Descriptive Statistics

Graphing, Central Tendency, & Dispersion

The Process of Quantitative Analysis

1. Collect quantitative data
2. **Conduct descriptive statistics**
   - Preparation for inferential statistics
     - Missing values?
     - “Odd” values?
     - Violation of test assumptions?
   - Select appropriate analyses
     - Data (variable) type to analysis
     - Distribution to analysis
3. Conduct desired inferential statistics
4. Draw conclusions based on results of descriptive and inferential statistics
Descriptive Statistics

- Often, students consider descriptive statistics to be “optional”
  - Rarely answer the experimental questions we’re interested in asking, so they are ignored
  - Do NOT skip this process.
    - Essential for determining how and if further inferential statistics may be conducted
    - Violations of inferential test assumptions lead to incorrect conclusions
      - Descriptive statistics often allow us to test these assumptions

Types of Descriptive Statistics

- Graphical data representations
  - Frequency distributions
  - Histograms
  - “Other” (stem & leaf, boxplots, etc.)
- Measures of central tendency
- Measures of dispersion
Graphical Representations

• Frequency distributions
  – A way of organizing the data in which each type or category of response is listed by the number of times it appeared in the data set.
  • Identify out of place values (a 6 on a 1 to 5 scale)
  • Identify extreme high or low values = outliers
  • Provides a “rough” idea of how participants responded to our question(s)

<table>
<thead>
<tr>
<th>Age</th>
<th>Frequency</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>1</td>
<td>4.5</td>
<td>4.5</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>4.5</td>
<td>9.1</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>9.1</td>
<td>18.2</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>4.5</td>
<td>22.7</td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td>9.1</td>
<td>31.8</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>18.2</td>
<td>50.0</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>9.1</td>
<td>59.1</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>4.5</td>
<td>63.6</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>9.1</td>
<td>72.7</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>4.5</td>
<td>77.1</td>
</tr>
<tr>
<td>37</td>
<td>1</td>
<td>4.5</td>
<td>81.8</td>
</tr>
<tr>
<td>39</td>
<td>1</td>
<td>4.5</td>
<td>86.4</td>
</tr>
<tr>
<td>43</td>
<td>1</td>
<td>4.5</td>
<td>90.9</td>
</tr>
<tr>
<td>45</td>
<td>1</td>
<td>4.5</td>
<td>95.5</td>
</tr>
<tr>
<td>82</td>
<td>1</td>
<td>4.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>100.0</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Graphical Representations

- Frequency Distribution for Gender

- Frequency Distribution for Height

<table>
<thead>
<tr>
<th>Gender</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>4</td>
<td>18.2%</td>
<td>18.2%</td>
<td>18.2%</td>
</tr>
<tr>
<td>Female</td>
<td>18</td>
<td>81.8%</td>
<td>81.8%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Height</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>60.0</td>
<td>1</td>
<td>4.5%</td>
<td>4.5%</td>
<td>4.5%</td>
</tr>
<tr>
<td>62.0</td>
<td>1</td>
<td>4.5%</td>
<td>4.5%</td>
<td>9.1%</td>
</tr>
<tr>
<td>63.0</td>
<td>4</td>
<td>18.2%</td>
<td>18.2%</td>
<td>27.3%</td>
</tr>
<tr>
<td>64.0</td>
<td>3</td>
<td>13.6%</td>
<td>13.6%</td>
<td>40.9%</td>
</tr>
<tr>
<td>65.0</td>
<td>2</td>
<td>9.1%</td>
<td>9.1%</td>
<td>50.0%</td>
</tr>
<tr>
<td>66.0</td>
<td>5</td>
<td>22.7%</td>
<td>22.7%</td>
<td>72.7%</td>
</tr>
<tr>
<td>67.0</td>
<td>2</td>
<td>9.1%</td>
<td>9.1%</td>
<td>81.8%</td>
</tr>
<tr>
<td>68.0</td>
<td>1</td>
<td>4.5%</td>
<td>4.5%</td>
<td>86.4%</td>
</tr>
<tr>
<td>69.0</td>
<td>2</td>
<td>9.1%</td>
<td>9.1%</td>
<td>95.5%</td>
</tr>
<tr>
<td>70.5</td>
<td>1</td>
<td>4.5%</td>
<td>4.5%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>100.0%</td>
<td>100.0%</td>
<td></td>
</tr>
</tbody>
</table>

Graphical Representations

- Creating Frequency Distributions
  - Take a set of values: 1, 2, 1, 5, 3, 4, 2, 1, 3, 4, 1, 2, 4, 3, 2, 1, 2, 3
  - Create a table reflecting how frequently they occur
  - Rounding error can produce totals slightly greater than 100%

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>28%</td>
<td>28%</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>28%</td>
<td>56%</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>22%</td>
<td>78%</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>17%</td>
<td>95%</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>6%</td>
<td>101%</td>
</tr>
<tr>
<td>Totals</td>
<td>18</td>
<td>101%</td>
<td></td>
</tr>
</tbody>
</table>
Graphical Representations

- Histograms
  - Provide a way of clumping observations together to provide a simplified graph of the results
    - Observe the general structure of the data
    - Demonstrate “trends”
    - Obscure “error”

Graphical Representations

- Frequency Distribution of GPA → Bar Chart
  - Hard to observe trends or systematic relationships in data
• So, we use a histogram
  – There were 16 distinct values of GPA
  – Reduce number of values to a series of “intervals”
    • *Intervals are ranges of values that we plot*
      – GPA of 3.00 and 3.10, for example, become 1 interval instead of two distinct values
      – Histograms ALWAYS have fewer intervals than values in the data set
      – The number of intervals used is arbitrary and up to the researcher
        » Typically, 5 to 10 intervals

### Graphical Representations

• GPA, before and after
Graphical Representations

- The number of intervals chosen can distort the observed findings

Describing Distributions

- Generally, we describe distributions depending on their shape
  - Distributions are described in terms of their similarity to the prototypical distribution, the normal distribution
    - Scores distributed symmetrically around a central peak—the peak representing the mean, or the most frequent value.
      - The “bell curve”
Describing Distributions

• Parts of a distribution
  – Center (1)
  – Shoulders (2)
  – Tails (2)

• Normal distributions have the correct number of scores/observation in the center, shoulders, and tails

Describing Distributions

• Kurtosis
  – Reflects the relative distribution of scores in the center, shoulders, and tails of the distribution
    • Mesokurtic distributions have the correct number of scores in all of their parts – normal distribution
    • Leptokurtic distributions have distributions where scores are shifted from the shoulders to the center and tails – peaked center and thick tails
    • Platykurtic distributions have scores shifted from the tails and shoulders to the center – creating a “plateau-like” appearance
  – Kurtosis is not “all or nothing”
    • Distributions are more or less like one of these descriptors
Describing Distributions

• Modality
  – Refers to the number of “peaks” present
    • 1 peak = unimodal
      – The normal distribution is unimodal
    • 2 peaks = bimodal
    • 3 or more peaks = multimodal
  – Theoretically, a distribution can have any number of peaks
    • Very rare to see anything other than unimodal or bimodal distributions

Describing Distributions

• Skew
  – Refers to an asymmetrical distribution of values around the peak
    • No skew
      – The normal distribution
    • Positive skew – “tail” points towards high values
      – “Low base-rate” behaviors
        » Schizophrenia, suicide, lottery winning
    • Negative skew – “tail” points towards low values
      – “High base-rate” behaviors
        » High self-esteem
Describing Distributions

• Skew
  – Can be measured mathematically
    • No skew = skew statistic of 0
    • Positive skew = positive values
    • Negative skew = negative values
  – We will not use mathematical values of skew in this class
    • Used only to describe shape of distribution

• Examples
…The Road to Central Tendency

• Numerical representations of data from this point on
• However, before we begin, we need to agree on some basic terminology

Notation

• Consider the following sets of data
  1. 1, 1, 4, 4, 5, 6, 8
  2. 56, 78, 88, 89, 91, 123, 145
  3. A, B, C, D, E, F

  – Each set of data represents a different variable (1 = X, 2 = Y, & 3 = Z)
  – X & Y are continuous and quantitative
  – Z is discrete and categorical
Notation

1. 1, 1, 4, 4, 5, 6, 8
2. 56, 78, 88, 89, 91, 123, 145
3. A, B, C, D, E, F

• For variable $X$, we refer to each piece of data in that variable with a subscript
  – $X_1 = 1$, $X_2 = 1$, $X_3 = 4$, $X_4 = 4$, $X_5 = 5$, etc…
  – Here, $Z_4 = D$
  – $Y_3$, $Y_6$, $Z_6$?
  – We use this notation for all items in a variable, $X_i$

Notation

• Another very common symbol we will use is upper-case sigma, $\Sigma$
  – We translate sigma to mean “add it up”
  – So, if we see $\Sigma X$, we know to add up ALL the values for variable $X$
    • $\Sigma X = 1 + 1 + 4 + 4 + 5 + 6 + 8 = 29$
  – So, if we see $\Sigma X^2$, we know to square all of the $X$ values and then to add all the squared values
    • $\Sigma X = 1^2 + 1^2 + 4^2 + 4^2 + 5^2 + 6^2 + 8^2$
    • $\Sigma X = 1 + 1 + 16 + 16 + 25 + 36 + 64 = 159$
Notation

• $\sum$ (continued)
  – So, if we see $(\sum X)^2$: we know to add up ALL the values for variable $X$ and then square them
    $$(\sum X)^2 = (1 + 1 + 4 + 4 + 5 + 6 + 8)^2 = 29^2 = 841$$
  – Similarly, if you see $\sum (Y - X)$, we know subtract the $X$ values from $Y$ and to then add the results
    $$\sum (Y - X) = (56 - 1) + (78 - 1) + (88 - 4) + (89 - 4) + (91 - 5) + (123 - 6) + (145 - 8)$$
    $$= 55 + 77 + 84 + 85 + 86 + 117 + 137 = 641$$

  – So, if we see $[\sum (Y - X)]^2$: we know to simply square the result we just obtained
    $$[\sum (Y - X)]^2 = 641^2 = 410,881$$
  – Finally, we resolve $\sum XY$ by first multiplying matching values of $X$ and $Y$ and summing the result
    $$\sum XY = (56*1) + (78*1) + (88*4) + (89*4) + (91*5) + (123*6) + (145*8)$$
    $$= 56 + 78 + 352 + 356 + 455 + 738 + 1160 = 3195$$
Notation

• Remember: if you’re confused about the order in which to calculate values
  – “Pretty-please my dear Aunt Sally.”
    • Powers
    • Parentheses
    • Multiply
    • Divide
    • Add
    • Subtract

Central Tendency

• Central tendency
  – *Refers to a set of measures that reflect where a distribution is located on a given scale* 
    • Central tendency is often referred to as a measure of "location"
    • Measures of Central tendency
      – Mean
      – Median
      – Mode
Central Tendency

• **Mode**
  – *The most common score*
    • Graphically, reflected as the “peak” of a distribution
    • Often considered to be the least useful measure of central tendency
    • Most common score
      – 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 7, 8, 9: Mode = 4
    • When there is more than 1 common, adjacent score, use the average of the two values
      – 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 6, 7, 8, 9: Mode = 4.5

Central Tendency

• **Mode**
  – *The most common score*
    • When there are 2 or more common, *non-adjacent* scores, the distribution is multimodal
      – 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 7, 8, 8, 8, 8, 9: Mode = 4,8
    • We might still consider a distribution multimodal if there are two common values with slightly different frequencies
      – 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 7, 8, 8, 8, 9: Mode = 4,8
Central Tendency

• Median
  – The point at which 50% of the scores are above and below—The 50th percentile
    • For odd numbers of scores, simply the middle value
      – 3, 5, 8, 9, 12, 15, 18: Median = 9
    • For even numbers of scores, the average of the two middle values
      – 3, 5, 8, 9, 12, 15: MEDIAN = (8+9)/2 or 8.5

• Median
  – The point at which 50% of the scores are above and below—The 50th percentile
    • One easy way to find where the median is located is to use the aptly named Median location formula
      – MEDIAN LOCATION = (N+1)/2
      – Where N= total number of data points
    • Thus, applying this formula to our first example:
      – MEDIAN LOCATION = (7+1)/2 = 4
      – 3, 5, 8, 9, 12, 15, 18
    • Our second example
      – MEDIAN LOCATION = (6+1)/2 = 3.5
      – 3, 5, 8, 9, 12, 15
Central Tendency

• Median
  – The point at which 50% of the scores are above and below—The 50th percentile
    • Often, we use the median to split a continuous variable into a two-value categorical variable
      – Continuous depression measure → Hi vs. Lo depression
        » This procedure is referred to as a median split
        » Hi depression above median
        » Lo depression below median
    – We will return to this concept throughout the semester

Central Tendency

• Mean
  – The average or
    • Population mean (parameter) = mu (µ)
    • Sample mean (statistic) = x-bar (\(\bar{x}\))

  • 3, 5, 8, 9, 12, 15, 18

  \[
  \bar{x} = \frac{\sum x}{n} = \frac{3 + 5 + 8 + 9 + 12 + 15 + 18}{7} = \frac{70}{7} = 10
  \]
Dispersion

• Dispersion or Variability
  – *The degree to which values around the mean, median, or mode vary*
    • How much values *disperse* from the mean
      – Are values tightly packed around the mean?
        » Low variability
      – Are values spread out away from the mean?
        » High variability
      – Graphical examples

Dispersion

• Measures of Dispersion or Variability
  – Range
  – Variance
  – Standard deviation
Dispersion

• Range
  – *Measure of distance between the highest and lowest score*
    • The least frequently used measure of dispersion
      – 1, 1, 2, 3, 3, 3, 4, 5, 6, 7, 8, 8, 8
      – Range = Highest score – Lowest score
      – Range = 8 – 1 = 7
    • Greatly affected by outliers
      – 1, 1, 2, 3, 3, 3, 4, 5, 6, 7, 8, 8, 255
      – Range = 255 – 1 = 254
    • Consequently, we must be cautious about interpreting the results of the range

Dispersion

• Variance
  – *We define the variance as the average\(^3\) of the squared\(^2\) deviations\(^1\) within our data*
    • A case of defining a value with an undefined value
    • Population variance (parameter) = sigma-squared (\(\sigma^2\))
    • Sample variance (statistic) = s-squared (s\(^2\))

1. Deviation
  – *The difference between a value within a data set and the mean of all the data within the data set*
    • Deviation score = \(x - \bar{x}\)
Dispersion

- Deviation\(^1\) scores
  - 3, 5, 8, 9, 12, 15, 18
  - Mean = 10

- Notice anything about deviation scores?

<table>
<thead>
<tr>
<th>Score</th>
<th>Mean</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>-7</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>-5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>-2</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>8</td>
</tr>
</tbody>
</table>

Dispersion

- Variance
  - Average of the squared deviations within our data

- Squared Deviations\(^2\)

<table>
<thead>
<tr>
<th>Score</th>
<th>Mean</th>
<th>Dev.</th>
<th>Dev. Sq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>10</td>
<td>-7</td>
<td>49</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>-5</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>-2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>10</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>10</td>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>18</td>
<td>10</td>
<td>8</td>
<td>64</td>
</tr>
</tbody>
</table>
Dispersion

- Variance
  - *Average of the squared deviations within our data*

- Average deviation\(^3\)
  - Once you have the squared deviations, sum them and divide by the total number of data points
    - \(\sigma^2 = \frac{49+25+4+1+4+25+64}{7} = \frac{172}{7} = 24.57\)
      - Note: this is the POPULATION value of the variance
    - \(s^2 = \frac{49+25+4+1+4+25+64}{(7-1)} = \frac{172}{6} = 28.67\)
      - Note: this is the SAMPLE value of the variance

- Why the difference? Stay tuned…

**Dispersion**

- Variance
  - \(\sigma^2 = \frac{\sum(x - \bar{x})^2}{n} = \frac{172}{7} = 24.57\)
  - \(s^2 = \frac{\sum(x - \bar{x})^2}{(n-1)} = \frac{172}{6} = 28.67\)
  - These are the equations we used.
Dispersion

• So, why the change?
  – Remember that x-bar is an estimate of μ
    • Whenever we estimate a value and expect to come up with a given value, we lose what is called a degree of freedom
      – We know that the mean of 1, 4, & 7 is 4
      – How many values can we freely change and still obtain the same mean?
        » Since we need to continue with the same mean, we can only freely change 2 values. The third is determined for us—it is not free to vary

Dispersion

• So, why the change?
  – Remember that x-bar is an estimate of μ
    • If we’re working in the population, we have the population mean μ—nothing needs to be estimated
    • However, if we’re working in a sample, we lose one degree of freedom because we expect the sample mean to equal the population mean
      – In practice, the sample and population means may not be equal—this is simply a mathematical practice that allows us to arrive at the most accurate estimates possible
Dispersion

• Variance
  – \( \sigma^2 = \frac{\sum (x - \bar{x})^2}{n} \)
  – \( s^2 = \frac{\sum (x - \bar{x})^2}{n-1} \)

\[
\sigma^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{172}{7} = 24.57
\]
\[
s^2 = \frac{\sum (x - \bar{x})^2}{n-1} = \frac{172}{6} = 28.67
\]
– The second form of the equations is a quicker, computational form.

Dispersion

• Standard deviation
  – Defined as the positive square root of the variance

\[
\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n}}
\]
\[
s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n-1}}
\]