EMORY UNIVERSITY | Office of the Senior Vice President for Research

Introduction to Descriptive Research Administration Statistics using Excel

Research Analytics Summit
March 2024

Alex Wagner,
Director of Research Administration Analytics
Emory University

Agenda

- Emory University's Research Enterprise
- Research Data Analytics Team and Roles
- NSF HERD Survey Quick Overview
- Data Management and Quality (always important!)
- Descriptive Statistics in Excel
  - Central Tendency: Mean vs Median
  - Variation: Standard Deviation
  - Correlation
  - Outlier detection
- Sample charts and best practices
- Q&A

Emory University's Research Enterprise

- Atlanta, GA
- R1:Very high research activity
- FY23 NSF HERD TRE >$1B
- About 61% Total Federal
- About 26% Total Institutional
- About 66% Health Sciences
- Most of that is from NIH

Research Data Analytics Team and Roles

- Formed in October 2020 and build out since
- 4 Full-Time, 1 Temporary Part-Time, 1 Intern
- Support ORA and SVPR
- Vision: Best Research Administration Unit in the nation
  - #1: Develop Strong and Supported Workforce
  - #2: Pursue and Reach Operational Efficiency
  - #3: Built Robust and Resilient Infrastructure
- Strategic and Competitive Analysis
- Everyone is required to train in Research Administration
- Tools reach from Excel, SQL Developer, Oracle Analytics, SPSS, JMP, PowerBI to Tableau

NSF HERD Survey Quick Overview

- Higher Education Research and Development (HERD) Survey
- Primary source of data on R&D expenditures in higher ED
- More than 900 institutions annually (>$150,000
- Most recent publicly available survey is from FY22
- https://ncses.nsf.gov/surveys/higher-education-research-development/2022#data
- Table 22. Higher education R&D expenditures, ranked by all R&D expenditures, by source of funds: FY 2022

Data Management and Quality (always important!)

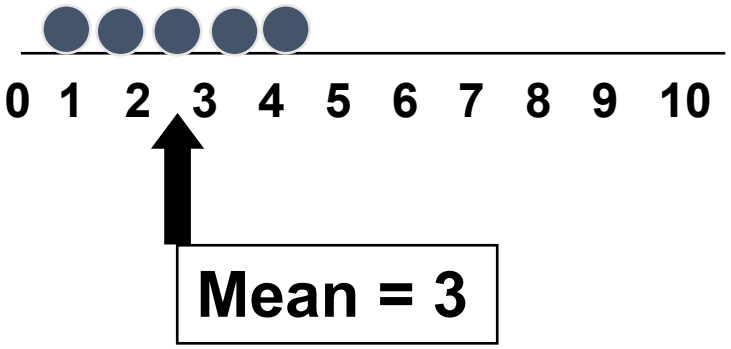
- Data rarely comes in a clean form for analysis
- Always carefully review data and context
- Data cleaning and transformations can be 80% of the work
- See examples in NSF HERD Survey Table 22
    - Dollar amounts are in thousands of dollars
    - Headers and footnotes
    - Extra and blank columns
    - NA as a data point

Descriptive Statistics in Excel: Central Tendency

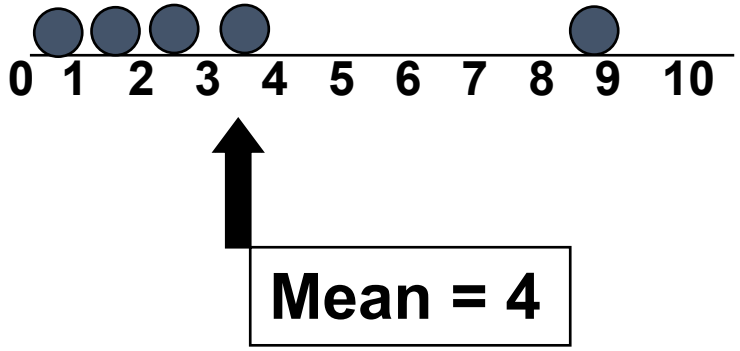
- Mean, most common measure, simple average
- Median, the middle number 50% above and below
- How to do this in Excel
  - More than 1 option
  - Long way
  - Pivot Tables for means only
  - Data Analysis Add-In

The Arithmetic Mean

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers)



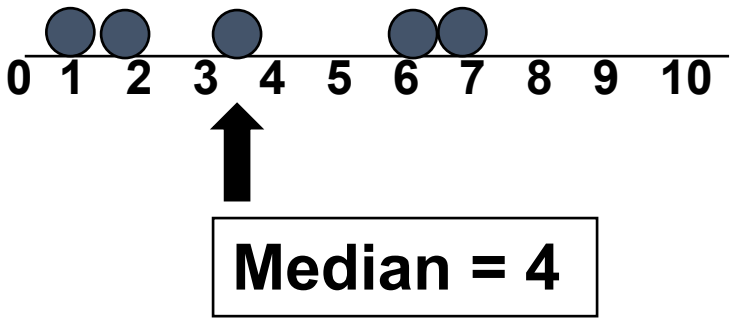
$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

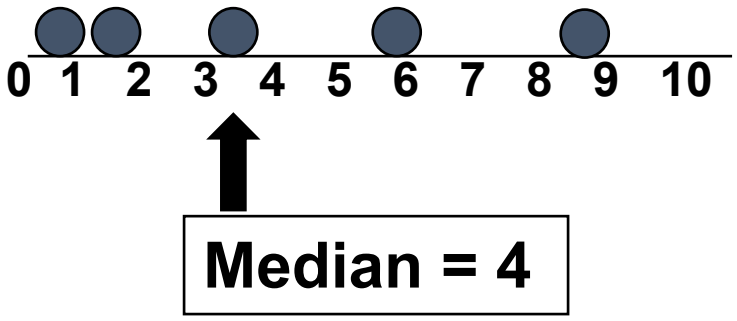


$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

The Median

- In an ordered array, the median is the “middle” number (50% above, 50% below)





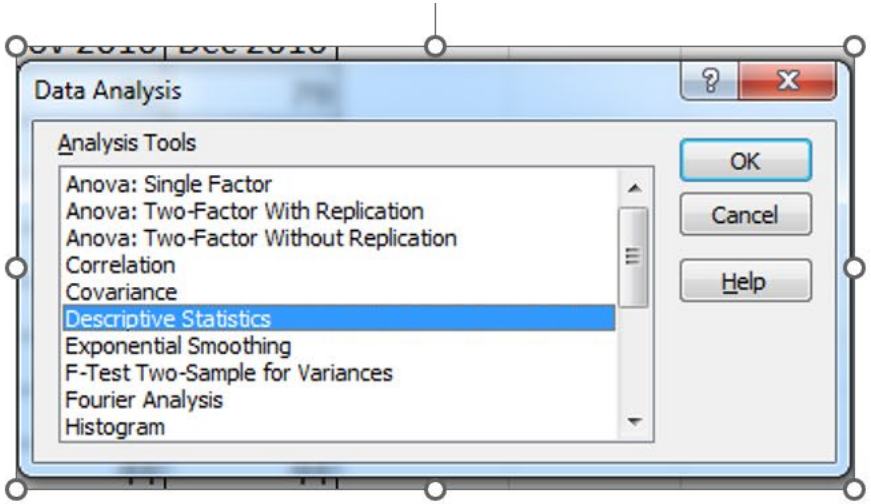
- Less affected by extreme values

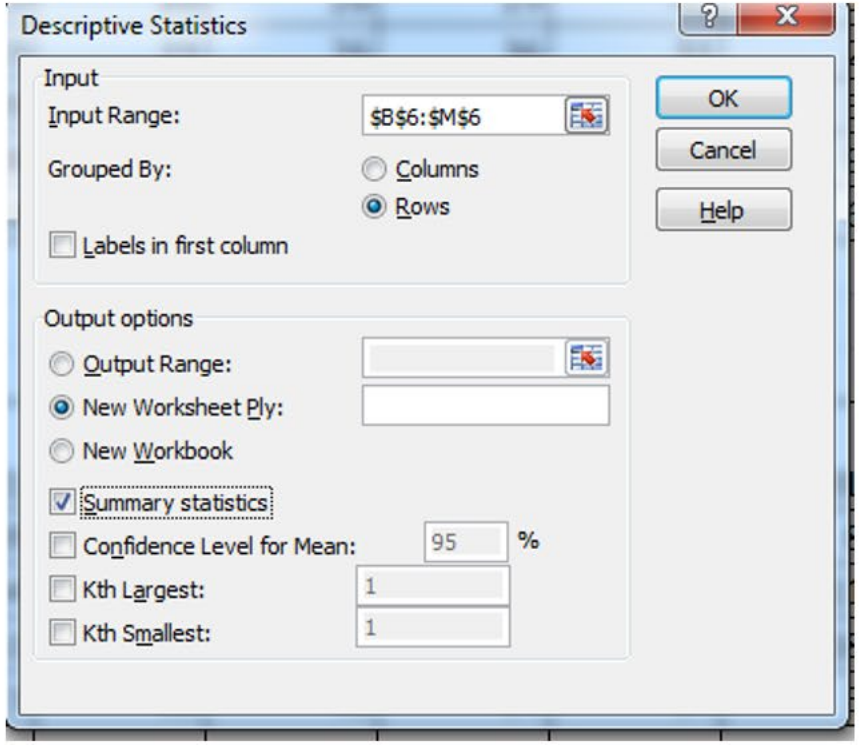
Which measure to choose?

- The **mean** is generally used, unless extreme values (outliers) exist.
- Then **median** is often used, since the median is not sensitive to extreme values. For example, median stolen property values may be reported for a city; it is less sensitive to outliers.

MS Excel Statistics Tools

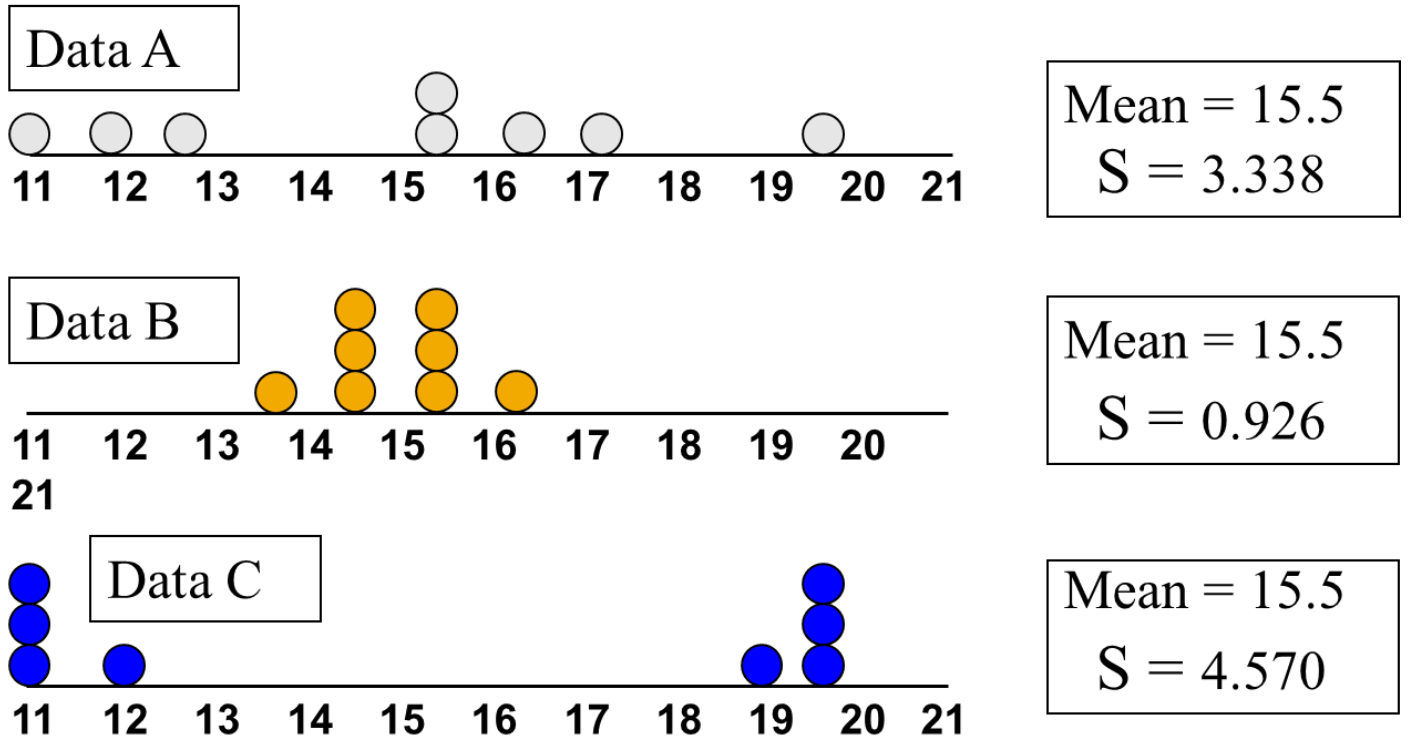
- You could enter the formula, or use the formula tool in Excel … OR….
- Use the Analysis tool Pak





Descriptive Statistics in Excel: Variation Standard Deviation

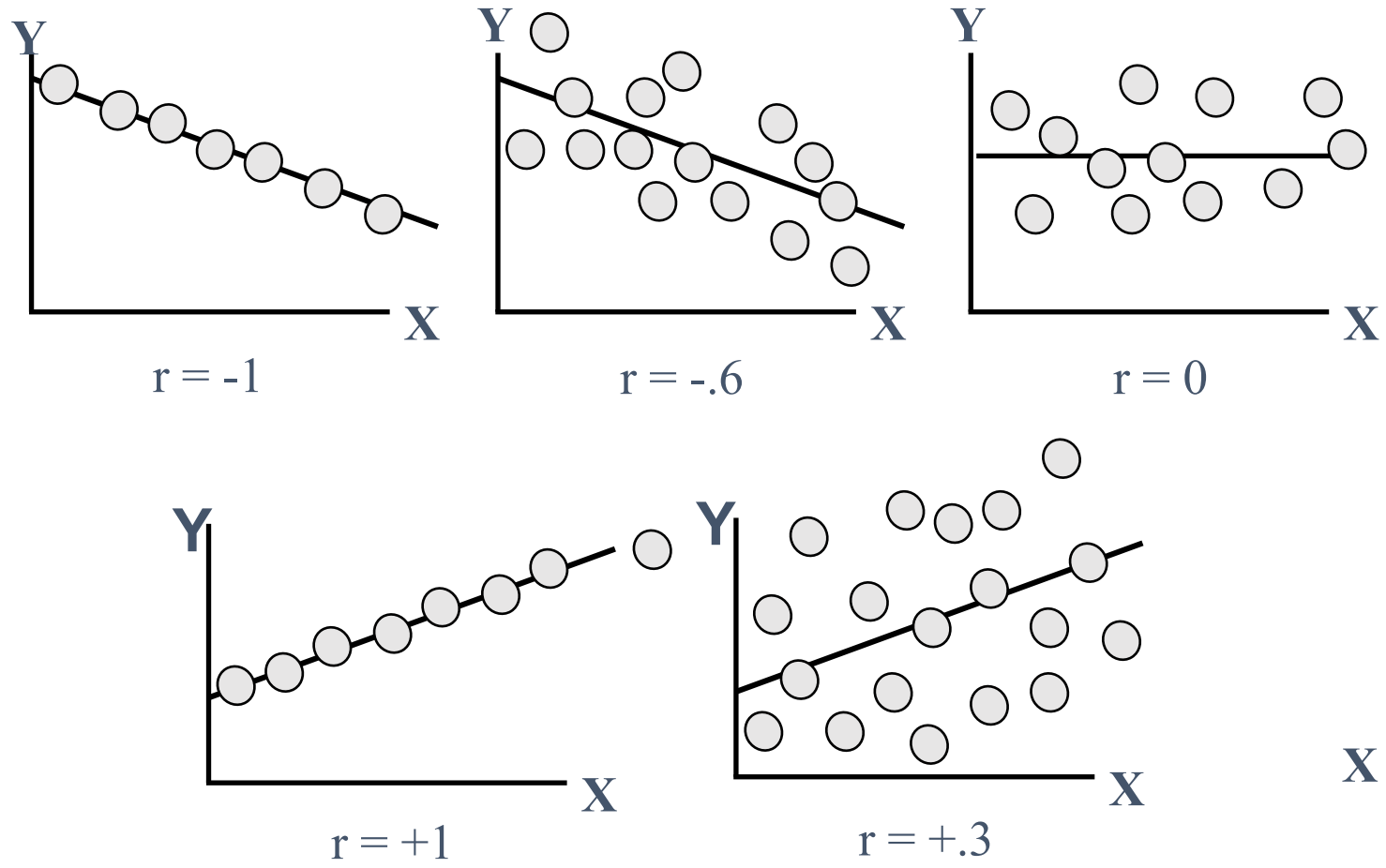
- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data



Correlation

- Correlations measures the relative strength of the ***linear*** relationship between two variables.
- Unit free
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any linear relationship

The Correlation Coefficient



What does it mean?

- Typically:
- greater than .7 = strong correlation
- between .4 and .7 = moderate correlation
- between .2 and .4 = weak correlation
- below .2 = no correlation (similar for negative values)

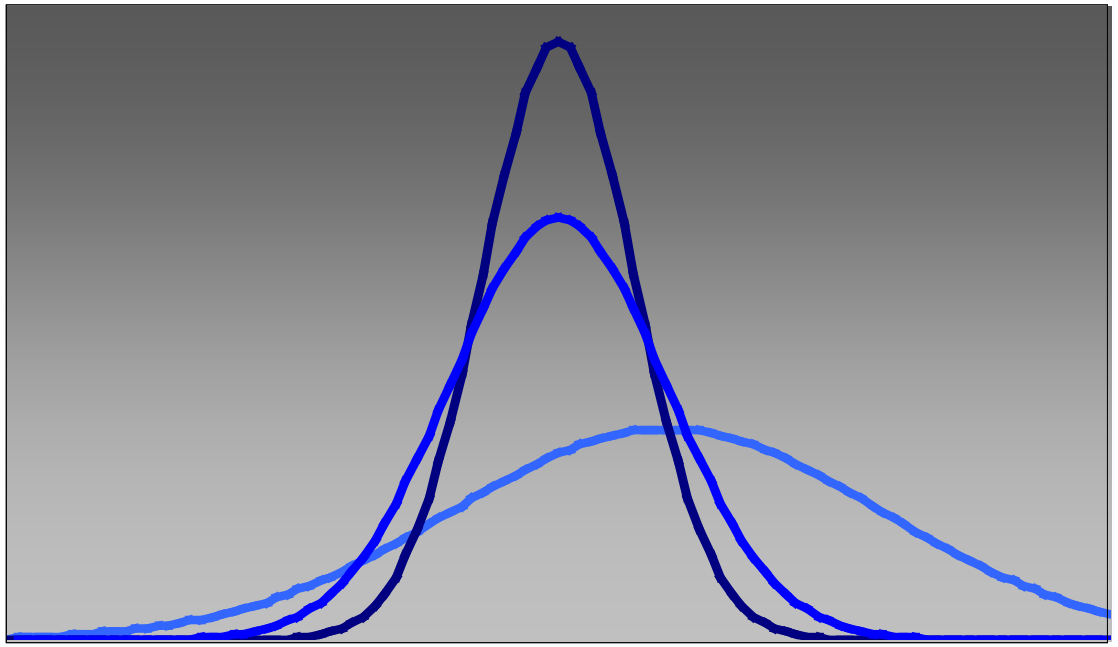
	A	B	C
1		*Row 1*	*Row 2*
2	Row 1	1	
3	Row 2	0.47162	1
4			

Descriptive Statistics in Excel: Outlier Detection

- Outlier Detection
- Why does that matter
- How do to this in Excel

The Normal Distribution aka Bell Curve

- ‘Bell Shaped’ or symmetrical
- Mean, Median and Mode are equal
- Spread is measured by standard deviation
- Infinite number of normal distributions



The Standardized Normal Distribution

- Translate from X to the standardized normal (the “Z” distribution) by subtracting the mean of X and dividing by its standard deviation:

$$Z = \frac{X - \mu}{\sigma}$$

Locating Extreme Outliers Z-Score

$$Z = \frac{X - \overline{X}}{S}$$

where X represents the data value

$\overline{X}$ is the mean

S is the standard deviation

Locating Extreme Outliers Z-Score

- To compute the Z-**score** of a data value, subtract the mean and divide by the standard deviation.
- The Z-score is the number of standard deviations a data value is from the mean.
- A data value is considered an extreme outlier if its Z-score is less than -3.0 or greater than +3.0.
  Note: Applied research -2 and +2
- The larger the absolute value of the Z-score, the farther the data value is from the mean.

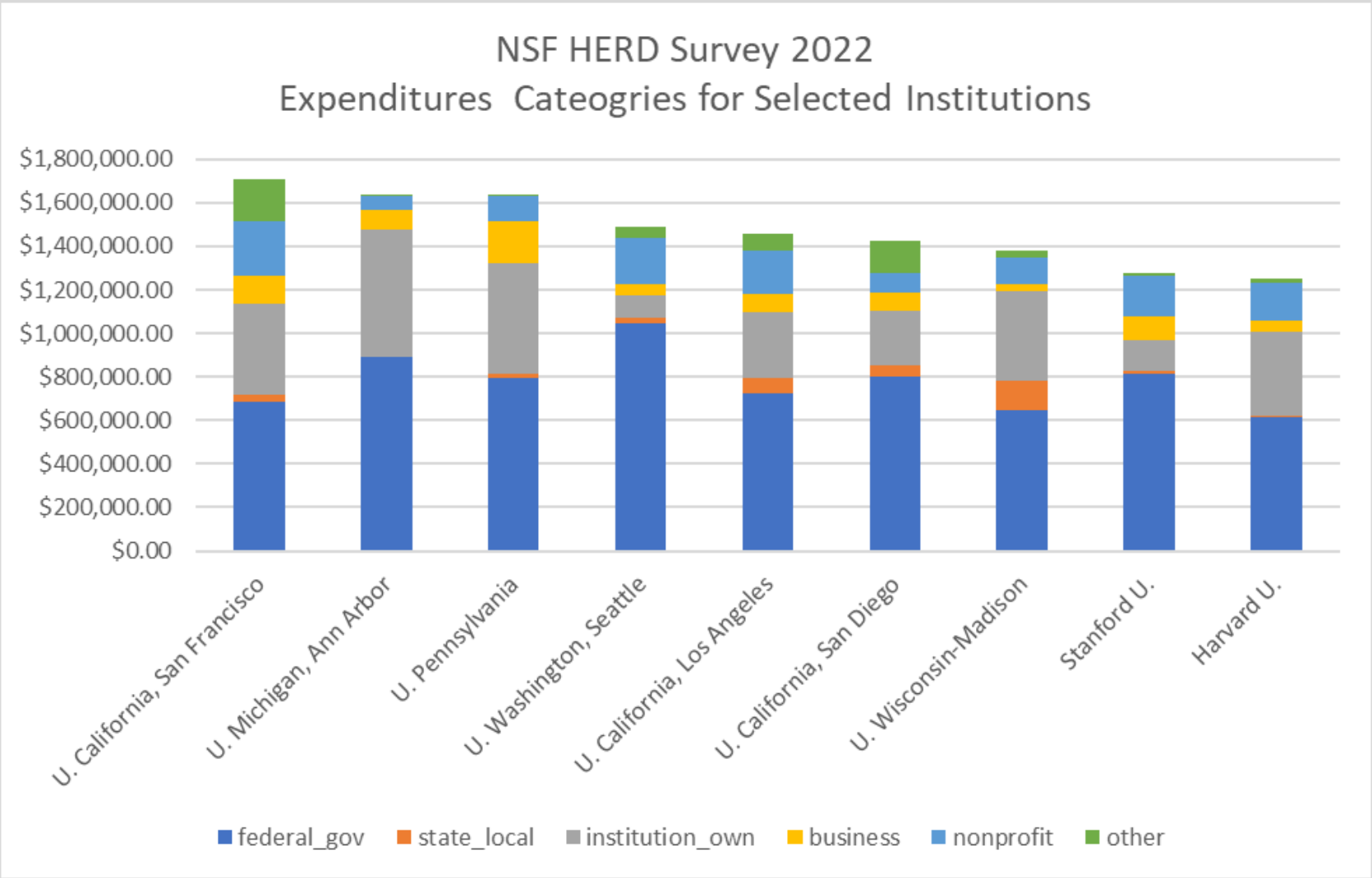
Locating Extreme Outliers Z-Score

- Suppose the mean of the number of proposals submitted is 490 for a month in your fiscal year, with a standard deviation of 100 proposals.
- You just had a month with 620 proposal submissions. Was that a true outlier?
- Compute the z-score for that month.

$$Z = \frac{X - \overline{X}}{S} = \frac{620 - 490}{100} = \frac{130}{100} = 1.3$$

- A score of 620 is 1.3 standard deviations above the mean and would not be considered a true outlier.

Descriptive Statistics in Excel: Sample Charts



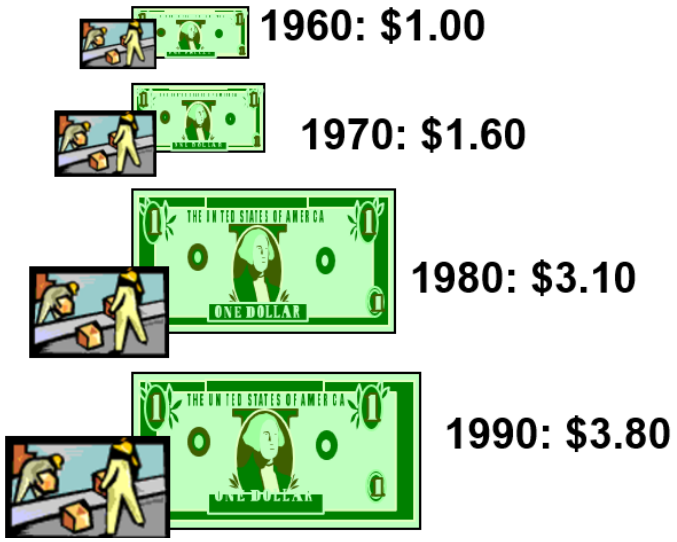
Best Practices for Charts

- The graph should not distort the data.
- The graph should not contain unnecessary adornments (sometimes referred to as chart junk).
- The scale on the vertical axis should begin at zero.
- All axes should be properly labeled.
- The graph should contain a title.
- The simplest possible graph should be used for a given set of data.

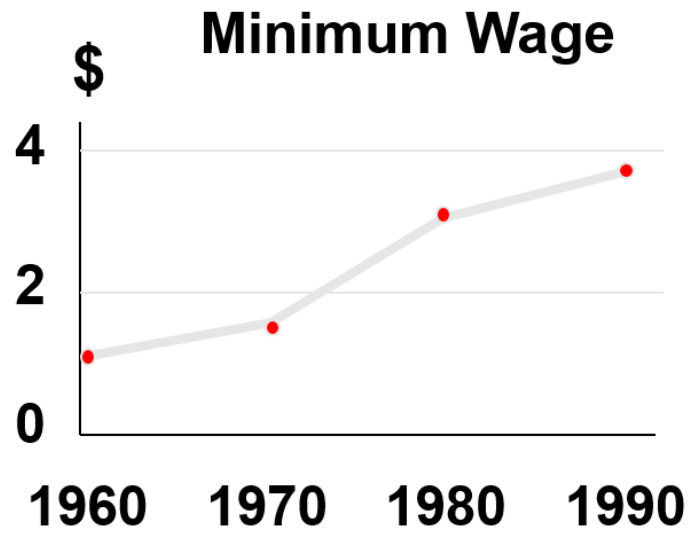
Best Practices for Charts

Graphical Errors: Chart Junk

Bad Presentation



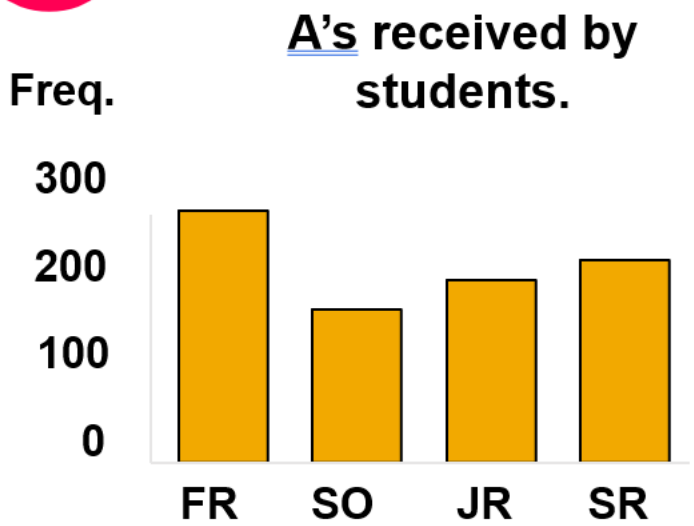
Good Presentation



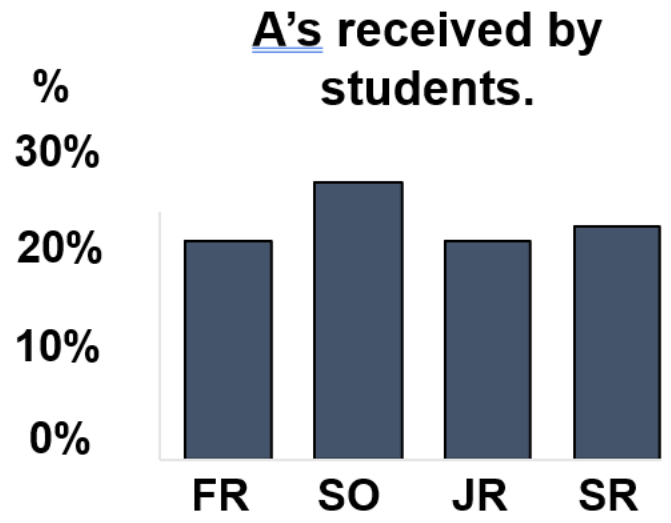
Best Practices for Charts

Graphical Errors: No Relative Basis

Bad Presentation



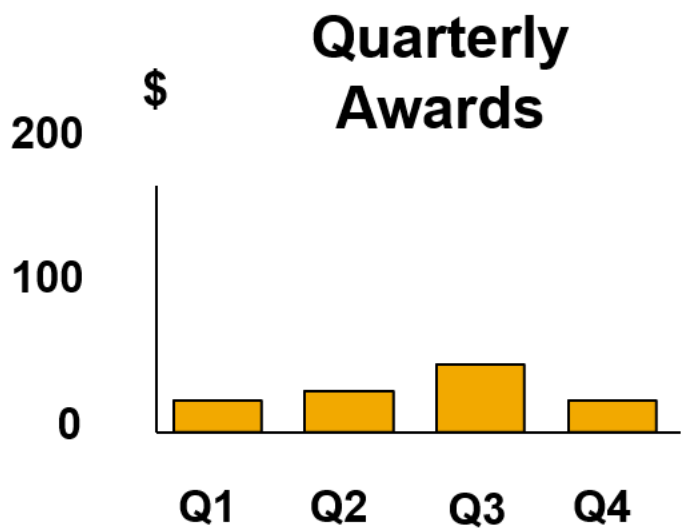
Good Presentation



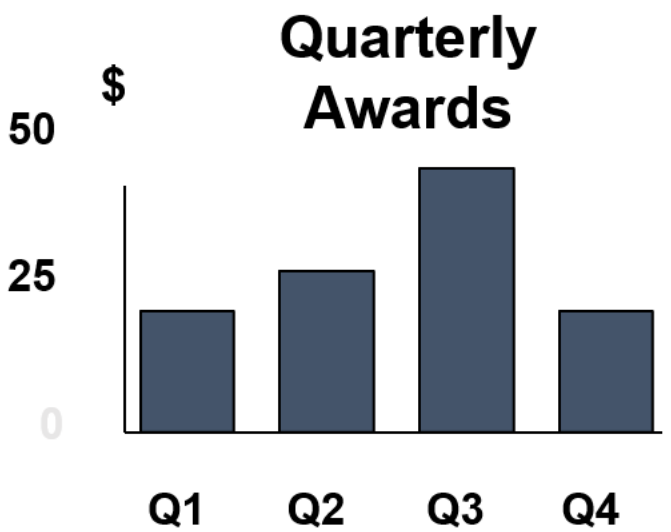
Best Practices for Charts

Graphical Errors: Compressing the Vertical Axis

Bad Presentation

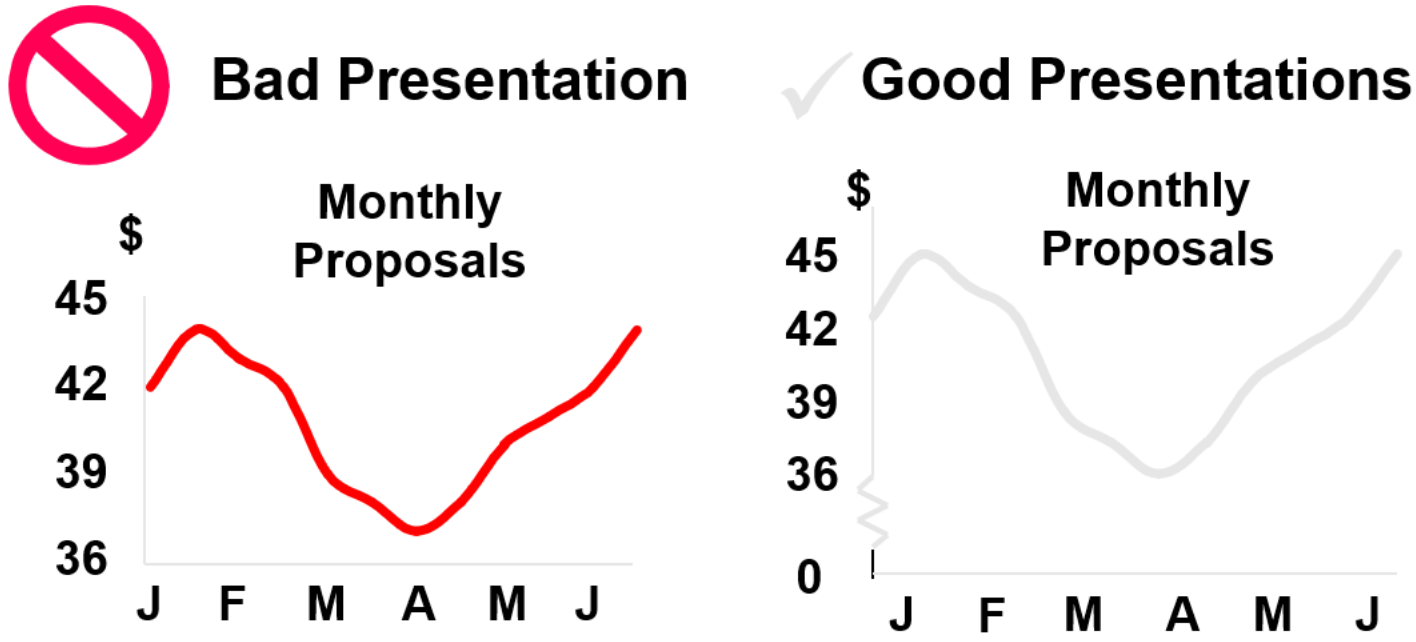


Good Presentation



Best Practices for Charts

Graphical Errors: No Zero Point on the Vertical Axis



EMORY UNIVERSITY | Office of the Senior Vice President for Research

Questions?

Alex Wagner
alexander.wagner@emory.edu
Emory University