Numerical Descriptive Measures for Quantitative Data I

Dr. Tom Ilvento
FREC 408

Measures of Central Tendency

- The central tendency of a variable is the tendency of the data to cluster or center about certain numerical values
- The variability is the spread of the data
- For central tendency we will focus on the mean, the mode, and the median

We need some tools

- Sigma Notation (p. 131)
- This relates to the rules of summation
- We start with the Greek Sigma $\Sigma$
  $$\sum_{i=1}^{n} x_i$$
  $$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4 + x_5 \ldots x_n$$

Sigma Notation

- For a variable $X$ with the set
  $$\{5, 7, 4, 3, 2, 5\}$$
  $$\sum_{i=1}^{n} x_i = 5 + 7 + 4 + 3 + 2 + 5 = 26$$

Rules of Sigma Notation

- The sum of a constant $c$
  $$\sum_{i=1}^{n} c = n \cdot c$$

Rules of Sigma Notation

- The sum of a constant times $x$
  $$\sum_{i=1}^{n} (c \cdot x_i) = c \sum_{i=1}^{n} x_i$$
**Rules of Sigma Notation**

Let \( c = 5 \) and \( x \) be the set \( \{2, 4, 5, 2\} \)

\[
\sum_{i=1}^{n} 5 \cdot x_i = (5 \cdot 2) + (5 \cdot 4) + (5 \cdot 5) + (5 \cdot 2) = 65
\]

\[
5 \sum_{i=1}^{n} x_i = 5 \cdot (2 + 4 + 5 + 2) = 5 \cdot 13 = 65
\]

**Rules of Sigma Notation**

Let:
- \( x \) be the set \( \{2, 4, 5, 2\} \)
- \( y \) be the set \( \{5, 3, 2, 1\} \)

\[
\sum_{i=1}^{n} (x_i + y_i) = (2 + 5) + (4 + 3) + (5 + 2) + (2 + 1) = 24
\]

\[
\sum_{i=1}^{n} x_i = (2 + 4 + 5 + 2)
\]

\[
\sum_{i=1}^{n} y_i = 13 + 11 = 24
\]

**Rules of Sigma Notation**

- The sum of the addition of two variables, \( x \) and \( y \)

\[
\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i
\]

**Rules of Sigma Notation**

Let \( a \) and \( c \) represent constants

\[
\sum_{i=1}^{n} (a \cdot x_i + c) = n \sum_{i=1}^{n} x_i + n \cdot c
\]

**Rules of Sigma Notation**

- The sum of the addition of two variables squared

\[
\sum_{i=1}^{n} (x_i + y_i)^2 = \sum_{i=1}^{n} (x_i^2 + 2x_i y_i + y_i^2) = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} (x_i \cdot y_i) + \sum_{i=1}^{n} y_i^2
\]

**Rules of Sigma Notation**

- Note

\[
\sum_{i=1}^{n} (x_i + y_i)^2 \neq \sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2
\]

\[
\sum_{i=1}^{n} x_i^2 \neq \left( \sum_{i=1}^{n} x_i \right)^2
\]
The Mean

- The arithmetic mean or mean (Def 3.1 p133) is the sum of the measurements divided by the number of measurements contained in the data set.
- For a sample we use \( x \) with a bar over it: \( \bar{x} \).
- For a population, we use the Greek \( \mu \).

Two ways to express the mean:

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

The sum of all the values, divided by the number of values.

\[
\bar{x} = \frac{\sum_{i=1}^{n} (x_i / n)}
\]

The sum of values weighted by the number of values.

Properties of the Mean

- As a measure of central tendency the mean has several advantages:
  - The first is that the mean uses information of all the values in a variable.

The mean has two important mathematical properties:

- The sum of the deviations about the mean equals zero.
- The sum of squared deviations about the mean is a minimum.

Properties of the mean

1. \( \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \)
   
   Here's the proof!

2. \( \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = 0 \)

3. \( \sum_{i=1}^{n} x_i - n \cdot \bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0 \)

Properties of the mean

- Sum of squared deviations about the mean is a minimum – Least Squares property

\[
\sum_{i=1}^{n} (x_i - \bar{x})^2
\]

There is no other value or constant we could substitute in the equation for the mean that would result in a lower sum of squares.
### Properties of the Mean

- We can make inferences from a sample to a population for the mean.
- The mean forms the basis for a number of other statistics known as Product Moment Statistics.
- But, the mean is sensitive to outliers and extremes in the data. It is not as resistant as other measures of central tendency.

### Marriage Rate Data

<table>
<thead>
<tr>
<th>Marriage Rate data for 1996</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 50 (50 states)</td>
</tr>
<tr>
<td><code>\sum x</code> = 523.36</td>
</tr>
<tr>
<td>Mean = 10.47</td>
</tr>
<tr>
<td>Rounded to Mean = 10.5</td>
</tr>
</tbody>
</table>

### Example using the Mean

**Marriage Rate data for 1996**

- n = 50 (50 states)
- `\sum x` = 523.36
- Mean = 10.47
- Rounded to Mean = 10.5

### Stem and Leaf Plots

<table>
<thead>
<tr>
<th>Stem Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
</tbody>
</table>

*Note: Stems are whole numbers, leafs are decimal places.*

### Marriage Data Example

**Marriage Data Example**

- Note: by removing Nevada the figures are:
  - N = 49
  - `\sum x` = 435.12
  - Mean = 8.88
- Compared to 10.5 for all 50 states

### The Median

- The median (Def 3.2 p135) is the middle value when the measurements are arranged in ascending order.
- It is a *positional* measure because it is based on the middle case in a variable.
- In order to find the median value, we first must sort the data in ascending or descending order.
Finding the Median

- First Sort the data
- If N is an odd number, the median is the \((N+1)/2\)th value in the ordered data.
- EXAMPLE: If \(N=99\), then the median value is the value of the \((99+1)/2 = 50\)th case.

Finding the Median

- But, if \(N\) is an even number, the median value will fall between the \(N/2\)th and the \(N/2+1\)th cases.
- EXAMPLE: If \(N = 100\) Median value is between the 50th and 51st cases.
  - In this case, we usually take the average of the two values to find the median value.

Properties of the Median

- The median is an intuitive measure of central tendency - the middle.
- However, it is actually difficult to compute because it requires you to sort the data.
- Spreadsheets and software packages will now calculate the median for us.

Properties of the Median

- The median has limited inferential properties
- But, it is not as sensitive to outliers and thus is used in data with extreme values
  - Income
  - Company size
  - Weather data

Median Example

- Marriage rate data for 1996
- \(N = 50\)
- Median value is between 25 value and the \((51+1)/2 = 26\)th value in an ordered sort of the data
- This is the average of 25th (Iowa) and 26th (New Hampshire) value both of which is 8.4

Median Example

- If we dropped Nevada, the median would be
  - \(N = 49\)
  - Median = 25th case
  - Median = 8.4
- This is identical to the 50 state data example
**Median**
- The median is also referred to as the 50th percentile.
- There are other ordered measures:
  - Quartiles
    - Q1 is the 25th percentile and Q3 is the 75th percentile
    - The median is also second quartile or Q2
  - Percentiles, deciles, and quintiles
- (We'll discuss more about percentile and quartile next class)

**The Mode**
- The mode (Def 3.3 p135) is the most frequent occurring value in a variable.
- In a qualitative variable, we refer to the Modal Class or Category.
- The mode may make more sense in reference to categorical data.

**Mode**
- In continuous level data, there may not be any single value that is the most frequent.
- We may also have the experience of multiple “modes” as in Bi-Modal or Tri-Modal.

**Marriage Data**

```
<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaf</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>6</td>
<td>1 5 6 7 9</td>
</tr>
<tr>
<td>7</td>
<td>0 1 1 1 6 6 7 8 8 8</td>
</tr>
<tr>
<td>8</td>
<td>0 0 1 4 4 4 4 5 6 8</td>
</tr>
<tr>
<td>9</td>
<td>0 0 0 1 4 4 8</td>
</tr>
<tr>
<td>10</td>
<td>1 2 4 9</td>
</tr>
<tr>
<td>11</td>
<td>0 1 5</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td>89</td>
<td>2</td>
</tr>
</tbody>
</table>
```

Note: Stems are whole numbers, leaffs are decimal places

- The median is 8.4
- The mode is undefined
- The mean is 10.5

**Comparing the Mean, Median, and Mode**
- In a symmetrical, bell shaped curve depicting the distribution of a variable, the mean, median, and mode would the same value
- The normal curve is a very special bell shaped curve where the mean, median, and mode are all equal.
### EPA Mileage Ratings

<table>
<thead>
<tr>
<th>#</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>31</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>33</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>34</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>35</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- **N** = 100
- Mean = 36.99
- Median = 37.0
- Mode = 37.0

### Excel Spreadsheet of EPA Data

<table>
<thead>
<tr>
<th>FREC 408</th>
<th>EPA Mileage Ratings on 100 Cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>N MPG</td>
<td></td>
</tr>
<tr>
<td>1 30.0</td>
<td>Sum =SUM(B5:B104) 3699.40</td>
</tr>
<tr>
<td>2 31.8</td>
<td>Count =COUNT(B5:B104) 100.00</td>
</tr>
<tr>
<td>3 32.5</td>
<td>Mean =AVERAGE(B5:B104) 36.99</td>
</tr>
<tr>
<td>4 32.7</td>
<td>Minimum =MIN(B5:B104) 30.00</td>
</tr>
<tr>
<td>5 32.9</td>
<td>Maximum =MAX(B5:B104) 44.60</td>
</tr>
<tr>
<td>6 32.9</td>
<td>Median =MEDIAN(B5:B104) 37.00</td>
</tr>
<tr>
<td>7 33.1</td>
<td>Mode =MODE(B5:B104) 37.00</td>
</tr>
<tr>
<td>8 33.2</td>
<td></td>
</tr>
<tr>
<td>9 33.8</td>
<td></td>
</tr>
<tr>
<td>10 33.8</td>
<td></td>
</tr>
<tr>
<td>11 33.8</td>
<td></td>
</tr>
</tbody>
</table>

### Go to the Spreadsheet of EPA Data