

Survey Data Analysis & Statistics

- Many Statistics are Available
- Which one to use is based on:
 - ☞ Number of variables analyzed
 - ☞ Level of measurement
 - ☞ Research questions
 - ☞ Your knowledge, time, budget...

Always use the most powerful statistic at the lowest level of measurement.

Good Guide: Table on Page 149

Outline:

1. Stats for **One Variable**
...at Nominal, ordinal, interval, ratio.
2. Stats for **Two Variables**
...at Nominal, ordinal, interval, ratio.
3. Stats for **Three or more Variables**
...at Nominal, ordinal, interval, ratio.

Statistics for Analyzing...

One Variable

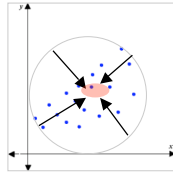
Statistics for Analyzing **One** Variable

1. Measures of Central Tendency
2. Measures of Dispersion
3. Frequency Distributions
4. Rank Ordering Data

Statistics for Analyzing **One** Variable

1. Measures of Central Tendency

- Mode: **nominal** and above
- Median: **ordinal** and above
- Mean: **interval** and above



NOTE: Mean most powerful, then Median, then Mode.

Statistics for Analyzing **One** Variable

1. Measures of Central Tendency

- Mode: **nominal** and above
 - Score over 100 = {1, 2, 3, 3, 4, 4, 4, 5, 8, 20, 99}
 - Mode = 4 (most reoccurring number)
- Median: **ordinal** and above
 - {1, 3, 5, 8, 98} then Median = 5 (center point)
 - {1, 3, 5, 6, 8, 9} then Median = 5.5 (average cp)
- Mean: **interval** and above
 - {2, 4, 6, 8, 10}
 - Mean = $(2+4+6+8+10)/5 = 30/5 = 6$ (average)

Statistics for Analyzing **One** Variable

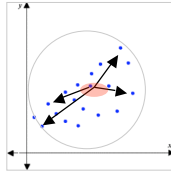
1. Measures of Central Tendency

- They have Problems!
 - Biased and do not give an accurate summary of the values
 - Mean = 50
 - for {48, 49, 51, 52}
 - and for {0, 1, 99, 100}
 - We need more tools to help interpret the info

Statistics for Analyzing **One** Variable

2. Measures of Dispersion

- **Range:** ordinal and above
- **Variance:** interval and above
- **Standard deviation:** interval and above



NOTE: Variance most powerful

Statistics for Analyzing **One** Variable

2. Measures of Dispersion

- **Range** (ordinal and above)
 - Difference between highest and lowest
 - {4, 20, 30, 60, 100}
 - Range = 100-4 = 96
 - Problems! May be misleading due to extreme cases and because it doesn't take the middle numbers into consideration:
 - {1, 48, 49, 50, 51, 52, 53, 100}
 - {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 100}
 - {1, 96, 97, 98, 99, 100}

Statistics for Analyzing **One** Variable

2. Measures of Dispersion

- **Variance** (interval and above)
 - Dispersion around the mean
 - How far the scores are spread from/around the mean
 - Measured in square units (S^2)
 - **Large (S^2) = numbers are more spread out**
 - **Small (S^2) = numbers are less spread out**

$$S^2 = \frac{\sum(X - M)^2}{N}$$

Statistics for Analyzing **One** Variable

2. Measures of Dispersion

- **Standard deviation** (interval and above)
 - Same concept as variance
 - Square root of the variance
 - Standardized dispersion from/around the mean
 - Measured in the same units of the variable
 - **Large (S) = numbers are more spread out**
 - **Small (S) = numbers are less spread out**
 - Demonstration of Normal Curve
<http://www.stat.sc.edu/~sweat/applets/normaldemo1.html>

$$S.D. = \sqrt{S^2} = S$$

Statistics for Analyzing **One** Variable

3. Frequency Distribution: **nominal & above**

- Shows two main things:
 - what frequency each response represents
 - what percentage of the total each response represents

Highest degree received

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	LT High school	400	27.2	27.2	27.2
	High school	764	51.9	52.0	79.2
	Junior college	54	3.7	3.7	82.9
	Bachelor	175	11.9	11.9	94.8
	Graduate	77	5.2	5.2	100.0
	Total	1470	99.8	100.0	
Missing	Missing data	3	.2		
	Total	1473	100.0		

Statistics for Analyzing **One** Variable

3. Frequency Distribution: **nominal & above**

- Shows two main things:
 - what frequency each response represents
 - what percentage of the total each response represents
- The Valid Percent Column
 - When to ignore missing values?
 - depends on what "missing" means
 - because of a contingency question? (use valid percent)
 - because there was no answer? (use valid percent)
 - because of "don't know"? (up to you what to use)
 - Be consistent

Statistics for Analyzing **One** Variable

3. Frequency Distribution: **nominal & above**

Highest degree received

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	LT High school	400	27.2	27.2	27.2
	High school	764	51.9	52.0	79.2
	Junior college	54	3.7	3.7	82.9
	Bachelor	175	11.9	11.9	94.8
	Graduate	77	5.2	5.2	100.0
	Total	1470	99.8	100.0	
Missing	Missing data	3	.2		
Total		1473	100.0		

Statistics for Analyzing **One** Variable

3. Frequency Distribution: **nominal & above**

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	5	100.0	100.0	100.0
	4	0	0.0	0.0	100.0
	3	0	0.0	0.0	100.0
	2	0	0.0	0.0	100.0
	1	0	0.0	0.0	100.0
Missing	Missing data	0	0.0		
Total		5	100.0		

Statistics for Analyzing...

Two Variables

Statistics for Analyzing **Two** Variables

📊 **Statistical Measures of Association**

📊 **Testing for Difference**

Statistics for Analyzing **Two** Variables

📊 **Statistical Measures of Association**

- Indicate the presence or absence of an association between **two or more** variables

- Tells about the association:
 - Direction & strength for **ordinal, interval & ratio**
 - Only strength for **Nominal** (no direction!)

- Measures Linear Associations only- CAUTION!

Statistics for Analyzing **Two** Variables

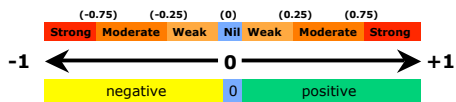
Statistical Measures of Association

- Magnitude of association range {-1 to +1} for **ordinal** and above
 - **+1** strong positive relationship
 - **-1** strong negative relationship
 - **0** No relationship
- For **nominal** variables {0 to +1}
 - **1** strong relationship
 - **0** No relationship

Statistics for Analyzing **Two** Variables

Statistical Measures of Association

- Magnitude of association range {-1 to +1} for **ordinal** and above
 - **+1** strong positive relationship
 - **-1** strong negative relationship
 - **0** No relationship



Statistics for Analyzing **Two** Variables

Statistical Measures of Association

- For **nominal** variables {0 to +1}
 - **1** strong relationship
 - **0** No relationship



Statistics for Analyzing **Two** Variables

Statistical Measures of Association

- Magnitude of association range $\{-1 \text{ to } +1\}$ for **ordinal** and above
 - **+1** strong positive relationship
 - **-1** strong negative relationship
 - **0** No relationship
- For **nominal** variables $\{0 \text{ to } +1\}$
 - **1** strong relationship
 - **0** No relationship

Statistics for Analyzing **Two** Variables

Statistical Measures of Association

For **Nominal** variables:

- Phi: if variables are **both nominal** (range 0 to +1)
- Cramer's V: if **one nominal one ordinal** (range 0 to +1)

Statistics for Analyzing **Two** Variables

Statistical Measures of Association

For **Ordinal** variables:

- Kendall's tau (range -1 to +1)
- Spearman's rho: (range -1 to +1)

Statistics for Analyzing **Two** Variables

📊 Statistical Measures of Association

For **Interval** or above:

- Product-Moment Correlation (r) or Pearson's Correlation Coefficient (r) developed by Karl Pearson and referred to as – **Pearson's r**
 - Interval or ratio level (range -1 to +1)
 - P.144 – guide to interpreting Pearson's r

Statistics for Analyzing **Two** Variables

📊 Testing for Difference:

- **T-tests compare means**
- **Others...**

Statistics for Analyzing **Two** Variables

📊 Testing for Difference:

- **T-tests compare means (Interval or above)**
 - It tells if **two** means are significantly different
 - You can compare means on the same questions across groups OR
 - On questions with identical (or very similar) response categories in a single instrument
 - Why? It is the same measure and scale
 - More about t-tests later (experiments)

Statistics for Analyzing...

Three of More Variables

Statistics for Analyzing **Three or more** Variables

- 📊 **Multivariate analysis**
- 📊 **Factor Analysis***
- 📊 **Multiple Regression***
- 📊 **Analysis of Variance or ANOVA***
- 📊 **Analysis of Covariance or ANCOVA**
- 📊 **Many more...**

Statistics for Analyzing **Three or more** Variables

- 📊 **Factor Analysis (Interval or above)**
 - Is a form of data reduction where SPSS groups/combines different questions into one variable
 - Basically grouping certain variables

Statistics for Analyzing **Three or more Variables**

- **Multiple Regression (Interval or above)**
 - Tool used to predict how three or more variables will influence another variable
 - (Can also be used with **two** variables)
- **ANOVA (Interval or above)**
 - Analysis of variance
 - Similar to t-test where it tells if 3 or more means are significantly different

Survey Data Analysis & Statistics

- **Many Statistics are Available**
- **Which one to use is based on:**
 - ☞ Number of variables analyzed
 - ☞ Level of measurement
 - ☞ Research questions
 - ☞ Your knowledge, time, budget...

Always use the statistic at the lowest level of measurement and the most powerful

Good Guide: Table on Page 149

Survey Data Analysis & Statistics

- **Always use the statistic at the lowest level of measurement and the most powerful**
- **Example:**
 - Score of quizzes in class: {8, 7, 3, 1, 1}
 - Mean= 4
 - Median= 3
 - Mode= 1
 - Which statistic do we use?

Coding/Processing Survey Data

- Pre-coding and Computers
- The Master Code Book (p. 115)
 - Close-ended questions
 - Open-ended questions
 - Create a list of the responses and find patterns in them
 - Apply codes to the patterns by creating categories
 - Not too many categories
 - Must follow mutually exclusive and exhaustive guidelines

Coding/Processing Survey Data

- Written Instructions and Coder Training Produce Reliable Coding
- Become familiar with the stat. software (SPSS)
- Cleaning and Recoding May Be Necessary Before Final Computer Runs
 - *Cleaning "Dirty" Data*, ALWAYS necessary
 - Ethics! Don't cook the numbers...

Coding/Processing Survey Data

- *Recoding Decisions*
 - Often have too much detail
 - Convert Ratio/interval level variables to ordinal
 - Not enough responses from one group or another, weighting...
 - Always based on the distribution of data and your research questions

Coding/Processing Survey Data

- A Frequency Printout
- Interpreting the Table (p. 125)
 - The Valid Percent Column
- When to ignore missing values
 - Depends on what "missing" means
 - because of a contingency question? (use valid percent)
 - because there was no answer? (use valid percent)
 - because of "don't know"? (up to you what to use)
 - Be consistent
