Review & Synthesis, Part 1

Joseph J. Ferrare, Ph.D.
School of Interdisciplinary Arts & Sciences
University of Washington Bothell
www.josephferrare.com
What are statistics?

Statistics are tools we can use to collect and analyze data

1. **Description**: summarizing or exploring patterns and relationships in data

2. **Inference**: making generalizations or predictions about some phenomena based on the available data
Populations and Samples

- **A population** (or universe) is the total set of entities that are of descriptive or inferential interest.

- **A sample** is the subset of entities drawn from the population that serve as the basis for describing or inferring knowledge about the population.

Source: Mr. Hays has flipped
Parameters and Statistics

- A **parameter** is a summary characteristic of a population (e.g., the mean ACT among test-takers).

- A **statistic** is a summary of the sample (e.g., the mean ACT score among a random sample of 1,000 Washington State students).

Source: Quara.com
What are variables and why do we care?

**Variables** are qualities of an object (e.g. person, event) that can take on multiple values.

Examples are endless:
- Income
- Education
- Religious beliefs
- Temperature
- Blood type
What are variables and why do we care?

- **Independent** variable (X): the manipulated variable in an experiment
  - Also referred to as explanatory variables

- **Dependent** variable (Y): the variable the measures the outcome of an experiment
  - Also referred to as outcome variables
  - The outcome is “dependent” on the independent variable
Types of scales for variables (measurement)

- **Nominal**: a set of categories that simply name a qualitative attribute
  - No ordering to the categories
  - Sometimes called categorical variables
  - E.g., Variable - mode of transportation: car, bike, bus, train, jetpack

Source: Popular Science
Types of scales for variables (measurement)

- **Ordinal**: a set of categories that order a variable, but without defined interval distances
  - E.g., Variable - degree: high school diploma, associate, bachelor, graduate

Source: Someone on the internet
Types of scales for variables (measurement)

- **Interval**: a set of quantitative values that order a variable with defined interval distances
  - E.g., Variable - SAT score

- If an interval variable has a non-arbitrary zero point, it is considered a **ratio** scale (e.g., Fahrenheit v. Kelvin)

Source: Carson et al. (1993)
1. **Frequency**: number of cases observed at each value
2. **Relative frequency**: the frequency expressed as the proportion (P) of the total (i.e., \( \frac{N_i}{N} \)), where \( N_i \) = the observed number of cases and \( N \) = the total number of cases
   - Often expressed as a percentage (i.e. \( P \times 100 \)) to express the frequency as a per 100 ratio.
3. **Valid frequency**: in SPSS, the frequency among non-missing or otherwise excluded cases (see next slide)
4. **Cumulative frequency**: the sum of the cases observed from one value to the next (most useful with ordinal or interval vars)
Frequency Distributions

1. Frequency
2. Relative frequency
3. Valid frequency
4. Cumulative frequency

Relative frequency = \( \frac{N_i}{2348} \times 100 \)

Valid frequency = \( \frac{N_i}{2347} \times 100 \)

### Living with parents when 16 yrs old

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Valid</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>OTHER</td>
<td>65</td>
<td>2.8</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>MOTHER &amp; FATHER</td>
<td>1499</td>
<td>63.8</td>
<td>63.9</td>
<td>66.6</td>
</tr>
<tr>
<td></td>
<td>FATHER &amp; STPMOTHER</td>
<td>37</td>
<td>1.6</td>
<td>1.6</td>
<td>68.2</td>
</tr>
<tr>
<td></td>
<td>MOTHER &amp; STPFATHER</td>
<td>150</td>
<td>6.4</td>
<td>6.4</td>
<td>74.6</td>
</tr>
<tr>
<td></td>
<td>FATHER</td>
<td>74</td>
<td>3.2</td>
<td>3.2</td>
<td>77.8</td>
</tr>
<tr>
<td></td>
<td>MOTHER</td>
<td>411</td>
<td>17.5</td>
<td>17.5</td>
<td>95.3</td>
</tr>
<tr>
<td></td>
<td>MALE RELATIVE</td>
<td>9</td>
<td>.4</td>
<td>.4</td>
<td>95.7</td>
</tr>
<tr>
<td></td>
<td>FEMALE RELATIVE</td>
<td>46</td>
<td>2.0</td>
<td>2.0</td>
<td>97.6</td>
</tr>
<tr>
<td></td>
<td>M AND F RELATIVES</td>
<td>56</td>
<td>2.4</td>
<td>2.4</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2347</td>
<td>100.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>NA</td>
<td>1</td>
<td>.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2348</td>
<td>100.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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, $\bar{X}$)
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:

- \((2, 6, 27)\), median = 6
- Mean = 11.67

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

- \((2, 6, 19, 27)\), median = 12.5
  (i.e., \([6+19]/2=12.5\))
- Mean = 13.5

Even if there are two or more median values:

- \((2, 6, 6)\), median = 6
- Mean = 4.67
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., $50^{th}$ percentile) then we have the **median**.

The $25^{th}$ percentile is called the **lower quartile** and the $75^{th}$ percentile is called the **upper quartile**.
# Mode

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

<table>
<thead>
<tr>
<th>Suppose you have the following data:</th>
<th>In some cases there may be more than one mode:</th>
<th>In other cases there may not be a mode:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 1, 1, 4, 5, 8, 9, 9</td>
<td>1, 1, 1, 4, 5, 8, 9, 9</td>
<td>1, 4, 5, 8, 9, 10, 11</td>
</tr>
<tr>
<td>Mode = 1</td>
<td>Mode = 1 &amp; 9</td>
<td>Mode = Undefined</td>
</tr>
</tbody>
</table>
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.
Measuring Spread: Range

The spread of the distribution, as measured by range, indicates that U.S. adults in this sample spend anywhere from 0 to 100 hours per week on email.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>N</th>
<th>Range</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Email hours per week</td>
<td>1419</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>7.15</td>
</tr>
<tr>
<td>Valid N (listwise)</td>
<td>1419</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measuring Spread: Inter-Quartile Range (IQR)

\[
IQR = 75^{th} - 25^{th} \text{ Percentile}
\]

The spread of the distribution, as measured by IQR, indicates that the middle 50% of U.S. adults in this sample spend anywhere from 0 to 10 hours per week on email.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Email hours per week</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>Valid 1419</td>
</tr>
<tr>
<td></td>
<td>Missing 929</td>
</tr>
<tr>
<td>Mean</td>
<td>7.15</td>
</tr>
<tr>
<td>Percentiles</td>
<td>25 .00</td>
</tr>
<tr>
<td></td>
<td>50 2.00</td>
</tr>
<tr>
<td></td>
<td>75 10.00</td>
</tr>
</tbody>
</table>
Measuring Spread: Variance

The **variance** is a ratio that represents the average distance between the mean and the all observed values in the distribution. Formally, in the population, the variance is defined by the parameter,

$$
\sigma^2 = \frac{\sum(X_i-\mu)^2}{N} ,
$$

where the $X_i$’s refer to the individual values, $\mu$ refers to the population mean, and $N$ refers to the total number in the population.
Measuring Spread: Variance

The (unbiased) sample variance is defined by the statistic,

\[ S^2 = \frac{\sum (X_i - M)^2}{n-1}, \]

where the \( X_i \)'s refer to the individual values in the sample, \( M \) refers to the sample mean, and \( n \) refers to the total number in the sample.
The standard deviation is simply the square root of the variance. Formally, in the population, the standard deviation is defined by the parameter,

$$\sigma = \sqrt{\frac{\sum(X_i-\mu)^2}{N}},$$

where the $X_i$'s refer to the individual values, $\mu$ refers to the population mean, and $N$ refers to the total number in the population.
Sample Standard Deviation

The **sample standard deviation** is simply the square root of the **sample variance**. Formally, in the population, the standard deviation is defined by the parameter,

$$s = \sqrt{\frac{\sum (X_i - M)^2}{n-1}},$$

where the $X_i$'s refer to the individual values, $M$ refers to the sample mean, and $n$ refers to the total number in the sample.
The Shape of a Variable’s Distribution

Symmetric: the shape of the distribution is the same about the middle
➢ Later we will explore this through the concept of normal probability distributions
The Shape of a Variable’s Distribution

**Positive skew**: tail of the distribution stretches further into the positive values of the distribution

- Also called “skewed right”
The Shape of a Variable’s Distribution

**Negative skew**: tail of the distribution stretches further into the negative values of the distribution

- Also called “skewed left”
Some (Even Less) Deep Thoughts

1. All of the measures of dispersion discussed require ranked data, thus none of them are appropriate for nominal variables.
2. For ordinal variables, use the range or IQR.
3. For interval/ratio variables, all three are appropriate. However, as with the mean, the variance and standard deviation are susceptible to outliers. In such cases, the IQR may be preferable when describing the spread (or all 3 at the same time).