STAT 155 Introductory Statistics

Lecture 3: Displaying Distributions with Numbers
Exploratory Data Analysis (EDA)

- **Graphical Visualization: Shape**
  - Bar Graph
  - Pie Chart
  - Stem plot
  - Histogram
  - Time plot *(not for distribution, but for changing pattern over time)*

- **Numerical Summary: Center and Spread**
  - Center:
    - Mean, Median and Mode
  - Spread:
    - Quartiles, Five-number summary and Boxplot
    - Standard Deviation
  - Choose one from each category.
What is the average highway (city) mileage?
What is the "middle value" of highway (city) mileage?

Table 1.10
Fuel economy (miles per gallon) for model year 2004 vehicles

<table>
<thead>
<tr>
<th>Two-seater cars</th>
<th>Minicompact cars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>City</td>
</tr>
<tr>
<td>Acura NSX</td>
<td>17</td>
</tr>
<tr>
<td>Audi TT Roadster</td>
<td>20</td>
</tr>
<tr>
<td>BMW Z4 Roadster</td>
<td>20</td>
</tr>
<tr>
<td>Cadillac XLR</td>
<td>17</td>
</tr>
<tr>
<td>Chevrolet Corvette</td>
<td>18</td>
</tr>
<tr>
<td>Dodge Viper</td>
<td>12</td>
</tr>
<tr>
<td>Ferrari 360 Modena</td>
<td>11</td>
</tr>
<tr>
<td>Ferrari Maranello</td>
<td>10</td>
</tr>
<tr>
<td>Ford Thunderbird</td>
<td>17</td>
</tr>
<tr>
<td>Honda Insight</td>
<td>60</td>
</tr>
<tr>
<td>Lamborghini Gallardo</td>
<td>9</td>
</tr>
<tr>
<td>Lamborghini Murcielago</td>
<td>9</td>
</tr>
<tr>
<td>Lotus Esprit</td>
<td>15</td>
</tr>
<tr>
<td>Maserati Spyder</td>
<td>12</td>
</tr>
<tr>
<td>Mazda Miata</td>
<td>22</td>
</tr>
<tr>
<td>Mercedes-Benz SL500</td>
<td>16</td>
</tr>
<tr>
<td>Mercedes-Benz SL600</td>
<td>13</td>
</tr>
<tr>
<td>Nissan 350Z</td>
<td>20</td>
</tr>
<tr>
<td>Porsche Boxster</td>
<td>20</td>
</tr>
<tr>
<td>Porsche Carrera 911</td>
<td>15</td>
</tr>
<tr>
<td>Toyota MR2</td>
<td>26</td>
</tr>
</tbody>
</table>
Measuring center: the mean $\bar{x}$

- **Mean** = Average value
- **The sample mean** $\bar{x}$: If the $n$ observations in a sample are $x_1, x_2, \ldots, x_n$, then their mean is

\[
\bar{x} = \frac{1}{n} \sum x_i
\]
Measuring center: the median

THE MEDIAN \( M \)

The median \( M \) is the midpoint of a distribution, the number such that half the observations are smaller and the other half are larger. To find the median of a distribution:

1. Arrange all observations in order of size, from smallest to largest.

2. If the number of observations \( n \) is odd, the median \( M \) is the center observation in the ordered list. Find the location of the median by counting \( (n + 1)/2 \) observations up from the bottom of the list.

3. If the number of observations \( n \) is even, the median \( M \) is the mean of the two center observations in the ordered list. The location of the median is again \( (n + 1)/2 \) from the bottom of the list.
Example: Fuel economy (miles per gallon) for 2004 two-seater cars

- Look at the **Highway mileage (w/o Honda Insight):**
  - Mean
  - Median
- **How about with Honda Insight?**
  - Mean
  - Median
- **What can you say?**

<table>
<thead>
<tr>
<th>Two-seater cars</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>City</td>
<td>Highway</td>
</tr>
<tr>
<td>Acura NSX</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>Audi TT Roadster</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>BMW Z4 Roadster</td>
<td>20</td>
<td>28</td>
</tr>
<tr>
<td>Cadillac XLR</td>
<td>17</td>
<td>25</td>
</tr>
<tr>
<td>Chevrolet Corvette</td>
<td>18</td>
<td>25</td>
</tr>
<tr>
<td>Dodge Viper</td>
<td>12</td>
<td>20</td>
</tr>
<tr>
<td>Ferrari 360 Modena</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>Ferrari Maranello</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>Ford Thunderbird</td>
<td>17</td>
<td>23</td>
</tr>
<tr>
<td>Honda Insight</td>
<td>60</td>
<td>66</td>
</tr>
<tr>
<td>Lamborghini Gallardo</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Lamborghini Murcielago</td>
<td>9</td>
<td>13</td>
</tr>
<tr>
<td>Lotus Esprit</td>
<td>15</td>
<td>22</td>
</tr>
<tr>
<td>Maserati Spyder</td>
<td>12</td>
<td>17</td>
</tr>
<tr>
<td>Mazda Miata</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>Mercedes-Benz SL500</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>Mercedes-Benz SL600</td>
<td>13</td>
<td>19</td>
</tr>
<tr>
<td>Nissan 350Z</td>
<td>20</td>
<td>26</td>
</tr>
<tr>
<td>Porsche Boxster</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>Porsche Carrera 911</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Toyota MR2</td>
<td>26</td>
<td>32</td>
</tr>
</tbody>
</table>
Example: Salary Survey of UNC Graduates

- Survey a certain number of graduates from UNC.
- A lot of departments are surveyed.
- Question:
  - Which department produces students that earn the most on average ten years after they got their degrees?
- Answer:
  - Geography!!!!???????
  - Michael Jordan
Mean vs. Median

- **Mean:**
  - easy to calculate
  - easy to work with algebraically
  - highly affected by outliers
  - Not a resistant measure

- **Median:**
  - can be time consuming to calculate
  - more resistant to a few extreme observations (sometimes outliers)
  - robust
Mode

- The most frequent value in the data
- Important for categorical data
- Possible to have more than one mode
Mean, Median and Mode

- If the distribution is exactly symmetric and unimodal, the mean, the median and the mode are exactly the same.
- If the distribution is skewed, the three measures differ.
Which one to use?

• Different by definition
  – Mean and median are unique, and only for quantitative variables.
  – Mode is not unique.
  – Mode is defined for categorical variables also.

• The choice depends on the shape of the distribution, the type of data and the purpose of your study
  – Skewed: median
  – Categorical: mode
  – Total quantity: mean
  – …
Numerical Summary for Distributions

• Center
  – Mean
  – Median
  – Mode

• Spread
  – Quartiles, Five-number summary and Boxplot
  – Standard Deviation
Why do we need “Spread”?

• Knowing the center of a distribution alone is not a good enough description of the data.
  – Two basketball players with the same shooting percentage may be very different in terms of consistency.
  – Two companies may have the same average salary, but very different distributions.

• We need to know the spread, or the variability of the values.
A raw measure: Range

- Range = maximum - minimum
- Depends only on two values
- Tends to increase with larger samples
- Affected by outliers
  - Not robust
Percentiles

• Percentiles are derived from the ordered data values.

• The $p$th percentile is the value such that $p$ percent of the observations fall at or below it.

• The median = the 50$^{th}$ percentile.
Quartiles

• The sample quartiles are the values that divide the sorted sample into quarters, just as the median divides it into half.

• The most commonly used quantiles are
  – The median $M = 50^{th}$ percentile
  – The 1st (lower) quartile $Q_1 = 25^{th}$ percentile
  – The 3rd (upper) quartile $Q_3 = 75^{th}$ percentile
Calculations of Quartiles

THE QUARTILES $Q_1$ AND $Q_3$

To calculate the quartiles:

1. Arrange the observations in increasing order and locate the median $M$ in the ordered list of observations.

2. The first quartile $Q_1$ is the median of the observations whose position in the ordered list is to the left of the location of the overall median.

3. The third quartile $Q_3$ is the median of the observations whose position in the ordered list is to the right of the location of the overall median.
Examples: 2004 Gasoline-powered Two-seater Cars

• The highway mileages of the 20 gasoline-powered two-seater cars:
  13 15 16 16 17 19 20 22 23 23 | 23 24 25 25 26 28 28 28 29 32

• \( Q_1 = \text{Median of } \{13 \ 15 \ 16 \ 16 \ 17 \ 19 \ 20 \ 22 \ 23 \ 23\} \)
  = 

• \( Q_3 = \text{median of } \{23 \ 24 \ 25 \ 25 \ 26 \ 28 \ 28 \ 28 \ 29 \ 32\} \)
  = 

Interquartile Range: IQR

• IQR = Q3 – Q1
  – The range of the center half of the data
  – A resistant measure for spread

• IQR can be used to identify suspected outliers.

• Rule-of-thumb:
  – An observation is called a suspected outlier if it falls more than 1.5*IQR above Q3 or below Q1.
Examples: 2004 Gasoline-powered Two-seater Cars

• The highway mileages of the 20 gasoline-powered two-seater cars:
  13 15 16 16 17 19 20 22 23 23 | 23 24 25 25 26 28 28 28 29 32

• IQR = Q3 – Q1=
• 1.5*IQR=
• Q3+1.5*IQR=
• Q1-1.5*IQR=
• Any suspected outliers?
Examples: 2004 Two-Seater Cars

- The highway mileages of the 21 two-seater cars:
  13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32 66

- Q1 =
- Q3 =
- IQR = Q3 – Q1 =
- 1.5*IQR =
- Q3+1.5*IQR =
- Q1-1.5*IQR =
- Any suspected outliers?
The five-number summary

- To get a quick summary of both center and spread, use the following five-number summary:

  Minimum Q1 $M$ Q3 Maximum
Example: HWY Gas Mileage of 2004 Two-seater/Mini Cars

• Two-seater
  – Five-number summary:
    • 13, 18, 23, 27, 32

• Mini-compact
  – Five-number summary:
    • 19, 23, 26, 29, 32
Boxplots

- a visual representation of the five-number summary.
- A boxplot consists of
  - A central box spans the quartiles Q1 and Q3.
  - A line inside the box marks the median M.
  - Lines extend from the box out to the smallest and largest observations.
Boxplots of highway/city gas mileages (Two-seaters/minicompacts)
Pros and cons of Boxplots

- Location of the median line in the box indicates symmetry/asymmetry.
- Best used for side-by-side comparison of more than one distribution at a glance.

- Less detailed than histograms or stem plots.
- The box focuses attention on the central half of the data.
Income for different Education Level
Modified Boxplot

• The current boxplot can not reveal those possible outliers.

• To modify it,
  – the two lines extend out from the central box only to the smallest and largest observations that are not suspected outliers.
  – Observations more than 1.5*IQR outside the box are plotted as individual points.
### Call length (seconds)

#### Table 1.1
Service times (seconds) for calls to a customer service center

<p>| | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>77</td>
<td>289</td>
<td>128</td>
<td>59</td>
<td>19</td>
<td>148</td>
<td>157</td>
<td>203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td>118</td>
<td>104</td>
<td>141</td>
<td>290</td>
<td>48</td>
<td>3</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>372</td>
<td>140</td>
<td>438</td>
<td>56</td>
<td>44</td>
<td>274</td>
<td>479</td>
<td>211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>179</td>
<td>1</td>
<td>68</td>
<td>386</td>
<td>2631</td>
<td>90</td>
<td>30</td>
<td>57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>116</td>
<td>225</td>
<td>700</td>
<td>40</td>
<td>73</td>
<td>75</td>
<td>51</td>
<td></td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>9</td>
<td>115</td>
<td>19</td>
<td>76</td>
<td>138</td>
<td>178</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>67</td>
<td>102</td>
<td>35</td>
<td>80</td>
<td>143</td>
<td>951</td>
<td>106</td>
<td>55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>137</td>
<td>367</td>
<td>277</td>
<td>201</td>
<td>52</td>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>182</td>
<td>73</td>
<td>199</td>
<td>325</td>
<td>75</td>
<td>103</td>
<td>64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>11</td>
<td>9</td>
<td>88</td>
<td>1148</td>
<td>2</td>
<td>465</td>
<td>25</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1-1  
*Introduction to the Practice of Statistics, Fifth Edition*  
© 2005 W.H. Freeman and Company
Call length (seconds)

7.6% of all calls are \( \leq 10 \) seconds long

Figure 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2003 W.H. Freeman and Company
Take Home Message

• **Center**
  – Mean
  – Median
  – Mode

• **Spread**
  – Quartiles:
    • Q1, Q3 and the median
    • IQR
      – Identify possible outliers
  – Five-number summary
  – Boxplots and modified boxplots
    • Pros and cons