2.1: Descriptive Statistics and Distributions

Learning Objectives

- Define and distinguish between statistics and parameters
- Define and calculate proportions
- Introduce and distinguish between frequency and relative frequency distributions
- Introduce and use summation notation
- Introduce grouped distributions
- Introduce bar charts and histograms
- Identify the skew of a distribution

Section 2.1 Excel File (contains all of the data sets for this section)

Population Parameters and Sample Statistics

We gain a better understanding of the world around us by collecting and analyzing data. Recall that, most of the time, it is not possible or practical to collect all the data around a certain topic. For this reason, we often rely on inferential statistics to make informed guesses about the population using data from a random sample. It is important for our analyses to differentiate facts about the population from facts about a sample. Facts about sample data are called **statistics** while facts about populations are called **parameters**. It is common, but not universal, to use Greek letters (such as μ, σ) when referring to population parameters and Latin letters (such as $\bar{x}, s$) when referring to sample statistics. Our first example highlights one of the exceptions to this practice.

Both sample statistics and population parameters fall under the umbrella of descriptive statistics; they are numbers that are used to summarize and describe data. A commonly used descriptive statistic is the proportion. A **proportion** is the percentage of observations that have a certain characteristic. Many important issues rely on estimating proportions. What proportion of customers are satisfied with our services? What proportion of people who take some medicine experience negative side effects? What proportion of voters support this political candidate? The symbol for population proportion is p, and the symbol for sample proportion is $\hat{p}$ (read: p-hat). For example, suppose a survey was given out a month before a local election. Out of the 100 people surveyed, 54 supported a particular candidate. However, on the actual election day, that candidate only got 48% of the votes. We would say $\hat{p} = \frac{54}{100} = 0.54$ and $p = 0.48$.

Text Exercise 2.1.1

A husband (Adam) and wife (Betsy) have three children (Cathy, Damon, and Erin). Adam, Betsy, and Cathy wear glasses. We are interested in studying this particular family.

1. Compute the population proportion p of family members that wear glasses.

Answer

To find the percentage of family members that wear glasses, we need to know the total number of family members (the population size N) and the number of family members that wear glasses (the number of observations with the characteristic x). $N = 5$ and $x = 3$. Thus

$$p = \frac{x}{N} = \frac{3}{5} = 0.6$$

Thus 60% of the family wears glasses.

2. Construct the different samples and show that there is no sample such that $p = \hat{p}$.

Answer

To show that there is no sample such that $\hat{p} = p$, we must consider all possible samples from the population. The sample size n could be any number $\{1, 2, 3, 4\}$. If $n = 1$, then we could have someone with glasses or someone without glasses. Thus

$$\hat{p} = \frac{x}{n} = \begin{cases} \frac{0}{1} = 0 \\ \frac{1}{1} = 1 \end{cases}$$

If you are unfamiliar with this notation, we are saying that $\hat{p}$ could be 0 or 1. If $n = 2$, then we could have 0, 1, or 2 people with glasses. Thus

$$\hat{p} = \begin{cases} \frac{0}{2} = 0 \\ \frac{1}{2} = 0.5 \\ \frac{2}{2} = 1 \end{cases}$$

If $n = 3$, then we could have 1, 2, or 3 people with glasses (since there are only two people without glasses). Thus

$$\hat{p} = \begin{cases} \frac{1}{3} = 0.\overline{3} \\ \frac{2}{3} = 0.\overline{6} \\ \frac{3}{3} = 1 \end{cases}$$

If $n = 4$, then we could still only have 2, or 3 people with glasses. Thus

$$\hat{p} = \begin{cases} \frac{2}{4} = 0.5 \\ \frac{3}{4} = 0.75 \end{cases}$$

We notice that none of the possible $\hat{p}$ values match the calculated p value.

3. Suppose they had another child, Frank, show that it is now possible to have a sample such that $p = \hat{p}$

Answer

With the addition of Frank, $N = 6$ and $x = 3$ or 4 since we do not know whether Frank wears glasses or not. Thus

$$p = \begin{cases} \frac{3}{6} = 0.5 \\ \frac{4}{6} = 0.\overline{6} \end{cases}$$

Since we are showing that it is possible, finding particular samples will be sufficient. If $n = 2$, we could have someone with glasses and someone without glasses, and $\hat{p} = \frac{1}{2}$. If $n = 3$, we could have 2 people with glasses and 1 person without glasses, and $\hat{p} = \frac{2}{3}$. Thus, in this situation it is possible to have a sample such that $p = \hat{p}$.

Notice all of the proportions calculated throughout the example fell between 0 and 1. Proportions are the percentage of observations that have a certain characteristic; it is impossible to have negative numbers of observations just as it is impossible to have more observations with a certain characteristic than the total. Knowing what values are possible helps us identify when we make mistakes. We must always ask if our results are reasonable.

Distributions

Getting a firm grasp on a set of data generally requires several descriptive statistics and a method of visualization. Two very different data sets may have the same values for certain descriptive statistics while differing for others. A good place to start is to see how the data is distributed. We will build our understanding through examples.

A recently purchased bag of Plain M&M's contained six different colors of candy. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. Consider Table 2.1.1 below.

Table 2.1.1: Frequencies and Relative Frequencies of Sampled M&M's

Color	Frequency	Relative Frequency
Brown	17	$\frac{17}{55} \approx 0.309$
Red	18	$\frac{18}{55} \approx 0.327$
Yellow	7	$\frac{7}{55} \approx 0.127$
Green	7	$\frac{7}{55} \approx 0.127$
Blue	2	$\frac{2}{55} \approx 0.036$
Orange	4	$\frac{4}{55} \approx 0.073$

This table describes both the **frequency distribution** and the **relative frequency distribution** of M&M's by color. The colors form what we call **classes**. Since, every M&M must belong to a class, we say the classes are **exhaustive**. Since any particular M&M cannot be classified in more than one class, the classes are **mutually exclusive**. These two properties are important to guarantee that we count each observation only once. Distributions are often shown graphically with **bar graphs** as in Figure 2.1.1.

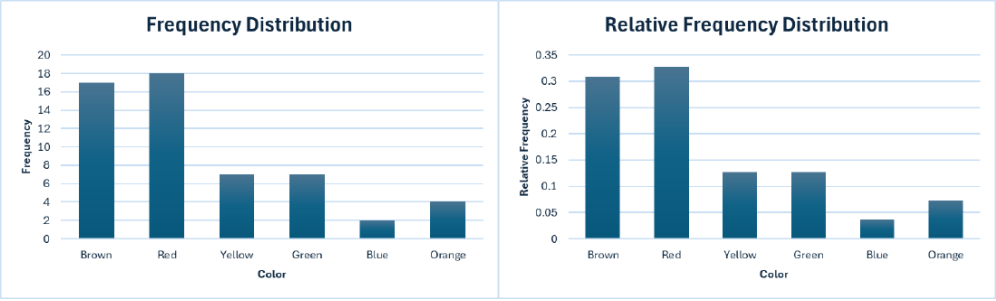

Figure 2.1.1: Frequency and Relative Frequency Distributions of 55 M&M's.

Notice how the two distributions show essentially the same information about where the data falls. Most of the M&M's were either brown or red. Yellow and green appeared equally often. Blue occurred the fewest number of times. What is the difference between the two distributions? If we looked closely at the table, we saw how the relative frequency column was computed; the frequency of the particular color was divided by the total number of M&M's. Hopefully, this reminds us of the computation for proportions. The **relative frequency** is simply the percentage of observations that have the characteristic defining the class.

Text Exercise 2.1.2

Using Table 2.1.1, determine the sum of all the relative frequencies in the relative frequency distribution of M&M's. Explain why your result must be true for all relative frequency distributions.

Answer

$\frac{17}{55}+\frac{18}{55}+\frac{7}{55}+\frac{7}{55}+\frac{2}{55}+\frac{4}{55}=\frac{55}{55}=1$. This is true for all relative frequency distributions. Since classes must be exhaustive and mutually exclusive, each observation must be in one and only one class. This means that the sum of all frequencies must add up to the total number of observations. Now relative frequencies are just frequencies divided by the total number of observations. In adding up all the relative frequencies, we could factor out the total number of observations in the denominator $\frac{1}{55}(17+18+7+7+2+4)=\frac{55}{55}=1$ and arrive at the sum of all frequencies divided by the total number of observations which is 1.

Note: Summation Notation

The explanation in the previous exercise is a bit tedious: writing down all 6 numbers repeatedly is inconvenient, and imagine repeating this exercise if there were 100 colors! Mathematicians have developed a notation to help express such arguments and computations easily. We call it summation notation. The capital Greek letter sigma $\sum$ is what we use to denote a summation. We then name all the terms that we are adding together. In our M&M's example, there were six classes and we were interested in the frequency of each class. We might refer to our frequencies by row (from the top) as f_i for $i \in \{1,2,3,4,5,6\}$. The following expression is the sum of all the relative frequencies of M&M's:

$$\sum_{i=1}^{6} \frac{f_i}{n} = \frac{f_1}{n}+\frac{f_2}{n}+\frac{f_3}{n}+\frac{f_4}{n}+\frac{f_5}{n}+\frac{f_6}{n}=1$$

For instance, $f_1 = 17$ and $f_2 = 18$. The following expression is the sum of all the relative frequencies of M&M's:

$$\sum_{i=1}^{6} \frac{f_i}{n} = \frac{17}{55}+\frac{18}{55}+\frac{7}{55}+\frac{7}{55}+\frac{2}{55}+\frac{4}{55}=\frac{55}{55}=1$$

Note that the $i = 1$ at the bottom tells us where to begin (the first class in this case) and the 6 up top tells us to add each subsequent term through that index value (the sixth class in this case).

With this notation, we can clean up our argument from the previous exercise. Suppose that we have a sample of size n with k classes. If we let f_i be the frequency in the i^{th} class, then $\sum_{i=1}^{k} f_i = n$. The sum of all of the relative frequencies would be

$$\sum_{i=1}^{k} \frac{f_i}{n} = \frac{f_1}{n}+\frac{f_2}{n}+\frac{f_3}{n}+\ldots+\frac{f_{k-1}}{n}+\frac{f_k}{n}=\frac{1}{n}\sum_{i=1}^{k} f_i=\frac{n}{n}=1$$

We encourage the reader to be familiar with summation notation, as it will be used throughout the text. If not understood, many equations given later may be confusing. It would be good to practice using the notation now while the examples are relatively simple.

Text Exercise 2.1.3

Using the frequencies from our bag of M&M's, compute the following summations:

1. $\sum_{i=3}^{5} f_i$

Answer

$\sum_{i=3}^{5} f_i = f_3+f_4+f_5 = 7+7+2 = 16$

2. $\sum_{i=2}^{4} 2f_i$

Answer

$$\sum_{i=2}^{4} 2f_i = 2f_2 + 2f_3 + 2f_4 = 2(f_2 + f_3 + f_4) = 2(18 + 7 + 7) = 2(32) = 64$$

3. $\sum_{i=1}^{3} f_i^2$

Answer

$$\sum_{i=1}^{3} f_i^2 = f_1^2 + f_2^2 + f_3^2 = 17^2 + 18^2 + 7^2 = 289 + 324 + 49 = 662$$

4. $\left(\sum_{i=1}^{3} f_i\right)^2$

Answer

$$\left(\sum_{i=1}^{3} f_i\right)^2 = (f_1 + f_2 + f_3)^2 = (17 + 18 + 7)^2 = 42^2 = 1764$$

5. $\sum_{i=1}^{6} f_i f_{7-i}$

Answer

$$\sum_{i=1}^{6} f_i f_{7-i} = f_1 f_6 + f_2 f_5 + f_3 f_4 + f_4 f_3 + f_5 f_2 + f_6 f_1 = 17 \cdot 4 + 18 \cdot 2 + 7 \cdot 7 + 7 \cdot 7 + 2 \cdot 18 + 4 \cdot 17 = 68 + 36 + 49 + 49 + 36 + 68 = 306$$

Text Exercise 2.1.4

One class of 30 students had a 10 point assignment. The student scores (raw data) were tabulated in the following set. Use the set to construct a frequency distribution in a table.

$$\{3, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 8, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 10, 10, 10\}$$

Answer

Looking at the set of scores, our classes can consist of $\{3, 4, 5, 6, 7, 8, 9, 10\}$All that is left is to count the number of observations in each class and put them in a table.

Table 2.1.2: Grouped Frequency Distribution

Student Score	Frequency
3	1
4	1
5	3
6	5
7	5
8	7
9	5
10	3

The distribution shown in Figure 2.1.1 concerns just the one bag of M&M's. We might expand our study to the distribution of colors for all regular Plain M&M's. Only the manufacturer of M&M's could provide this sort of information, but they do not tell us exactly how many M&M's of each color were ever produced. Instead, they only report the relative frequencies. See Figure 2.1.2.

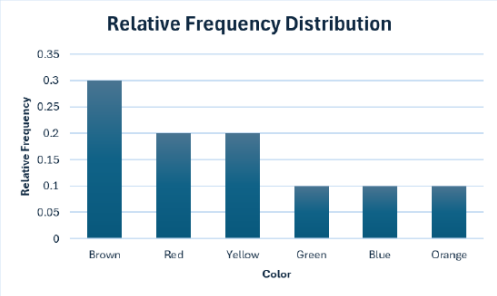

Figure 2.1.2: Distribution of all M&M's.

Notice that the relative frequency distributions in Figures 2.1.1 and 2.1.2 are not identical. Figure 2.1.1 portrays the distribution in a sample of 55 M&M's. Figure 2.1.2 shows the distribution of all M&M's. Chance factors involving the machines used by the manufacturer introduce random variation into the different bags produced. Some bags will have a distribution of colors that is close to Figure 2.1.2; others will be much different. This reinforces an important concept; sample data most often do not produce the exact same distribution/measures as what is happening in population data. We must remember this important concept to interpret our findings properly and to draw appropriate conclusions.

Distributions of Continuous Variables

Using the color of M&M's for classes seems natural, but that was not the only set of classes that could have been used. We could sort the colors as warm (red, yellow, and orange) or cool (brown, green, and blue). We could use any variable regarding M&Ms as a basis for our classes, such as weight. If we did not have a precise enough scale to differentiate weights between individual candies, weight might not been helpful. On the other hand, if our scale was too exact, we might not have had many measurements that were precisely the same. In either case, our frequency distribution would have been uninformative. Having precise measurements is a good thing; we do not want to "fix" this issue by settling for lower-quality data, but instead, we can address the problem by how we define classes. Rather than having a singular value determine our classes, we define them using a range of values. The classes must still be exhaustive and mutually exclusive. When we group values to build classes, we describe the frequency and relative frequency distributions as **grouped distributions** (this term applies as long as various values are grouped together to form classes regardless of whether we have discrete or continuous data).

The data shown in Table 2.1.3 are the times (in milliseconds) it took to move the mouse over a tiny target in a series of 20 trials. The times are sorted from shortest to longest. The variable "time to respond" is a continuous variable. With time measured so precisely, no two response times were the same; creating a grouped frequency distribution is in order.

Table 2.1.3: Response Times

568	645	720	824
577	657	728	825
581	673	729	865
640	696	777	875
641	703	808	1007

Table 2.1.4 shows one of many possible choices we could have made for a grouped frequency distribution of these 20 times. It is important to note that there is flexibility in the number of classes and where they start. Constructing multiple tables and graphs with various class numbers, sizes, and starting places helps us understand the data. We can select the most enlightening version.

Table 2.1.4: Grouped frequency distribution

Class	Class	Frequency
500 to 600	$(500, 600]$	3
600 to 700	$(600, 700]$	6
700 to 800	$(700, 800]$	5
800 to 900	$(800, 900]$	5
900 to 1000	$(900, 1000]$	0
1000 to 1100	$(1000, 1100]$	1

Notice that the classes cover all numbers from 500 to 1100, and each has the same length. To ensure that classes are mutually exclusive, we need to clarify where $600, 700, 800, 900$, and 1000 belong. While getting such a value is unlikely, paying attention to the details is essential. We set the lower bounds to be exclusive and the upper bounds to be inclusive. For example, an observation of precisely 900 milliseconds would not be assigned to $900 - 1000$ but rather be assigned to $800 - 900$. Nothing is objective about this choice; we could have decided to do the reverse. What's important is that it is consistent across all classes and that the classes remain mutually exclusive.

Note: Interval Notation

We use interval notation as a way to describe a continuous set of numbers and how we include (or not include) the endpoints.

The use of (a, b) implies that we are including all possible numbers between a and b, but we would not include either number a or b. We can choose numbers very close to a and b but never equal to a and b.

The use of $[a, b]$ implies that we are including all possible numbers between a and b including the endpoints a and b.

We can also use both in a single interval; $(1000, 1100]$says we are taking all of the numbers from 1000 up to 1100 including the number 1100 but not 1000.

As with our previous distributions, grouped frequency distributions can be portrayed graphically. Figure $2.1.3$ shows a graphical representation of the grouped frequency distribution in Table $2.1.3$. Notice there are no longer gaps between all of the bars. We do this to emphasize that this is a grouped distribution of a continuous quantitative variable. The graph of a frequency or relative frequency distribution of a continuous quantitative variable is called a **histogram** (note that some statisticians extend this name to the graphs of distributions of quantitative variables in general, others just to grouped distributions).

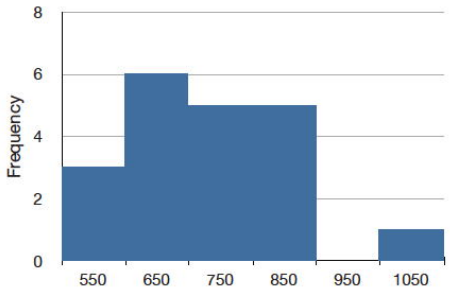

Figure $2.1.3$: A histogram of the grouped frequency distribution shown in Table $2.1.3$. The labels on the horizontal axis are the middle values of the range they represent.

Shapes of Distributions

The order that colors were presented in our frequency distribution for the M&M's did not matter, because color is a nominal variable. When we examine variables on the ordinal, interval, or ratio scales, we construct the distributions following the natural order. If we have a quantitative variable (on interval or ratio scale), the meaningful arithmetic differences in values allow us to describe the distribution by its general shape through a graph. We must be careful when describing the distribution (graph) of grouped data because the shape depends on how the classes were defined. We will develop a greater ability to describe distributions; for now, we will focus on three descriptions of bar graphs and histograms: symmetric, positively skewed, and negatively skewed.

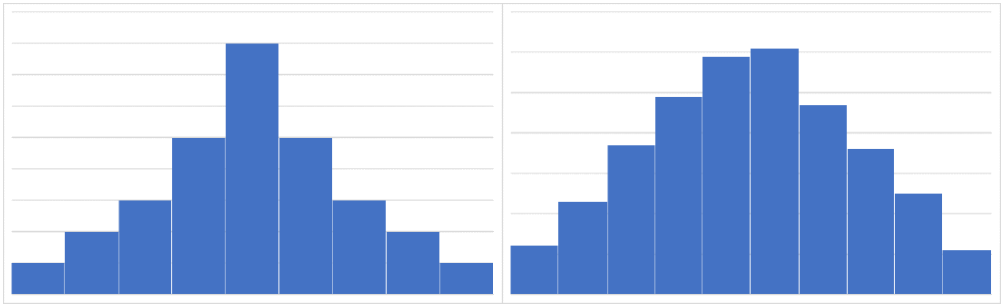

Figure $2.1.4$: Histograms of two hypothetical distributions with varying degrees of symmetry

The graphs of the distributions shown in Figure $2.1.4$, are **symmetric**; if we folded each graph in half, the two sides would match. The histogram on the left is perfectly symmetric; perfect symmetry is not likely to occur using experimental data. The histogram on the right is not perfectly symmetric (see how the four central bars would not match across the middle), but this is closer to what we would expect from observational data coming from a symmetric variable.

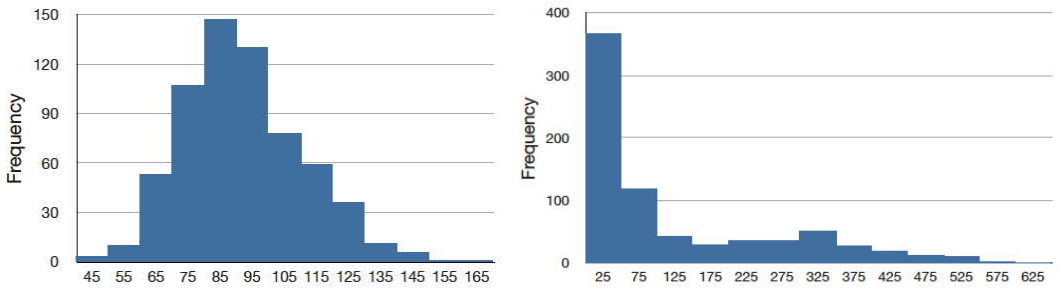

Figure 2.1.5: Two histograms (scores on a psychology test and 1974 MLB salaries (in thousands of dollars)) with varying degrees of positive skew.

Figure 2.1.5 shows two histograms that are not symmetric. Notice the ends of the graphs (called tails) in the positive direction extend further than the tails in the negative direction. A graph of a quantitative variable (bar graph or histogram) with the longer tail extending in the positive direction is said to be **positively skewed** or skewed to the right. A graph of a distribution can be **negatively skewed** or skewed to the left. These graphs have the tails in the negative direction extending further than the tails in the positive direction.
