Lecture 3: Descriptive Statistics

Carolyn Anderson & Youngshil Paek
(Slide contributors: Shuai Wang, Yi Zheng, Michael Culbertson, & Haiyan Li)
Department of Educational Psychology
University of Illinois at Urbana-Champaign
Topic 1: Describing the Center of Quantitative Data
Key Points

1. Calculating the mean
2. Calculating the median
3. Comparing the mean & median
4. Identify the mode of a distribution
5. Definition of resistant
Mean

- The mean is the sum of the observations divided by the number of observations.
- It is the center of mass.

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

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

- The median is the midpoint of the observations when they are ordered from the smallest to the largest (or from the largest to smallest).

- If the number of observations is:
  - Odd, then the median is the middle observation
  - Even, then the median is the average of the two middle observations
# Median

1) Sort observations by size.

\[ n = \text{number of observations} \]

<table>
<thead>
<tr>
<th>Order</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>78</td>
</tr>
<tr>
<td>2</td>
<td>91</td>
</tr>
<tr>
<td>3</td>
<td>94</td>
</tr>
<tr>
<td>4</td>
<td>98</td>
</tr>
<tr>
<td>5</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>101</td>
</tr>
<tr>
<td>7</td>
<td>103</td>
</tr>
<tr>
<td>8</td>
<td>105</td>
</tr>
<tr>
<td>9</td>
<td>114</td>
</tr>
</tbody>
</table>

2.a) If \( n \) is **odd**, the median is the observation \( \frac{n+1}{2} \) down the list.

\[ n = 9 \]

\[ \frac{n+1}{2} = \frac{10}{2} = 5 \]

Median = 99

2.b) If \( n \) is **even**, the median is the mean of the two middle observations.

\[ n = 10 \]

\[ \frac{n+1}{2} = 5.5 \]

Median = \( \frac{99+101}{2} = 100 \)
Find the mean and median

$\text{CO}_2$ Pollution levels in 8 largest nations measured in metric tons per person:

$2.3\ 1.1\ 19.7\ 9.8\ 1.8\ 1.2\ 0.7\ 0.2$

a. Mean = $4.6$    Median =  $1.5$

b. Mean = $4.6$    Median =  $5.8$

c. Mean = $1.5$    Median =  $4.6$
In 2004 one of the questions on the GSS survey asked respondents, “To how many service clubs do you belong?” Find the median from the data listed in the table below.

<table>
<thead>
<tr>
<th>No. of Organizations</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 organization</td>
<td>111</td>
</tr>
<tr>
<td>2 organizations</td>
<td>14</td>
</tr>
<tr>
<td>3 organizations</td>
<td>14</td>
</tr>
<tr>
<td>4 organizations</td>
<td>3</td>
</tr>
<tr>
<td>7 or more organizations</td>
<td>1</td>
</tr>
</tbody>
</table>

a) 0 organizations
b) 1 organization
c) 2 organizations
d) 3 organizations
e) Cannot be determined
Comparing the Mean and Median

- The mean and median of a symmetric distribution are close together.

- For symmetric distributions, the mean is typically preferred because it takes the values of all observations into account.
Comparing the Mean and Median

- In a skewed distribution, the mean is farther out in the long tail than is the median.
  - For highly skewed distributions, the median is preferred because it is better representative of a typical observation.
Mode

- Mode
  - Value that occurs most often
  - Highest point on a distribution
  - The mode is most often used with categorical data
In 2006 the GSS asked, “How often do you read a traditional newspaper?” The answers are displayed in the table below. What is the mode?

a) Every Day
b) A few times a week
c) Once a week
d) Less than once a week
e) Never

![Bar chart showing frequency of newspaper reading]

- Every Day
- A few times a week
- Once a week
- Less than once a week
- Never
Resistant Measures

- A numerical summary measure is resistant if extreme observations (outliers) have little, if any, influence on its value
  - The Median and Mode are resistant to outliers
  - The Mean is not resistant to outliers
Key Points Revisited

1. Calculating the mean
2. Calculating the median
3. Comparing the mean & median
4. Identify the mode of a distribution
5. Definition of resistant
Topic 2: Describing the Variability of Quantitative Data
Key Points

1. Calculate the range
2. Calculate the standard deviation
3. Properties of the standard deviation
4. Interpret the magnitude of $s$ : The empirical rule
Range

- One way to measure the spread is to calculate the range.
- The range is the difference between the largest and smallest values in the data set.

\[ \text{Range} = \text{max} - \text{min} \]

- The range is strongly affected by outliers.
Deviation

- Each data value has an associated deviation from the mean, $x - \bar{x}$
- A deviation is positive if it falls above the mean and negative if it falls below the mean
- The sum of the deviations is always zero
**Standard Deviation**

- The standard deviation represents a typical distance or a type of “average distance” of an observation from the mean.

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

- The variance is the square of standard deviation.

  \[ s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \]
Calculating the Standard Deviation

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

- Find the mean
- Find the deviation of each value from the mean
- Square the deviations
- Sum the squared deviations
- Divide the sum by \( n - 1 \)

*(gives typical squared deviation from mean)*
Calculating the Standard Deviation

Metabolic rates of 7 men (cal./24hr.):
1792  1666  1362  1614  1460  1867  1439

\[
\bar{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7}
\]

= \frac{11,200}{7}

= 1600
Calculating the Standard Deviation

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations</th>
<th>Squared deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_i$</td>
<td>$x_i - \bar{x}$</td>
<td>$(x_i - \bar{x})^2$</td>
</tr>
<tr>
<td>1792</td>
<td>1792 - 1600 = 192</td>
<td>(192)^2 = 36,864</td>
</tr>
<tr>
<td>1666</td>
<td>1666 - 1600 = 66</td>
<td>(66)^2 = 4,356</td>
</tr>
<tr>
<td>1362</td>
<td>1362 - 1600 = -238</td>
<td>(-238)^2 = 56,644</td>
</tr>
<tr>
<td>1614</td>
<td>1614 - 1600 = 14</td>
<td>(14)^2 = 196</td>
</tr>
<tr>
<td>1460</td>
<td>1460 - 1600 = -140</td>
<td>(-140)^2 = 19,600</td>
</tr>
<tr>
<td>1867</td>
<td>1867 - 1600 = 267</td>
<td>(267)^2 = 71,289</td>
</tr>
<tr>
<td>1439</td>
<td>1439 - 1600 = -161</td>
<td>(-161)^2 = 25,921</td>
</tr>
<tr>
<td>sum = 0</td>
<td>sum = 214,870</td>
<td></td>
</tr>
</tbody>
</table>

$s^2 = \frac{214,870}{7} \cdot \frac{1}{6} = 35,811.67$

$s = \sqrt{35,811.67} = 189.24$ calories
Properties of the Standard Deviation

- $s$ measures the spread of the data
- $s = 0$ only when all observations have the same value, otherwise $s > 0$. As the spread of the data increases, $s$ gets larger.
- $s$ has the same units of measurement as the original observations. The variance $= s^2$ has units that are squared.
- $s$ is not resistant. Strong skewness or a few outliers can greatly increase $s$. 
EMPIRICAL RULE

If a distribution of data is bell-shaped, then approximately

- 68% of the observations fall within 1 standard deviation of the mean, that is, between $\bar{x} - s$ and $\bar{x} + s$ (denoted $\bar{x} \pm s$).
- 95% of the observations fall within 2 standard deviations of the mean ($\bar{x} \pm 2s$).
- All or nearly all observations fall within 3 standard deviations of the mean ($\bar{x} \pm 3s$).
Suppose that the height of college males has a bell shaped distribution with a mean of 70 inches and a standard deviation of 2 inches. Approximately what percentage of college males are between 66 and 74 inches?

a) 68%
b) 90%
c) 95%
d) 99.7%
e) 100%
Key Points

1. Calculate the range
2. Calculate the standard deviation
3. Properties of the standard deviation
4. Interpret the magnitude of $s$: The empirical rule
Topic 3: Measures of Positions
Key Points

1. Percentile, quartile
2. Calculating interquartile range and detecting potential outliers
3. Drawing boxplots
4. Comparing distributions
5. Calculating a z-score
The $p^{th}$ percentile is a value such that $p$ percent of the observations fall below or at that value.
Finding Quartiles

- Splits the data into four parts
  - Arrange the data in order
  - The median is the second quartile, $Q_2$
  - The first quartile, $Q_1$, is the median of the lower half of the observations
  - The third quartile, $Q_3$, is the median of the upper half of the observations
Finding Quartiles

Quartiles divide a ranked data set into four equal parts.

The first quartile, Q₁, is the value in the sample that has 25% of the data at or below it and 75% above.

The second quartile, Q₂, is the same as the median of a data set. 50% of the data are above the median and 50% are below.

The third quartile, Q₃, is the value in the sample that has 75% of the data at or below it and 25% above.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>2</td>
<td>1.2</td>
</tr>
<tr>
<td>3</td>
<td>1.6</td>
</tr>
<tr>
<td>4</td>
<td>1.9</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>2.1</td>
</tr>
<tr>
<td>7</td>
<td>2.3</td>
</tr>
<tr>
<td>8</td>
<td>2.3</td>
</tr>
<tr>
<td>9</td>
<td>2.5</td>
</tr>
<tr>
<td>10</td>
<td>2.8</td>
</tr>
<tr>
<td>11</td>
<td>2.9</td>
</tr>
<tr>
<td>12</td>
<td>3.3</td>
</tr>
<tr>
<td>13</td>
<td>3.4</td>
</tr>
<tr>
<td>14</td>
<td>3.6</td>
</tr>
<tr>
<td>15</td>
<td>3.7</td>
</tr>
<tr>
<td>16</td>
<td>3.8</td>
</tr>
<tr>
<td>17</td>
<td>3.9</td>
</tr>
<tr>
<td>18</td>
<td>4.1</td>
</tr>
<tr>
<td>19</td>
<td>4.2</td>
</tr>
<tr>
<td>20</td>
<td>4.5</td>
</tr>
<tr>
<td>21</td>
<td>4.7</td>
</tr>
<tr>
<td>22</td>
<td>4.9</td>
</tr>
<tr>
<td>23</td>
<td>5.3</td>
</tr>
<tr>
<td>24</td>
<td>5.6</td>
</tr>
<tr>
<td>25</td>
<td>6.1</td>
</tr>
</tbody>
</table>

\[
M = \text{median} = 3.4
\]

\[
Q₁ = \text{first quartile} = 2.2
\]

\[
Q₃ = \text{third quartile} = 4.35
\]
Quartile Example

Find the first and third quartiles

Prices per share of 10 most actively traded stocks on NYSE (rounded to nearest $)

2  4  11  13  14  15  31  32  34  47

a.  $Q_1 = 2 \quad Q_3 = 47$

b.  $Q_1 = 12 \quad Q_3 = 31$

c.  $Q_1 = 11 \quad Q_3 = 32$

d.  $Q_1 = 12 \quad Q_3 = 33$
Calculating Interquartile range

• The interquartile range is the distance between the third quartile and first quartile:
  • \( IQR = Q_3 - Q_1 \)
  • \( IQR \) gives spread of middle 50% of the data
Criteria for identifying an outlier

- An observation is a potential outlier if it falls more than $1.5 \times IQR$ below the first quartile or more than $1.5 \times IQR$ above the third quartile.
Below are some descriptive statistics about the median household income per state. Is Puerto Rico, with $20,107 dollars as its median income, a potential outlier?

a) Yes, it is below Q1 – 1.5*IQR.
b) No, it is not below Q1 – 1.5*IQR.
c) Yes, it is the minimum so it is an outlier.
d) It cannot be determined from the given information.

\[
\begin{align*}
&\bar{x} = 55,062 \\
&s = 9899 \\
&\text{min} = 20107 \\
&Q_1 = 48916 \\
&Q_3 = 61401 \\
&\text{max} = 75541
\end{align*}
\]
5 Number Summary

The five-number summary of a dataset consists of the
- Minimum value
- First Quartile
- Median
- Third Quartile
- Maximum value
Boxplot

- A box goes from the Q1 to Q3
- A line is drawn inside the box at the median
- A line goes from the lower end of the box to the smallest observation that is not a potential outlier and from the upper end of the box to the largest observation that is not a potential outlier
- The potential outliers are shown separately
Comparing Distributions Using Box Plots

Box Plots do not display the shape of the distribution as clearly as histograms, but are useful for making graphical comparisons of two or more distributions.
A personal trainer was interested in studying the effects of different types of diets (liquid diet, prepared meals, and low carb) on total weight loss in two months. What description below best describes the boxplots below?

a) The range of the data sets are very similar.

b) The median weight loss is similar for the diets.

c) The third quartiles weight loss is similar for the diets.
Z-Score

• The z-score for an observation is the number of standard deviations that it falls from the mean

\[ z = \frac{\text{observation} - \text{mean}}{\text{standard deviation}} \]

• An observation from a bell-shaped distribution is a potential outlier if its z-score < −3 or > +3
Key Points Revisited

1. Percentile, quartile
2. Calculating interquartile range and detecting potential outliers
3. Drawing boxplots
4. Comparing distributions
5. Calculating a z-score