Describing Data With Measures of Location

Joseph J. Ferrare, Ph.D.
School of Interdisciplinary Arts & Sciences
University of Washington Bothell
www.josephferrare.com
Measures of Location: Where are the data?

- Frequency distributions are helpful to understand possible and observed values in our data.
- Another important piece of information concerns the location of specific values in these distributions.
Where is the Minimum and Maximum?

- **Minimum**: smallest observed value in the distribution
- **Maximum**: largest value
- **Range**: max - min

DESCRIPTIVES VARIABLES=SEI10
/STATISTICS=RANGE MIN MAX.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>R’s socioeconomic index (2010)</td>
<td>2248</td>
<td>82.2</td>
<td>10.6</td>
<td>92.8</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Double-click to activate
What is typical?

We often ask questions concerning what is typical. These questions also concern the location of a specific value in a distribution.

1. What is the weather typically like during the winter in Seattle?
2. How many inches of snow fall during a typical winter in the Cascades?
3. How many students typically take BIS 315 each year?
Measures of Central Tendency

(Arithmetic) **Mean**: the sum of measurements divided by the total number of measurements

In the population, the mean is defined by the parameter, 

$$
\mu = \frac{\sum X}{N}, \text{ or } \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N},
$$

where the $X_i$’s refer to the individual values and $N$ refers to the total number in the population.


**Sample Mean**

(Arithmetic) **Mean**: the sum of measurements divided by the total number of measurements

In the sample, the mean is defined by the statistic (M),

\[ M = \frac{\sum X}{N}, \text{ or } \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}, \]

but the computation is effectively identical (note: the sample mean is often denoted by the symbol X bar, \( \bar{X} \))
Sample Mean from SEI10 variable in GSS

\[ \bar{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_{2,248}}{2,248} \]

\[ \bar{X} = 105,531.2 / 2,248 \]

\[ \bar{X} = 46.94 \]
Algebraic Property of the Mean

The mean has an intuitive property such that the sum of deviations of each value from the mean will always equal zero:

\[ \sum_{i=1}^{N} (X_i - \bar{X}) = 0 \]

That is, the mean, by definition, finds the point in the distribution around which the sum of deviations is zero.
# Median

The **median** refers to the value in the middle of an ordered distribution.

This is simple with an odd number of data points:

<table>
<thead>
<tr>
<th>Data Points</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 6, 27)</td>
<td>6</td>
<td>11.67</td>
</tr>
</tbody>
</table>

With an even number, take the midpoint between the two middle points:

<table>
<thead>
<tr>
<th>Data Points</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 6, 19, 27)</td>
<td>12.5 (i.e., [6+19]/2=12.5)</td>
<td>13.5</td>
</tr>
</tbody>
</table>

Even if there are two or more median values:

<table>
<thead>
<tr>
<th>Data Points</th>
<th>Median</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(2, 6, 6)</td>
<td>6</td>
<td>4.67</td>
</tr>
</tbody>
</table>
Median of SEI10 variable in GSS

Median = 41.00
\[ \bar{X} = 46.94 \]
Median for Ordinal Data

Simply find the category where the 50% response falls

<table>
<thead>
<tr>
<th>R's highest degree</th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LT HIGH SCHOOL</td>
<td>262</td>
<td>11.2</td>
<td>11.2</td>
<td>11.2</td>
</tr>
<tr>
<td>HIGH SCHOOL</td>
<td>1178</td>
<td>50.2</td>
<td>50.2</td>
<td>61.3</td>
</tr>
<tr>
<td>JUNIOR COLLEGE</td>
<td>196</td>
<td>8.3</td>
<td>8.3</td>
<td>69.7</td>
</tr>
<tr>
<td>BACHELOR</td>
<td>465</td>
<td>19.8</td>
<td>19.8</td>
<td>89.5</td>
</tr>
<tr>
<td>GRADUATE</td>
<td>247</td>
<td>10.5</td>
<td>10.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>2348</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>
Percentiles

The median is a specific type of measure of location that falls in a more general class of measures called percentiles.

The \( p \)th percentile identifies the \( p\% \) of observations falling below and \( (100-p) \) falling above.

If \( p=50 \) (i.e., 50th percentile) then we have the median.

The 25th percentile is called the lower quartile and the 75th percentile is called the upper quartile.
The middle half of the observations – i.e., between the lower and upper quartile – refers to the **inter-quartile range**.

Note: this anticipates the concept of **variability**, discussed next time.
Percentiles for SEI10 Variable (incl. Min/Max)

Min: 10.6
25th: 25.2
50th: 41.0
75th: 65.5
Max: 92.8
Using a Box Plot (AKA Box & Whisker Plot)
How SPSS Calculates Percentiles

E.g., 25th percentile
1. Compute Rank (R)
   \[ R = \frac{P}{100} \times (N+1) \]
   \[ = 0.25 \times 2249 \]
   \[ = 562.25 \]
2. Find the 562nd and 563rd ranked observations
   \[ = 25.2 \text{ & } 25.2 \]
3. Multiply R by the difference in the two values and add the lower value
   \[ = 0.25(25.2 - 25.2) + 25.2 \]
   \[ = 25.2 \]
Mode

The **mode** is the most frequently occurring measurement in a distribution.

Suppose you have the following data:

1, 1, 1, 4, 5, 8, 9, 9

Mode = 1

In some cases there may be more than one mode:

1, 1, 1, 4, 5, 8, 9, 9, 9

Mode = 1 & 9

In other cases there may not be a mode:

1, 4, 5, 8, 9, 10, 11

Mode = Undefined
Mode of SEI10 variable in GSS

Mode = 38.8
Median = 41.00
\( \bar{X} = 46.94 \)
Mode of a Nominal Variable

Type of place lived in when 16 yrs old

Percent

Type of place lived in when 16 yrs old

COUNTRY, NONFARM
FARM
TOWN LT 50000
50000 TO 250000
BIG-CITY SUBURB
CITY GT 250000
Some (Not So) Deep Thoughts

1. When a distribution is unimodal and symmetric, the mean, median, and mode will be identical.
2. The mean is influenced by outliers, while the median is not. The median may be more appropriate for highly skewed data.
3. The mean can only be used with quantitative data, the median can be used with ordinal or interval/ratio, and the mode can be used with all scales (though not necessarily in an informative way).
4. The median may be less informative than the mean when a variable has few values.
5. The mean, median, and mode can paint very different pictures of what is typical for the same data. This can lead to misuse, whether intentional or not.