Using Excel in Educational Research: An Introduction

Brittney Hernandez

Housekeeping

- Start recording
- Presentation: https://drive.google.com/drive/folders/1SDH9OZJhiSZSAOeSaIyUxxDAEDL4WMe-?usp=sharing
- Sign in: https://docs.google.com/forms/d/e/1FAIpQLSdUzlDSTxvi-UTHE95oiI-bBhqZErBlbaQyrcSFQkdm5VHySw/viewform?usp=sf_link

Agenda

- Data Collection
- Data Cleaning in Excel
- Descriptive Analysis in Excel
- Statistical Tests in Excel

Introduction

- Why do we collect data?
  - To gain information!
- What types of data can we collect?
  - Qualitative: verbal information
  - Quantitative: numbers and quantities

Introduction

- How do we collect data? Some typical steps include:
  - Survey or Instrument Design
  - Data Collection
  - **Data Entry** or Transcription
  - **Data Cleaning**
  - **Data Analysis (i.e., descriptives and inferential)**

Hands-on Activities

Activity 1: Setting up a data set (Data Entry)

Activity 2: Handling imperfect data (Data Cleaning)

Activity 3: Getting to know your data (Data Analysis: Descriptives)

Activity 4: Making inferences from your data (Data Analysis: Inferences)

Data Collection

- There are two main ways to get data:
  - Collect your own data
  - Acquire an existing dataset
- If you are collecting your own data, you'll want to set up the variables in that will go in your dataset before entering and cleaning data
- If you acquire an existing dataset, it is likely already set up (and hopefully cleaned!)

Data Entry

- I’ve collected my own data, how do I create a dataset?
  - Two main components (rows and columns):
    - Individual: any person who is providing data
    - Variables: any item, factor, or condition that can be controlled or changed
- E.g., How might we create a variable for an item that says, “select all that apply?”
- Recoding and dummy coding categorical variables

Demo

Importing an Existing Dataset:

- Delimitations
  - File types: .csv, .txt
- Transposing datasets:

  Home > Paste drop down menu > Transpose

- Modifying data:

  Data > Text to Columns

  Text to Columns based on width

Activity 1: Importing a data set

In the Google Drive folder there is a file called **spi_matches_latest.txt**. This data contains club soccer prediction data from fivethirtyeight.com. With your group, attempt to import the dataset into Excel.

Once the data is in Excel, look at the file:

- What does each column represent?
- What does each row represent?
- What does each cell represent?

Demo

- Freezing panes
  - View > Freeze Top Row
- When you insert a new row or column it will insert wherever the selection is
  - It'll look like it inserts to the left or above
- Formatting Data

Demo

- Format the date
- Sortingdata
  - Data > sort > sort by group > sort on value > A to Z
- Look at unique values of a variable
  - Copy to a new sheet > Data > Remove duplicates

If Statements

=IF(Something is True, then do something, otherwise do something else)

=IF('Sheet1'!B2="Group 1", "1", IF('Sheet1'!B2="Group 2", "2",
IF('Sheet1'!B2="Group 3", "3", IF('Sheet1'!B2="Group 4", "4",
IF('Sheet1'!B2="Group 5", "5", IF('Sheet1'!B2="Group 6", "6"))))))

Activity 2: Setting up a data set

In order to do any quantitative analyses we need numeric data. So any responses in our dataset that have names instead of numbers need to be changed.

- With your group, attempt to recode Season to 1, 2, and 3

Demo

- Copy and paste values, formulas, etc
- Countif statements:
  - Count if an exact value
    =COUNTIFS(C2:C56, "22")
  - Count if a value appears at all
    =COUNTIFS(D2:D56, "*Cooking*")

Countif Statements

- Some useful uses of countifs for finding imperfect data:

	Command	**Example**
Count numeric values	=countifs(range, “argument”)	=countifs(E2:E25, “2”) =countifs(E2:E25, “=2”) =countifs(E2:E25, “<2”) =countifs(E2:E25, “>2”) =countifs(B2:B5,"=2",C2:C5,"=Yes")
Count text values	=countifs(range, “argument”) =countifs(range, “*argument*”)	=countifs(E2:E25, “Cooking”) =countifs(E2:E25, “*Cooking*”)
Count blank values	=countifs(range, “”)	=countifs(E2:E25, “”)

- The Find and Replace feature can also be useful for replacing missing values

Demo

- Count missing data
  =COUNTIFS(E:E, “”)

  =COUNTBLANK(E:E)
- Conditional formatting
  - Select range
  - Home > conditional formatting > greater than
- Find and replace missing values Edit > find > leave blank > replace with -999

Data Cleaning: Handling imperfect data

At some point after receiving data, likely either during data entry or in the descriptives, you might find some responses that stray from the prompt(s) you gave. For example....

- Someone marked two answers, what do I put?
- There's missing data, how do I code it?
- There's an impossible answer, how do I code it?
- Someone left notes about a response, what do I do with that information?

What is the issue?	During what stage might you see this issue?	How should you handle this issue?
Someone marked two answers	Data entry	Randomly choose one of the two answers (e.g., flip a coin and H=higher number, T= lower number). This can be done pre- or post- data entry
Someone left notes about a response	Data entry	Don't include it unless it may change the response
There is missing data	Data entry or descriptive analysis	Mark it as a response that is very different from any possible values (e.g., -999 is a conventional way to denote missing data)
There is an impossible answer	Data entry or descriptive analysis	If it is truly impossible (and not just an outlier) mark it as missing

Activity 3: Handling imperfect data

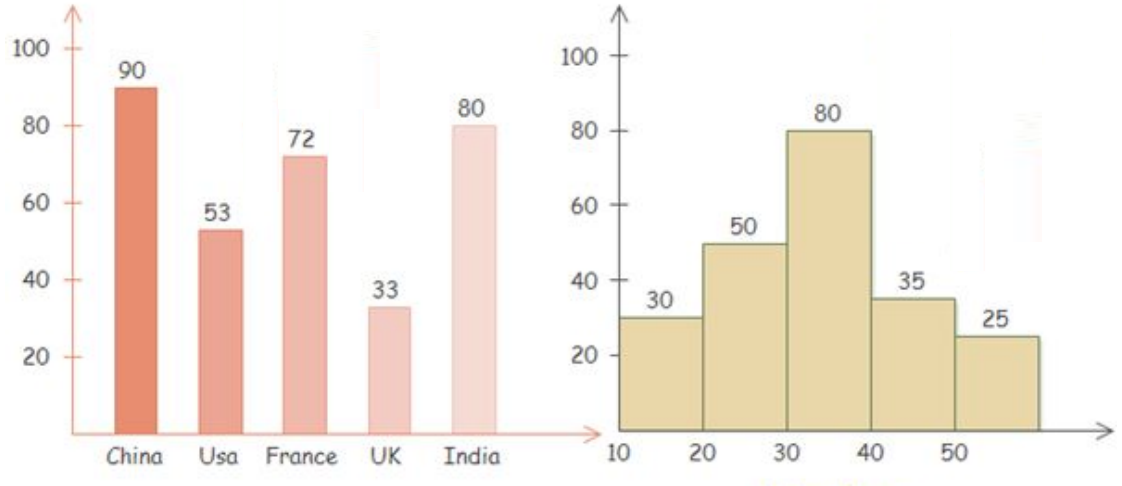
With your group, spi2 has missing values. How many missing values are there?

Getting to know your data

- Once you have a clean dataset and descriptive statistics will help you better understand it
- What types of information might you want to know?
  - Frequencies of responses
  - Central tendency (mean, median, mode)
  - Variability (variance, standard deviation, outliers, etc.)

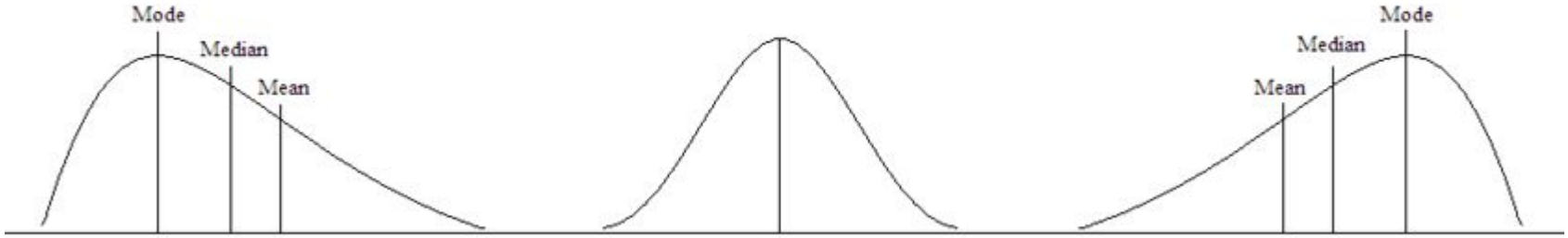
Descriptive Statistics

- Frequencies describe the number of occurrences of a given value
  - Relative vs cumulative frequency
  - Bar chart vs histogram



Descriptive Statistics

- Central tendency describes the
  - Mean: Sum of scores divided by number of scores
  - Median: Middle number in an ordered data distribution
  - Mode: Most common value in a distribution



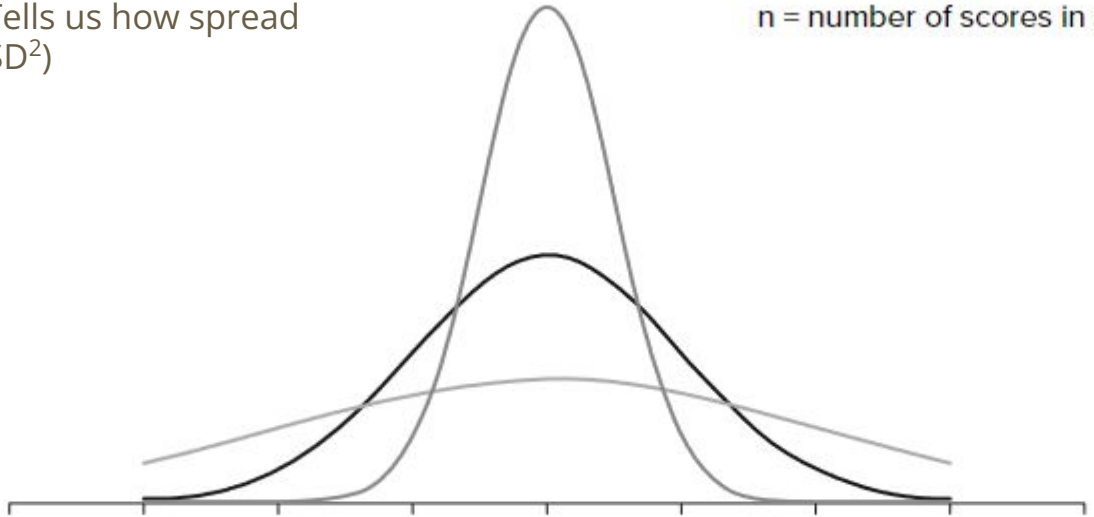
Descriptive Statistics

- Variability
  - Standard deviation (σ Represents the average (or standardized) distance of the scores from their mean
  - Variability (^{2}): Tells us how spread out our data is (SD^2)

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{n}}$$

where,

σ = population standard deviation
$\sum$ = sum of...
μ = population mean
n = number of scores in sample.



Activity 4: Getting to know your data

- Calculate summary statistics for the variables spi1 and importance1

Mean	= average ( )
Median	= median ( )
Mode	= mode ( )
Standard deviation	= stdev ( )

- The following link has more information about formulas in Excel

Demo

Charts

- Chart types
- Creating frequency charts
- Scatter plots

Activity 5: Getting to know your data

- In your group, create a bar chart of season
  - First need to create a chart of the frequency of each possible value
    - What are all the possible values of season?
    - Count how many were in each season?
    - Insert a bar chart
    - Which season did most of the entries come from?

Introduction to inferential statistics

For the purpose of this workshop we're going to think about variables as **categorical** or **continuous** (though it's still important to consider if the variables you're dealing with are nominal, ordinal, interval, or ratio).

- Continuous variables can be made categorical but categorical variables cannot be made continuous.

Introduction to inferential statistics

- Let’s take the following example: How does sleep (or lack of sleep) affect test scores.
- Ways we can measure sleep:
  - Continuous: Number of hours slept at night
  - Categorical (2-levels): High (over 8 hours) or Low (under 8 hours)
  - Categorical (3-levels): High (over 9 hours), Medium (6-9 hours), or Low (under 6 hours)
- Ways we can measure GPA
  - Continuous: raw GPA score
  - Categorical (2-levels): High (over 3.0) or Low (under 3.0)
  - Categorical (3-levels): High (over 3.5), Medium (2.5-3.5), or Low (under 2.5)

Inferential statistics

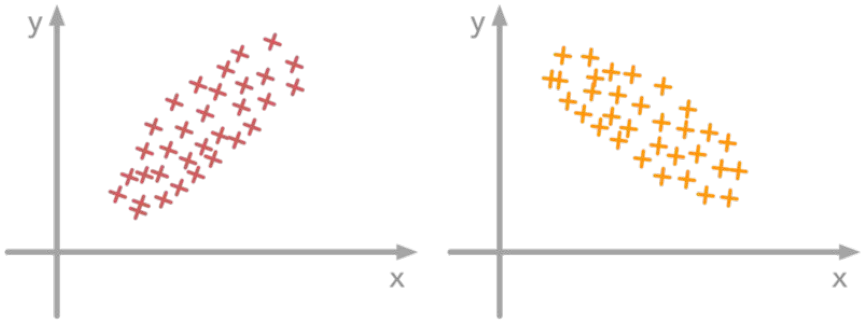
	Chi-Square	Correlation	Independent Samples T-Test	ANOVA
Predictor (IV)	Categorical	Continuous	Categorical (2 levels)	Categorical (2+ levels)
Outcome (DV)	Categorical	Continuous	Continuous	Continuous
Inference Made	The frequency of membership in a sub group	Strength and direction of the association	The mean difference between two groups across an outcome variable	The mean difference between two or more groups across an outcome variable
Sample Interpretation	There were more people who got high sleep and high test scores than low sleep and high test scores	The more hours of sleep a person gets, the higher their test score	The high sleep group has a higher mean test score than the low sleep group	The high sleep group has a higher mean test score than both the medium and low sleep groups

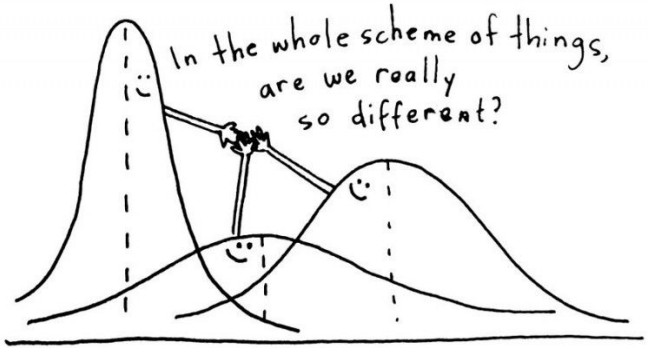
Interpreting inferential analyses

- What to include when interpreting analyses
  - Correlations
    - The magnitude of the correlation compared to “industry” standards (Cohen 1992)
      - Small: .1
      - Medium: .3
      - Large: .5
    - Create a scatterplot of the data
  - T-Tests
    - The result of the t-test (*t*=)
    - The p-value of the test (*p* =)
- A good resource for finding information about running inferential analysis
- A good resource for reporting statistics
- A good resource for formatting statistics in APA format

Inferential analysis in excel

Correlation	Independent Samples T-Test	ANOVA
● = correl(var1, var2)	● = t.test(var1, var2, tails= , test= 2) or ToolPak	● ToolPak





Activity 6: Making inferences from your data

- With your group, calculate the correlation between spi1 and importance1
- Create a scatter plot of this data
- How would you interpret this correlation? What is the magnitude and direction of this correlation?

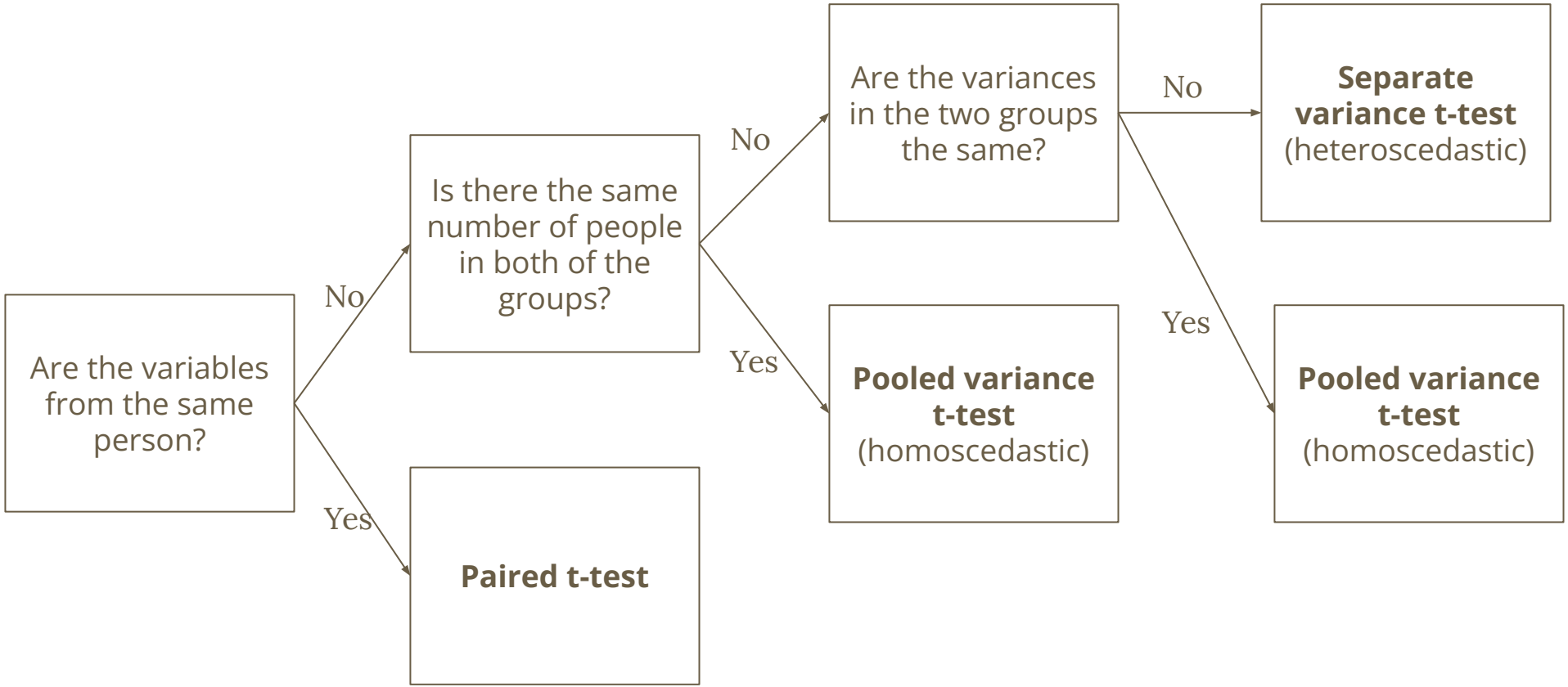
Install

Data analysis toolpak

More on t-tests

- An **independent samples t-test** compares the means for two groups.
  - There are two types we'll consider today:
    - Two-sample **equal variance (homoscedastic) t-test**
    - Two-sample **unequal variance (heteroscedastic) t-test**
- A **paired sample t-test** compares means from the same group at different times (say, one year apart).
- Which t-test should I use?

Which t-test should I use?



Statistical Tests

- P-value: the probability you would observe a difference of the magnitude obtained if there was really no difference
  - The probability we incorrectly find an association when there really isn't one
- If it is less than a certain threshold (alpha) we call the difference statistically significant (less than .05 or .01)
- Statistical significance tells us the difference we found probably isn't just by chance