Statistics

Michael DeDonno Ph.D.

1. *Frequencies and Descriptives*
2. *Statistical Significance*
3. *Power, & Effect Size*
4. *Association Research*
5. *Comparison Research*

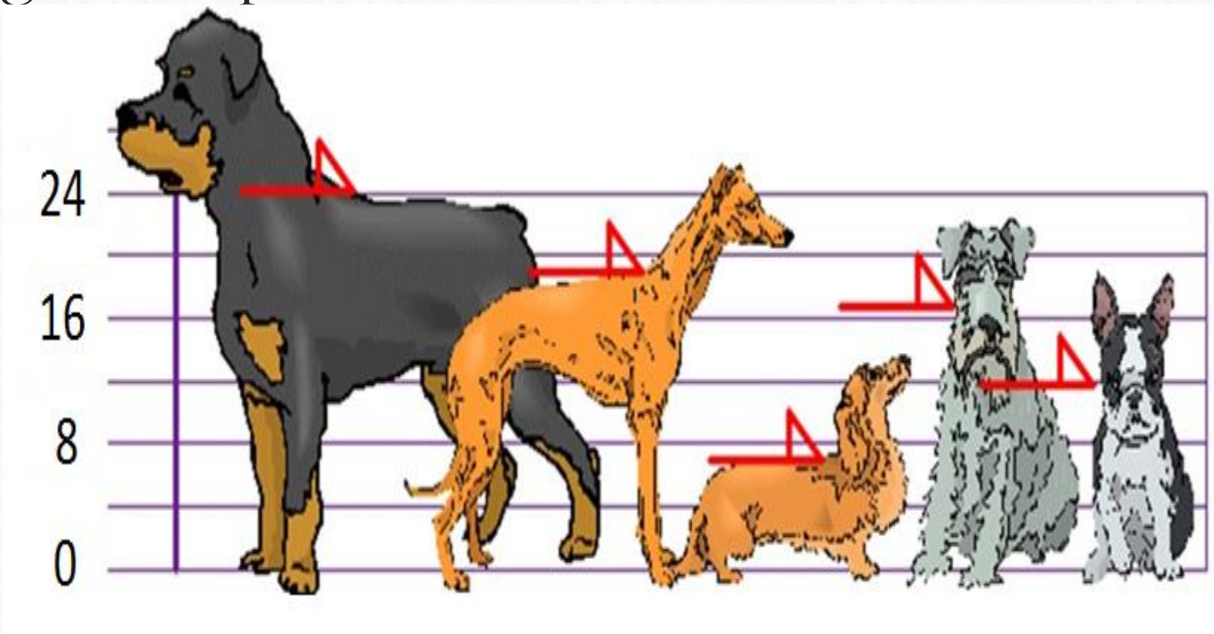
Frequencies

- **<u>Frequencies</u>**: Sum the number of instances within a particular category
  - 56 males, 37 females
  - 25 psychotics, 15 non-psychotics
  - 15 Hispanics, 22 Native Americans, 8 Asians

Number of Cups of Coffee	Tally	Frequency
0 - 3	//	2
4 - 7	///	3
8 - 11	~~////~~ ///	8
12 - 15	///	3
16 - 19	//	2

Descriptives

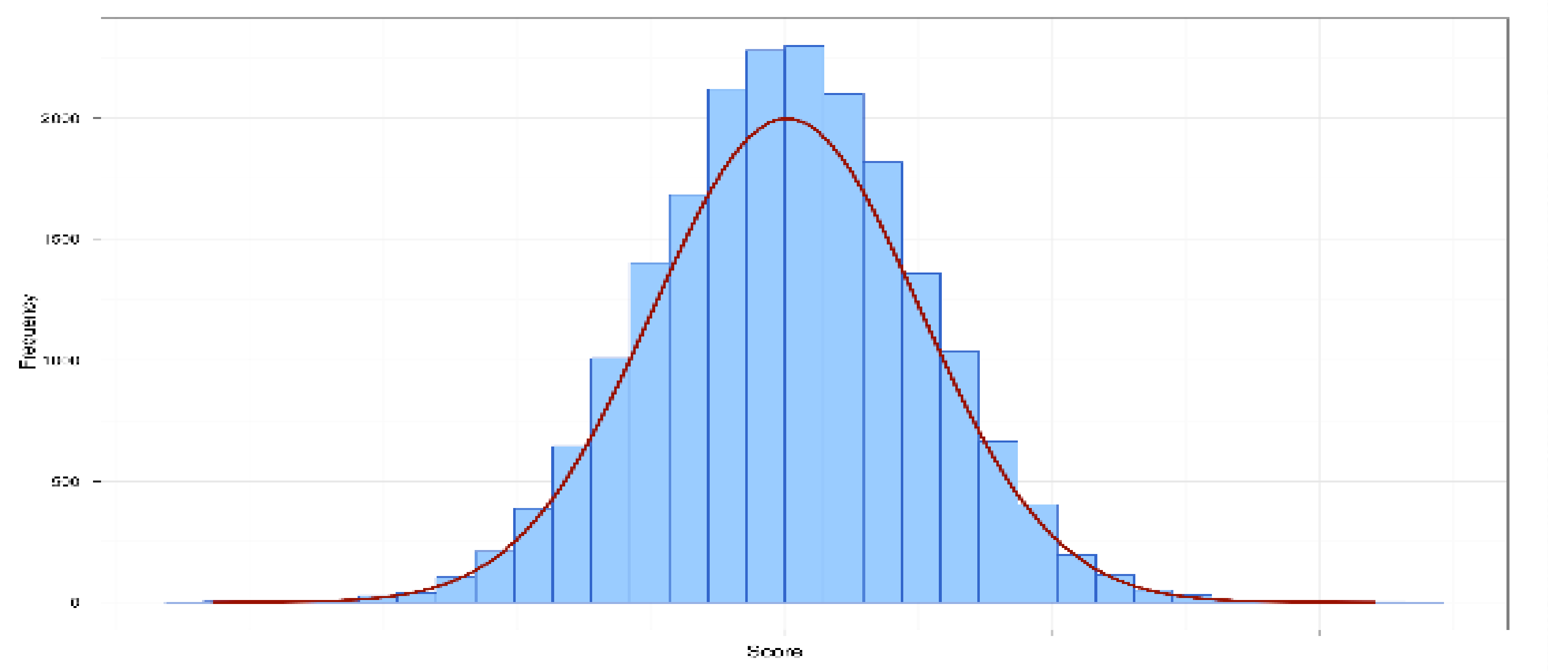
- **Descriptive statistics**: Designed to provide information about distributions of variables
  - Mean, Median, Mode
  - Maximum, Minimum, Range
  - Standard deviation, variance



Variability

- The degree to which scores in a distribution are spread out or dispersed
  - **Homogeneity**—little variability
  - **Heterogeneity**—great variability

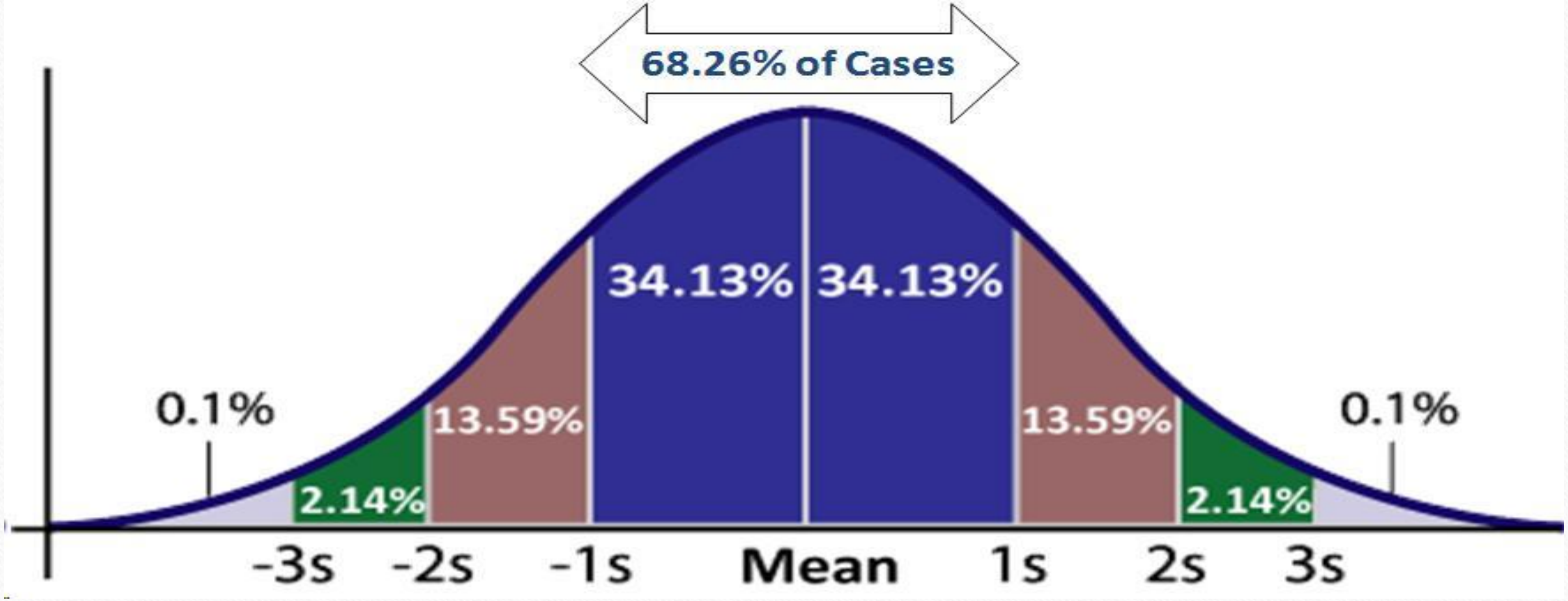
The Normal Distribution



The curve shows the idealized shape.

Variability in a Normal Distribution

- **Variance** – Deviation from the mean in squared units.
- **Standard Deviation** – Deviation from the mean in original units



Properties of Frequency Distributions

- Skew
  - The symmetry of the distribution.
  - Positive skew (scores bunched at low values with the tail pointing to high values).
  - Negative skew (scores bunched at high values with the tail pointing to low values).
- Kurtosis
  - The 'heaviness' of the tails.
  - Leptokurtic = heavy tails.
  - Platykurtic = light tails.

Skew

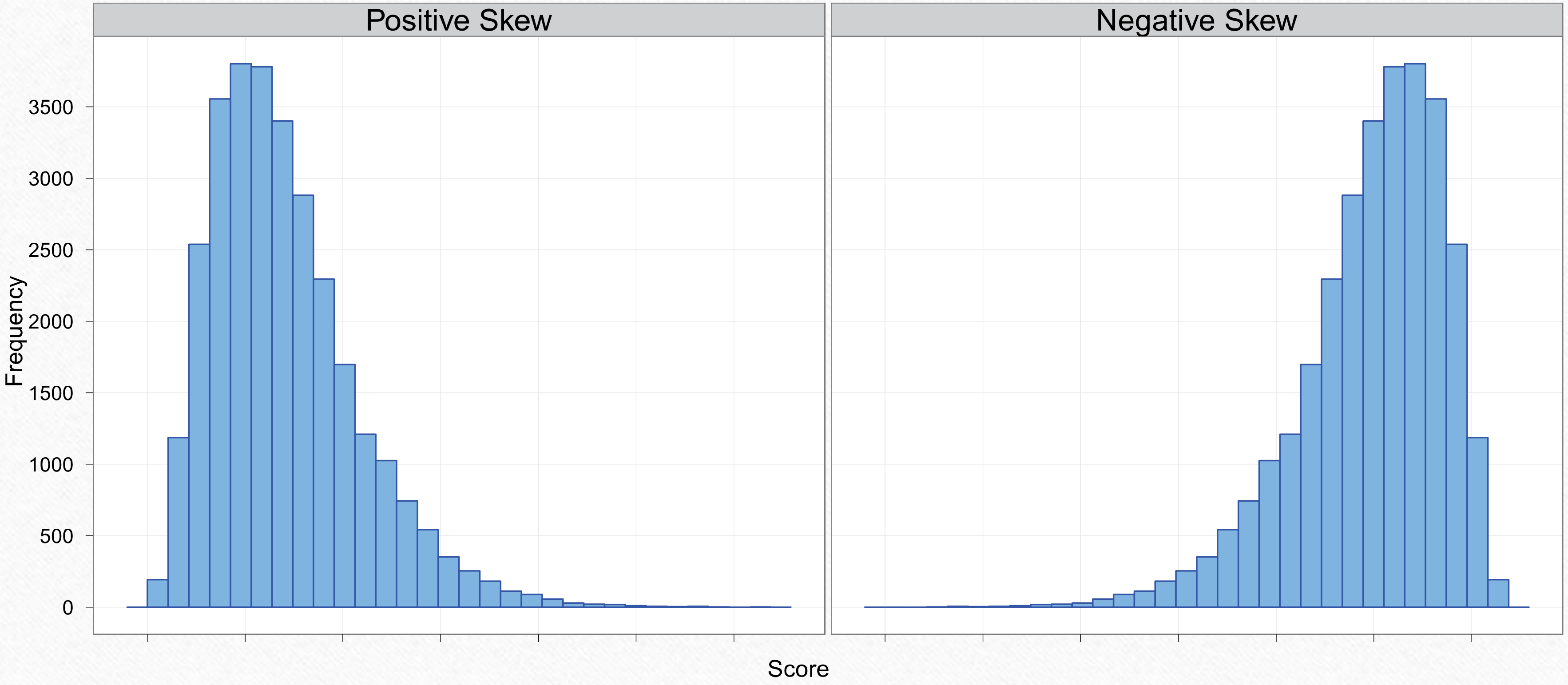


FIGURE 1.4 A positively (left) and negatively (right) skewed distribution

Kurtosis

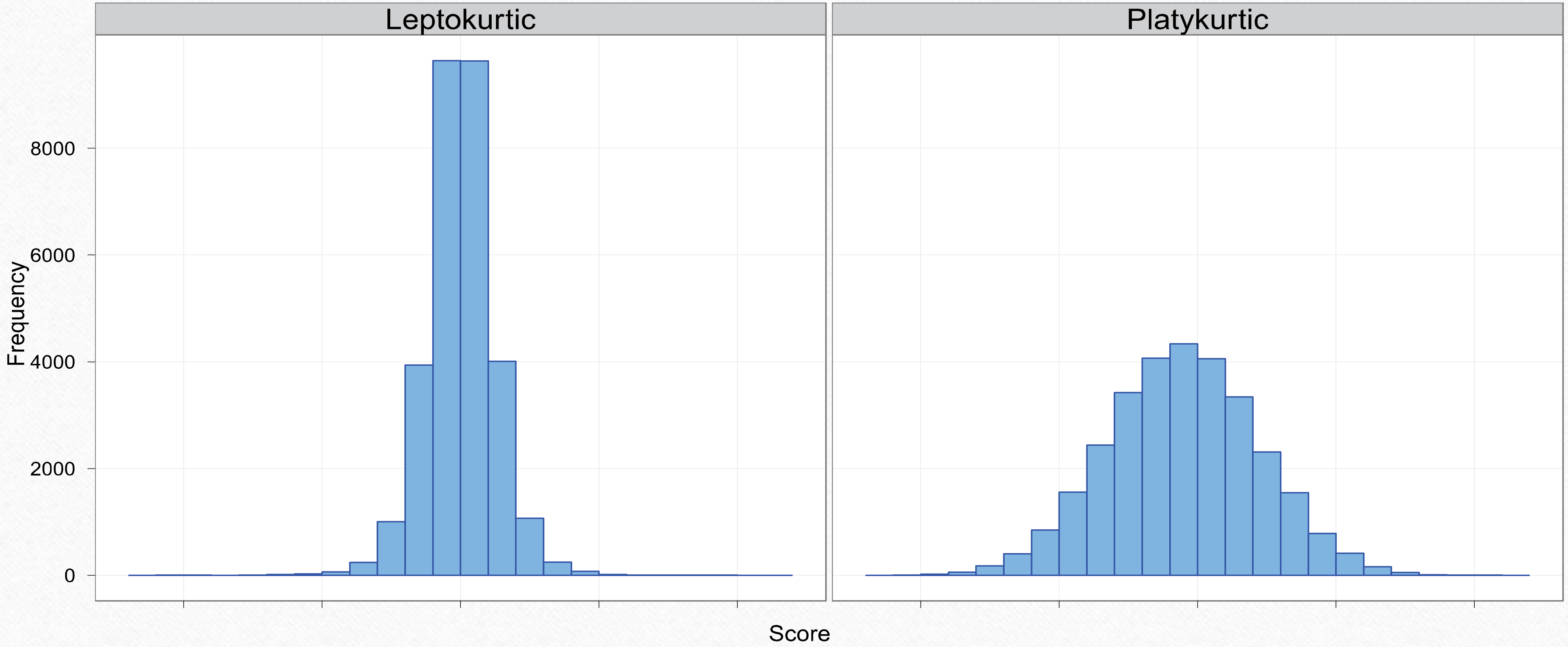


FIGURE 1.5 Distributions with positive kurtosis (leptokurtic, left) and negative kurtosis (platykurtic, right)

Bimodal and Multimodal Distributions

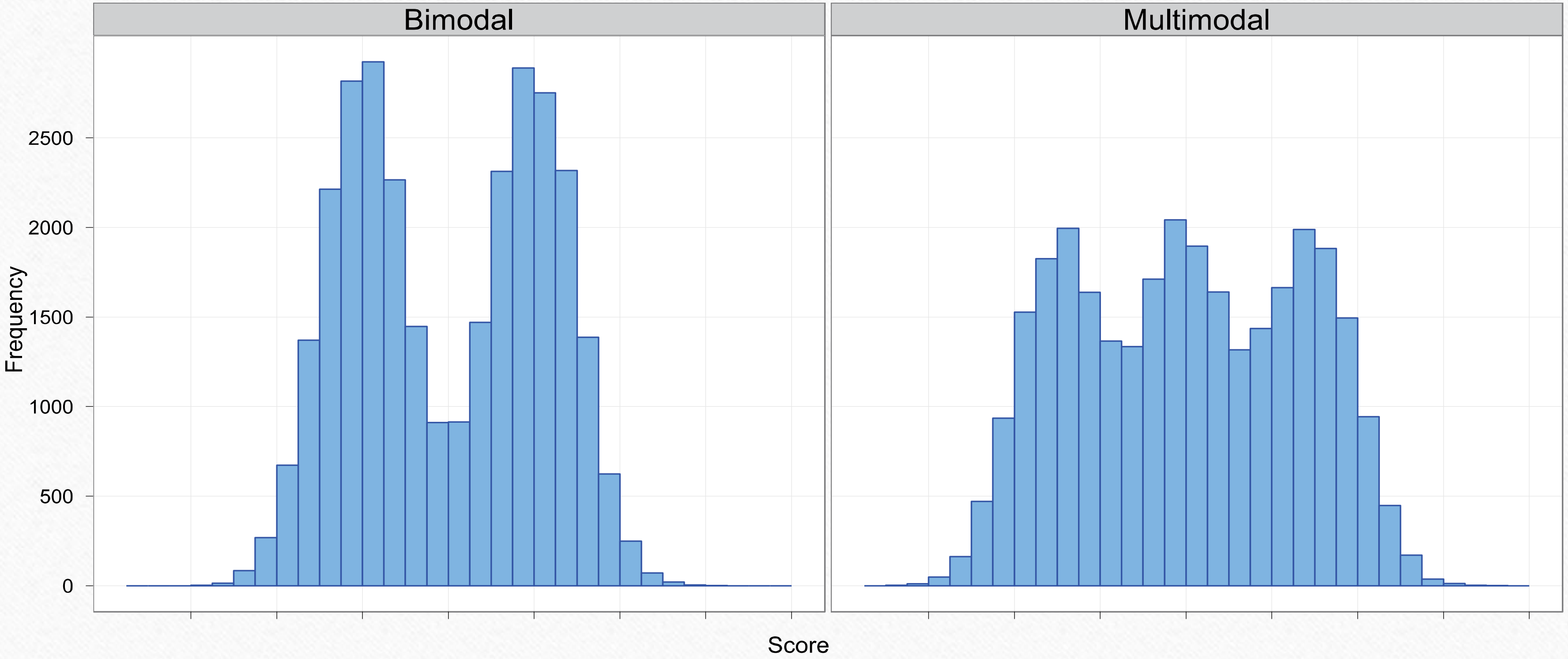


FIGURE 1.6
Examples of bimodal (left) and multimodal (right) distributions

??

$p < .05$

Statistical Significance

- **Probability "p" value** – Identifies the likelihood a particular outcome may have occurred by chance
  - Statistically significant: $p < .05$
  - Marginally significant: $.05 < p < .10$

What is a *p*-value?

- 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
      - $p = .60$ = There is a 60% 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 lbs 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 lbs 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?

- The effect of the treatment
  - Larger reduction (10 lbs.) in weight by the treatment vs. a smaller reduction (5lbs.)
- Variation in data
  - Larger variation can result in larger p-value
  - Source of variation
    - Between-subject variation
    - Measurement error
  - *and what else?*

What impacts a p-value?

"I climbed all this way, and you tell me That's the meaning of p-value?"

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

What sample size do I need?

- Too small – Risk non-significant results
- Too large – Time consuming (increased risk of error), costly(finite dollars), finite resources (countless hours collecting data)
- Is it even ethical to expose an unnecessary large number of participants?

- Need a sample size estimation!

Why do we need a *p-value*?

- Validates a hypothesis
  - Null Hypothesis (H_o): There is no difference between the treatment and control groups
    - e.g., (H_o): There is no difference in using Crest or water alone in preventing tooth decay.
  - Two possible results
    - Crest is more effective than water alone in preventing tooth decay. *Reject the null ($p<.05$)*
    - No evidence Crest is better than water in preventing tooth decay. *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*?

- Two possible results (post research):
  - Crest is more effective than water alone in preventing tooth decay. *Reject the null ($p<.05$)*
    - This could be an accurate finding
    - This could be an inaccurate finding – *false positive*
  - No evidence Crest is better than water in preventing tooth decay. *Fail to reject the null ($p>.05$)*
    - This could be an accurate finding
    - This could be an inaccurate finding – *false negative*

Types of Error

- Type 1 error (α): *False Positive*: Falsely concluding a drug is effective when the drug actually has no effect. $\alpha = 0\%$ to 5% ($\alpha = p$)
  - 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 a drug has no effect when the drug is actually effective. $\beta = 0\%$ to 20%
  - Pregnancy test shows negative, but in realty pregnant
  - Fire alarm does not sound, but there is a fire
  - Innocent verdict, but actually guilty

This is the p value, typically .05

Types of Error

- Type 1 error (α): *False Positive*: Falsely concluding a drug is effective when the drug actually has no effect. α = 0% to 5% ($\alpha = p$)
  - 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 a drug has no effect when the drug is actually effective. β = 0% to 20%
  - Pregnancy test s… negative, but in realty pregnant
  - Fire alarm … ire
  - Innocent ve…

This relates to Power

Types of Error

Measured / Perceived		Reality: True	Reality: False
	True	Correct ☺	Type I False Positive
	False	Type II False Negative	Correct ☺

Adjusting Power

- **Power of a test** $(1 - \beta)$: **the probability of correctly concluding the drug is effective when it is actually effective.** ($\beta = 0$ to $.20$)
  - $(1 - \beta) = (1 - .20) = .80$
  - $(1 - \beta) = (1 - .10) = .90$
  - $(1 - \beta) = (1 - .05) = .95$
  - Increasing power

Adjusting Power

- **Power of a test** $(1 - \beta)$: **the probability of correctly concluding the drug is effective when it is actually effective.** ($\beta = 0$ to $.20$)
  - $(1 - \beta) = (1 - .20) = .80$
  - $(1 - \beta) = (1 - .10) = .90$
  - $(1 - \beta) = (1 - .05) = .95$
  - Increasing power

Decreasing beta, increases Power and the probability of concluding drug is effective when it is actually effective

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

- Reducing the risk of a Type I error – Reduce the significance level
  - $p = .05 \rightarrow p = .025$
  - 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
    - Reducing Type I error (Reducing odds of a false positive, but increasing odds of a false negative).
      - Cancer treatment – Important to verify treatment actually works. A false positive could be disastrous in treating patients (thinking it helps when it does nothing.) $p = .025$
      - Training program for dyslexia – Better to show any improvement in performance. $p = .05$.

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

- Reducing the risk of a Type II error – Decrease β increases power (1 – β)
  - $(1 - \beta) = (1 - .20) = .80 \rightarrow (1 - \beta) = (1 - .10) = .90$
  - Lowers the chance of a false negative (Increased sample size = increased probability of finding a significant difference)
  - Increases chance of a false positive
    - New education program being considered for all public schools in the U.S. A false positive would run the risk of wasting dollars on a program that does not improve educational system
  - Increased sample size is not always possible due to cost, time and resource restraints.

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

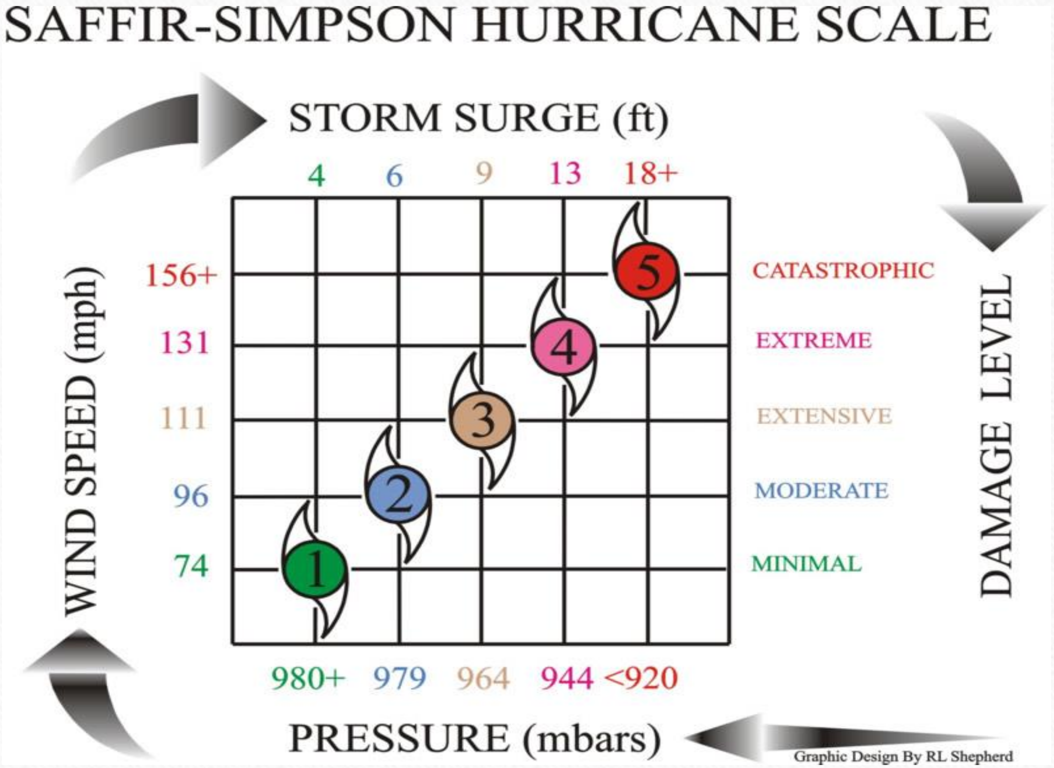
- Researchers can choose to adjust variables when a Type I or Type II error is preferred.
  - Prefer Type I error
    - Cancer diagnosis
    - Fire Alarm
  - Prefer Type II error
    - Judiciary system

Beyond *p-value*, what is important? Effect size!

- Effect size – Measurements that tell us the relative magnitude of the experimental treatment.
  - Tell us the ***size*** of the experimental ***effect***
- Allow us to compare the magnitude of experimental treatments from one experiment to another

Beyond *p-value*, what's important? Effect size!

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



Common Indices of Effect Size

- Comparison Studies
  - Cohen's d
  - Odds ratio (OR)
  - Relative risk or risk ratio (RR)
- 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

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

- 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)

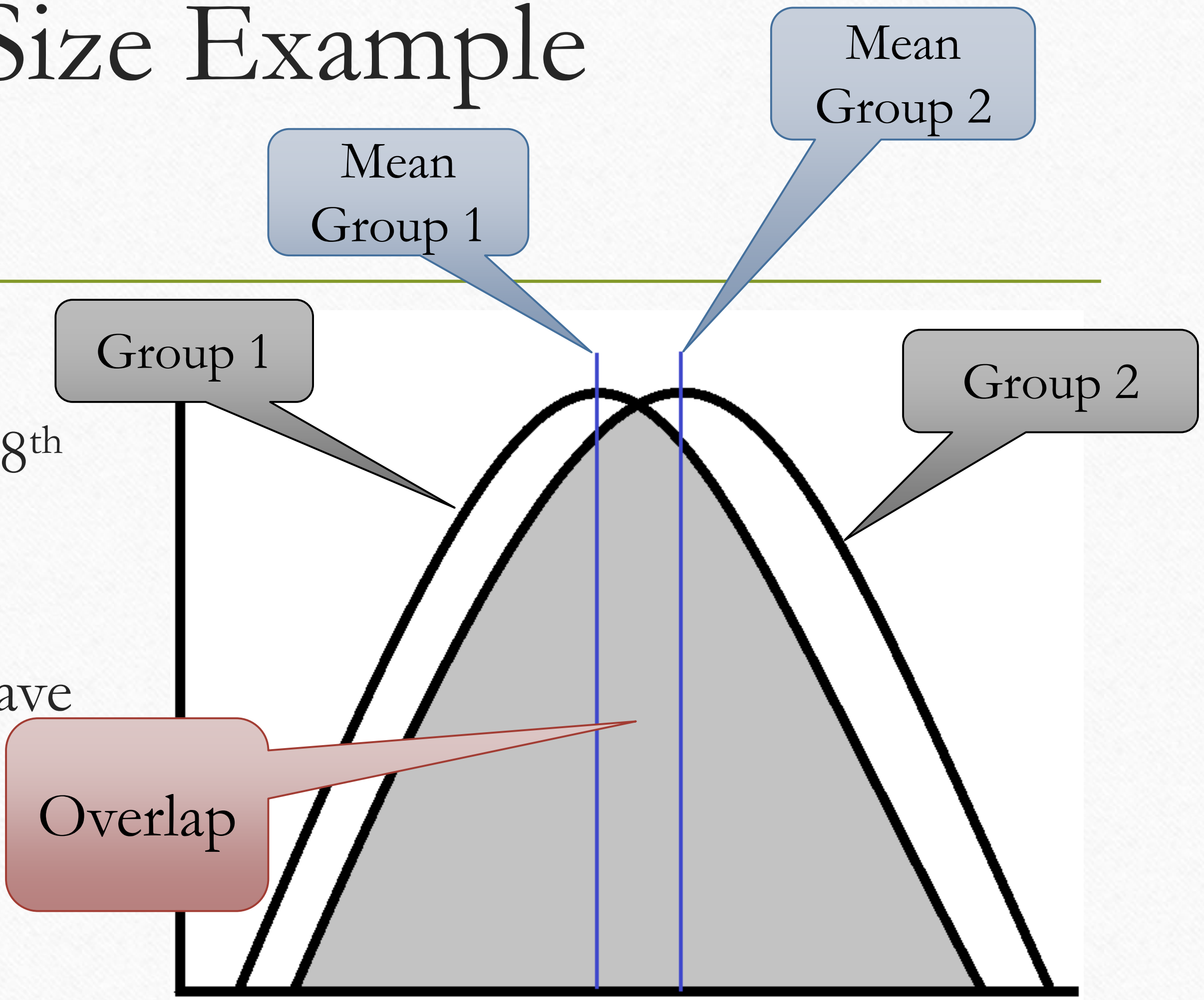
Cohen's *d*

- 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.

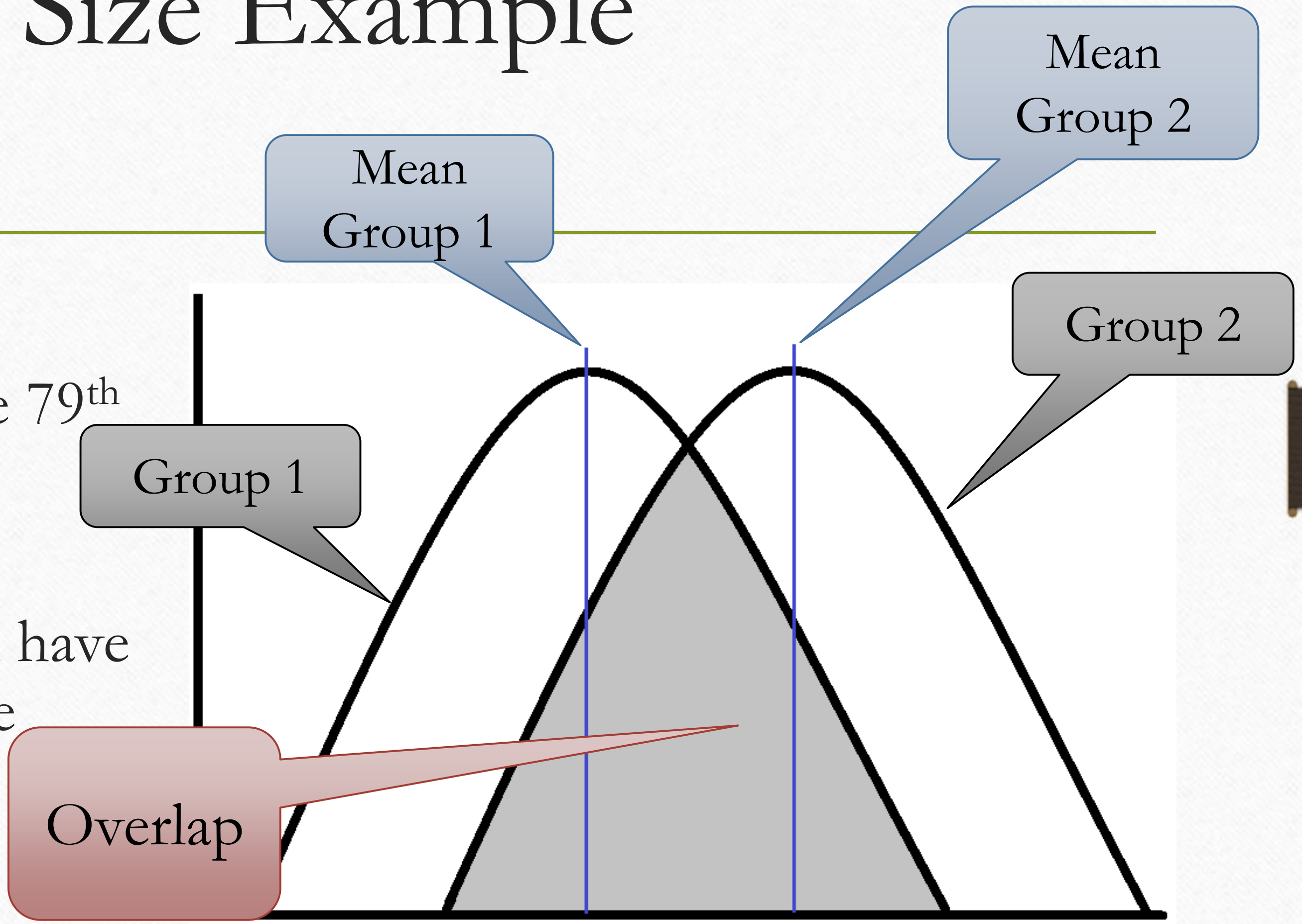
Effect Size Example

- Effect size = 0.2
- The mean of Group 2 is at the 58th percentile of Group 1
- 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

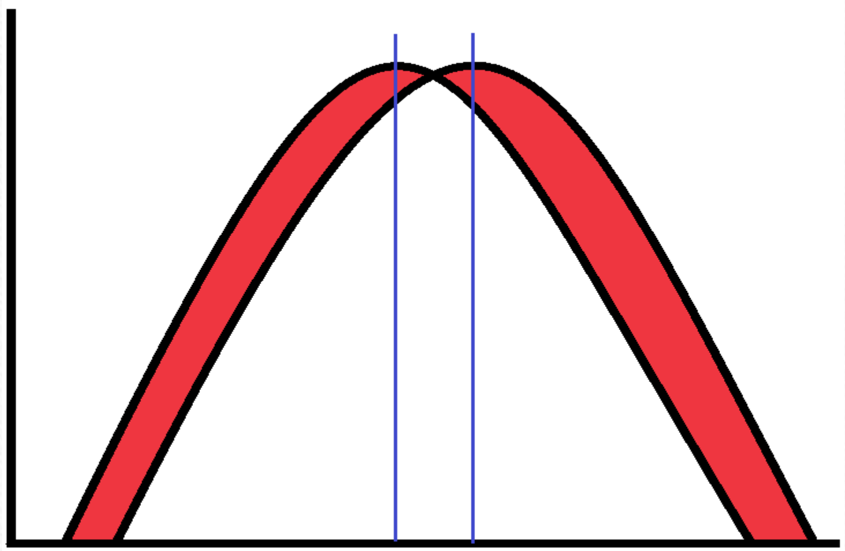
- Effect size = 0.8
- The mean of Group 2 is at the 79th percentile of Group 1
- Someone in Group 2 with an average score (ie, mean) would have a higher score than 79% of the people in Group 1
- 53% Overlap of participants



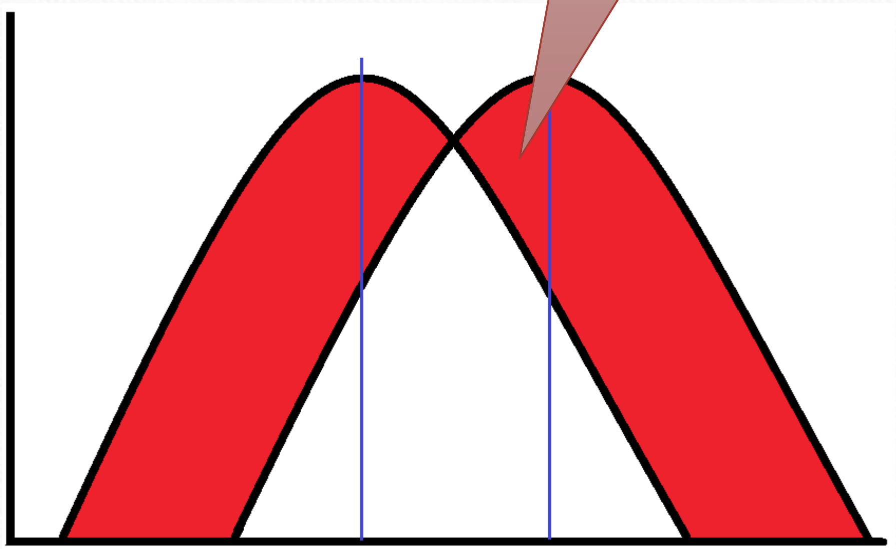
Effect Size Example

As Effect Size increases, magnitude of difference increases

Effect Size = 0.2



Effect Size = 0.8



Eta-squared (η^2)

- Eta-squared is a measure of effect size typically for use in ANOVA
- Proportion of variance in Y explained by X
- Interpret η^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.***

Odd Ratio (OR)

- Appropriate when both variables are binary
- Research example - Influence of positive priming on passing a class
  - 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 $6/1 = 6$)
  - Effect size computed by noting the odds of passing in Treatment group are three times higher than the Control group ($6 \div 2 = 3$)
  - OR = 3 (Note: not comparable to Cohen's d)

Relative Risk (RR) aka risk ratio

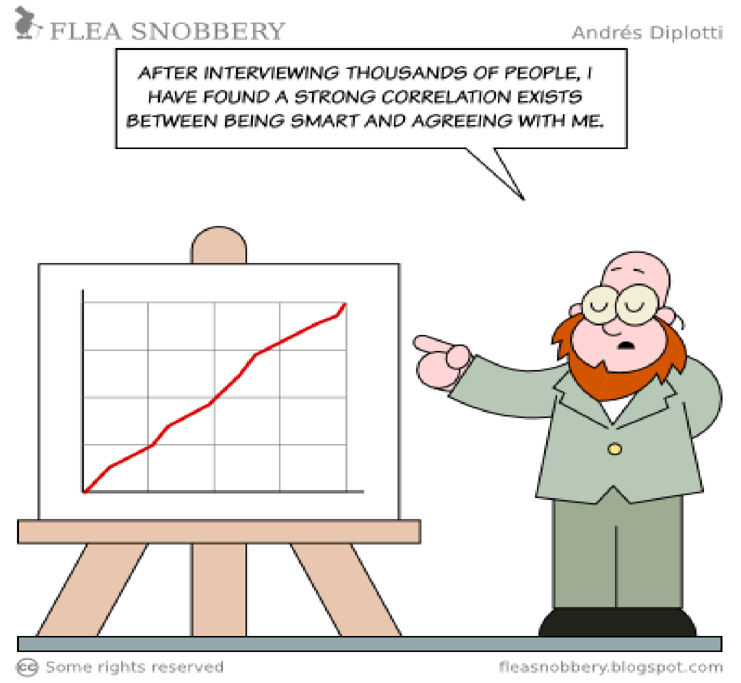
- The risk (probability) of an event relative to some independent variable
  - Different from odds ratio (OR) in that it compares probabilities instead of odds
- Research example - Influence of positive priming on passing a class
    - Control: 2 students pass for every 1 that fails
      - Probability 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)

Correlation

- Correlation analysis is used when you have measured two continuous variables and want to quantify how consistently they vary together
- The stronger the correlation, the more likely to accurately estimate the value of one variable from the other

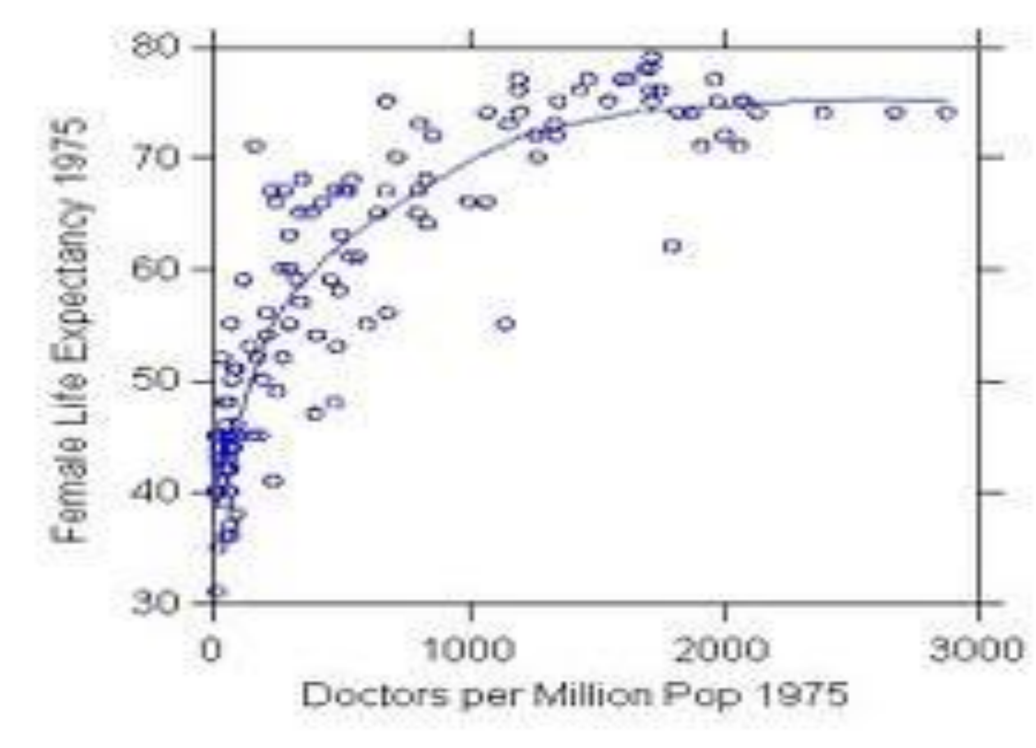
Patterns of Correlation

- Linear correlation
- Curvilinear correlation
- No correlation
- Positive correlation
- Negative correlation



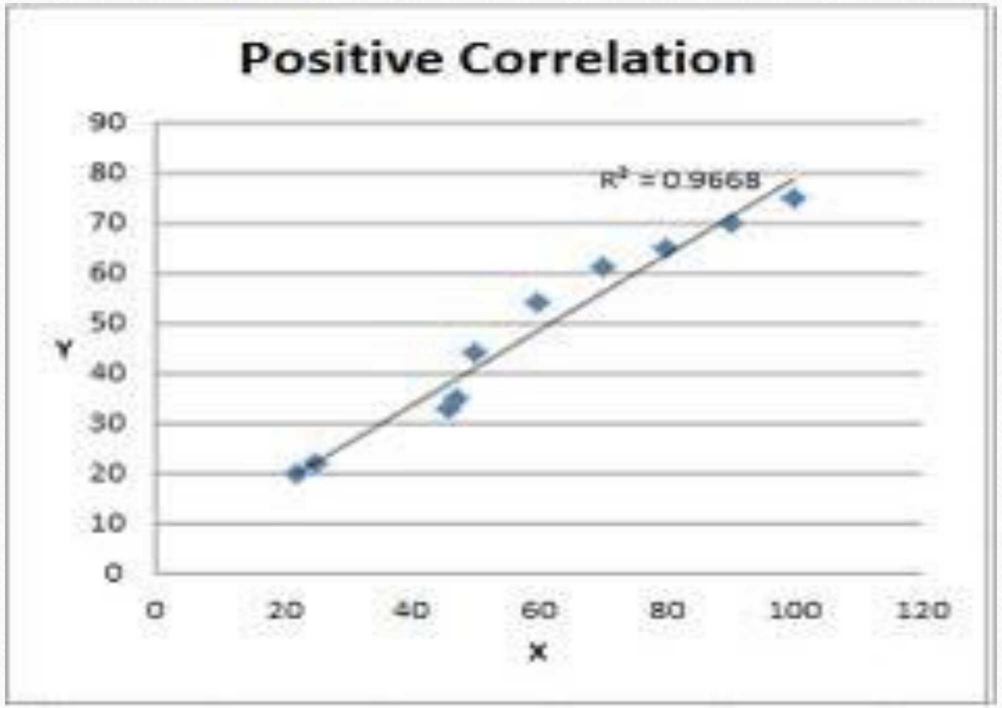
Curvilinear Correlation

- Female life expectancy and Doctors per million of population
- Test performance and average hours of sleep per night



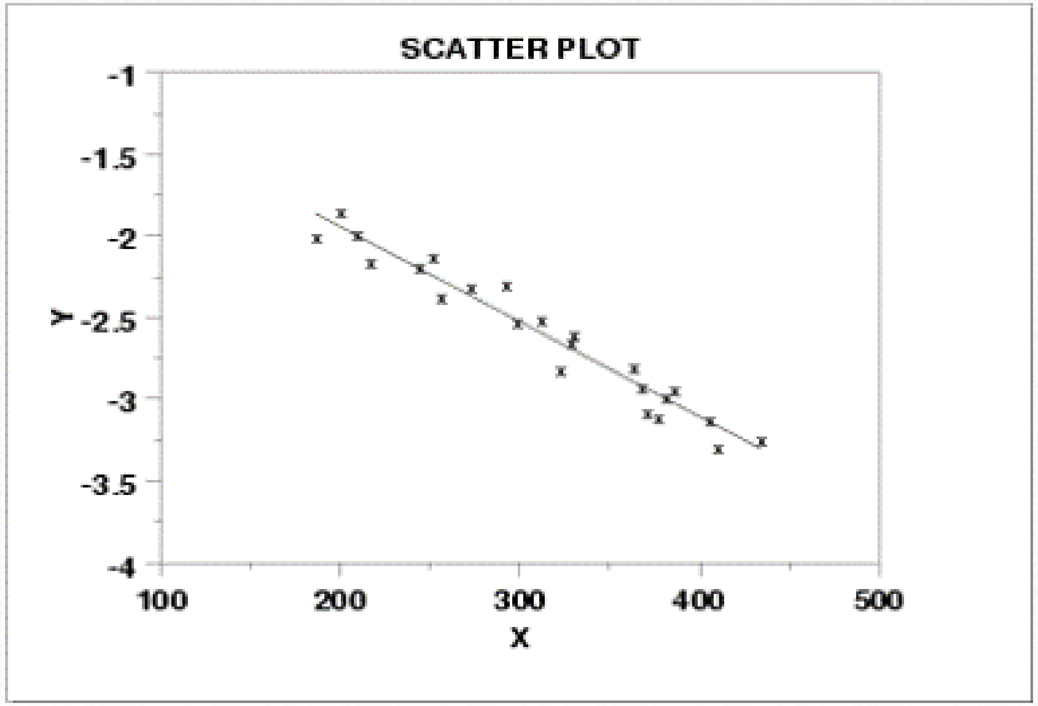
Positive Correlation (+1)

- Income and education
- Weight and height
- Extraversion and size of social network



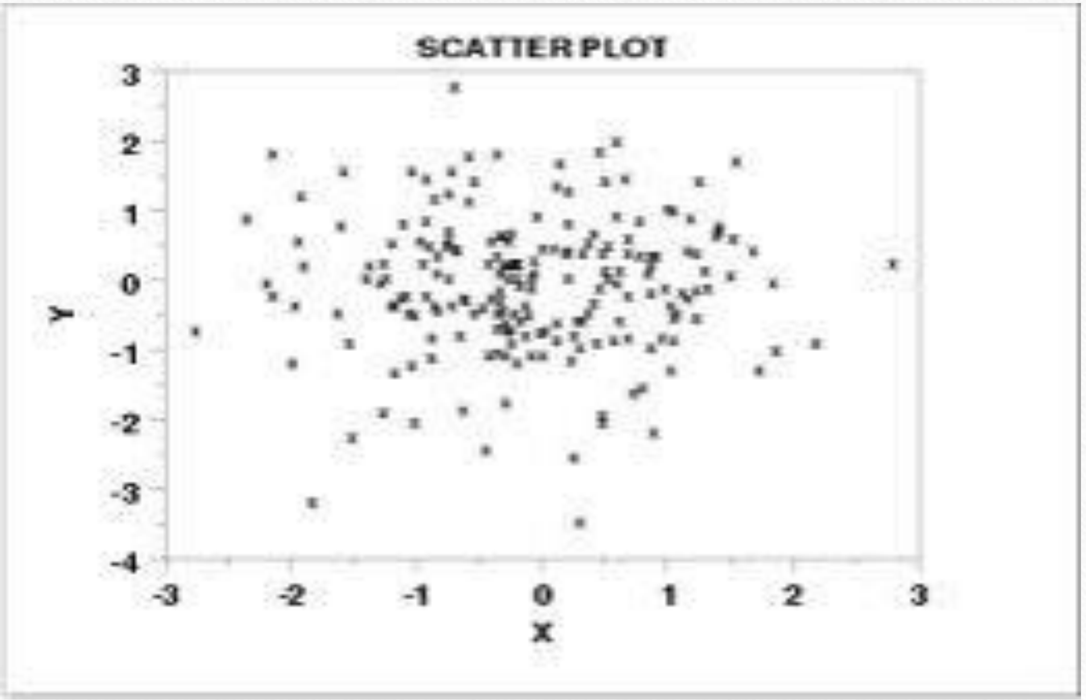
Negative Correlation (-1)

- Time incarcerated and education
- Temperature and amount of clothing
- Speed and average miles per gallon



No Correlation

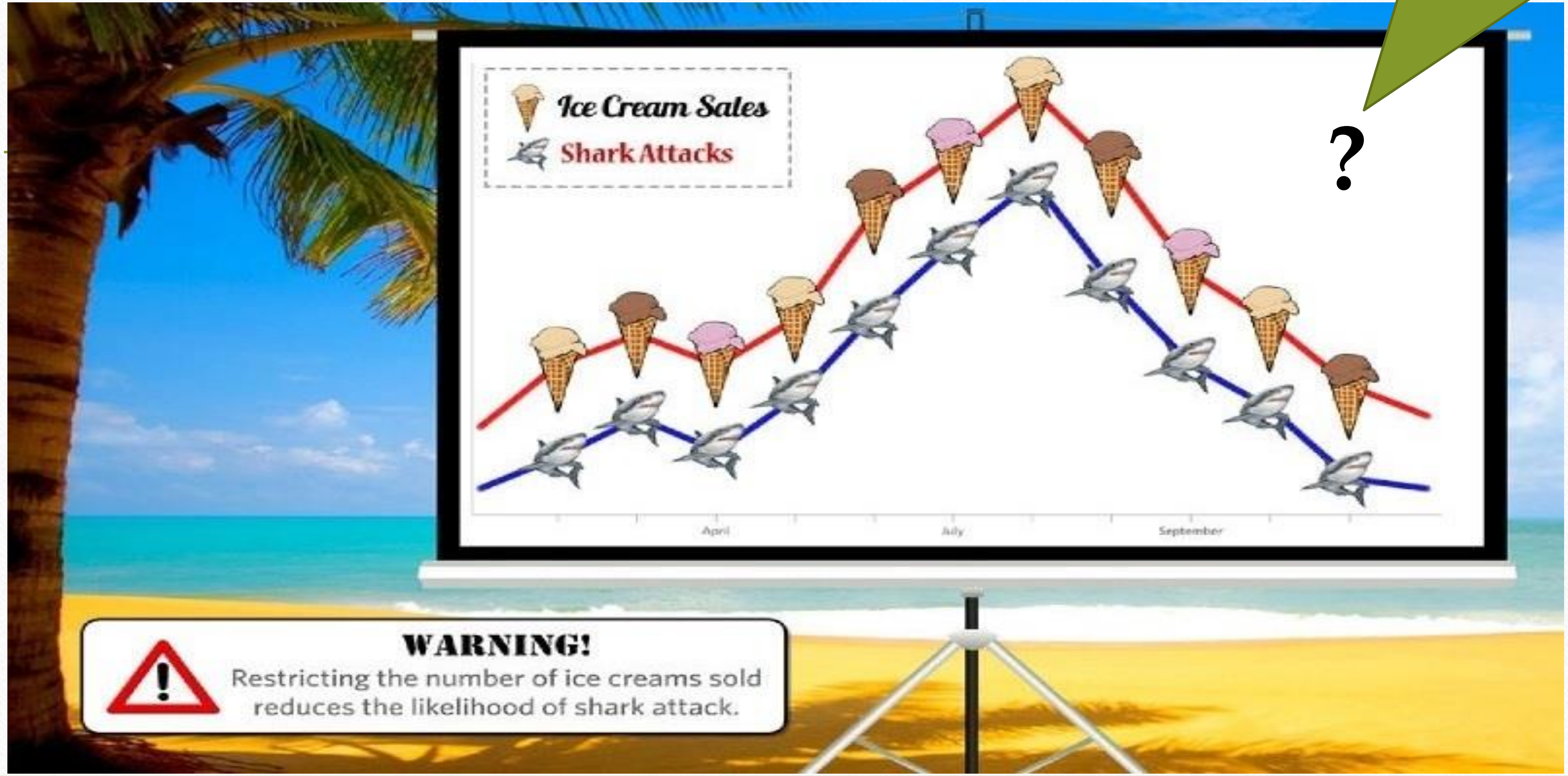
- Favorite color and IQ
- Hair color and personality
- Height and education curriculum



Correlation ≠ Causation

Confounding Variable

?



Confounding Variable Example

- 1940 polio epidemic
- Increased polio diagnoses during summer months
- Children eat a lot of ice cream during the summer months
- A connection was made…

Confounding Variable Example

- Dr. Benjamin Sandler published a paper "The production of neuronal injury and necrosis with the virus of poliomyelitis in rabbits during insulin hypoglycemia" The American Journal of Pathology.

- Results of article - cut out ice cream.

Confounding Variable Example

- Reality
  - Lack of harmless immunizing infections during infancy due to better sanitation that attributed to the epidemic.
  - Not ice cream – flush toilets

Linear Regression

- Attempt to predict one variable from another variable
  - Criterion variable (DV) = Exam score
  - Predictor variable (IV) = Score on anxiety scale
  - Want to predict exam score from anxiety scale score

Linear Regression

- Linear Regression examples
  - Predict adult height and weight from height and weight at 2 years old
  - Predict first year college success from SAT
  - Predict first year law school success from LSAT

Linear Regression

- Linear Regression examples
  - Predict average rain fall from geographical information
  - Predict criminal behavior from housing foreclosure rate
  - Predict job performance from letters of recommendation

Chi-Square (χ^2)

- Compare expected value with observed value (categorical variables)
  - Sample of 100 students
  - 30 males, 70 females
  - 10 Asians
  - It would be expected that among the Asians, 3 would be male and 7 female
  - If this was the case, χ^2 would have a p value above .05

Chi-Square (χ^2)

- If there is a large discrepancy between the observed values and the expected values
  - X^2 statistic would be large, suggesting a significant difference
  - p value of .05 ($p < .05$) demonstrates significant difference

Chi-Square (χ^2)

- Examples
  - Compare ethnicity percentiles of a school based on local ethnic population
    - City of Boston
    - Harvard
  - Compare ethnic population at FAU vs. ethnic population in south Florida

The *t-test*

- ***t-test***: Compare two means, see if there is sufficient evidence to infer that the means of corresponding population distributions also differ
    - Does treatment A yield a higher rate of recovery than treatment B?
    - Does one advertising technique produce higher sales than another technique?
    - Do men or women score higher on a measure of empathic tendency?

The *t-test*

- The key word is two: *t tests* always compare two different means or values
  - One independent variable
    - Males vs. females,
    - Treatment A vs. treatment B
    - Married vs. unmarried

One-Way Analysis of Variance (ANOVA)

- ANOVA most easily explained by contrasting it with *t tests*
  - *t tests* compare two distributions
    - Male vs. female on a quiz
  - ANOVA able to compare many distributions
    - Native Americans, African Americans, Hispanics, and Asians on a quiz

Two-Way ANOVA

- One-Way ANOVA compares multiple groups on a single variable
  - IV = Ethnicity, DV = quiz score
- Two-Way ANOVA compares multiple groups (multiple IV's)
  - IV(1) = Ethnicity, IV(2) = gender, DV = quiz score

Two-Way ANOVA (cont.)

- Two-Way ANOVA allows us to explore:
  - A main effect for ethnicity
    - Does ethnicity have an effect on quiz scores?
  - A main effect for gender
    - Does gender have an effect on quiz scores?
  - A ethnicity by gender interaction
    - Does the interaction between ethnicity and gender have an effect on quiz scores?

Two-Way ANOVA (cont.)

- Interaction example
  - Antipsychotic drug Thorazine
    - Reduces symptoms such as delusions and hallucinations
  - Alcohol
    - Typical depressant
  - Thorazine and alcohol interaction
    - Breathing difficulties – Potentially fatal

Two-Way ANOVA (cont.)

- Interaction example
  - Headache
    - Physiological discomfort
  - Kids screaming
    - Increased irritability
  - Headache and kids screaming interac
    - Loss of temper

Two-Way ANOVA

- Misc. Terminology
  - Within group effects
    - Effects within a group such as ethnicity or gender
  - Between group effects
    - Between groups such as interaction between ethnicity and gender

Tests and Measures

- Scales of Measurement
- Mean, Median, Mode
- Variance and Standard Deviation
- z-scores

Scales of Measurement

	Magnitude	Equal Interval	Absolute Zero
<u>N</u>ominal			
<u>O</u>rdinal	X		
<u>I</u>nterval	X	X	
<u>R</u>atio	X	X	X

Standard Scores

- **Standard Score** – Characterizing the relative position of scores in a distribution
- **Z-score** – a standard score in a normal distribution
  - Z score = -1 ~ Score is 1 standard deviation below the mean
  - Z score = 0 ~ Score is equal to the mean score
  - Z score = 1 ~ Score is 1 standard deviation above the mean
  - Z score = 1.5 ~ Score is 1.5 standard deviations above the mean

Z Score Examples

- Mean Score = 80, Standard Deviation = 5
  - Z score = 1 ~ Score is 85 (1 SD above the mean)
  - Z score = -1 ~ Score is 75 (1 SD below the mean)
  - Z score = 1.5 ~ Score is 87.5 (1.5 SD above the mean
  - Z score = -0.5 ~ Score is 77.5 (.5 SD below the mean)

End of Review

Michael A. DeDonno Ph.D.

www.michaeldedonno.com