1

Summarizing Data

(Ch 1.1, 1.3, 1.10-1.13, 2.4.3, 2.5)
Populations and Samples

An investigation of some characteristic of a population of interest.

Example: You want to study the average GPA of juniors who are engineering majors.

Population:
All engineering majors who are juniors.

Characteristic of interest:
Average GPA.
Populations and Samples

What statisticians need to do:

1) Learn about the distribution of the characteristic in the population.
2) Do this by taking a sample from the population. Why?
3) Sample statistics.
Graphics: Histograms

A histogram is a graphical representation of the distribution of numerical data.

Construct a histogram:

1. “Bin” the range of values. (The bins are usually consecutive, non-overlapping, and are usually equal size.)

2. Frequency histogram: count how many values fall into each bin/interval and draw accordingly.

3. Density histogram: count how many values fall into each bin, and adjust the height such that the sum of the area of each bin equals 1.
Graphics: Histograms

Examples:
- Drawing a frequency histogram by hand.
- Drawing a density histogram by hand.
Example

Charity is a big business in the United States. The Web site charitynavigator.com gives information on roughly 5500 charitable organizations.

Some charities operate very efficiently, with fundraising and administrative expenses that are only a small percentage of total expenses, whereas others spend a high percentage of what they take in on such activities.
Here are the data on fundraising expenses as a percentage of total expenditures for a random sample of 60 charities:

<table>
<thead>
<tr>
<th>6.1</th>
<th>12.6</th>
<th>34.7</th>
<th>1.6</th>
<th>18.8</th>
<th>2.2</th>
<th>3.0</th>
<th>2.2</th>
<th>5.6</th>
<th>3.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>3.1</td>
<td>1.3</td>
<td>1.1</td>
<td>14.1</td>
<td>4.0</td>
<td>21.0</td>
<td>6.1</td>
<td>1.3</td>
<td>20.4</td>
</tr>
<tr>
<td>7.5</td>
<td>3.9</td>
<td>10.1</td>
<td>8.1</td>
<td>19.5</td>
<td>5.2</td>
<td>12.0</td>
<td>15.8</td>
<td>10.4</td>
<td>5.2</td>
</tr>
<tr>
<td>6.4</td>
<td>10.8</td>
<td>83.1</td>
<td>3.6</td>
<td>6.2</td>
<td>6.3</td>
<td>16.3</td>
<td>12.7</td>
<td>1.3</td>
<td>0.8</td>
</tr>
<tr>
<td>8.8</td>
<td>5.1</td>
<td>3.7</td>
<td>26.3</td>
<td>6.0</td>
<td>48.0</td>
<td>8.2</td>
<td>11.7</td>
<td>7.2</td>
<td>3.9</td>
</tr>
<tr>
<td>15.3</td>
<td>16.6</td>
<td>8.8</td>
<td>12.0</td>
<td>4.7</td>
<td>14.7</td>
<td>6.4</td>
<td>17.0</td>
<td>2.5</td>
<td>16.2</td>
</tr>
</tbody>
</table>
Example

We can see that a substantial majority of the charities in the sample spend less than 20% on fundraising:
Graphics: Histograms

Histograms come in a variety of shapes.
• **Unimodal** histogram: single peak
• **Bimodal** histogram: two different peaks
• **Multimodal** histogram: many different peaks

**Bimodality**: Can occur when the data set consists of observations on two quite different kinds of individuals or objects.

**Multimodality**

**Symmetric** histograms
**Positively skewed** histograms
**Negatively skewed** histograms
Sample Statistics

• Histograms and other *visual summaries* of samples are excellent tools for informal learning about population characteristics.

• The calculation and interpretation of certain summarizing numbers are required for a deeper understanding of the data.

• These sample numerical summaries are called “Sample Statistics”
Sample Statistics: Measures of Centrality

Summarizing the center of the sample data is a popular and important characteristic of a set of numbers.

3 popular types of center:
1. Mean
2. Median
3. Mode
The Sample Mean

For a given set of numbers $x_1, x_2, \ldots, x_n$, the most familiar measure of the center is the *mean* (arithmetic average).

**Sample mean** $\bar{x}$ of observations $x_1, x_2, \ldots, x_n$:

$$
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
$$
The Sample Mean

For a given set of numbers \( x_1, x_2, \ldots, x_n \), the most familiar measure of the center is the \textit{mean} (arithmetic average).

**Sample mean** \( x \) of observations \( x_1, x_2, \ldots, x_n \):

\[
\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

Disadvantage?
The Sample Median

**Median**: Middle value when observations are ordered smallest to largest.
The Sample Median

**Median**: Middle value when observations are ordered smallest to largest.

To calculate: Order the $n$ observations smallest to largest (repeated values included and find the middle one.

$$\tilde{x} = \begin{cases} 
\text{The single middle value if } n \\
\text{is odd} \\
\text{The average of the two middle values if } n \\
\text{is even} 
\end{cases} = \left(\frac{n + 1}{2}\right)^{\text{th}} \text{ ordered value} = \text{average of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2} + 1\right)^{\text{th}} \text{ ordered values}$$
The Mean vs. the Median

The population mean $\mu$ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as pictured below, then $\mu \neq \tilde{\mu}$.

(a) Negative skew

(b) Symmetric

(c) Positive skew

Three different shapes for a population distribution

Which population characteristic is most important?
The population mean $\mu$ and median $\tilde{\mu}$ will not generally be identical. If the population distribution is positively or negatively skewed, as pictured below, then $\mu \neq \tilde{\mu}$.

(a) Negative skew

(b) Symmetric

(c) Positive skew

Three different shapes for a population distribution
Other Sample Measures

- **Quartiles**: divide the data set into four equal parts (how is this calculated?)
- **Percentiles**: A data set can be even more finely divided. What does “percentile” mean?

Example calculations of the median and quartiles.
A boxplot is a convenient way of graphically depicting groups of numerical data through the five number summary: minimum, first quartile, median, third quartile, and maximum.

Example: Drawing a boxplot by hand.
Variability

So far, we’ve learned techniques for visualizing our data and measures of center. What about how far apart the data is spread out?

Samples with identical measures of center but different amounts of variability
Variability

Simplest measure of variability: The range.

Samples with identical measures of center but different amounts of variability
Variability

Simplest measure of variability: The range.

Samples with identical measures of center but different amounts of variability

What are the disadvantages of the range?
Variability

Can we combine the deviations into a single quantity by finding the average deviation?

A more robust measure of variation takes into account deviations from the mean

\[ x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}. \]
The **sample variance**, denoted by $s^2$, is given by

$$s^2 = \frac{\sum(x_i - \overline{x})^2}{n - 1}$$

The **sample standard deviation**, denoted by $s$, is the (positive) square root of the variance:

$$s = \sqrt{s^2}$$

Note that $s^2$ and $s$ are both nonnegative. The unit for $s$ is the same as the unit for each of the $x_i$.

Example: Calculation of the SD.
Summarizing Data in R

- Summary statistics
- Graphics (boxplots, histograms)