4.2: Graphing Quantitative Variables

As discussed in the section on variables in Chapter 1, quantitative variables are variables measured on a numeric scale. Height, weight, response time, subjective rating of pain, temperature, and score on an exam are all examples of quantitative variables. Quantitative variables are distinguished from categorical (sometimes called qualitative) variables such as favorite color, religion, city of birth, favorite sport in which there is no ordering or measuring involved.

There are many types of graphs that can be used to portray distributions of quantitative variables. The upcoming sections cover the following types of graphs:

- histograms
- frequency polygons
- bar charts
- line graphs
- dot plots
- scatter plots (discussed in a different chapter)

Some graph types are best-suited for small to moderate amounts of data, whereas others such as histograms are best suited for large amounts of data. Graph types such as box plots are good at depicting differences between distributions. Scatter plots are used to show the relationship between two variables.

Histograms

A histogram is a graphical method for displaying the shape of a distribution. It is particularly useful when there are a large number of observations. We begin with an example consisting of the scores of 642 students on a psychology test. The test consists of 197 items each graded as “correct” or “incorrect.” The students' scores ranged from 46 to 167.

The first step is to create a frequency table. Unfortunately, a simple frequency table would be too big, containing over 100 rows. To simplify the table, we group scores together as shown in Table 4.2.1.

Table 4.2.1: Grouped Frequency Distribution of Psychology Test Scores

Interval's Lower Limit	Interval's Upper Limit	Class Frequency
39.5	49.5	3
49.5	59.5	10
59.5	69.5	53
69.5	79.5	107
79.5	89.5	147
89.5	99.5	130
99.5	109.5	78
109.5	119.5	59
119.5	129.5	36
129.5	139.5	11
139.5	149.5	6
149.5	159.5	1
159.5	169.5	1

To create this table, the range of scores was broken into intervals, called class intervals. The first interval is from 39.5 to 49.5, the second from 49.5 to 59.5, etc. Next, the number of scores falling into each interval was counted to obtain the class frequencies. There are three scores in the first interval, 10 in the second, etc.

Class intervals of width 10 provide enough detail about the distribution to be revealing without making the graph too "choppy." More information on choosing the widths of class intervals is presented later in this section. Placing the limits of the class intervals midway between two numbers (e.g., 49.5) ensures that every score will fall in an interval rather than on the boundary between intervals.

In a histogram, the class frequencies are represented by bars. The height of each bar corresponds to its class frequency. A histogram of these data is shown in Figure 4.2.8.

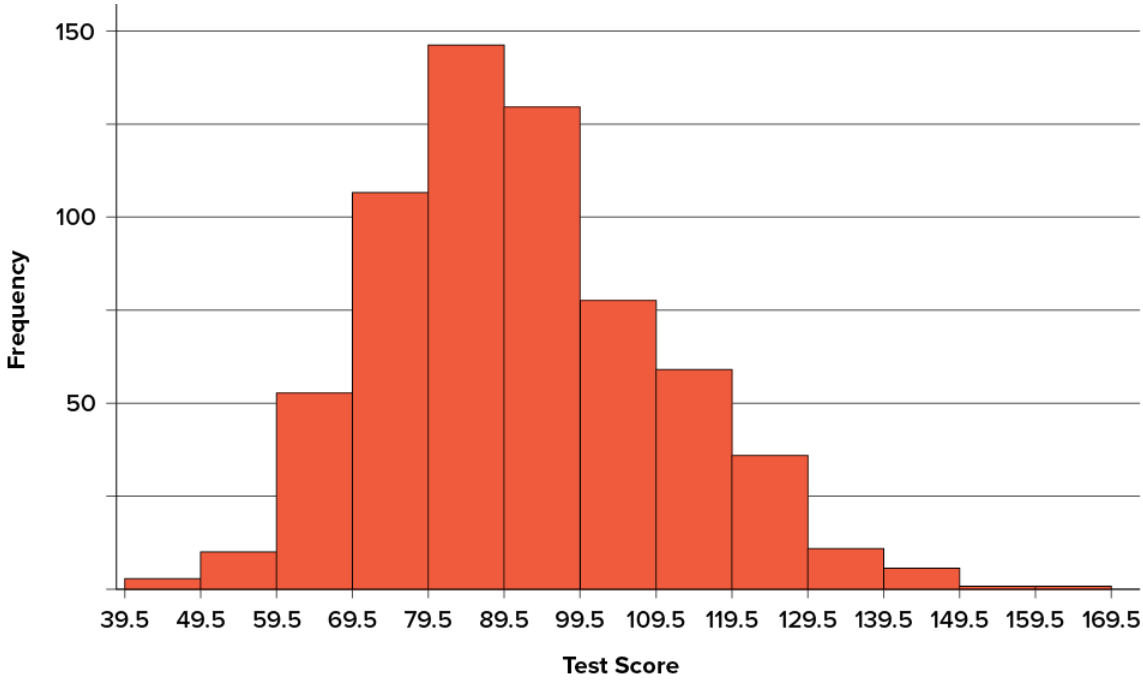


Figure 4.2.8: Histogram of scores on a psychology test.

Image Credit: Judy Schmitt, from Cote et al, 2021.

The histogram makes it plain that most of the scores are in the middle of the distribution, with fewer scores in the extremes. You can also see that the distribution is not symmetric: the scores extend to the right farther than they do to the left. The distribution is therefore said to be skewed. (We'll have more to say about shapes of distributions in Chapter 3.)

In our example, the observations are whole numbers. Histograms can also be used when the scores are measured on a more continuous scale such as the length of time (in milliseconds) required to perform a task. In this case, there is no need to worry about fence sitters since they are improbable. (It would be quite a coincidence for a task to require exactly 7 seconds, measured to the nearest thousandth of a second.) We are therefore free to choose whole numbers as boundaries for our class intervals, for example, 4000, 5000, etc. The class frequency is then the number of observations that are greater than or equal to the lower bound, and strictly less than the upper bound. For example, one interval might hold times from 4000 to 4999 milliseconds. Using whole numbers as boundaries avoids a cluttered appearance, and is the practice of many computer programs that create histograms. Note also that some computer programs label the middle of each interval rather than the end points.

Histograms can be based on relative frequencies instead of actual frequencies. Histograms based on relative frequencies show the proportion of scores in each interval rather than the number of scores. In this case, the Y-axis runs from 0 to 1 (or somewhere in between if there are no extreme proportions). You can change a histogram based on frequencies to one based on relative frequencies by (a) dividing each class frequency by the total number of observations, and then (b) plotting the quotients on the Y-axis (labeled as proportion).

There is more to be said about the widths of the class intervals, sometimes called bin widths. Your choice of bin width determines the number of class intervals. This decision, along with the choice of starting point for the first interval, affects the shape of the

histogram. The best advice is to experiment with different choices of width, and to choose a histogram according to how well it communicates the shape of the distribution.

Frequency Polygons

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions.

To create a frequency polygon, start just as for histograms, by choosing a class interval. Then draw an X-axis representing the values of the scores in your data. Mark the middle of each class interval with a tick mark, and label it with the middle value represented by the class. Draw the Y-axis to indicate the frequency of each class. Place a point in the middle of each class interval at the height corresponding to its frequency. Finally, connect the points. You should include one class interval below the lowest value in your data and one above the highest value. The graph will then touch the X-axis on both sides.

A frequency polygon for 642 psychology test scores shown in Figure 4.2.8 was constructed from the frequency table shown in Table 4.2.2.

Table 4.2.2: Frequency Distribution of Psychology Test Scores

Lower Limit	Upper Limit	Count	Cumulative Count
29.5	39.5	0	0
39.5	49.5	3	3
49.5	59.5	10	13
59.5	69.5	53	66
69.5	79.5	107	173
79.5	89.5	147	320
89.5	99.5	130	450
99.5	109.5	78	528
109.5	119.5	59	587
119.5	129.5	36	623
129.5	139.5	11	634
139.5	149.5	6	640
149.5	159.5	1	641
159.5	169.5	1	642
169.5	170.5	0	642

The first label on the X-axis is 35. This represents an interval extending from 29.5 to 39.5. Since the lowest test score is 46, this interval has a frequency of 0. The point labeled 45 represents the interval from 39.5 to 49.5. There are three scores in this interval. There are 147 scores in the interval that surrounds 85.

You can easily discern the shape of the distribution from Figure 4.2.9. Most of the scores are between 65 and 115. It is clear that the distribution is not symmetric inasmuch as good scores (to the right) trail off more gradually than poor scores (to the left). In the terminology of Chapter 3 (where we will study shapes of distributions more systematically), the distribution is skewed.

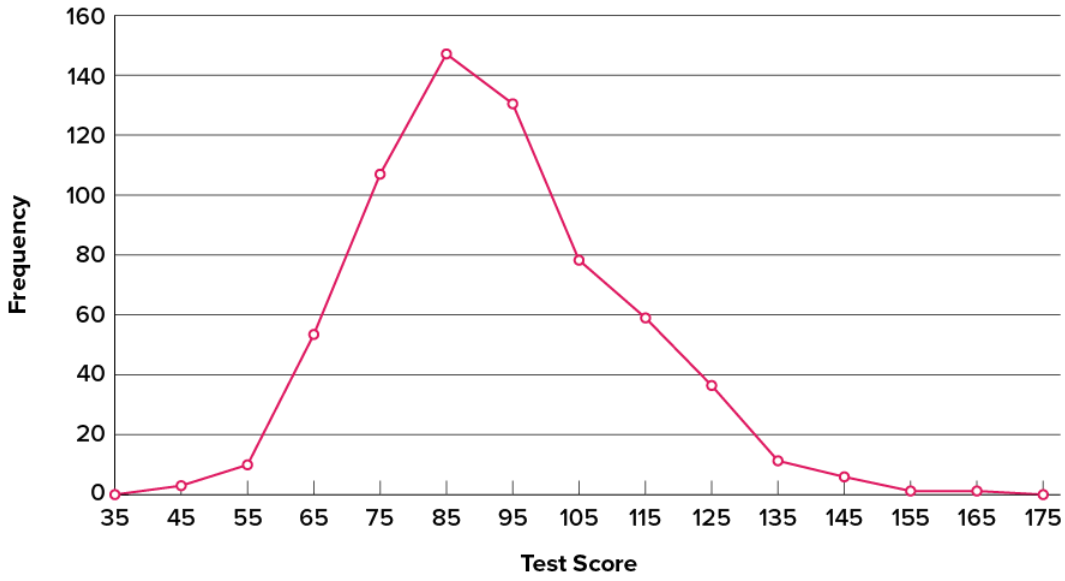


Figure 4.2.9: Frequency polygon for the psychology test scores.

Image Credit: Judy Schmitt, from Cote et al, 2021.

A cumulative frequency polygon for the same test scores is shown in Figure 4.2.10. The graph is the same as before except that the Y value for each point is the number of students in the corresponding class interval plus all numbers in lower intervals. For example, there are no scores in the interval labeled "35," three in the interval "45," and 10 in the interval "55." Therefore, the Y value corresponding to "55" is 13. Since 642 students took the test, the cumulative frequency for the last interval is 642.

Figure 4.2.10: Cumulative frequency polygon for the psychology test scores.

Frequency polygons are useful for comparing distributions. This is achieved by overlaying the frequency polygons drawn for different data sets. Figure 2.1.3 provides an example. The data come from a task in which the goal is to move a computer cursor to a target on the screen as fast as possible. On 20 of the trials, the target was a small rectangle; on the other 20, the target was a large rectangle. Time to reach the target was recorded on each trial. The two distributions (one for each target) are plotted together in Figure 4.2.11. The figure shows that, although there is some overlap in times, it generally took longer to move the cursor to the small target than to the large one.

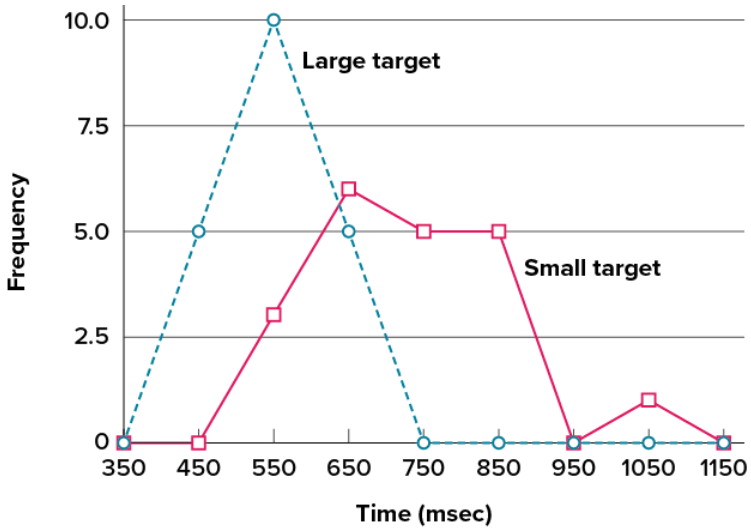


Figure 4.2.11: Overlaid frequency polygons.

It is also possible to plot two cumulative frequency distributions in the same graph. This is illustrated in Figure 4.2.12 using the same data from the cursor task. The difference in distributions for the two targets is again evident.

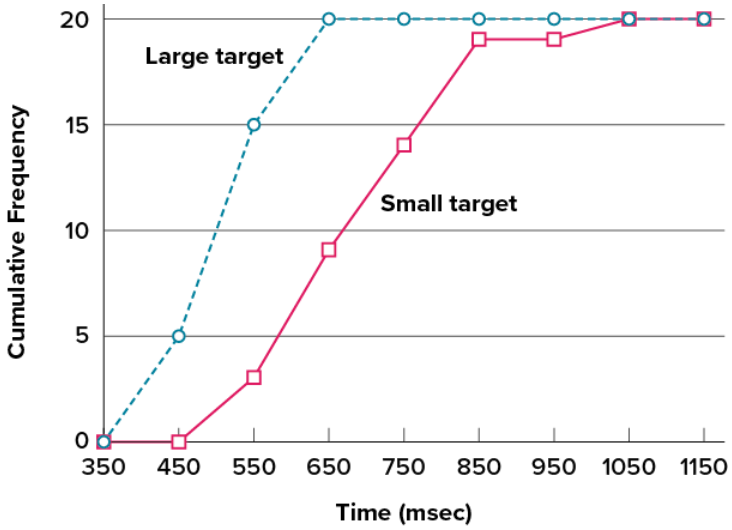


Figure 4.2.12: Overlaid cumulative frequency polygons.

Bar Charts

In the section on qualitative variables, we saw how bar charts could be used to illustrate the frequencies of different categories. For example, the bar chart shown in Figure 4.2.19 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

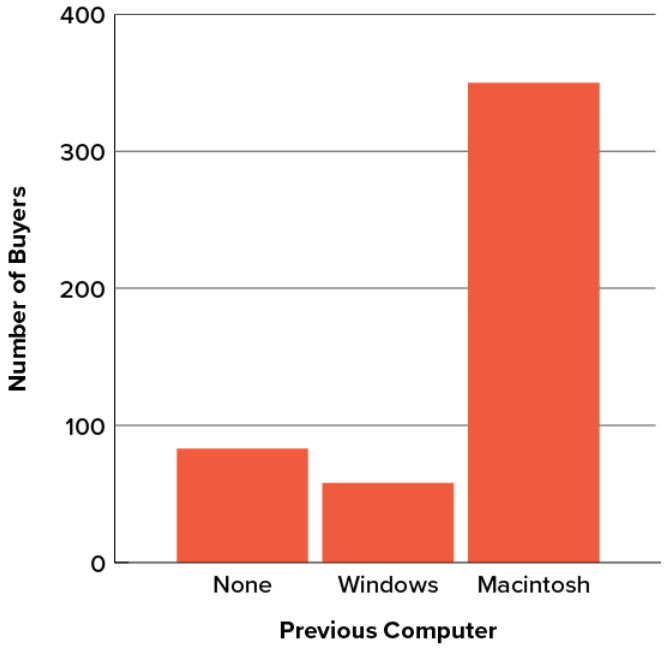


Figure 4.2.19: iMac buyers as a function of previous computer ownership.

Image Credit: Judy Schmitt, from Cote et al, 2021.

In this section we show how bar charts can be used to present other kinds of quantitative information, not just frequency counts. The bar chart in Figure 4.2.20 shows the percent increases in the Dow Jones, Standard and Poor 500 (S & P), and Nasdaq stock indexes from May 24th 2000 to May 24th 2001. Notice that both the S & P and the Nasdaq had "negative increases" which means that they decreased in value. In this bar chart, the Y-axis is not frequency but rather the signed quantity percentage increase.

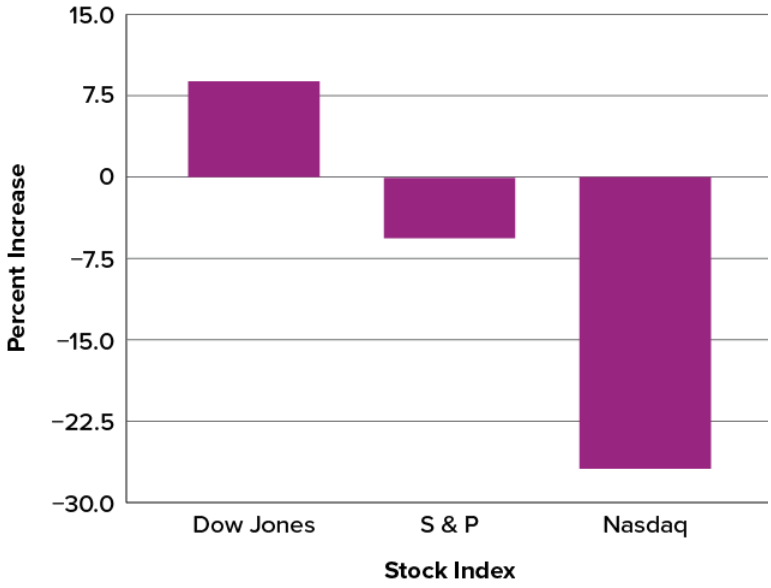


Figure 4.2.20: Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Bar charts are particularly effective for showing change over time. Figure 4.2.21, for example, shows the percent increase in the Consumer Price Index (CPI) over four three-month periods. The fluctuation in inflation is apparent in the graph.

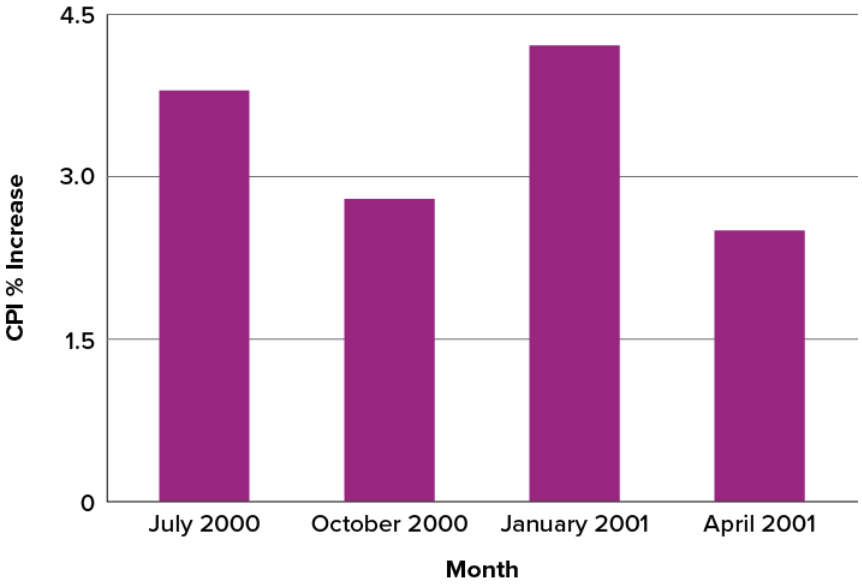


Figure 4.2.21: Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Bar charts are often used to compare the means of different experimental conditions. Figure 2.1.4 shows the mean time it took one of us (DL) to move the cursor to either a small target or a large target. On average, more time was required for small targets than for large ones.

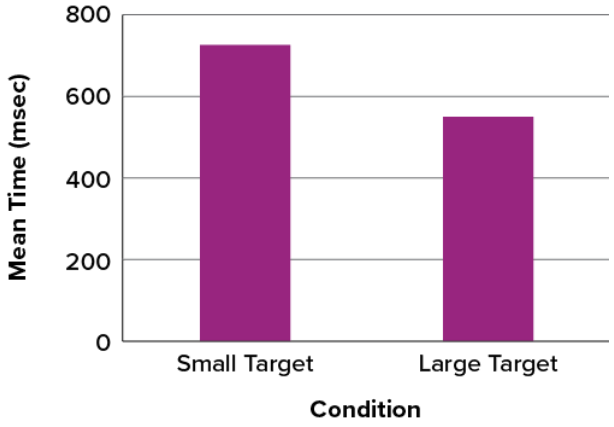


Figure 4.2.22: Bar chart showing the means for the two conditions.

Image Credit: Judy Schmitt, from Cote et al, 2021.

The section on qualitative variables presented earlier in this chapter discussed the use of bar charts for comparing distributions. Some common graphical mistakes were also noted. The earlier discussion applies equally well to the use of bar charts to display quantitative variables.

Line Graphs

A line graph is a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). For example, Figure 4.2.24was presented in the section on bar charts and shows changes in the Consumer Price Index (CPI) over time.

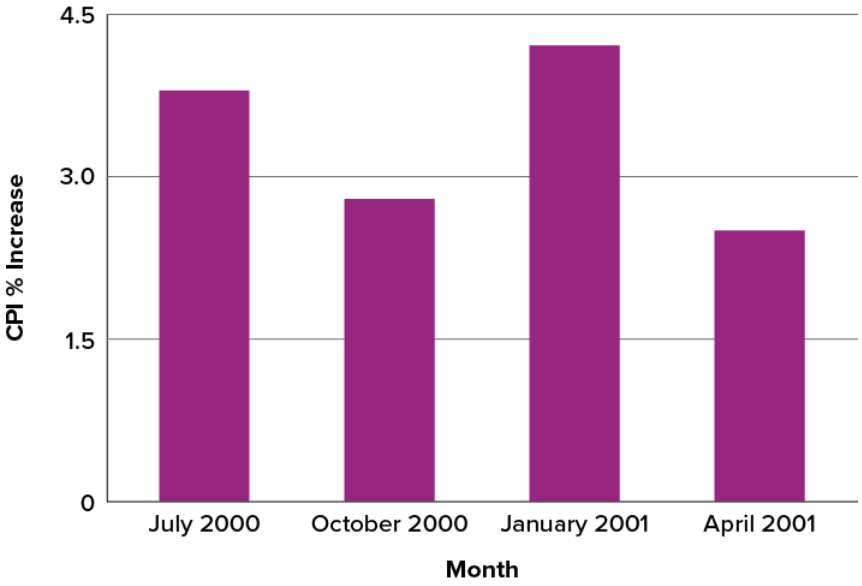


Figure 4.2.24: A bar chart of the percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

A line graph of these same data is shown in Figure 4.2.25. Although the figures are similar, the line graph emphasizes the change from period to period.

Figure 4.2.25: A line graph of the percent change in the CPI over time. Each point represents percent increase for the three months ending at the date indicated.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Line graphs are appropriate only when both the X- and Y-axes display ordered (rather than qualitative) variables. Although bar charts can also be used in this situation, line graphs are generally better at comparing changes over time. Figure 4.2.26, for example, shows percent increases and decreases in five components of the CPI. The figure makes it easy to see that medical costs had a steadier progression than the other components. Although you could create an analogous bar chart, its interpretation would not be as easy.

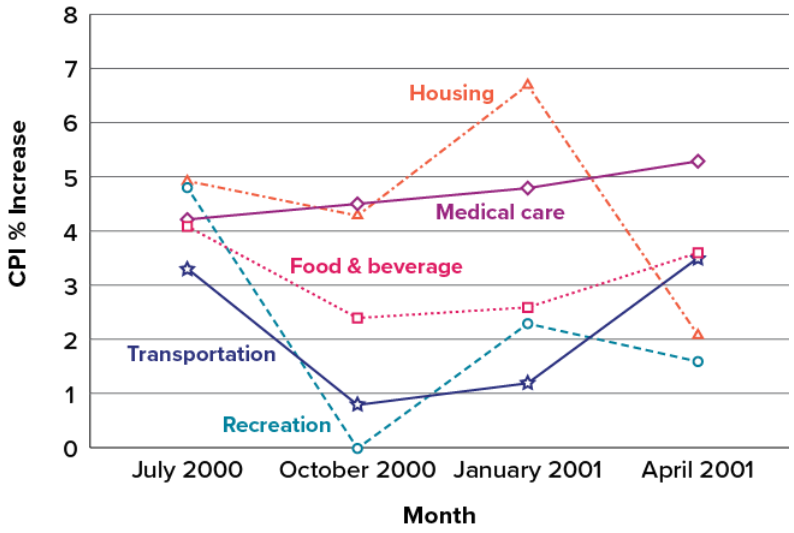


Figure 4.2.26: A line graph of the percent change in five components of the CPI over time.

Image Credit: Judy Schmitt, from Cote et al, 2021.

Let us stress that it is **misleading** to use a line graph when the X-axis contains merely qualitative variables.

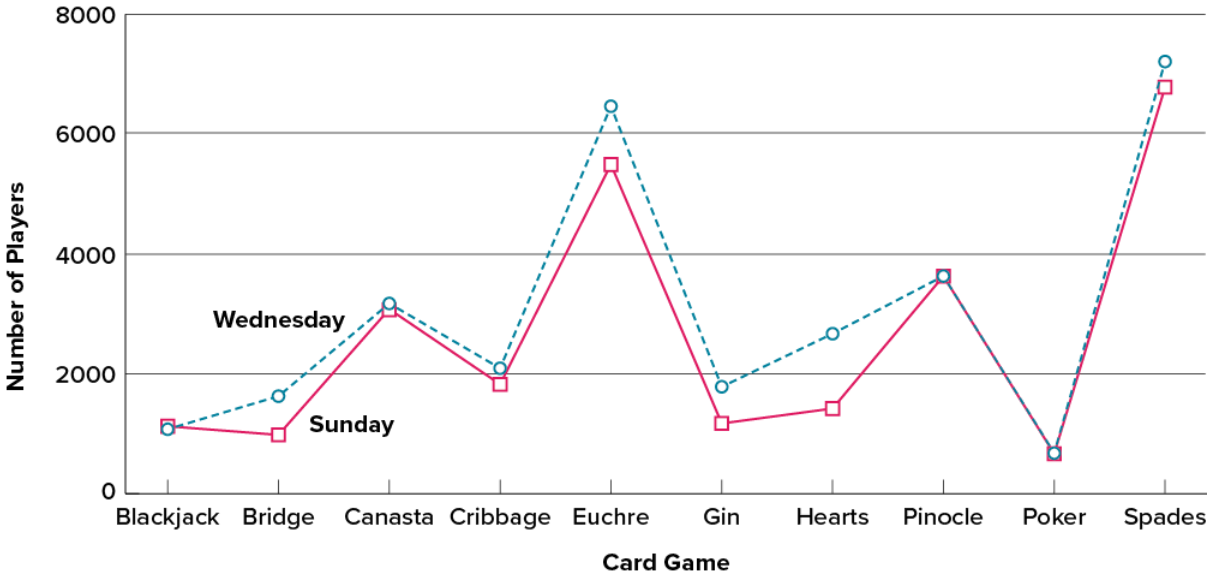


Image Credit: Judy Schmitt, from Cote et al, 2021.

The Shape of Distribution

Finally, it is useful to present discussion on how we describe the shapes of distributions, which we will revisit in the next chapter to learn how different shapes affect our numerical descriptors of data and distributions.

The primary characteristic we are concerned about when assessing the shape of a distribution is whether the distribution is symmetrical or skewed. A symmetrical distribution, as the name suggests, can be cut down the center to form 2 mirror images. Although in practice we will never get a perfectly symmetrical distribution, we would like our data to be as close to symmetrical as possible for reasons we delve into in Chapter 3. Many types of distributions are symmetrical, but by far the most common and pertinent distribution at this point is the normal distribution, shown in Figure 4.2.28. Notice that although the symmetry is not perfect (for instance, the bar just to the right of the center is taller than the one just to the left), the two sides are roughly the same shape. The normal distribution has a single peak, known as the center, and two tails that extend out equally, forming what is known as a bell shape or bell curve.

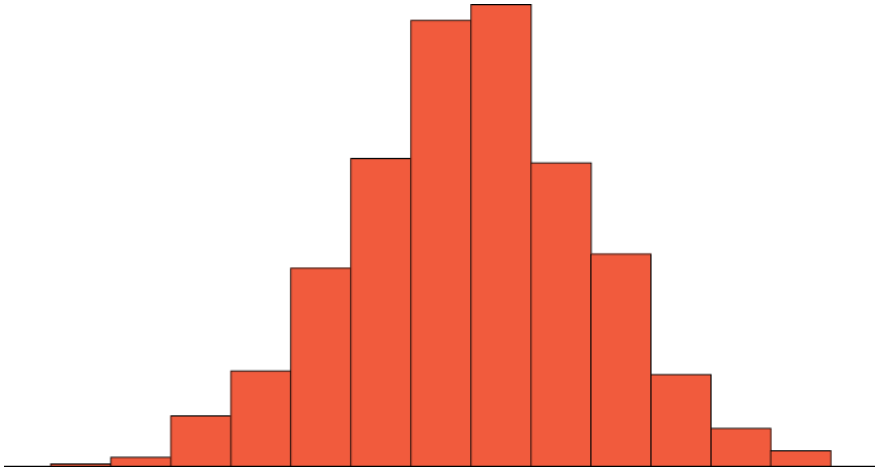


Figure 4.2.28: A symmetrical distribution

Image Credit: Judy Schmitt, from Cote et al, 2021.

Symmetrical distributions can also have multiple peaks. Figure 4.2.29 shows a bimodal distribution, named for the two peaks that lie roughly symmetrically on either side of the center point. As we will see in the next chapter, this is not a particularly desirable characteristic of our data, and, worse, this is a relatively difficult characteristic to detect numerically. Thus, it is important to visualize your data before moving ahead with any formal analyses.

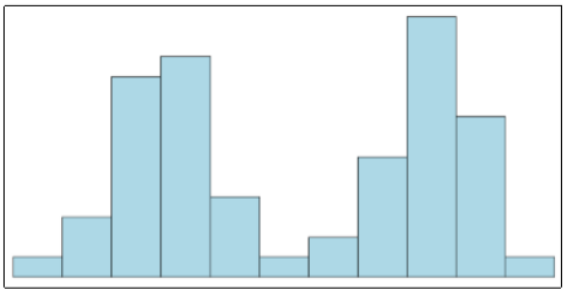
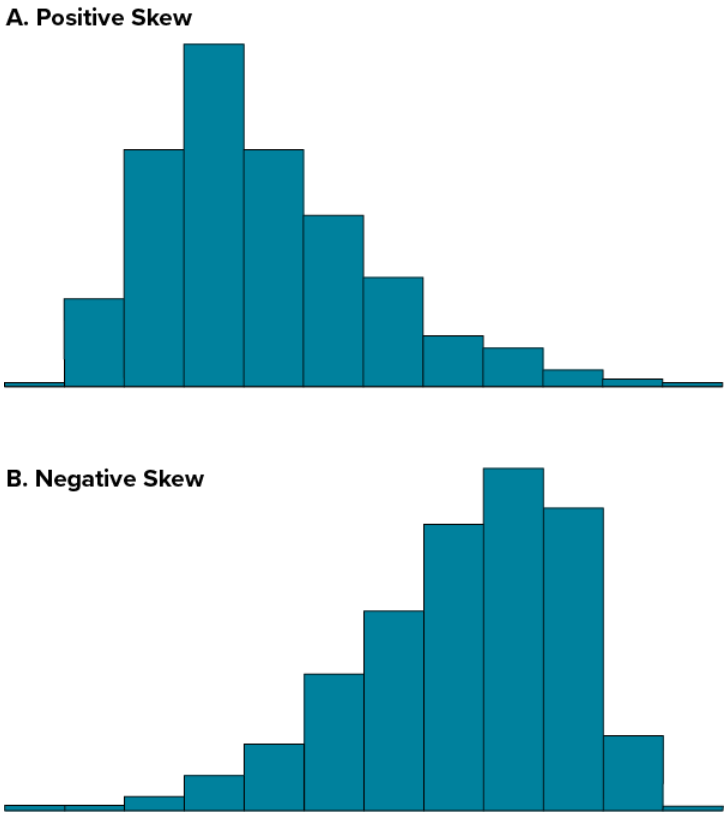


Figure 4.2.29: A bimodal distribution.

Distributions that are not symmetrical also come in many forms, more than can be described here. The most common asymmetry to be encountered is referred to as skew, in which one of the two tails of the distribution is disproportionately longer than the other. This property can affect the value of the averages we use in our analyses and make them an inaccurate representation of our data, which causes many problems.

Skew can either be positive or negative (also known as right or left, respectively), based on which tail is longer. It is very easy to get the two confused at first; many students want to describe the skew by where the bulk of the data (larger portion of the histogram, known as the body) is placed, but the correct determination is based on which tail is longer. You can think of the tail as an arrow: whichever direction the arrow is pointing is the direction of the skew. Figures 4.2.30 and 4.2.31 show positive (right) and negative (left) skew, respectively.



Figures 4.2.30 and 4.2.31

Image Credit: Judy Schmitt, from Cote et al, 2021.
