Description of Samples and Populations

Random Variables

Data are generated by some underlying random process or phenomenon. Any datum (data point) represents the outcome of a random variable. We represent random variables with capital letters, usually X, Y, and Z.

Example: Let X = weight of a newborn baby. Let Y = weight of baby at one week old

Types of Random Variables

Qualitative

- Categorical (or nominal)

- Ordinal

Quantitative

- Discrete

- Continuous

Definition: An observation is a recorded outcome of a variable from a random sample. We represent observations with lower case letters. For example, suppose we are measuring the outcome of a random variable X = weight of 10 newborn babies. Our observations would be denoted by \( x_1, x_2, \ldots, x_{10} \). Notation: \( x_1 \) is the first observation.

Definition: The observational unit is the type of subject being sampled. It is the smallest unit upon which we make our observations. Observational units could be a baby, moth, Petri dish, etc...
Example: For the following setting identify the (i) variable(s) in the study, (ii) type of variable (iii) observational unit (iv) sample size

(From example 1.1.4) In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 42 patients. The results were recorded as nmols benzyaldehyde product per 108 platelets.

**Definition:** A frequency distribution is a summary display of the frequencies of occurrence of each value in a sample.

**Definition:** A relative frequency is a raw frequency divided by n (sample size).

Example 2.2.4 $Y =$ number of piglets surviving 21 days (litter size at 21 days) What is the sample size?

What is the relative frequency for a litter size at 21 days of 10?

<table>
<thead>
<tr>
<th>Number of piglets</th>
<th>Frequency (number of sowss)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>

**Graphical Displays**

After you collect data, we hope that it tells us something. We look at it in order to learn something about the process that generated it. One way to summarize (or describe) the data is through a graphical display.

A graphical display should always be as clear as possible. It should be well labeled with a title, key (if necessary), labels on the graphical display itself, units should be clear from your display, and the sample size should be clear. Do not over-label your graphical display!

Nicely formatting a graphical display can be more difficult than it sounds….we’ll try one in a minute and you’ll see they require some thought.
A **dot plot** is a graphical display where dots indicate observed data in a sample.

Figure 2.2.4 is an illustration of the frequency distribution of example 2.2.4.

Surviving Piglets at 21 Days   
n=36 sows

A **histogram** is a graphical display where bars (or bins) replace the dots from a dotplot.

Figure 2.2.5 is a histogram of the frequency distribution of example 2.2.4.

**Surviving Piglets at 21 Days   (n=36 sows)**

Let’s make a histogram in R

First, a trick.  `rep(10,9)` means “repeat 10 9 times” or list 9 10’s in a row

We can quickly enter the Litter Size data of table 2.2.4

```r
Lsize<-c(5,7,7,8,8,9,9,rep(10,9),rep(11,8),rep(12,5),13,13,13,14,14)
```

Look at what `hist(Lsize)` does. What is wrong with the display?
Turns out this is a glitch in R.

Type ?hist

This gives all the options available for the histogram function. We’ll use “breaks” to fix the histogram.

Type hist(Lsize,breaks=4:14)

The breaks argument in the histogram function tells where the edges of the bars (bins) of the histogram are defined. The 4:14 statement is the same as saying “list the numbers from 4 up to 14”.

Now, notice this is not well labeled.

Hist(Lsize,breaks=4:14,main=“insert title here”,sub=“insert subtitle here”) puts a title and subtitle on the display.

**Warning:** A histogram can be a very misleading display – we’ll see an example of this in a minute.

A **stemplot** or **stem and leaf plot** is a lot like a dot plot, usually turned on its side. We use a stemplot when we have more detailed information to replace the dots with.

The “stems” are the core values of the data and the “leaves” are the last values of the data points. We’ll put the “leaves” in numerical order in this class and the resulting plot is called an ordered stemplot … but, we’ll never use the unordered kind, so we’ll refer to ours simply as a stemplot. **A stemplot should always include a key with units!**

Example (data from a portion of example 1.1.4)

In a study of schizophrenia, researchers measured the activity of the enzyme monoamine oxidase (MAO) in the blood platelets of 18 patients. The results were recorded as nmols benzylaldehyde product per 108 platelets.

4.1  5.2  6.8  7.3  7.4  7.8  7.8  8.4  8.7  9.7  9.9  10.6  10.7  11.9  12.7  14.2  14.5  18.8

Create a stemplot for these data using R. stem() is the stemplot function in R. Let’s figure it out together.
Describing the “shape” of a frequency distribution

We can see the shape of a frequency distribution by looking at an appropriate graphical display. The following are some basic words we use to describe the shape of a frequency distribution

- symmetric / asymmetric
- bell-shaped / skewed left / skewed right
- unimodal / bimodal

Some examples...
Example

(From Exercise 2.13 in 3rd ed. of text) Trypanosomes are parasites which cause disease in humans and animals. In an early study of trypanosome morphology, researchers measured the lengths (μm) of 500 individual trypanosomes taken from the blood of a rat. The results are summarized in the accompanying frequency distribution.

The following is the default histogram returned by a statistical software package (not well labeled yet!) Describe the shape of the frequency distribution.

![Default Histogram](image1)

This next histogram is returned by the same software package for the trypanosome data, with the bin width changed. How would you describe the shape of the distribution now? Discuss the changes.

![Modified Histogram](image2)
We learned to describe data sets graphically. We can also describe a data set numerically.

**Definition** A statistic is a sample quantity used to *estimate* a population parameter. A statistic is a numerical quantity calculated from the sample data.

**Measures of Location**

The **sample median** is the value of the data nearest their middle. We call the median \( Q_2 \) or \( M \) and your text denotes it as \( \hat{y} \). To find the median of a data set

- Put the data in order
- The median is
  - The middle value if \( n \) is odd
  - The average of the two middle values if \( n \) is even

**Example** Exercise 2.3.1 \( Y = \) weight gain (lb) of lambs on a special diet for \( n = 6 \) lambs. The ordered values are \( y(1)=1, \ y(2)=2, \ y(3)=10, \ y(4)=11, \ y(5)=13, \ y(6)=19 \)

Find \( Q_2 \).

Notice the new notation. A lower case letter denoting the outcome of a random variable with parenthesis in the subscript \( y(i) \) denotes the \( i^{th} \) observation in order from smallest to largest. \( i=(1) \) denotes the smallest observation (minimum) and \( i=(n) \) denotes the largest (maximum).

**Definition** The **sample mean** is the arithmetic average of \( n \) values. We denote the sample mean by

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \cdots + y_n}{n}
\]

**Example** 2.3.3 \( Y = \) weight gain (lb) of lambs on a special diet for \( n = 6 \) lambs. Compute the sample mean for the resulting data set.

11 13 19 2 10 1
The following two figures illustrate that the sample mean can be viewed as a balancing point in the data, whereas the sample median is the point which divides the data set into the upper half and lower half.

**Figure 2.3.2** Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

**Figure 2.3.3** Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

**Question:** Which measure of location (or measure of center) do we report? Mean or median? To answer this, explore what happens on certain data sets to the relation between the mean and median.

Consider two data sets. Find the mean and the median for both.

1 2 3 4 5
1 2 3 4 20

What happened and why?

So...the comparison of the mean and median can indicate skewness.

Data skewed right, mean median
Data skewed left, mean median
Data symmetric, mean median

**Measures of Dispersion (IQR, Range, and Standard Deviation)**

The quartiles of a data set are points that separate the data into quarters (or fourths).

Q₁ separates the lower quarter (25%) from the upper three quarters (75%)
Q₂ separates the lower two quarters (50%) from the upper two quarters (50%)
Q₃ separates the lower three quarters (75%) from the upper quarter (25%).
Notice the median is the second quartile.

One way to report the dispersion (or spread) of a data set is to report the inter-quartile range.

**Definition** The inter-quartile range is \( \text{IQR} = Q₃ - Q₁ \)

**Definition** The sample range is \( y_{(n)} - y_{(1)} = \text{max} - \text{min} \)

**Definition** The five number summary is \( \{y_{(1)}, Q₁, Q₂, Q₃, y_{(n)}\} \)

Example (from Example 2.22) In a common biology experiment, radishes were grown in total darkness and the length (mm) of each radish shoot was measured at the end of three days. Find the five number summary for these data. 
8 10 11 15 15 15 20 20 22 25 29 30 33 35 37

**Definition** An outlier is an observation that differs dramatically from the rest of the data.
Formally \( yᵢ \) is an outlier if

\[
Yᵢ < Q₁ - (1.5 \times \text{IQR}) \quad \text{or} \quad Yᵢ > Q₃ + (1.5 \times \text{IQR})
\]

“lower fence” “upper fence”

Example 2.4.5 \( Y = \) radish growth in full light condition. The data are 
3 5 5 7 7 8 9 10 10 10 14 20 21
Find any outliers.
A boxplot (a.k.a. box and whisker plot) is a graphical display of the five number summary.

The “box” spans the quartiles and the “whiskers” extend from the quartiles to the min / max.

Figure 2.4.2 Dotplot and boxplot of radish height after three days in full light. Notice the outliers are plotted beyond the whiskers.

Boxplots are often used for comparative purposes as in figure 2.5.3, below. Compare the three distributions on shape, center, and spread.

Radish Length at Three Days
Grown Under Three Conditions
**Definition**  The sample variance is  \( s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \)

**Definition**  The sample standard deviation is  \( s = \sqrt{s^2} \)

**Example 2.28**  In an experiment on chrysanthemums, a botanist measured the stem elongation (mm in 7 days) of five plants grown on the same greenhouse bench:

| 76 | 72 | 65 | 70 | 82 |

Find the sample standard deviation.

**Empirical Rule**

For unimodal, not too skewed data sets, the empirical rule states the following:

\(~68\%\) of the data lie between \( \bar{Y} + s \) and \( \bar{Y} - s \)

\(~95\%\) of the data lie between \( \bar{Y} + 2s \) and \( \bar{Y} - 2s \)

\(>99\%\) of the data lie between \( \bar{Y} + 3s \) and \( \bar{Y} - 3s \)

**Example from 3rd ed.**  Suppose \( Y \) = pulse rate after 5 minutes of exercise.  For \( n = 28 \) subjects, we find \( \bar{Y} = 98 \) (beats/min) and \( s = 13.4 \) (beats/min) and the data are unimodal and not too skewed.

Thus, from the empirical rule we expect \(~95\%\) of the data to lie between

\[ 98 - (2)(13.4) = 98 - 26.8 = 71.2 \text{ beats/min} \]

and

\[ 98 + (2)(13.4) = 98 + 26.8 = 124.8 \text{ beats/min} \]
Figure 2.5.3, comparing the distribution of radish growth under the three lighting conditions, alludes to the idea that often, comparisons are made among groups and we like to analyze the relationship between variables.

Figure 2.5.5 illustrates a common way to explore the relationship between two quantitative variables – a scatterplot. This plot is depicting tooth selenium concentration versus liver selenium concentration for 20 beluga whales.

![Scatterplot of tooth selenium concentration versus liver selenium concentration for 20 beluga whales.](image)

<p>| Table 2.5.3 Liver and tooth selenium concentrations of twenty belugas |
|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|</p>
<table>
<thead>
<tr>
<th>Whale</th>
<th>Liver Se (µg/g)</th>
<th>Tooth Se (µg/g)</th>
<th>Whale</th>
<th>Liver Se (µg/g)</th>
<th>Tooth Se (µg/g)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.23</td>
<td>140.16</td>
<td>11</td>
<td>15.28</td>
<td>112.63</td>
</tr>
<tr>
<td>2</td>
<td>6.79</td>
<td>133.32</td>
<td>12</td>
<td>18.68</td>
<td>245.07</td>
</tr>
<tr>
<td>3</td>
<td>7.92</td>
<td>135.34</td>
<td>13</td>
<td>22.08</td>
<td>140.48</td>
</tr>
<tr>
<td>4</td>
<td>8.02</td>
<td>127.82</td>
<td>14</td>
<td>27.33</td>
<td>177.93</td>
</tr>
<tr>
<td>5</td>
<td>9.34</td>
<td>108.67</td>
<td>15</td>
<td>32.83</td>
<td>160.73</td>
</tr>
<tr>
<td>6</td>
<td>10.00</td>
<td>146.22</td>
<td>16</td>
<td>26.04</td>
<td>227.60</td>
</tr>
<tr>
<td>7</td>
<td>10.57</td>
<td>131.18</td>
<td>17</td>
<td>37.74</td>
<td>177.69</td>
</tr>
<tr>
<td>8</td>
<td>11.04</td>
<td>145.51</td>
<td>18</td>
<td>40.00</td>
<td>174.23</td>
</tr>
<tr>
<td>9</td>
<td>12.36</td>
<td>163.24</td>
<td>19</td>
<td>41.23</td>
<td>206.30</td>
</tr>
<tr>
<td>10</td>
<td>14.53</td>
<td>136.55</td>
<td>20</td>
<td>45.47</td>
<td>141.31</td>
</tr>
</tbody>
</table>

How would you describe the relationship between liver and tooth selenium concentrations?

Does it look like there are any possible outliers (are there any extreme observations that might qualify as outliers - quantifying an extreme observation as an outlier when comparing two variables is totally different from our univariate definition)?
**Where are we headed?** The goal of a statistical analysis is to ultimately describe a population.

**Definition** The population is the larger group of subjects (organisms, plots, regions, ecosystems, etc.) on which we wish to draw inferences.

**Definition** A parameter is a quantified population characteristic. It is usually unknown and often assumed to be a fixed value.

**Definition** A statistic is a sample quantity used to estimate a population parameter. It is known and can vary from sample to sample.

**Question:** When we use a statistic to estimate a parameter, is that statistic going to hit the target (is the statistic going to exactly equal the population parameter)?

The process of using a statistic to estimate a parameter and then quantifying the uncertainty in our estimate (i.e. making a statement about how far off our estimate is, in general) is called **statistical inference**.

**Definition** The population proportion is the proportion of subjects exhibiting a particular trait or outcome in the population. (It generalizes to the probability that any population element will exhibit the trait.) **NOTATION:** $p$

**Definition** The SAMPLE PROPORTION is the number of sample elements exhibiting the trait, divided by the sample size, $n$. **NOTATION:** $\hat{p}$
<table>
<thead>
<tr>
<th>Measure</th>
<th>Sample value (statistic)</th>
<th>Population value (parameter)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion</td>
<td>$\hat{p}$</td>
<td>$p$</td>
</tr>
<tr>
<td>Mean</td>
<td>$\bar{y}$</td>
<td>$\mu$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$s$</td>
<td>$\sigma$</td>
</tr>
</tbody>
</table>