STAT 145 (Notes)

Al Nosedal
anosedal@unm.edu
Department of Mathematics and Statistics
University of New Mexico

Fall 2013
CHAPTER 1

PICTURING DISTRIBUTIONS WITH GRAPHS.
Definitions

- Statistics is the science of data.
Definitions

- Statistics is the science of data.
- Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
Definitions

- Statistics is the science of data.
- Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things.
- A variable is any characteristic of an individual. A variable can take different values for different individuals.
Descriptive Statistics

Most of the statistical information in newspapers, magazines, company reports, and other publications consists of data that are summarized and presented in a form that is easy for the reader to understand. Such summaries of data, which may be tabular, graphical, or numerical, are referred to as descriptive statistics.
Many situations require information about a large group of elements. But, because of time, cost, and other considerations, data can be collected from only a small portion of the group. The larger group of elements in a particular study is called the population, and the smaller group is called the sample. As one of its major contributions, Statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population through a process referred to as statistical inference.
Categorical and Quantitative Variables

- A categorical variable places an individual into one of several groups or categories.
Categorical and Quantitative Variables

- A categorical variable places an individual into one of several groups or categories.
- A quantitative variable takes numerical values for which arithmetic operations such as adding and averaging makes sense. The values of a quantitative variable are usually recorded in a unit of measurement such as seconds or kilograms.
Example. Fuel economy

Here is a small part of a data set that describes the fuel economy (in miles per gallon) of model year 2010 motor vehicles:

<table>
<thead>
<tr>
<th>Make and Model</th>
<th>Type</th>
<th>Transmission</th>
<th>Cylinders</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aston Martin Vantage</td>
<td>Two-seater</td>
<td>Manual</td>
<td>8</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>Subcompact</td>
<td>Automatic</td>
<td>4</td>
</tr>
<tr>
<td>Toyota Prius</td>
<td>Midsize</td>
<td>Automatic</td>
<td>4</td>
</tr>
<tr>
<td>Chevrolet Impala</td>
<td>Large</td>
<td>Automatic</td>
<td>6</td>
</tr>
</tbody>
</table>
Example. Fuel economy (cont.)

Here is a small part of a data set that describes the fuel economy (in miles per gallon) of model year 2010 motor vehicles:

<table>
<thead>
<tr>
<th>Make and Model</th>
<th>City mpg</th>
<th>Highway mpg</th>
<th>Carbon footprint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aston Martin Vantage</td>
<td>12</td>
<td>19</td>
<td>13.1</td>
</tr>
<tr>
<td>Honda Civic</td>
<td>25</td>
<td>36</td>
<td>6.3</td>
</tr>
<tr>
<td>Toyota Prius</td>
<td>51</td>
<td>48</td>
<td>3.7</td>
</tr>
<tr>
<td>Chevrolet Impala</td>
<td>18</td>
<td>29</td>
<td>8.3</td>
</tr>
</tbody>
</table>

The carbon footprint measures a vehicle’s impact on climate change in tons of carbon dioxide emitted annually.

a) What are the individuals in this data set?
b) For each individual, what variables are given? Which of these variables are categorical and which are quantitative?
Fuel economy (solution)

a) The individuals are the car makes and models.
b) For each individual, the variables recorded are Vehicle Type (categorical), Transmission Type (categorical), Number of cylinders (quantitative), City mpg (quantitative), Highway mpg (quantitative), and Carbon footprint (tons, quantitative).
The distribution of a variable tells us what values it takes and how often it takes these values. The values of a categorical variable are labels for the categories. The distribution of a categorical variable lists the categories and gives either the count or the percent of individuals that fall in each category.
Example. Never on Sunday?

Births are not, as you might think, evenly distributed across the days of the week. Here are the average numbers of babies born on each day of the week in 2008:

<table>
<thead>
<tr>
<th>Day</th>
<th>Births</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>7,534</td>
</tr>
<tr>
<td>Monday</td>
<td>12,371</td>
</tr>
<tr>
<td>Tuesday</td>
<td>13,415</td>
</tr>
<tr>
<td>Wednesday</td>
<td>13,171</td>
</tr>
<tr>
<td>Thursday</td>
<td>13,147</td>
</tr>
<tr>
<td>Friday</td>
<td>12,919</td>
</tr>
<tr>
<td>Saturday</td>
<td>8,617</td>
</tr>
</tbody>
</table>
Example. Never on Sunday? (cont.)

Present these data in a well-labeled bar graph. Would it also be correct to make a pie chart? Suggest some possible reasons why there are fewer births on weekends.

Solution.
It would be correct to make a pie chart but a pie chart would make it more difficult to distinguish between the weekend days and the weekdays. Some births are scheduled (e.g., induced labor), and probably most are scheduled for weekdays.
Example. Never on Sunday? Bar chart.
Example. What color is your car?

The most popular colors for cars and light trucks vary by region and over time. In North America white remains the top color choice, with black the top choice in Europe and silver the top choice in South America. Here is the distribution of the top colors for vehicles sold globally in 2010.

<table>
<thead>
<tr>
<th>Color</th>
<th>Popularity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silver</td>
<td>26</td>
</tr>
<tr>
<td>Black</td>
<td>24</td>
</tr>
<tr>
<td>White</td>
<td>16</td>
</tr>
<tr>
<td>Gray</td>
<td>16</td>
</tr>
<tr>
<td>Red</td>
<td>6</td>
</tr>
<tr>
<td>Blue</td>
<td>5</td>
</tr>
<tr>
<td>Beige, brown</td>
<td>3</td>
</tr>
<tr>
<td>Other colors</td>
<td></td>
</tr>
</tbody>
</table>
What color is your car? (cont.)

a) Fill in the percent of vehicles that are in other colors.
b) Make a graph to display the distribution of color popularity.
Solution

a) Other = 100 − (26 + 24 + 16 + 16 + 6 + 5 + 3) = 4.
Graph
A common graphical representation of quantitative data is a histogram. This graphical summary can be prepared for data previously summarized in either a frequency, relative frequency, or percent frequency distribution. A histogram is constructed by placing the variables of interest on the horizontal axis and the frequency, relative frequency, or percent frequency on the vertical axis.
Example

Consider the following data
14 21 23 21 16 19 22 25 16 16
24 24 25 19 16 19 18 19 21 12
16 17 18 23 25 20 23 16 20 19

a. Develop a frequency distribution using classes of 12-14, 15-17, 18-20, 21-23, and 24-26.
b. Develop a relative frequency distribution and a percent frequency distribution using the classes in part (a).
c. Make a histogram.
## Example (solution)

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Freq.</th>
<th>Percent Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 -14</td>
<td>2</td>
<td>2/40</td>
<td>0.05</td>
</tr>
<tr>
<td>15 - 17</td>
<td>8</td>
<td>8/40</td>
<td>0.20</td>
</tr>
<tr>
<td>18 - 20</td>
<td>11</td>
<td>11/40</td>
<td>0.275</td>
</tr>
<tr>
<td>21 - 23</td>
<td>10</td>
<td>10/40</td>
<td>0.25</td>
</tr>
<tr>
<td>24 - 26</td>
<td>9</td>
<td>9/40</td>
<td>0.225</td>
</tr>
</tbody>
</table>
### Modified classes (solution)

<table>
<thead>
<tr>
<th>Class</th>
<th>Frequency</th>
<th>Relative Freq.</th>
<th>Percent Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$12 \leq x &lt; 15$</td>
<td>2</td>
<td>$2/40$</td>
<td>0.05</td>
</tr>
<tr>
<td>$15 \leq x &lt; 18$</td>
<td>8</td>
<td>$8/40$</td>
<td>0.20</td>
</tr>
<tr>
<td>$18 \leq x &lt; 21$</td>
<td>11</td>
<td>$11/40$</td>
<td>0.275</td>
</tr>
<tr>
<td>$21 \leq x &lt; 24$</td>
<td>10</td>
<td>$10/40$</td>
<td>0.25</td>
</tr>
<tr>
<td>$24 \leq x &lt; 27$</td>
<td>9</td>
<td>$9/40$</td>
<td>0.225</td>
</tr>
</tbody>
</table>
Histogram of data

Frequency

Data

15 16 17 18 19 20 21 22 23 24 25

2 8 11 10 9
Symmetric and Skewed Distributions

A distribution is symmetric if the right and left sides of the histogram are approximately mirror images of each other. A distribution is skewed to the right if the right side of the histogram (containing the half of the observations with larger values) extends much farther out than the left side. It is skewed to the left if the left side of the histogram extends much farther out than the right side.
Symmetric Distribution
Distribution Skewed to the Right

Skewed to the right

0 2 4 6 8 10 12 14
0 100 200 300
Distribution Skewed to the Left
Examining a histogram

In any graph of data, look for the overall pattern and for striking deviations from that pattern. You can describe the overall pattern of a histogram by its shape, center, and spread. An important kind of deviation is an outlier, and individual value that falls outside the overall pattern.
Quantitative Variables: Stemplots

To make a stemplot:
1. Separate each observation into a stem, consisting of all but the final (rightmost) digit, and a leaf, the final digit. Stems may have as many digits as needed, but each leaf contains only a single digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.
Example: Making a stemplot

Construct stem-and-leaf display (stemplot) for the following data: 70 72 75 64 58 83 80 82 76 75 68 65 57 78 85 72.
<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>4</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Health care spending.

The table below shows the 2009 health care expenditure per capita in 35 countries with the highest gross domestic product in 2009. Health expenditure per capita is the sum of public and private health expenditure (in international dollars, based on purchasing-power parity, or PPP) divided by population. Health expenditures include the provision of health services, for health but exclude the provision of water and sanitation. Make a stemplot of the data after rounding to the nearest $100 (so that stems are thousands of dollars, and leaves are hundreds of dollars). Split the stems, placing leaves 0 to 4 on the first stem and leaves 5 to 9 on the second stem of the same value. Describe the shape, center, and spread of the distribution. Which country is the high outlier?
<table>
<thead>
<tr>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>1387</td>
<td>India</td>
<td>132</td>
<td>Saudi Arabia</td>
<td>1150</td>
</tr>
<tr>
<td>Australia</td>
<td>3382</td>
<td>Indonesia</td>
<td>99</td>
<td>South Africa</td>
<td>862</td>
</tr>
<tr>
<td>Austria</td>
<td>4243</td>
<td>Iran</td>
<td>685</td>
<td>Spain</td>
<td>3152</td>
</tr>
<tr>
<td>Belgium</td>
<td>4237</td>
<td>Italy</td>
<td>3027</td>
<td>Sweden</td>
<td>3690</td>
</tr>
<tr>
<td>Brazil</td>
<td>943</td>
<td>Japan</td>
<td>2713</td>
<td>Switzerland</td>
<td>5072</td>
</tr>
<tr>
<td>Canada</td>
<td>4196</td>
<td>Korea, South</td>
<td>1829</td>
<td>Thailand</td>
<td>345</td>
</tr>
<tr>
<td>China</td>
<td>308</td>
<td>Mexico</td>
<td>862</td>
<td>Turkey</td>
<td>965</td>
</tr>
<tr>
<td>Denmark</td>
<td>4118</td>
<td>Netherlands</td>
<td>4389</td>
<td>U. A. E.</td>
<td>1756</td>
</tr>
<tr>
<td>Finland</td>
<td>3357</td>
<td>Norway</td>
<td>5395</td>
<td>U. K.</td>
<td>3399</td>
</tr>
<tr>
<td>France</td>
<td>3934</td>
<td>Poland</td>
<td>1359</td>
<td>U. S. A.</td>
<td>7410</td>
</tr>
<tr>
<td>Germany</td>
<td>4129</td>
<td>Portugal</td>
<td>2703</td>
<td>Venezuela</td>
<td>737</td>
</tr>
<tr>
<td>Greece</td>
<td>3085</td>
<td>Russia</td>
<td>1038</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table, after rounding to the nearest $100

<table>
<thead>
<tr>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>1400</td>
<td>India</td>
<td>100</td>
<td>Saudi Arabia</td>
<td>1200</td>
</tr>
<tr>
<td>Australia</td>
<td>3400</td>
<td>Indonesia</td>
<td>100</td>
<td>South Africa</td>
<td>900</td>
</tr>
<tr>
<td>Austria</td>
<td>4200</td>
<td>Iran</td>
<td>700</td>
<td>Spain</td>
<td>3200</td>
</tr>
<tr>
<td>Belgium</td>
<td>4200</td>
<td>Italy</td>
<td>3000</td>
<td>Sweden</td>
<td>3700</td>
</tr>
<tr>
<td>Brazil</td>
<td>900</td>
<td>Japan</td>
<td>2700</td>
<td>Switzerland</td>
<td>5100</td>
</tr>
<tr>
<td>Canada</td>
<td>4200</td>
<td>Korea, South</td>
<td>1800</td>
<td>Thailand</td>
<td>300</td>
</tr>
<tr>
<td>China</td>
<td>300</td>
<td>Mexico</td>
<td>900</td>
<td>Turkey</td>
<td>1000</td>
</tr>
<tr>
<td>Denmark</td>
<td>4100</td>
<td>Netherlands</td>
<td>4400</td>
<td>U. A. E.</td>
<td>1800</td>
</tr>
<tr>
<td>Finland</td>
<td>3400</td>
<td>Norway</td>
<td>5400</td>
<td>U. K.</td>
<td>3400</td>
</tr>
<tr>
<td>France</td>
<td>3900</td>
<td>Poland</td>
<td>1400</td>
<td>U. S. A.</td>
<td>7400</td>
</tr>
<tr>
<td>Germany</td>
<td>4100</td>
<td>Portugal</td>
<td>2700</td>
<td>Venezuela</td>
<td>700</td>
</tr>
<tr>
<td>Greece</td>
<td>3100</td>
<td>Russia</td>
<td>1000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table, rounded to units of hundreds

<table>
<thead>
<tr>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
<th>Country</th>
<th>Dollars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>14</td>
<td>India</td>
<td>1</td>
<td>Saudi Arabia</td>
<td>12</td>
</tr>
<tr>
<td>Australia</td>
<td>34</td>
<td>Indonesia</td>
<td>1</td>
<td>South Africa</td>
<td>9</td>
</tr>
<tr>
<td>Austria</td>
<td>42</td>
<td>Iran</td>
<td>7</td>
<td>Spain</td>
<td>32</td>
</tr>
<tr>
<td>Belgium</td>
<td>42</td>
<td>Italy</td>
<td>30</td>
<td>Sweden</td>
<td>37</td>
</tr>
<tr>
<td>Brazil</td>
<td>9</td>
<td>Japan</td>
<td>27</td>
<td>Switzerland</td>
<td>51</td>
</tr>
<tr>
<td>Canada</td>
<td>42</td>
<td>Korea, South</td>
<td>18</td>
<td>Thailand</td>
<td>3</td>
</tr>
<tr>
<td>China</td>
<td>3</td>
<td>Mexico</td>
<td>9</td>
<td>Turkey</td>
<td>10</td>
</tr>
<tr>
<td>Denmark</td>
<td>41</td>
<td>Netherlands</td>
<td>44</td>
<td>U. A. E.</td>
<td>18</td>
</tr>
<tr>
<td>Finland</td>
<td>34</td>
<td>Norway</td>
<td>54</td>
<td>U. K.</td>
<td>34</td>
</tr>
<tr>
<td>France</td>
<td>39</td>
<td>Poland</td>
<td>14</td>
<td>U. S. A.</td>
<td>74</td>
</tr>
<tr>
<td>Germany</td>
<td>41</td>
<td>Portugal</td>
<td>27</td>
<td>Venezuela</td>
<td>7</td>
</tr>
<tr>
<td>Greece</td>
<td>31</td>
<td>Russia</td>
<td>10</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Stemplot

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 1 3 3</td>
</tr>
<tr>
<td>0</td>
<td>7 7 9 9 9</td>
</tr>
<tr>
<td>1</td>
<td>0 0 2 4 4</td>
</tr>
<tr>
<td>1</td>
<td>8 8</td>
</tr>
<tr>
<td>2</td>
<td>7 7</td>
</tr>
<tr>
<td>2</td>
<td>7 7</td>
</tr>
<tr>
<td>3</td>
<td>0 1 2 4 4 4 4</td>
</tr>
<tr>
<td>3</td>
<td>7 9</td>
</tr>
<tr>
<td>4</td>
<td>1 1 2 2 2 4</td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1 4</td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Shape, Center and Spread

This distribution is somewhat right-skewed, with a single high outlier (U.S.A.). There are two clusters of countries. The center of this distribution is around 27 ($2700 spent per capita), ignoring the outlier. The distribution’s spread is from 1 ($100 spent per capita) to 74 ($7400 spent per capita).
Time Plots

A time plot of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

When you examine a time plot, look once again for an overall pattern and for strong deviations from the pattern. A common overall pattern in a time plot is a trend, a long-term upward or downward movement over time. Some time plots show cycles, regular up-and-down movements over time.
Example. The cost of college

Below you will find data on the average tuition and fees charged to in-state students by public four-year colleges and universities for the 1980 to 2010 academic years. Because almost any variable measured in dollars increases over time due to inflation (the falling buying power of a dollar), the values are given in "constant dollars" adjusted to have the same buying power that a dollar had in 2010.

a) Make a time plot of average tuition and fees.
b) What overall pattern does your plot show?
c) Some possible deviations from the overall pattern are outliers, periods when changes went down (in 2010 dollars), and periods of particularly rapid increase. Which are present in your plot, and during which years?
<table>
<thead>
<tr>
<th>Year</th>
<th>Tuition ($)</th>
<th>Year</th>
<th>Tuition ($)</th>
<th>Year</th>
<th>Tuition ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>2119</td>
<td>1991</td>
<td>3373</td>
<td>2002</td>
<td>4961</td>
</tr>
<tr>
<td>1982</td>
<td>2305</td>
<td>1993</td>
<td>3827</td>
<td>2004</td>
<td>5900</td>
</tr>
<tr>
<td>1983</td>
<td>2505</td>
<td>1994</td>
<td>3974</td>
<td>2005</td>
<td>6128</td>
</tr>
<tr>
<td>1984</td>
<td>2572</td>
<td>1995</td>
<td>4019</td>
<td>2006</td>
<td>6218</td>
</tr>
<tr>
<td>1985</td>
<td>2665</td>
<td>1996</td>
<td>4131</td>
<td>2007</td>
<td>6480</td>
</tr>
<tr>
<td>1986</td>
<td>2815</td>
<td>1997</td>
<td>4226</td>
<td>2008</td>
<td>6532</td>
</tr>
<tr>
<td>1987</td>
<td>2845</td>
<td>1998</td>
<td>4338</td>
<td>2009</td>
<td>7137</td>
</tr>
<tr>
<td>1988</td>
<td>2903</td>
<td>1999</td>
<td>4397</td>
<td>2010</td>
<td>7605</td>
</tr>
<tr>
<td>1989</td>
<td>2972</td>
<td>2000</td>
<td>4426</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1990</td>
<td>3190</td>
<td>2001</td>
<td>4626</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Time Plot of Average Tuition and Fees (1980-2010)
Answers to b) and c)

b) Tuition has steadily climbed during the 30-year period, with sharpest absolute increases in the last 10 years.

c) There is a sharp increase from 2000 to 2010.
CHAPTER 2

DESCRIBING DISTRIBUTIONS WITH NUMBERS.
Problem

How much do people with a bachelors degree (but no higher degree) earn? Here are the incomes of 15 such people, chosen at random by the Census Bureau in March 2002 and asked how much they earned in 2001. Most people reported their incomes to the nearest thousand dollars, so we have rounded their responses to thousands of dollars. 110 25 50 50 55 30 35 30 4 32 50 30 32 74 60.

How could we find the "typical" income for people with a bachelors degree (but no higher degree)?
Measuring center: the mean

The most common measure of center is the ordinary arithmetic average, or mean. To find the mean of a set of observations, add their values and divide by the number of observations. If the \( n \) observations are \( x_1, x_2, \ldots, x_n \), their mean is

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}
\]

or in more compact notation,

\[
\bar{x} = \sum_{i=1}^{n} x_i
\]
\[ \bar{x} = \frac{110 + 25 + 50 + 50 + 55 + 30 + \ldots + 32 + 74 + 60}{15} = 44.466 \]

Do you think that this number represents the "typical" income for people with a bachelors degree (but no higher degree)?
Measuring center: the median

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 the distribution:

Arrange all observations in order of size, from smallest to largest. 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 $\frac{n+1}{2}$ observations up from the bottom of the list. If the number of observations $n$ is even, the median $M$ is the mean of the two center observations in the ordered list. Find the location of the median by counting $\frac{n+1}{2}$ observations up from the bottom of the list.
Income Problem (Median)

We know that if we want to find the median, $M$, we have to order our observations from smallest to largest: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. Let's find the location of $M$

location of $M = \frac{n+1}{2} = \frac{15+1}{2} = 8$

Therefore, $M = x_8 = 35$ ($x_8$ = 8th observation on our ordered list).
Another measure of location is the mode. The mode is defined as follows. The mode is the value that occurs with greatest frequency. Note: situations can arise for which the greatest frequency occurs at two or more different values. In these instances more than one mode exists.
Using the definition of mode, we have that:
\[mode_1 = 30\]
and
\[mode_2 = 50\]
Note that both of them have the greatest frequency, 3.
Example: New York travel times.

Here are the travel times in minutes of 20 randomly chosen New York workers:
10 30 5 25 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45.
Compare the mean and median for these data. What general fact does your comparison illustrate?
Solution

Mean:
\[
\bar{x} = \frac{10 + 30 + 5 + \ldots + 60 + 60 + 40 + 45}{20} = 31.25
\]

Median:
First, we order our data from smallest to largest
5 10 10 15 15 15 15 20 20 20 25 30 30 40 40 45 60 60 65 85.
location of \( M \) = \( \frac{n+1}{2} = \frac{20+1}{2} = 10.5 \)
Which means that we have to find the mean of \( x_{10} \) and \( x_{11} \).
\[
M = \frac{x_{10} + x_{11}}{2} = \frac{20 + 25}{2} = 22.5
\]
Comparing the mean and the median

The mean and median of a symmetric distribution are close together. In a skewed distribution, the mean is farther out in the long tail than is the median. Because the mean cannot resist the influence of extreme observations, we say that it is not a resistant measure of center.
The quartiles $Q_1$ and $Q_3$

To calculate the quartiles:
Arrange the observations in increasing order and locate the median $M$ in the ordered list of observations.
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.
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.
Income Problem ($Q_1$)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_1 = 7$ observations.
Let us find the location of $Q_1$:
location of $Q_1 = \frac{n_1+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_1 = x_4 = 30$. 
Income Problem ($Q_3$)

Data:
4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

From previous work, we know that $M = x_8 = 35$.
This implies that the first half of our data has $n_2 = 7$ observations.
Let us find the location of $Q_3$:
location of $Q_3 = \frac{n_2+1}{2} = \frac{7+1}{2} = 4$.
This means that $Q_3 = 55$. 

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. In symbols, the five-number summary is \( \text{min} \ Q_1 \ M \ Q_3 \ \text{MAX} \).
Income Problem (five-number summary)

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110. The five-number summary for our income problem is given by:
4 30 35 55 110
Boxplot

A boxplot is a graph of the five-number summary. A central box spans the quartiles $Q_1$ and $Q_3$. A line in the box marks the median $M$. Lines extended from the box out to the smallest and largest observations.
Boxplot for income data
The 2010 roster of the Pittsburgh Steelers professional football team included 7 defensive linemen and 9 offensive linemen. The weights in pounds of the defensive linemen were 305 325 305 300 285 280 298 and the weights of the offensive linemen were 338 324 325 304 344 315 304 319 318

a) Make a stemplot of the weights of the defensive linemen and find the five-number summary.
b) Make a stemplot of the weights of the offensive linemen and find the five-number summary.
c) Does either group contain one or more clear outliers? Which group of players tends to be heavier?
Solution

a) Defensive linemen.

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>29</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
b) Offensive linemen.

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>32</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>33</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>
Five-number summary

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>$Q_1$</th>
<th>Median</th>
<th>$Q_3$</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offensive line (lbs)</td>
<td>304</td>
<td>309.5</td>
<td>319</td>
<td>331.5</td>
<td>344</td>
</tr>
<tr>
<td>Defensive line (lbs)</td>
<td>280</td>
<td>285</td>
<td>300</td>
<td>305</td>
<td>325</td>
</tr>
</tbody>
</table>

c) Apparently, neither of these two groups contain outliers. It seems that the offensive line players are heavier.
Measures of Variability: Range

The simplest measure of variability is the range.  
Range = Largest value - smallest value  
Range = MAX - min
Measures of Variability: IQR

A measure of variability that overcomes the dependency on extreme values is the interquartile range (IQR).

\[ IQR = Q_3 - Q_1 \]
1.5 IQR Rule

Identifying suspected outliers. Whether an observation is an outlier is a matter of judgement: does it appear to clearly stand apart from the rest of the distribution? When large volumes of data are scanned automatically, however, we need a rule to pick out suspected outliers. The most common rule is the 1.5 IQR rule. A point is a suspected outlier if it lies more than 1.5 IQR below the first quartile $Q_1$ or above the third quartile $Q_3$. 
A high income.

In our income problem, we noted the influence of one high income of $110,000 among the incomes of a sample of 15 college graduates. Does the 1.5 IQR rule identify this income as a suspected outlier?
Solution

Data: 4 25 30 30 30 32 32 35 50 50 50 55 60 74 110.

$Q_1$ and $Q_3$ are given by:

$Q_1 = 30$ and $Q_3 = 55$

$Q_3 + 1.5 \text{ IQR} = 55 + 1.5(25) = 92.5$

Since $110 > 92.5$ we conclude that 110 is an outlier.
Problem

Ebby Halliday Realtors provide advertisements for distinctive properties and estates located throughout the United States. The prices listed for 22 distinctive properties and estates are shown here. Prices are in thousands.

1500 895 719 619 625 4450 2200
1280 700 619 725 739 799 2495
1395 2995 880 3100 1699 1120 1250 912.

a) Provide a five-number summary.

b) The highest priced property, $4,450,000, is listed as an estate overlooking White Rock Lake in Dallas, Texas. Should this property be considered an outlier?
Solution

a) $\text{min} = 619$
$Q_1 = 725$
$M = 1016$
$Q_3 = 1699$
$\text{MAX} = 4450$

b) $\text{IQR} = 1699 - 725 = 974$
$Q_3 + 1.5 \times \text{IQR} = 1699 + 1.5 \times (974) = 1699 + 1461 = 3160.$
Since $4450 > 3160$, we conclude that $4450$ is an outlier.
The variance $s^2$ of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of $n$ observations $x_1, x_2, ..., x_n$ is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$

or, more compactly,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$
Measures of Variability: Standard Deviation

The standard deviation $s$ is the square root of the variance $s^2$:

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

Consider a sample with data values of 10, 20, 12, 17, and 16. Compute the variance and standard deviation.
First, we have to calculate the mean, $\bar{x}$:

$$\bar{x} = \frac{10 + 20 + 12 + 17 + 16}{5} = 15.$$ 

Now, let's find the variance $s^2$:

$$s^2 = \frac{(10-15)^2 + (20-15)^2 + (12-15)^2 + (17-15)^2 + (16-15)^2}{5-1}.$$

$$s^2 = \frac{64}{4} = 16.$$ 

Finally, let's find the standard deviation $s$:

$$s = \sqrt{16} = 4.$$
Radon is a naturally occurring gas and is the second leading cause of lung cancer in the United States. It comes from the natural breakdown of uranium in the soil and enters buildings through cracks and other holes in the foundations. Found throughout the United States, levels vary considerably from state to state. There are several methods to reduce the levels of radon in your home, and the Environmental Protection Agency recommends using one of these if the measured level in your home is above 4 picocuries per liter. Four readings from Franklin County, Ohio, where the county average is 9.32 picocuries per liter, were 5.2, 13.8, 8.6 and 16.8.

a) Find the mean step-by-step.

b) Find the standard deviation step-by-step.

c) Now enter the data into your calculator and use the mean and standard deviation buttons to obtain \( \bar{x} \) and \( s \). Do the results agree with your hand calculations?
Solution

First, we have to calculate the mean, $\bar{x}$:

$$\bar{x} = \frac{5.2+13.8+8.6+16.8}{4} = 11.1.$$  

Now, let’s find the variance $s^2$:

$$s^2 = \frac{(5.2-11.1)^2+(13.8-11.1)^2+(8.6-11.1)^2+(16.8-11.1)^2}{4-1}.$$  

$$s^2 = \frac{80.84}{3} = 26.9466$$

Finally, let’s find the standard deviation $s$:

$$s = \sqrt{26.9466} = 5.1910.$$
Choosing a summary

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use \( \bar{x} \) and \( s \) only for reasonably symmetric distributions that are free of outliers.
CHAPTER 3

THE NORMAL DISTRIBUTIONS.
Random Experiment: Rolling a fair die 300 times.

<table>
<thead>
<tr>
<th>Class</th>
<th>Expected Frequency</th>
<th>Expected Relative Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 \leq x &lt; 2$</td>
<td>50</td>
<td>1/6</td>
</tr>
<tr>
<td>$2 \leq x &lt; 3$</td>
<td>50</td>
<td>1/6</td>
</tr>
<tr>
<td>$3 \leq x &lt; 4$</td>
<td>50</td>
<td>1/6</td>
</tr>
<tr>
<td>$4 \leq x &lt; 5$</td>
<td>50</td>
<td>1/6</td>
</tr>
<tr>
<td>$5 \leq x &lt; 6$</td>
<td>50</td>
<td>1/6</td>
</tr>
<tr>
<td>$6 \leq x &lt; 7$</td>
<td>50</td>
<td>1/6</td>
</tr>
</tbody>
</table>
Histogram of Expected Frequencies

Histogram of expected frequencies

frequency
Histogram of Expected Relative Frequencies
Density Curve

A density curve is a curve that is always on or above the horizontal axis, and has area exactly 1 underneath it. A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range. Note. No set of real data is exactly described by a density curve. The curve is an idealized description that is easy to use and accurate enough for practical use.
Accidents on a bike path

Examining the location of accidents on a level, 5-mile bike path shows that they occur uniformly along the length of the path. The figure below displays the density curve that describes the distribution of accidents.

a) Explain why this curve satisfies the two requirements for a density curve.
b) The proportion of accidents that occur in the first mile of the path is the area under the density curve between 0 miles and 1 mile. What is this area?
c) There is a stream alongside the bike path between the 0.8-mile mark and the 1.3-mile mark. What proportion of accidents happen on the bike path alongside the stream?
d) The bike path is a paved path through the woods, and there is a road at each end. What proportion of accidents happen more than 1 mile from either road?
Solution

a) It is on or above the horizontal axis everywhere, and because it forms a $\frac{1}{5} \times 5$ rectangle, the area beneath the curve is 1.
Solution b)

Density Curve

Distance along bike path (miles)

proportion = 1 x 0.20 = 0.20
Solution c)

proportion = (1.3 - 0.8) \times 0.20 = 0.10
Solution d)

Density Curve

Distance along bike path (miles)

proportion = (4-1) x 0.20 = 0.60
Normal Distributions

A Normal Distribution is described by a Normal density curve. Any particular Normal distribution is completely specified by two numbers, its mean $\mu$ and standard deviation $\sigma$. The mean of a Normal distribution is at the center of the symmetric Normal curve. The standard deviation is the distance from the center to the change-of-curvature points on either side.
Normal Distribution
mean=0 and standard deviation=1
Two Different Standard Deviations
In the Normal distribution with mean $\mu$ and standard deviation $\sigma$:
Approximately 68% of the observations fall within $\sigma$ of the mean $\mu$.
Approximately 95% of the observations fall within $2\sigma$ of $\mu$.
Approximately 99.7% of the observations fall within $3\sigma$ of $\mu$. 
Problem

The national average for the verbal portion of the College Boards Scholastic Aptitude Test (SAT) is 507. The College Board periodically rescales the test scores such that the standard deviation is approximately 100. Answer the following questions using a bell-shaped distribution and the empirical rule for the verbal test scores.

a. What percentage of students have an SAT verbal score greater than 607?
b. What percentage of students have an SAT verbal score greater than 707?
c. What percentage of students have an SAT verbal score between 407 and 507?
d. What percentage of students have an SAT verbal score between 307 and 707?
Solution a)

\[
\begin{array}{c|c|c|c|c|c}
\text{SAT score} & 200 & 400 & 600 & 800 \\
\text{Probability} & 0.000 & 0.001 & 0.002 & 0.003 & 0.004 \\
\end{array}
\]

\[
16\% \quad 68\% \quad 16\%
\]
Solution b)

The diagram shows a normal distribution with the following points:

- 2.5% at 200
- 95% at 600
- 2.5% at 800

The SAT score range is from 200 to 800.
Solution c)

![Graph showing a normal distribution with SAT scores and percentages]
Solution d)
Fruit flies

The common fruit fly Drosophila melanogaster is the most studied organism in genetic research because it is small, easy to grow, and reproduces rapidly. The length of the thorax (where the wings and legs attach) in a population of male fruit flies is approximately Normal with mean 0.800 millimeters (mm) and standard deviation 0.078 mm. Draw a Normal curve on which this mean and standard deviation are correctly located.
Fruit flies

The length of the thorax in a population of male fruit flies is approximately Normal with mean 0.800 mm and standard deviation 0.078 mm. Use the 68-95-99.7 rule to answer the following questions.

a) What range of lengths covers almost all (99.7%) of this distribution?

b) What percent of male fruit flies have a thorax length exceeding 0.878 mm?
Solution a) Between 0.566 mm and 1.034 mm
Solution b) 16% of thorax lengths exceed 0.878 mm
Monsoon rains

The summer monsoon brings 80% of India’s rainfall and is essential for the country’s agriculture. Records going back more than a century show that the amount of monsoon rainfall varies from the year according to a distribution that is approximately Normal with mean 582 mm and standard deviation 82 mm. Use the 68-95-99.7 rule to answer the following questions.

a) Between what values do the monsoon rains fall in 95% of all years?
b) How small are the monsoon rains in the dryest 2.5% of all years?
Solution

a) In 95% of all years, monsoon rain levels are between $582 - 2(82)$ and $582 + 2(82)$ i.e. 688 mm and 1016 mm.

b) The driest 2.5% of monsoon rainfalls are less than 688 mm; this is more than two standard deviations below the mean.
The standard Normal distribution is the Normal distribution $N(0,1)$ with mean 0 and standard deviation 1. If a variable $x$ has any Normal distribution $N(\mu,\sigma)$ with mean $\mu$ and standard deviation $\sigma$, then the standardized variable

$$z = \frac{x - \mu}{\sigma}$$

has the standard Normal distribution.
In 2010, when she was a high school senior, Alysha scored 670 on the Mathematics part of the SAT. The distribution of SAT Math scores in 2010 was Normal with mean 516 and standard deviation 116. John took the ACT and scored 26 on the Mathematics portion. ACT Math scores for 2010 were Normally distributed with mean 21.0 and standard deviation 5.3. Find the standardized scores for both students. Assuming that both tests measure the same kind of ability, who had the higher score?
Solution

Alysha’s standardized score is

\[ z_A = \frac{670 - 516}{116} = 1.33. \]

John’s standardized score is

\[ z_J = \frac{26 - 21}{5.3} = 0.94. \]

Alysha’s score is relatively higher than John’s.
Men’s and women’s heights

The heights of women aged 20 to 29 are approximately Normal with mean 64.3 inches and standard deviation 2.7 inches. Men the same age have mean height 69.9 inches with standard deviation 3.1 inches. What are the z-scores for a woman 6 feet tall and a man 6 feet tall? Say in simple language what information the z-scores give that the original nonstandardized heights do not.
Solution

We need to use the same scale, so recall that 6 feet = 72 inches. A woman 6 feet tall has standardized score

\[ z_W = \frac{72 - 64.3}{2.7} = 2.85 \]

(quite tall, relatively).

A man 6 feet tall has standardized score

\[ z_M = \frac{72 - 69.9}{3.1} = 0.68. \]

Hence, a woman 6 feet tall is 2.85 standard deviations taller than average for women. A man 6 feet tall is only 0.68 standard deviations above average for men.
Using the Normal table

Use table A to find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that is the answer to the question.

a) $z < -1.42$

b) $z > -1.42$

c) $z < 2.35$

d) $-1.42 < z < 2.35$
Solution a) 0.0778

\[ z^* = -1.42 \]
Solution b) 0.9222
Solution c) 0.9906

z* = 2.35
Solution d) $0.9966 - 0.0778 = 0.9128$
The summer monsoon rains in India follow approximately a Normal distribution with mean 852 mm of rainfall and standard deviation 82 mm.

a) In the drought year 1987, 697 mm of rain fell. In what percent of all years will India have 697 mm or less of monsoon rain?

b) ”Normal rainfall” means within 20% of the long-term average, or between 683 and 1022 mm. In what percent of all years is the rainfall normal?
Solution a) 

1. State the problem. Let $x$ be the monsoon rainfall in a given year. The variable $x$ has the $N(852, 82)$ distribution. We want the proportion of years with $x \leq 697$.
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn $x$ into a standard Normal $z$. Hence $x \leq 697$ corresponds to $z \leq \frac{697-852}{82} = -1.89$.
3. Use the table. From Table A, we see that the proportion of observations less than $-1.89$ is 0.0294. Thus, the answer is 2.94%.
1. State the problem. Let \( x \) be the monsoon rainfall in a given year. The variable \( x \) has the \( N(852, 82) \) distribution. We want the proportion of years with \( 683 < x < 1022 \).

2. Standardize. Subtract the mean, then divide by the standard deviation, to turn \( x \) into a standard Normal \( z \).

\[
683 < x < 1022 \text{ corresponds to } \frac{683-852}{82} < z < \frac{1022-852}{82}, \text{ or } -2.06 < z < 2.07.
\]

3. Use the table. Hence, using Table A, the area is \( 0.9808 - 0.0197 = 96.11\% \).
Almost all medical schools in the United States require students to take the Medical College Admission Test (MCAT). The exam is composed of three multiple-choice sections (Physical Sciences, Verbal Reasoning, and Biological Sciences). The score on each section is converted to a 15-point scale so that the total score has a maximum value of 45. The total scores follow a Normal distribution, and in 2010 the mean was 25.0 with a standard deviation of 6.4. There is little change in the distribution of scores from year to year.

a) What proportion of students taking the MCAT had a score over 30?

b) What proportion had scores between 20 and 25?
Solution a)

1. State the problem. Let $x$ be the MCAT score of a randomly selected student. The variable $x$ has the $N(25, 6.4)$ distribution. We want the proportion of students with $x > 30$.

2. Standardize. Subtract the mean, then divide by the standard deviation, to turn $x$ into a standard Normal $z$. Hence $x > 30$ corresponds to $z > \frac{30-25}{6.4} = 0.78$.

3. Use the table. From Table A, we see that the proportion of observations less than 0.78 is 0.7823. Hence, the answer is $1 - 0.7823 = 0.2177$, or 21.77%. 
Solution b)

1. State the problem. Let $x$ be the MCAT score of a randomly selected student. The variable $x$ has the $N(25, 6.4)$ distribution. We want the proportion of students with $20 \leq x \leq 25$.  
2. Standardize. Subtract the mean, then divide by the standard deviation, to turn $x$ into a standard Normal $z$. $20 \leq x \leq 25$ corresponds to $\frac{20-25}{6.4} \leq z \leq \frac{25-25}{6.4}$, or $-0.78 \leq z \leq 0$.  
3. Use the table. Using Table A, the area is $0.5 - 0.2177 = 0.2833$, or 28.33%. 

Using a table to find Normal proportions

Step 1. State the problem in terms of the observed variable $x$. Draw a picture that shows the proportion you want in terms of cumulative proportions.

Step 2. Standardize $x$ to restate the problem in terms of a standard Normal variable $z$.

Step 3. Use Table A and the fact that the total area under the curve is 1 to find the required area under the standard Normal curve.
Use Table A to find the value $z^*$ of a standard Normal variable that satisfies each of the following conditions. (Use the value of $z^*$ from Table A that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of $z^*$ marked on the axis.

a) The point $z^*$ with 15% of the observations falling below it.
b) The point $z^*$ with 70% of the observations falling above it.
Solution a) $z^* = -1.04$
Solution b) $z^* = -0.52$
The Medical College Admission Test

The total scores on the Medical College Admission Test (MCAT) follow a Normal distribution with mean 25.0 and standard deviation 6.4. What are the median and the first and third quartiles of the MCAT scores?
Solution: Finding the median

Because the Normal distribution is symmetric, its median and mean are the same. Hence, the median MCAT score is 25.
Solution: Finding $Q_1$

1. State the problem. We want to find the MCAT score $x$ with area 0.25 to its left under the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 6.4$.
2. Use the table. Look in the body of Table A for the entry closest to 0.25. It is 0.2514. This is the entry corresponding to $z^* = -0.67$. So $z^* = -0.67$ is the standardized value with area 0.25 to its left.
3. Unstandardize to transform the solution from the $z^*$ back to the original $x$ scale. We know that the standardized value of the unknown $x$ is $z^* = -0.67$. So $x$ itself satisfies

$$\frac{x - 25}{6.4} = -0.67$$

Solving this equation for $x$ gives

$$x = 25 + (-0.67)(6.4) = 20.71$$
Solution: Finding $Q_3$

1. State the problem. We want to find the MCAT score $x$ with area 0.75 to its left under the Normal curve with mean $\mu = 25$ and standard deviation $\sigma = 6.4$.

2. Use the table. Look in the body of Table A for the entry closest to 0.75. It is 0.7486. This is the entry corresponding to $z^* = 0.67$. So $z^* = 0.67$ is the standardized value with area 0.75 to its left.

3. Unstandardize to transform the solution from the $z^*$ back to the original $x$ scale. We know that the standardized value of the unknown $x$ is $z^* = 0.67$. So $x$ itself satisfies

$$\frac{x - 25}{6.4} = 0.67$$

Solving this equation for $x$ gives

$$x = 25 + (0.67)(6.4) = 29.29$$
Finding a value when given a proportion

1. State the problem.
2. Use the table.
3. Unstandardize to transform the solution from the $z$* back to the original $x$ scale.