Sampling and Descriptive Statistics

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

Reference:
1. W. Navidi. *Statistics for Engineering and Scientists*. Chapter 1 & Teaching Material
Sampling (1/2)

• Definition: A population is the entire collection of objects or outcomes about which information is sought
  – All NTNU students

• Definition: A sample is a subset of a population, containing the objects or outcomes that are actually observed
  – E.g., the study of the heights of NTNU students
    • Choose the 100 students from the rosters of football or basketball teams (appropriate?)
    • Choose the 100 students living a certain dorm or enrolled in the statistics course (appropriate?)
Sampling (2/2)

- Definition: A simple random sample (SRS) of size \( n \) is a sample chosen by a method in which each collection of \( n \) population items is equally likely to comprise the sample, just as in the lottery.

- Definition: A sample of convenience is a sample that is not drawn by a well-defined random method.
  - Things to consider with convenience samples:
    - Differ systematically in some way from the population
    - Only use when it is not feasible to draw a random sample
More on SRS (1/3)

• Definition: A conceptual population consists of all the values that might possibly have been observed
  – It is in contrast to “tangible (可觸之的) population”
  – E.g., a geologist weighs a rock several times on a sensitive scale. Each time, the scale gives a slightly different reading
  – Here the population is conceptual. It consists of all the readings that the scale could in principle produce
More on SRS (2/3)

• A SRS is not guaranteed to reflect the population perfectly

• SRS’s always differ in some ways from each other, occasionally a sample is substantially different from the population

• Two different samples from the same population will vary from each other as well

• This phenomenon is known as sampling variation
More on SRS (3/3)

• The items in a sample are independent if knowing the values of some of the items does not help to predict the values of the others.

• (A Rule of Thumb) Items in a simple random sample may be treated as independent in most cases encountered in practice.
  – The exception occurs when the population is finite and the sample comprises a substantial fraction (more than 5%) of the population.

• However, it is possible to make a population behave as though it were infinite large, by replacing each item after it is sampled.
  – Sampling With Replacement.
Other Sampling Methods

• **Weighting Sampling**
  – Some items are given a greater chance of being selected than others
  – E.g., a lottery in which some people have more tickets than others

• **Stratified Sampling**
  – The population is divided up into subpopulations, called strata
  – A simple random sample is drawn from each stratum
  – Supervised (?)

• **Cluster Sampling**
  – Items are drawn from the population in groups or clusters
  – E.g., the U.S. government agencies use cluster sampling to sample the U.S. population to measure sociological factors such as income and unemployment
  – Unsupervised (?)
Types of Experiments

• One-Sample Experiment
  – There is only one population of interest
  – A single sample is drawn from it

• Multi-Sample Experiment
  – There are two or more populations of interest
  – A simple is drawn from each population
  – The usual purpose of multi-sample experiments is to make comparisons among populations
Types of Data

- **Numerical or quantitative** if a numerical quantity is assigned to each item in the sample
  - Height
  - Weight
  - Age

- **Categorical or qualitative** if the sample items are placed into categories
  - Gender
  - Hair color
  - Blood type

<table>
<thead>
<tr>
<th>Specimen</th>
<th>Torque (kN · m)</th>
<th>Failure Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>165</td>
<td>Wcld</td>
</tr>
<tr>
<td>2</td>
<td>237</td>
<td>Beam</td>
</tr>
<tr>
<td>3</td>
<td>222</td>
<td>Beam</td>
</tr>
<tr>
<td>4</td>
<td>255</td>
<td>Beam</td>
</tr>
<tr>
<td>5</td>
<td>194</td>
<td>Weld</td>
</tr>
</tbody>
</table>
Summary Statistics

- The **summary statistics** are sometimes called **descriptive statistics** because they describe the data
  - Numerical Summaries
    - Sample mean, median, trimmed mean, mode
    - Sample standard deviation (variance), range
    - Percentiles, quartiles
    - Skewness, kurtosis
    - ....
  - Graphical Summaries
    - Stem and leaf plot
    - Dotplot
    - Histogram (more commonly used)
    - Boxplot (more commonly used)
    - Scatterplot
    - ....
Numerical Summaries (1/4)

- **Definition: Sample Mean**  
  (the center of the data)
  
  - Let \( X_1, \ldots, X_n \) be a sample. The sample mean is
    \[
    \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i
    \]

  - It’s customary to use a letter with a bar over it to denote a sample mean

- **Definition: Sample Variance**  
  (how spread out the data are)

  - Let \( X_1, \ldots, X_n \) be a sample. The sample variance is
    \[
    s^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2
    \]

    Which is equivalent to
    \[
    s^2 = \frac{1}{n-1} \cdot \left( \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)
    \]

    \{20, 29, 30, 31, 32\}
    \{10, 20, 30, 40, 50\}
Numerical Summaries (2/4)

• Actually, we are interested in
  – Population mean
  – Population deviation: Measuring the spread of the population
    • The variations of population items around the population mean

• Practically, because population mean is unknown, we use sample mean to replace it

• Mathematically, the deviations around the sample mean tend to be a bit smaller than the deviations around the population mean
  – So when calculating sample variance, the quantity divided by $n - 1$ rather than $n$ provides the right correction
  – To be proved later on!
Numerical Summaries (3/4)

• Definition: Sample Standard Deviation
  – Let $X_1, \ldots, X_n$ be a sample. The sample deviation is

$$
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}
$$

Which is equivalent to

$$
s = \sqrt{\frac{1}{n-1} \left( \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \right)}
$$

– The sample deviation also measures the degree of spread in a sample (having the same units as the data)
Numerical Summaries (3/4)

- If \( X_1, \ldots, X_n \) is a sample, and \( Y_i = a + bX_i \), where \( a \) and \( b \) are constants, then \( \bar{Y} = a + b\bar{X} \)

- If \( X_1, \ldots, X_n \) is a sample, and \( Y_i = a + bX_i \), where \( a \) and \( b \) are constants, then \( s_y^2 = b^2 s_x^2 \) and \( s_y = |b| s_x \)

- Definition: Outliers
  - Sometimes a sample may contain a few points that are much larger or smaller than the rest (mainly resulting from data entry errors)
  - Such points are called outliers

---

Outlier
More on Numerical Summaries (1/2)

- Definition: The median is another measure of center of a sample $X_1, \ldots, X_n$, like the mean
  - To compute the median items in the sample have to be ordered by their values
  - If $n$ is odd, the sample median is the number in position $(n + 1)/2$
  - If $n$ is even, the sample median is the average of the numbers in positions $n/2$ and $(n/2)+1$
  - The median is an important (robust) measure of center for samples containing outliers
Definition: The trimmed mean of one-dimensional data is computed by

- First, arranging the sample values in (ascending or descending) order
- Then, trimming an equal number of them from each end, say, $p\%$
- Finally, computing the sample mean of those remaining
More on “mean”

• Arithmetic mean \( \bar{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \)

• Geometric mean \( \bar{X} = \left( \prod_{i=1}^{n} X_i \right)^{\frac{1}{n}} \)

• Harmonic mean \( \bar{X} = n \cdot \left( \sum_{i=1}^{n} \frac{1}{X_i} \right)^{-1} \)

• Power mean \( \bar{X} = \left( \frac{1}{n} \cdot \sum_{i=1}^{n} X_i^m \right)^{\frac{1}{m}} \)

• Arithmetic mean ≥ Geometric mean ≥ Harmonic mean

• Weighted arithmetic mean \( \bar{X} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \)

http://en.wikipedia.org/wiki/Mean

Statistics-Berlin Chen 17
Quartiles

• Definition: the quartiles of a sample $X_1, \ldots, X_n$ divides it as nearly as possible into quarters. The sample values have to be ordered from the smallest to the largest
  - To find the first quartile, compute the value $0.25(n+1)$
  - The second quartile found by computing the value $0.5(n+1)$
  - The third quartile found by computing the value $0.75(n+1)$

• Example 1.14: Find the first and third quartiles of the data in Example 1.12

<table>
<thead>
<tr>
<th>30</th>
<th>75</th>
<th>79</th>
<th>80</th>
<th>80</th>
<th>105</th>
<th>126</th>
<th>138</th>
<th>149</th>
<th>179</th>
<th>179</th>
<th>191</th>
</tr>
</thead>
<tbody>
<tr>
<td>223</td>
<td>232</td>
<td>232</td>
<td>236</td>
<td>240</td>
<td>242</td>
<td>254</td>
<td>247</td>
<td>254</td>
<td>274</td>
<td>384</td>
<td>470</td>
</tr>
</tbody>
</table>

• $n=24$
• To find the first quartile, compute $(n+1)25=6.25$
  $(105+126)/2=115.5$
• To find the third quartile, compute $(n+1)75=18.75$
  $(242+245)/2=243.5$
Percentiles

• Definition: The \( p \text{th percentile} \) of a sample \( X_1, \ldots, X_n \), for a number between 0 and 100, divide the sample so that as nearly as possible \( p\% \) of the sample values are less than the \( p \text{th percentile} \). To find:
  – Order the sample values from smallest to largest
  – Then compute the quantity \( (p/100)(n+1) \), where \( n \) is the sample size
  – If this quantity is an integer, the sample value in this position is the \( p \text{th percentile} \). Otherwise, average the two sample values on either side

• Note, the \textbf{first quartile} is the 25th percentile, the \textbf{median} is the 50th percentile, and the \textbf{third quartile} is the 75th percentile
Mode and Range

• Mode
  – The sample mode is the most frequently occurring values in a sample
  – Multiple modes: several values occur with equal counts

• Range
  – The difference between the largest and smallest values in a sample
  – A measure of spread that depends only on the two extreme values
Numerical Summaries for Categorical Data

• For categorical data, each sample item is assigned a category rather than a numerical value

• Two Numerical Summaries for Categorical Data
  – Definition: (Relative) Frequencies
    • The frequency of a given category is simply the number of sample items falling in that category
  – Definition: Sample Proportions (also called relative frequency)
    • The sample proportion is the frequency divided by the sample size
Sample Statistics and Population Parameters (1/2)

• A **numerical summary** of a sample is called a **statistic**
• A **numerical summary** of a population is called a **parameter**
  – If a population is finite, the methods used for calculating the numerical summaries of a sample can be applied for calculating the numerical summaries of the population (each value (or outcome) occurs with probability? See Chapter 2)
  – Exceptions are the variance and standard deviation (?)
    \[ \text{Normal} \ : \ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
• However, sample statistics are often used to estimate parameters (to be taken as estimators)
  – In practice, the entire population is never observed, so the population parameters cannot be calculated directly
Sample Statistics and Population Parameters (2/2)

• A Schematic Depiction

Population  \[\xrightarrow{\text{Inference}}\]  Sample

Parameters  \[\xrightarrow{\text{Statistics}}\]  Parameters
Graphical Summaries

• Recall that the mean, median and standard deviation, etc., are *numerical summaries* of a sample of a population.

• On the other hand, the *graphical summaries* are used as will to help visualize a list of numbers (or the sample items). Methods to be discussed include:
  - Stem and leaf plot
  - Dotplot
  - Histogram (more commonly used)
  - Boxplot (more commonly used)
  - Scatterplot
Stem-and-leaf Plot (1/3)

• A simple way to summarize a data set
• Each item in the sample is divided into two parts
  – *stem*, consisting of the leftmost one or two digits
  – *leaf*, consisting of the next significant digit
• The stem-and-leaf plot is a compact way to represent the data
  – It also gives us some indication of the shape of our data
Stem-and-leaf Plot (2/3)

• Example: Duration of dormant (靜止) periods of the geyser (間歇泉) Old Faithful in Minutes

| TABLE 1.3 Durations (in minutes) of dormant periods of the geyser Old Faithful |
|---------------------------------|---------------------------------|
| 42 45 49 50 51 51 51 53 53     | 55 55 56 56 57 58 60 66 67 67 |
| 68 69 70 71 72 73 73 74 75 75 | 75 75 76 76 76 76 79 79 80    |
| 80 80 80 81 82 82 82 83 83 84 | 84 84 85 86 86 86 88 90 91 93 |

**FIGURE 1.5 Stem-and-leaf plot for the geyser data in Table 1.3.**

– Let’s look at the first line of the stem-and-leaf plot. This represents measurements of 42, 45, and 49 minutes

– A good feature of these plots is that they display all the sample values. One can reconstruct the data in its entirety from a stem-and-leaf plot (however, the order information that items sampled is lost)
Another Example: Particulate matter (PM) emissions for 62 vehicles driven at high altitude

**TABLE 1.2** Particulate matter (PM) emissions (in g/gal) for 62 vehicles driven at high altitude

<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>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>7.59</td>
<td>6.28</td>
<td>6.07</td>
<td>5.23</td>
<td>5.54</td>
<td>3.46</td>
<td>2.44</td>
<td>3.01</td>
<td>13.63</td>
<td>13.02</td>
<td>23.38</td>
<td>9.24</td>
<td>3.22</td>
</tr>
<tr>
<td>2.06</td>
<td>4.04</td>
<td>17.11</td>
<td>12.26</td>
<td>19.91</td>
<td>8.50</td>
<td>7.81</td>
<td>7.18</td>
<td>6.95</td>
<td>18.64</td>
<td>7.10</td>
<td>6.04</td>
<td>5.66</td>
</tr>
<tr>
<td>8.86</td>
<td>4.40</td>
<td>3.57</td>
<td>4.35</td>
<td>3.84</td>
<td>2.37</td>
<td>3.81</td>
<td>5.32</td>
<td>5.84</td>
<td>2.89</td>
<td>4.68</td>
<td>1.85</td>
<td>9.14</td>
</tr>
<tr>
<td>8.67</td>
<td>9.52</td>
<td>2.68</td>
<td>10.14</td>
<td>9.20</td>
<td>7.31</td>
<td>2.09</td>
<td>6.32</td>
<td>6.53</td>
<td>6.32</td>
<td>2.01</td>
<td>5.91</td>
<td>5.60</td>
</tr>
<tr>
<td>5.61</td>
<td>1.50</td>
<td>6.46</td>
<td>5.29</td>
<td>5.64</td>
<td>2.07</td>
<td>1.11</td>
<td>3.32</td>
<td>1.83</td>
<td>7.56</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**FIGURE 1.6** Stem-and-leaf plot of the PM data in Table 1.2 in Section 1.2 as produced by MINITAB.
Dotplot

- A dotplot is a graph that can be used to give a rough impression of the shape of a sample
  - Where the sample values are concentrated
  - Where the gaps are
- It is useful when the sample size is not too large and when the sample contains some repeated values
- Good method, along with the stem-and-leaf plot to informally examine a sample
- Not generally used in formal presentations

Figure 1.7 Dotplot of the geyser data in Table 1.3
Histogram (1/3)

- A graph gives an idea of the *shape* of a sample
  - Indicate regions where samples are concentrated or sparse

- To have a histogram of a sample
  - The first step is to construct a frequency table
    - Choose boundary points for the class intervals
    - Compute the frequencies and relative frequencies for each class
      - *Frequency*: the number of items/points in the class
      - *Relative frequencies*: frequency/sample size
    - Compute the density for each class, according to the formula
      \[
      \text{Density} = \frac{\text{relative frequency}}{\text{class width}}
      \]
      - Density can be thought of as the *relative frequency* per unit
The second step is to draw a histogram for the table.

- Draw a rectangle for each class, whose height is equal to the density.

The total areas of rectangles is equal to 1.
Histogram (3/3)

- A common rule of thumb for constructing the histogram of a sample
  - It is good to have more intervals rather than fewer
  - But it also to good to have large numbers of sample points in the intervals
  - Striking the proper balance between the above is a matter of judgment and of trial and error
  - It is reasonable to take the number of intervals roughly equal to the square root of the sample size
Histogram with Equal Class Widths

- Default setting of most software package
- Example: an histogram with equal class widths for Table 1.4

![Histogram](image)

- Devoted to too many (more than half) of the class intervals to few (7) data points
- Compared to Figure 1.9, Figure 1.8 presents a smoother appearance and better enables the eye to appreciate the structure of the data set as a whole
Histogram, Sample Mean and Sample Variance (1/2)

• Definition: The center of mass of the histogram is

\[ \sum_i \text{CenterOfClassInterval}_i \times \text{DensityOfClassInterval}_i \]

– An approximation to the sample mean
– E.g., the center of mass of the histogram in Figure 1.8 is

\[ 2 \times 0.194 + 4 \times 0.177 + \cdots + 20 \times 0.065 = 6.730 \]

• While the sample mean is 6.596
– The narrower the rectangles (intervals), the closer the approximation (the extreme case \( \Rightarrow \) each interval contains only items of the same value)

\[
\begin{array}{cccc}
0.5 & 1, 1, 1 & 2, 3, 4 \\
1.5 & 1, 1, 1 & 2, 3, 4 \\
2.5 & 1, 1, 1 & 2, 3, 4 \\
3.5 & 1, 1, 1 & 2, 3, 4 \\
4.5 & 1, 1, 1 & 2, 3, 4 \\
\end{array}
\]

\[
\begin{align*}
2 \times \frac{5}{6 \times 3} + 4 \times \frac{1}{6 \times 1} &= \frac{22}{18} = 1.22 \\
1 \times \frac{3}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} &= \frac{12}{6} = 2
\end{align*}
\]
**Histogram, Sample Mean and Sample Variance (2/2)**

- **Definition:** The moment of inertia (力矩慣量) for the entire histogram is

  \[
  \sum_{i} \left( \text{CenterOfClassInterval}_i - \text{CenterOfMassOfHistogram} \right)^2 \\
  \times \text{DensityOfClassInterval}_i
  \]

- An approximation to the sample variance
- E.g., the moment of inertia for the entire histogram in Figure 1.8 is

  \[
  (2-6.730)^2 \times 0.194 + (4-6.730)^2 \times 0.177 + \cdots + (20-6.730)^2 \times 0.065 = 20.25
  \]

- While the sample mean is 20.42
- The narrower the rectangles (intervals) are, the closer the approximation is
Symmetry and Skewness (1/2)

• A histogram is perfectly **symmetric** if its right half is a mirror image of its left half
  – E.g., heights of random men

• Histograms that are not symmetric are referred to as **skewed**

• A histogram with a long right-hand tail is said to be **skewed to the right**, or **positively skewed**
  – E.g., incomes are right skewed (?)

• A histogram with a long left-hand tail is said to be **skewed to the left**, or **negatively skewed**
  – Grades on an easy test are left skewed (?)
Symmetry and Skewness (2/2)

- There is also another term called “kurtosis” that is also widely used for descriptive statistics
  - Kurtosis is the degree of peakedness (or contrarily, flatness) of the distribution of a population
More on Skewness and Kurtosis (1/3)

• Skewness can be used to characterize the symmetry of a data set (sample)

• Given a sample: \( X_1, X_2, \ldots, X_n \)
  
  – Skewness is defined by \( \text{Skewness} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{(n - 1)s^3} \)

  – If \( X_i \) follows a normal distribution or other distributions with a symmetric distribution shape => \( \text{Skewness} = 0 \)

  – \( \text{Skewness} > 0 \) : Skewed to the right

  – \( \text{Skewness} < 0 \) : Skewed to the left
More on Skewness and Kurtosis (2/3)

- Kurtosis can be used to characterize the flatness of a data set (sample)
- Given a sample: $X_1, X_2, \ldots, X_n$
  
  - Kurtosis is defined by
    \[
    Kurtosis = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^4}{s^4 (n - 1)}
    \]
  
  - A standard normal distribution has $Kurtosis = 3$

- A larger kurtosis value indicates a "peaked" distribution
- A smaller kurtosis value indicates a "flat" distribution
More on Skewness and Kurtosis (3/3)

Unimodal and Bimodal Histograms (1/2)

• Definition: Mode
  – Can refer to the most frequently occurring value in a sample
  – Or refer to a peak or local maximum for a histogram or other curves

A unimodal histogram

A bimodal histogram

• A bimodal histogram, in some cases, indicates that the sample can be divides into two subsamples that differ from each other in some scientifically important way.
Unimodal and Bimodal Histograms (2/2)

- Example: Durations of dormant periods (in minutes) and the previous eruptions of the geyser Old Faithful

<table>
<thead>
<tr>
<th>Dormant</th>
<th>Eruption</th>
<th>Dormant</th>
<th>Eruption</th>
<th>Dormant</th>
<th>Eruption</th>
<th>Dormant</th>
<th>Eruption</th>
</tr>
</thead>
<tbody>
<tr>
<td>76</td>
<td>Long</td>
<td>90</td>
<td>Long</td>
<td>45</td>
<td>Short</td>
<td>84</td>
<td>Long</td>
</tr>
<tr>
<td>80</td>
<td>Long</td>
<td>42</td>
<td>Short</td>
<td>88</td>
<td>Long</td>
<td>70</td>
<td>Long</td>
</tr>
<tr>
<td>84</td>
<td>Long</td>
<td>91</td>
<td>Long</td>
<td>51</td>
<td>Short</td>
<td>79</td>
<td>Long</td>
</tr>
<tr>
<td>50</td>
<td>Short</td>
<td>51</td>
<td>Short</td>
<td>80</td>
<td>Long</td>
<td>60</td>
<td>Long</td>
</tr>
<tr>
<td>93</td>
<td>Long</td>
<td>79</td>
<td>Long</td>
<td>49</td>
<td>Short</td>
<td>86</td>
<td>Long</td>
</tr>
<tr>
<td>55</td>
<td>Short</td>
<td>53</td>
<td>Long</td>
<td>82</td>
<td>Long</td>
<td>71</td>
<td>Long</td>
</tr>
<tr>
<td>76</td>
<td>Long</td>
<td>82</td>
<td>Long</td>
<td>75</td>
<td>Long</td>
<td>67</td>
<td>Short</td>
</tr>
<tr>
<td>58</td>
<td>Short</td>
<td>51</td>
<td>Short</td>
<td>73</td>
<td>Long</td>
<td>81</td>
<td>Long</td>
</tr>
<tr>
<td>74</td>
<td>Long</td>
<td>76</td>
<td>Long</td>
<td>67</td>
<td>Long</td>
<td>76</td>
<td>Long</td>
</tr>
<tr>
<td>75</td>
<td>Long</td>
<td>82</td>
<td>Long</td>
<td>68</td>
<td>Long</td>
<td>83</td>
<td>Long</td>
</tr>
<tr>
<td>80</td>
<td>Long</td>
<td>84</td>
<td>Long</td>
<td>86</td>
<td>Long</td>
<td>76</td>
<td>Long</td>
</tr>
<tr>
<td>58</td>
<td>Short</td>
<td>53</td>
<td>Short</td>
<td>72</td>
<td>Long</td>
<td>55</td>
<td>Short</td>
</tr>
<tr>
<td>80</td>
<td>Long</td>
<td>86</td>
<td>Long</td>
<td>75</td>
<td>Long</td>
<td>73</td>
<td>Long</td>
</tr>
<tr>
<td>69</td>
<td>Long</td>
<td>51</td>
<td>Short</td>
<td>75</td>
<td>Long</td>
<td>56</td>
<td>Short</td>
</tr>
<tr>
<td>57</td>
<td>Long</td>
<td>85</td>
<td>Long</td>
<td>66</td>
<td>Short</td>
<td>83</td>
<td>Long</td>
</tr>
</tbody>
</table>

long: more than 3 minutes
short: less than 3 minutes
Histogram with Height Equal to Frequency

• Till now, we refer the term “histogram” to a graph in which the heights of rectangles represent densities.

• However, some people draw histograms with the heights of rectangles equal to the frequencies.

• Example: The histogram of the sample in Table 1.4 with the heights equal to the frequencies.

![Figure 1.13](image1.png)  
**Figure 1.13**  
Exaggerate the proportion of vehicles in the intervals.

![Figure 1.8](image2.png)  
**Figure 1.8**  
Cf.
Boxplot (1/4)

- **A boxplot** is a graph that presents the median, the first and third quartiles, and any outliers present in the sample.

  - The **interquartile range (IQR)** is the difference between the **third** and **first** quartile. This is the distance needed to span the middle half of the data.
Boxplot (2/4)

• Steps in the Construction of a Boxplot
  – **Compute the median** and the first and third quartiles of the sample. Indicate these with horizontal lines. Draw vertical lines to complete the box
  – **Find the largest sample value** that is no more than 1.5 IQR above the third quartile, and the **smallest sample value** that is not more than 1.5 IQR below the first quartile. **Extend vertical lines (whiskers) from the quartile lines to these points**
  – Points more than 1.5 IQR above the third quartile, or more than 1.5 IQR below the first quartile are designated as outliers. Plot each outlier individually
Boxplot (3/4)

• Example: A boxplot for the geyser data presented in Table 1.5
  – Notice there are no outliers in these data
  – The sample values are comparatively densely packed between the median and the third quartile
  – The lower whisker is a bit longer than the upper one, indicating that the data has a slightly longer lower tail than an upper tail
  – The distance between the first quartile and the median is greater than the distance between the median and the third quartile
  – This boxplot suggests that the data are skewed to the left (?)
Boxplot (4/4)

- Another Example: Comparative boxplots for PM emissions data for vehicle driving at high versus low altitudes
Scatterplot (1/2)

- Data for which item consists of a pair of values is called bivariate
- The graphical summary for bivariate data is a scatterplot
- Display of a scatterplot (strength of Titanium (鈦) welds vs. its chemical contents)
• Example: Speech feature sample (Dimensions 1 & 2) of male (blue) and female (red) speakers after LDA transformation
Summary

• We discussed types of data

• We looked at sampling, mostly SRS

• We studied summary (descriptive) statistics
  – We learned about numeric summaries
  – We examined graphical summaries (displays of data)