Unit 2: Descriptive Statistics

Review
- Sampling design
  - Probability sampling
  - Nonprobability sampling
  - Sampling geographic data
  - Sample size, bias
- Measurement scale
  - Nominal
  - Ordinal
  - Interval/ratio
- Dataset creation, recoding, SPSS

RATs
- iRAT: 15 minutes
- tRAT: 15 minutes
- Missed questions/concepts?
- Appeal?

Ways to Summarize Data
- Tabulation
- Graph
- Mapping
- Summary statistics

Tabulations
- One variable: frequency table
- Two+ variables: cross-tabs

Group activity #1: Tabulations
- Frequency table is appropriate for
  A. Nominal, ordinal, ratio
  B. Nominal, ordinal
  C. Ratio, ordinal
  D. Ratio: continuous
  E. Ratio: discrete
Tabulations

- Frequency table
  - Table that shows how many cases take on a particular value, or fall in an interval
  - For nominal, ordinal, even ratio variables
  - Frequency (cases), percentage (relative frequency), cumulative percentage
  - Valid percentage
  - Excluding missing values

Frequency Table: Categorical

<table>
<thead>
<tr>
<th>Type of Family</th>
<th>Number (millions)</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Married couple with children</td>
<td>25.1</td>
<td>36.6</td>
</tr>
<tr>
<td>Married couple, no children</td>
<td>28.1</td>
<td>41.0</td>
</tr>
<tr>
<td>Single mother with children</td>
<td>7.6</td>
<td>11.1</td>
</tr>
<tr>
<td>Single father with children</td>
<td>1.3</td>
<td>1.9</td>
</tr>
<tr>
<td>Other families</td>
<td>6.4</td>
<td>9.3</td>
</tr>
<tr>
<td>Total</td>
<td>68.5</td>
<td>99.9</td>
</tr>
</tbody>
</table>


Frequency Table: Ratio (Continuous)

<table>
<thead>
<tr>
<th>Malaria Rate</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0-2.9</td>
<td>5</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>3.0-5.9</td>
<td>16</td>
<td>32.0</td>
<td>32.0</td>
</tr>
<tr>
<td>6.0-8.9</td>
<td>12</td>
<td>24.0</td>
<td>24.0</td>
</tr>
<tr>
<td>9.0-11.9</td>
<td>12</td>
<td>24.0</td>
<td>24.0</td>
</tr>
<tr>
<td>12.0-14.9</td>
<td>4</td>
<td>.08</td>
<td>8.0</td>
</tr>
<tr>
<td>15.0-17.9</td>
<td>0</td>
<td>.00</td>
<td>0.0</td>
</tr>
<tr>
<td>18.0-20.9</td>
<td>1</td>
<td>.02</td>
<td>2.0</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>1.00</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Source: Hypothetical Data

Table Style by Taylor & Francis

<table>
<thead>
<tr>
<th>Age group</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>1,340</td>
<td>1,250</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td>5-14</td>
<td>2,000</td>
<td>1,220</td>
<td>21.9</td>
<td></td>
</tr>
<tr>
<td>15-19</td>
<td>3,000</td>
<td>2,260</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>20-24</td>
<td>3,000</td>
<td>2,260</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>25-44</td>
<td>3,000</td>
<td>2,260</td>
<td>37.7</td>
<td></td>
</tr>
<tr>
<td>45-59</td>
<td>2,000</td>
<td>1,220</td>
<td>21.9</td>
<td></td>
</tr>
<tr>
<td>60-69</td>
<td>1,000</td>
<td>700</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>70-74</td>
<td>1,000</td>
<td>700</td>
<td>15.5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>12,150</td>
<td>9,790</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Source: Hypothetical Data, 2014 Taylor and Francis

Frequency Distribution of Number of Children: Ratio (discrete)

<table>
<thead>
<tr>
<th># of kids</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3-5</td>
<td>6</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>6-8</td>
<td>5</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>9-11</td>
<td>3</td>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>12-14</td>
<td>3</td>
<td>15</td>
<td>90</td>
</tr>
<tr>
<td>15-17</td>
<td>1</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>18-21</td>
<td>1</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>n = 20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Hypothetical Data

Exhibit 2.4: A Grouped Frequency Distribution

<table>
<thead>
<tr>
<th># of kids</th>
<th>Frequency</th>
<th>Percentage</th>
<th>Cumulative Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3-5</td>
<td>6</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>6-8</td>
<td>5</td>
<td>25</td>
<td>60</td>
</tr>
<tr>
<td>9-11</td>
<td>3</td>
<td>15</td>
<td>75</td>
</tr>
<tr>
<td>12-14</td>
<td>3</td>
<td>15</td>
<td>90</td>
</tr>
<tr>
<td>15-17</td>
<td>1</td>
<td>5</td>
<td>95</td>
</tr>
<tr>
<td>18-21</td>
<td>1</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>n = 20</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Hypothetical Data
Frequency Tables

Pros:
- Also easy
- Useful for large datasets
- Fairly rich description of data

Cons:
- Unlike a list, you can’t see which case is which or compare with other variables
- Not useful if all values are unique

Group activity #1: Tabulations

- Cross-tab is appropriate for:
  A. Nominal, ordinal, ratio
  B. Nominal, ordinal
  C. Ratio, ordinal
  D. Ratio: continuous
  E. Ratio: discrete

Cross-tab (Contingency table)

- 2 or 3 categorical variables
- For ratio variables, convert into categorical/ordinal
- Controlling one or two variables
- Frequency/relative frequency of subjects in each combined category
- 2 variables: gender, party affiliation
- Combined categories:
  - democratic women, democratic men, republican women, Republican men

Cross-tabulation

- Resulting table (“crosstab” or “joint contingency table”):

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrats</td>
<td>27</td>
<td>10</td>
</tr>
<tr>
<td>Republicans</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>25</td>
</tr>
</tbody>
</table>

This is the total N

Crosstabulation

- Tables may also have row and column marginals (i.e., totals)

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dem</td>
<td>27</td>
<td>10</td>
<td>37</td>
</tr>
<tr>
<td>Rep</td>
<td>16</td>
<td>15</td>
<td>31</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>25</td>
<td>68</td>
</tr>
</tbody>
</table>

Group Activity #2: Crosstabulation

- Women are more likely to be democrats than men.
  - A) True
  - B) False
Group Activity #2
Crosstabulation

<table>
<thead>
<tr>
<th></th>
<th>Women</th>
<th>Men</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dem</td>
<td>39.7%</td>
<td>14.7%</td>
<td>37</td>
</tr>
<tr>
<td>Rep</td>
<td>23.5%</td>
<td>22.1%</td>
<td>31</td>
</tr>
<tr>
<td>N</td>
<td>43</td>
<td>25</td>
<td>68</td>
</tr>
</tbody>
</table>

- Women are more likely to be democrats than men.
  - A) True
  - B) False

If we follow the placement rule, what kind of percentages should we have in cross-tabs?
- A) row percentages
- B) column percentages
- C) total percentages
- D) row and column percentages
- E) row, column, and total percentages

We are interested in demonstrating the relationship between gender and party identification using the above table.

Is this an effective cross-tab? Why?
Please list aspects that need to be revised

The choice of row/column percentage is determined by the nature of hypothesis.
- Are women more likely to be democrats? Row%
- Are democrats mostly women? Column%
- Conditional probability of being in each of the categories of the dependent variable (e.g., party identification) given that an individual is in a particular category of independent variable (gender)
Three Dimensional Tables

<table>
<thead>
<tr>
<th></th>
<th>Grammar School</th>
<th>High School</th>
<th>College</th>
</tr>
</thead>
<tbody>
<tr>
<td>Military</td>
<td>V</td>
<td>S</td>
<td>N</td>
</tr>
<tr>
<td>Militant</td>
<td>17%</td>
<td>22%</td>
<td>32%</td>
</tr>
<tr>
<td>Non-militant</td>
<td>83%</td>
<td>78%</td>
<td>68%</td>
</tr>
<tr>
<td>Total</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>N</td>
<td>100</td>
<td>2001</td>
<td>(44)</td>
</tr>
</tbody>
</table>

Source: Adapted from Mars, 1967a: Table 6.
*V=very religious, B=moderately religious, O=must vary religious or not at all religious.

2. Visual Representation (Graph)

- Pie chart
- Bar chart
- Histogram
- Frequency curve
- Box plot

When to use what kind of graph is the most effective/appropriate?

Group Activity#3

Graph

- For categorical variables (nominal, ordinal), which of the following is the most appropriate?
  - A) bar chart
  - B) pie chart
  - C) histogram
  - D) line chart
  - E) box plot

Group Activity#3

Graph

- For quantitative variables (ratio), which of the following is most appropriate?
  - A) bar chart
  - B) pie chart
  - C) histogram
  - D) line chart
  - E) box plot

Pie Chart

- For nominal and ordinal variables
- Size of the slice represent the share of cases in a specific category (sum=100%)

Bar Chart

- For Nominal & Ordinal Variables Only
- Height of bars represent number of cases, or share of cases
A Clustered Bar Graph

Happiness by Work Status

- Very Happy
- Pretty Happy
- Not Too Happy

Histogram

- For continuous measures, ratio
- Height of bar represents number of cases or % within a given range of values
- Data needs to be "grouped": some info is lost

Group Activity: how do you interpret these histograms?

Histogram (Example)
Histogram

- Interval width (bin): Histograms look very different depending on how wide you set the value for intervals.
  - Choose width carefully
  - Try multiple widths
  - Different bins may lead to different interpretations

Frequency Curve

- Connecting the mid-points of intervals
- Use a single line, reduce the visual emphasis on boundaries

Sample vs. Population Distribution

- For a continuous variable

Box Plot

- Measure both central tendency and "spread"
Box Plot

How do you interpret this box plot?

Mean
Lower quartile (25%)
Upper quartile (75%)
Median
Max
Min
Outlier

Outlier: > 1.5 IQR above upper quarter or below lower quartile

3. Mapping

- Spatial, 3-dimensional
- Use point, line or areal symbols
- SKIP

Individual activity:
Table and Graph in SPSS

- Dataset: housing sale
  - Create a frequency table for bedrooms
  - Create a frequency table for date
  - Recode the variable first!
  - Create a cross-tab for dprice by garage
    - Request appropriate %
- Create charts
  - Pie: bedrooms
  - Bar: bedroom
  - Boxplot: unempl * district
  - Histogram: price

Interpret the bar chart

A Graph that Lies:
Under-representing the change in the data

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Housing Foreclosures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>32</td>
</tr>
<tr>
<td>2010</td>
<td>79</td>
</tr>
</tbody>
</table>

A Graph that does not Lie:

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of Housing Foreclosures</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>32</td>
</tr>
<tr>
<td>2010</td>
<td>79</td>
</tr>
</tbody>
</table>

Lie Factor

- change shown in graphic/change shown in data
- 1: truthful representation of the data
- >1: over-representing, exaggerating
- <1: under representing
- Bar Chart 1:
  - Change in the data: (79-32)/32=147%
  - Change shown in graphic: height of the bars, (4.5-3)/3=50%
  - Lie factor=50%/147%=0.34
  - The chart under-representing the increase in number of foreclosure in 2010

An SPSS Graph that Somewhat Lies

Y-axis did not start at zero, lie factor=1.38, over-representing the effect of education on number of children
Recap:
Ways to Summarize Data
- Tabulation
- Graph
- Mapping
- Summary statistics
  - Measure central tendency
  - Measure variability
  - Measures of skewness, kurtosis
  - Measures of relative position
  - Geographic data

Measuring the Central Tendency
- The "center" of a distribution, "typical" case
  - Mean, median, mode

Central Tendency: Mean
- Arithmetic Mean, or "average", "Y-bar"
  - Sum of the Y for all cases divided by the number of subjects
  - Most frequently used measure

Variables
- Each column of a dataset is considered a variable, generally referred as "Y", or "X"

Calculating the Mean
- $Y_i$ represents $i^{th}$ case of variable Y
- $i$ goes from 1 to n
- $Y_1$ = value of Y for first case in spreadsheet
- $Y_2$ = value for second case, etc.
- $Y_n$ = value for last case

<table>
<thead>
<tr>
<th>Person</th>
<th># Guns owned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Summary Measures for Frequency Curve
- Figure 3.3 Summary measures for frequency distributions.
Calculating the Mean

\[ \sum_{i=1}^{5} Y_i = Y_1 + Y_2 + Y_3 + Y_4 + Y_5 \]

\[ = 0 + 3 + 0 + 1 + 1 = 5 \]

\[ \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{1}{5} \times 5 = 1 \]

<table>
<thead>
<tr>
<th>Person</th>
<th># Guns owned</th>
<th>( Y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>( Y_1 ) = 0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>( Y_2 ) = 3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>( Y_3 ) = 0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>( Y_4 ) = 1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>( Y_5 ) = 1</td>
</tr>
</tbody>
</table>

Mean of Groups

- Mean of groups is the weighted mean (average) of group means.
- Two groups of size \( n_1, n_2 \)
  \[ \bar{Y} = \frac{n_1 \bar{Y}_1 + n_2 \bar{Y}_2}{n_1 + n_2} \]
- More generally,
  \[ \bar{Y} = \frac{\sum_{i=1}^{k} M_i f_i}{n} \]

Team Activity (graded):

- Which of the following variables has a meaningful mean?
  - A) household income ($)
  - B) housing value ($)
  - C) education in highest degree (1: <hs; 2: hs; 3: college; 4: graduate)
  - D) age
  - E) gender (1: female; 0: male)

Properties of the Mean

- Pros:
  - Gives a sense of "typical" case
  - Useful for continuous data
  - Easy to calculate
  - Center of gravity

\[ \sum_{i=1}^{n} (Y_i - \bar{Y}^2) = 0 \]

\[ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \leq \sum_{i=1}^{n} (Y_i - A)^2 \]
Properties of the Mean

**Cons:**
- Every case influences outcome
- Extreme cases (outliers) affect results a lot. (e.g. Mean income is often not very meaningful)
- Doesn’t give you a full sense of the distribution
- Appropriate only for quantitative data

The Mean and Extreme Values

<table>
<thead>
<tr>
<th>Case</th>
<th>Num CD’s</th>
<th>Num CD’s^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>1000</td>
</tr>
<tr>
<td>Mean</td>
<td>32.5</td>
<td>265</td>
</tr>
</tbody>
</table>

Hypothetical Block After One Heck of a Remodel: Mean housing price/value is not very meaningful

Central Tendency: Median

- The middle measurement of a ranked sample: \((n+1)/2\).

Central Tendency: Median

- If \(n\) is odd, median is a single measurement; if \(n\) is even, median is the midpoint between the two middle measurements

Group Activity: What is the median education?

<table>
<thead>
<tr>
<th>Highest Degree Completed, for a Sample of Americans</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest Degree</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>A. Not a high school graduate</td>
</tr>
<tr>
<td>B. High school only</td>
</tr>
<tr>
<td>C. Some college, no degree</td>
</tr>
<tr>
<td>D. Associate’s degree</td>
</tr>
<tr>
<td>E. Bachelor’s degree</td>
</tr>
<tr>
<td>Master’s degree</td>
</tr>
<tr>
<td>Doctorate or professional</td>
</tr>
</tbody>
</table>
Central Tendency: Median

- Median: Appropriate for both ratio and ordinal data, but not for nominal data.
- Same as the mean for symmetric distributions; for skewed distribution, median lies toward the shorter tail related to the mean.

Team Activity: Median

- Median is not appropriate for_____.
  A) household income ($)
  B) housing value ($)
  C) education in highest degree (1: <hs; 2: hs; 3: college; 4: graduate)
  D) age
  E) gender (1: female; 0: male)

Central Tendency: Median

- Pros:
  - Unaffected by outliers (appropriate for variables such as income, housing price)
- Cons:
  - Insensitive to the distances of the measurements from the middle.
  - 1, 2, 10, 100, 500

Central Tendency: Mode

- The value that occurs most frequently -- the "Modal" value.
- Appropriate for all types of data.
  - Commonly used for categorical (nominal, ordinal) data
  - Only useful for continuous (interval/ratio) variables if you have grouped data
  - Otherwise, all values may very likely be unique.
- Modes = Peaks
  - Uni-modal distribution: One peak
  - Bi-modal distribution: Two peaks
  - Multi-modal distribution: Multiple peaks (usually more than two).

Central Tendency: Mode

- What is the mode?
- Why is the distribution bimodal?
Reasons for Multi-Modal Distributions

- The sample is heterogeneous (i.e., made up of more than one group)
  - Height forms a bell-shaped distribution for men and for women, but the peaks are different. A combined sample has two peaks.

- The sample reflects some exogenous structural ordering process
  - Years of education completed is peaked at 12 (high school), 16 (college).

Mode

- Pro: Easy, useful
- Con:
  - Do not necessarily close to the center
  - Not very helpful (even misleading) in certain circumstances, e.g. if there are many peaks, or a single unusual one; if the variable is distributed quite evenly.

Comparing Mean, Median, Mode

- Both mean and median can be easily calculated for grouped or ungrouped data. Mode is usually used for grouped data.
- Unequal class intervals in grouped data do not hinder the calculation of mean, median, but severely limit the calculation of mode.
- The presence of an open-ended class do not affect the median or mode, but severely limit the calculation of mean.

Group Activity: Central Tendency

- In Canada, based on the 2001 census, for religious affiliation (Catholic, Protestant, other Christian, Muslim, Jewish, None, others), the relative frequencies were 42%, 28% 4%, 2%, 1%, 16%, 7%.
  - A) the mean religion is Protestant
  - B) the mode religion is Catholic
  - C) The median religion is Protestant
  - D) only 2.5% of the subjects fall within one standard deviation of the mean
  - E) The Jewish response is an outlier.
Levels of Measurement and Measures of the Centre

<table>
<thead>
<tr>
<th></th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
<td>YES</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Median</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td>Mean</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
</tbody>
</table>

If appropriate, report all three. The differences between them tell something important about the distribution.

Summary statistics

- Measures of central tendency
  - Mean, median, mode
- Measures of variability
  - Describing how “spread out” a distribution is around its center

Variability

- Very different groups can have the same means:

<table>
<thead>
<tr>
<th></th>
<th>Number of CDs (Group 1)</th>
<th>Number of CDs (Group 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>75</td>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>125</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>175</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>Std. Dev</td>
<td>21.72</td>
<td>67.62</td>
</tr>
<tr>
<td>Mean</td>
<td>101</td>
<td>100</td>
</tr>
<tr>
<td>N</td>
<td>23</td>
<td>23</td>
</tr>
</tbody>
</table>

Variability

Measure of Variation

- Range ($y_{\text{max}} - y_{\text{min}}$)
  - Doesn’t tell you much about the middle cases
  - Influenced by extreme values... may not be representative
- Interpercentile (usually interquartile) range
  - Percentile: $p\%$ scores below it, $(100-p)\%$ above it
  - Lower quartile ($P_{25}$), upper quartile ($P_{75}$)
  - IQR = $P_{75} - P_{25}$
  - Not sensitive to extreme value
  - Outlier: $>1.5\text{IQR}$ above the upper quartile, or $1.5\text{IQR}$ below the lower quartile

Quartile and Interquartile Range

Which country would you prefer to live in?
Measures of Variation

- **Deviation**
  \[ d_i = Y_i - \bar{Y} \]

- **Variance \((s_Y)^2\)**
  \[ s_Y^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \]

- **Standard deviation \((s_Y)\)**
  \[ s_Y = \sqrt{s_Y^2} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}} \]

Example

<table>
<thead>
<tr>
<th>Case</th>
<th>Num CD's</th>
<th>Mean (Ybar)</th>
<th>Deviation ((Y_i-Ybar))</th>
<th>Square of deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>sum</td>
<td></td>
<td></td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>

\[ s_Y^2 = \frac{\sum_{i=1}^{n} d_i^2}{n-1} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \]

Variance\? Standard Deviation\?

Example

<table>
<thead>
<tr>
<th>Case</th>
<th>Num CD's</th>
<th>Mean (Ybar)</th>
<th>Deviation ((Y_i-Ybar))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>4</td>
<td>-3</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

Variance\? Standard Deviation\?

Properties of Variance \((s_Y)^2\)

- \(s_Y^2 \geq 0\)
  - Zero if all points cluster exactly on the mean
  - Larger for more “spread” distributions

- **Pros:**
  - Comparable across samples of different size

- **Cons:**
  - Values get fairly large, due to “squaring”
Properties of Standard Deviation

- $s \geq 0$
- $s = 0$ when all observations have the same value, grows larger if points are spread further from the mean
- $s$ is the average distance of an observation from the mean
- Most commonly used measure of dispersion
- Comparable across different sample sizes
- The Empirical Rule

Team Activity (graded)

- Data: 1, 1, 1, 2, 2, 3, 3, 5, 1, 1
- By hand, find the mean, median, mode, variance, and standard deviation

Team Activity: Standard Deviation

- SAT math score is approximately bell-shaped, with a mean 500 and standard deviation 100. How many people scored more than 700?
  - A) 0.5%
  - B) 1%
  - C) 2.5%
  - D) 4%
  - E) 5%

The Alternative

- Average Absolute Deviation (AAD)
  - Very intuitive interpretation
  - Has non-ideal statistical properties

Measure Central tendency and Variation

- Box plots

How to interpret this box plot?
Measuring Skewness

- Is the distribution symmetrical?
- Skewness measuring the degree of asymmetry around a measure of central tendency
- Zero = perfectly symmetrical
- Higher number = increasingly skew

Measureing Skewness

- A "tail" is referred to as "skewness"
- Tail on left = skewed to left = negative skew
- Tail on right = skewed to right = positive skew

Pearson’s Coefficient of Skewness

- Based on distance from Mean to Median
- Mean moves more if there are extreme cases, as when there is a "tail"

\[ \text{skew} = \frac{3(\bar{X} - \text{Mdn})}{s} \]

Measuring Skewness

- Pearson’s Coefficient of Skewness
- Quartile skewness
  - Measures distance between median and lower & upper quartiles
  - Extreme values move lower/upper quartiles further out, resulting in larger skewness

\[ \text{skew} = \frac{P_{25} + P_{75} - 2\text{Mdn}}{2} \]

Skewness in SPSS

\[ \text{skew} = \frac{n}{(n-1)(n-2)} \sum (X_i - \bar{X})^3 \]

Group Activity

- At a business, we find that the male employees’ salary distribution is positively skewed, whereas the female employees’ salary distribution is negatively skewed. Which of the following is true about this business?

  A) there are a small number of women have very low salary, and a small number of men have very high salary
  B) there are a small number of women have very high salary, and a small number of men have very low salary
  C) there are a large number of women have very low salary, and a large number of men have very high salary
  D) there are a large number of women have very high salary, and a large number of men have very low salary

Interpreting Skewness

Which way is it skewed? What is the social interpretation? What would be the interpretation if it were skewed in the opposite direction?

- Skewness provides information about inequality
  - Example: Economic wealth of nations
Interpreting Skewness

- Skewness may reflect “floor” or “ceiling” effects
  - Example: Number of crimes committed by individuals in a sample. Lower bound is zero. Mode is very low. A few cases are high.
  - Example: National secondary school enrollment ratio. Cannot exceed 100%

Notes on Skewness

- More often assessed informally “by eye” than calculated as a value.
  - Look at a histogram to identify skewness
- Some statistical techniques work properly only on variables that are not skewed (e.g. empirical rule).
  - It can be very important to identify highly skewed variables.
- Note: mode, skew sound like “jargon”, but are actually quite helpful in communicating descriptive information about your variables

Example:

- How would you describe this variable?

Measuring Relative Position

- Rank \( R_i \)
  - Sort the data, the position of score
- Cumulative frequency list/curve
  - Number of cases (percentage of cases) falling in or below a given interval

Cumulative Frequency List

<table>
<thead>
<tr>
<th>Years of Education (N=2904)</th>
<th>Value</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulat %</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 or less</td>
<td>21</td>
<td>1.4</td>
<td>3.9</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>82</td>
<td>5.3</td>
<td>9.3</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>51</td>
<td>3.3</td>
<td>12.6</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>70</td>
<td>4.6</td>
<td>17.2</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>95</td>
<td>6.2</td>
<td>23.4</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>489</td>
<td>31.8</td>
<td>55.2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>125</td>
<td>8.1</td>
<td>63.5</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>184</td>
<td>12.0</td>
<td>75.6</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>76</td>
<td>4.9</td>
<td>80.5</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>152</td>
<td>9.9</td>
<td>90.5</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>40</td>
<td>2.6</td>
<td>93.1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>61</td>
<td>4.0</td>
<td>97.1</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>1.2</td>
<td>98.2</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>61</td>
<td>4.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Indicates that 55% of students have 12 years of education or less
Cumulative % Graphs

Measuring Relative Position

- Rank \( (R_i) \)
- Cumulative frequency list/curve
- Quantile
  - Percentiles, quartiles, deciles, etc...
  - General term = quantile
  - Dividing cases up into fixed number of equal “chunks”:
    - 100 chunks = percentiles (1% each)
    - 10 chunks = deciles (10% each)
    - 5 = quintiles (20% each)
    - 4 = quartiles (25% each)

Is History Siding With Obama’s Economic Plan?

(NY Times)

Benefit of Quantiles

- Quantiles allow you to identify cases (or groups of cases) in relation to the larger group
  - Who is “high”, who is “low”
  - Regardless of the unit of measurement
- Advantage: Allows comparison between variables with different scales (or with different means)
  - Example: Reading test scored 1-100, Math test is scored 1-25. How do you know which you scored better on? Answer: percentile

Measuring Relative Position: Ratio

- Ratio
  - The position of an individual score in relation to some other score (e.g. mean, max, min...)
  - Standardized score (Z-score): deviation from mean divided by standard deviation
    \[ Z_i = \frac{Y_i - \bar{Y}}{s} \]
Z-Score Example

\[ Z_i = \frac{Y_i - \bar{Y}}{s} \]

- Number of CD's: Mean = 32.5, s = 29.8

<table>
<thead>
<tr>
<th>Case</th>
<th>CD's (Y)</th>
<th>Mean (Y bar)</th>
<th>Deviation (d)</th>
<th>Z-score (d/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>32.5</td>
<td>-12.5</td>
<td>-0.42</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
<td>32.5</td>
<td>7.5</td>
<td>+0.25</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>32.5</td>
<td>-32.5</td>
<td>-1.1</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>32.5</td>
<td>37.5</td>
<td>+1.3</td>
</tr>
</tbody>
</table>

Z-Score (Standardized Score)

- Unit of Z-scores is "standard deviation"
  - A Z-score of -1.1 indicates a case is nearly one standard deviation below the mean
  - \( Z=0.5 \rightarrow 0.5 \text{ St. Dev above the mean} \)
- You can convert any or all values of a variable to a common scale
  - mean = 0
  - negative = below mean
  - positive = above mean
  - range approximately from -3 to +3. WHY?

Z-Scores

- Z-scores can be compared across variables with different units or means
  - Examples: height and weight; a person is -0.3 on math, but 1.2 on income
  - Simple deviations can't be compared if units of measurement are different
- Convert an entire variable (all cases) to Z-scores, creating a new variable with useful properties
  - preserves the shape of the distribution, but unit is changed
  - Mean = zero, because it is based on deviations
  - Standard Deviation (s_y) = 1
  - Easier to compare different variables

Converting Variables to Z-scores

GSS Data, N=2904

Team Activity: Z-score

Here is a frequency distribution of number of household members less than 6 years old for respondents aged 20 to 29:

<table>
<thead>
<tr>
<th>Number of children (&lt;6 years old)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>216</td>
</tr>
<tr>
<td>1</td>
<td>59</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

For variable “number of children”, mean=0.49, s=0.82. Use z-score to describe the "uniqueness" of someone who has four young children.
### Summary statistics

- **Measures of central tendency**
  - Mean, median, mode
- **Measures of variability**
  - Range, IQR, variance, s.d., AAD
- **Measures of skewness**
- **Measures of relative position**
  - Rank, quantile, Z-score

### Special case: Dichotomous Variables

- **Mean, variance, and S.D. are generally NOT too useful for nominal variables**
- **Exception:** Mean of dichotomous variables
  - Dichotomous variable = nominal, w/ 2 categories, often called “dummy” variables
  - E.g.: Do you approve of gun control (yes/no)?
  - People saying “yes” assigned 1, no = 0

### Dichotomous (Dummy) Variables:

- **1 = Presence of something, 0 = absence of it**

<table>
<thead>
<tr>
<th>Person</th>
<th>View On Gun Control</th>
<th>Support? (Dummy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Favor</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Oppose</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>Favor</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Favor</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Oppose</td>
<td>0</td>
</tr>
</tbody>
</table>

### Dichotomous Variables

- **Interpretation:**
  - Mean = proportion indicating yes
- **Example:** "Do you approve of gun control?"
  - 14 yes, 24 no. (37% yes, 63% no)
  - Mean of variable = .37

### Geographic Data

- **Statistical and spatial distribution**
- **Summary statistics:**
  - Important to accessibility and dispersion
  - Centrality: mean center, median center
  - Dispersion: standard distance, relative distance
Geographic Data: Mean Center

- Minimize the sum of squared distance
  - Point data:
    \[(\bar{X}, \bar{Y})\]
  - Areal data:
    \[\bar{X} = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}, \bar{Y} = \frac{\sum_{i=1}^{n} W_i Y_i}{\sum_{i=1}^{n} W_i}\]

Mean Center

Why are they different?

Geographic Data: Median Center

- **NOT** the point defined by the medians of \(X\) and \(Y\)
- Minimum aggregated distance to all points
  \[\sqrt{\sum_{i=1}^{n} [(X_i - \bar{X})^2 + (Y_i - \bar{Y})^2]}\]
- Application in location theory

Median Center

Geographic Data: Standard Distance

- Spatial equivalent to standard deviation
- Average distance of observations to mean center, or radius around mean center
  \[SD = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n}}\]
  \[SD = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}\]
Geographic Data: Standard Distance
- Affected by the unit which distance is measured
- Affected by the study area

Geographic Data: Relative Distance
- Dividing SD by the radius of a circle with area equal to the size of the study area
  \[ RD = \frac{SD}{r} \]

Summary
- Measures of central tendency
  - Mean, median, mode
- Measures of variability, skewness, kurtosis
  - Variance, standard deviation, range, IQR
  - Pearson’s coefficient, Quartile skewness, kurtosis
- Measures of relative position
  - Rank, quantile, Z-score
- Dummy variables
- Measures for geographic data
  - Mean center, median center, standard distance, relative distance

Individual activity: SPSS application
- Dataset: crimedata
- Variable: crime rate; poverty rate
- Find mean, median, mode; variance, standard deviation; quartiles, IQR
- Interpret the results, and compare these two variables.
- Create a z-score for crime rate; what can you say about Washington DC?

SPSS: Descriptive Statistics