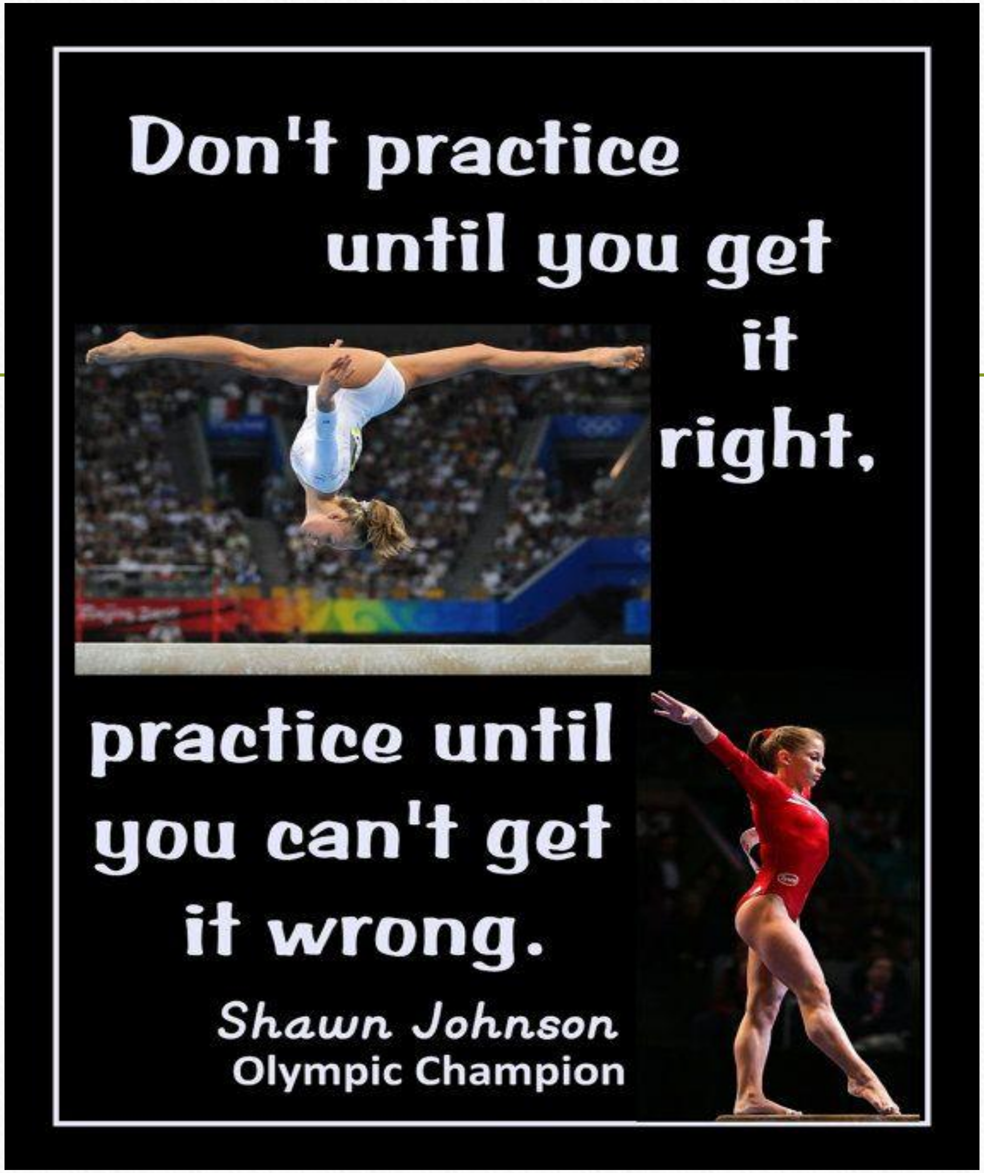
- IV: Sertraleen (Zoloft), Phenelzine (nardil), CBT (Cognitive Behavioral Therapy)
- DV: Score on a Pediatric Generalized Anxiety Disorder (GAD) scale
  - One *F* value
  - Post-hoc results (review only if overall model is significant)
    - Sertraleen vs. Phenelzine
    - Sertraleen vs. CBT
    - Phenelzine vs. CBT

# Two-way ANOVA

- Independent Variable (IV)
  - Thorazine (three levels) 10mg, 25mg, 50mg
  - Alcohol (three levels) BAC = .02, .05, .08
- Dependent Variable (DV)
  - Score on psychosis scale

# Two-way ANOVA

- Results:
  - Main effect of Thorazine on reducing symptoms of delusions and hallucinations
  - Main effect of Alcohol (a depressant) on reducing symptoms of delusions and hallucinations
  - Interaction effect: breathing difficulties, potentially fatal



Don't practice until you get it right,

practice until you can't get it wrong.

Shawn Johnson
Olympic Champion

# Other things being equal, which of the following actions will reduce the power of a hypothesis test?

- I. Increasing sample size.
- II. Reducing significance level (e.g., from 0.05 to 0.01).
- III. Increasing beta, the probability of a Type II error.
    - (A) I only
    - (B) II only
    - ☺ III only
    - (D) All of the above
    - (E) None of the above

Power = 1 – β, Increasing beta reduces power

# Suppose a researcher conducts an experiment to test a hypothesis. If she doubles her sample size, which of the following will increase?

- I. The power of the hypothesis test.
- II. The effect size of the hypothesis test.
- III. The probability of making a Type II error.
    - ☺ I only
    - (B) II only
    - (C) III only
    - (D) All of the above
    - (E) None of the above

Increasing sample size makes the hypothesis test more sensitive - more likely to reject the null hypothesis when it is, in fact, false. Thus, it increases the power of the test.

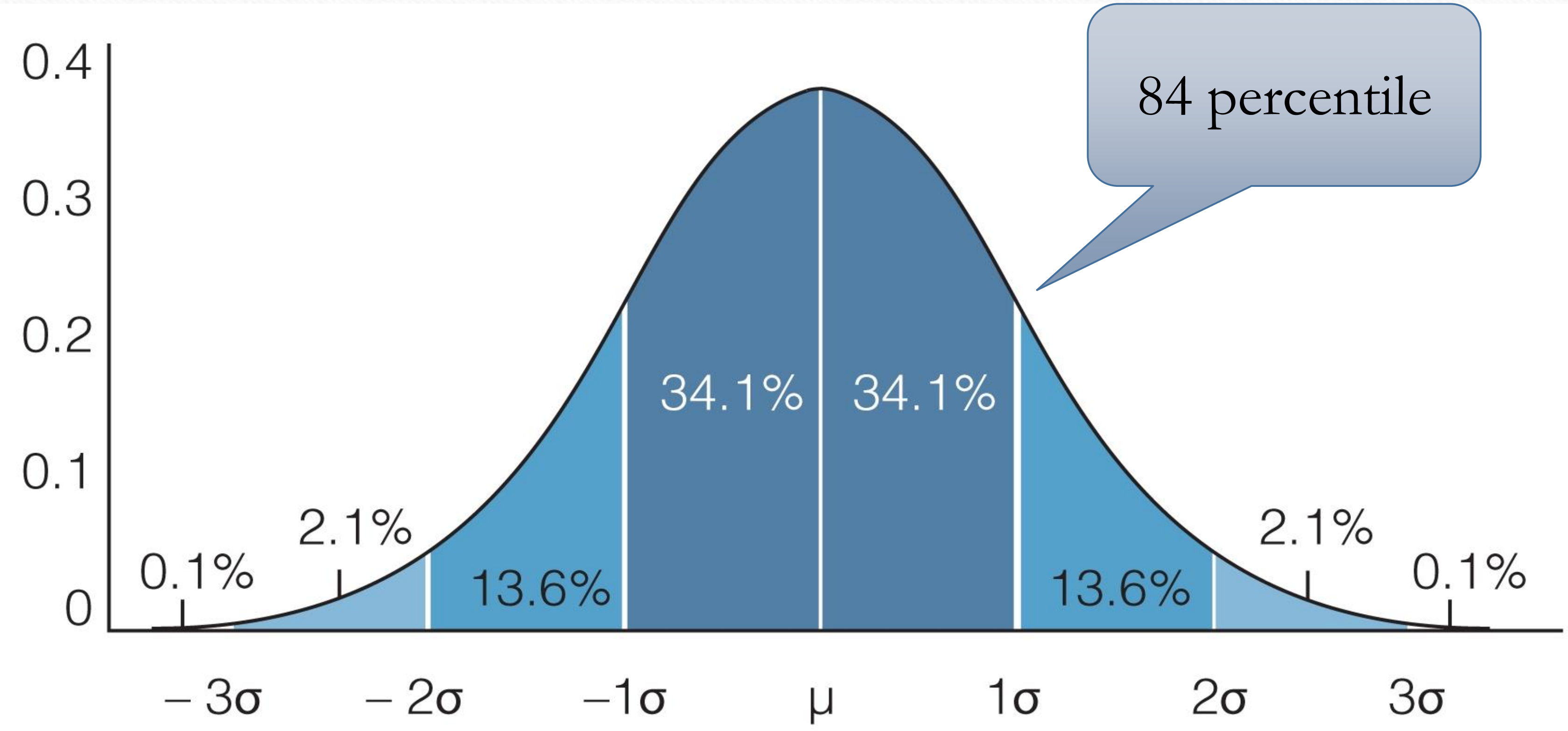
# You are conducting a research study and wish to know whether two different drugs are equally effective.

- H0 [null hypothesis]: Drugs A and B are equally effective.
- Ha [alternative hypothesis]: Drug B is more effective than A.
- Your study uses an alpha-level of alpha=0.05. The power of the test was 0.80.
- If the alternate hypothesis is actually true, what is the probability that the study will show a significant difference in efficacy between the two drugs?
  - A. 0.05
  - B. 0.20
  - [✓ 0.80]
  - D. 0.95
  - E. It is impossible to determine from the information given

the chance of correctly identifying a significant difference between the two drugs is equal to the power, which is 0.80

An aspiring medical student opens up her Step 1 score report and is elated to find that her score is 230, a 92 on the 1-100 scale! The score report indicates that the mean on the exam was 215 and the standard deviation is 20. Assuming a normal distribution, what is the most likely percentile of the student's score?

- A. 54.3%
- B. 72.6%
- [✓ 84.1%]
- D. 97.7%
- E. 99.9%



84 percentile

# If the variance is 9, then the standard deviation is

- A. 81
- B. 4.5
- C. 18
- ☺ 3

# Dr. Grey conducts an ANOVA and rejected the null hypothesis. The most likely *F* value is

- A. 1.01
- B. -3.12
- ☺ 3.67
- D. 0.05

# A characteristic of an F value is that

- A. The cutoff F equals the calculated F divided by 2
- ☺ It can never be less than 0
- C. It is negatively skewed
- D. Can be adjusted by increasing alpha

# Type II (false-negative) errors concern medical researchers because

☺. they could mean that useful treatments are not implemented

B. they could mean the experiments must be repeated to confirm positive results

C. rejecting the null hypothesis should only occur when the research hypothesis is true

D. future research might be based on results mistakenly declared significant

Dr. H. Pierce is investigating the impact of a treatment on post surgery infection.  If he can estimate the effect size and the desired power, *a priori*, he can determine

A. Minimum meaningful difference

☺ Number of participants needed

C. Population distribution

D. Alpha level of the experiment

# If the research hypothesis is true, but the study has low level of power, then

A. the probability that the study will have a significant results is high

[✓ the probability that the study will have a significant results is low]

C. the null hypothesis will almost certainly be rejected

D. beta is necessarily low

# Setting the significance level cutoff at .10 instead of the more usual .05 increases the likelihood of

- ☺ a Type I (false-positive) error
- B. a Type II (false-negative) error
- C. Failing to reject the null hypothesis
- D. accepting the null hypothesis when, in fact, it is false

# Review (1 of 3)

- **Statistical significance** - Probability ***p-value***: Identifies the likelihood a particular outcome may have occurred by chance
  - $p < .05$ = There is less than a 5% probability the findings occurred by chance

# Review (2 of 3)

- **Statistical significance** - Probability ***p-value***:
  - Considered to be confounded because of its dependence on sample size
    - Sometimes statistical significance means only that a huge sample size was used

# Review (3 of 3)

- **Effect size** – Measurements that tell us the relative magnitude of the experimental treatment.
  - Tells us the ***size*** of the experimental ***effect***
    - How much did Treatment A improve test performance vs. Treatment B
    - How much did a therapy improve function vs. normal activities
  - Effect size can be used to justify the costs of a new therapy
    - Cost vs. impact

# Helpful links

- http://rpsychologist.com/d3/cohend/
- http://www.gpower.hhu.de/en.html

# References

Altairi, M. A. (2014). The impact of mindfulness and test anxiety on academic performance. (Senior Honors Thesis). Paper 39. Retrieved from http://ir.library.louisville.edu/honors/39

Paulos, J. A. (1989). *Innumeracy: Mathematical Illiteracy and its Consequences.* New York, NY: Hill and Wang.

Prokscha, S. (2012). *Practical Guide to Clinical Data Management (3rd ed.).* Boca Raton, FL: Taylor & Francis.