1. Introduction

This chapter begins by discussing what statistics are and why the study of statistics is important. Subsequent sections cover a variety of topics all basic to the study of statistics. The only theme common to all of these sections is that they cover concepts and ideas important for other chapters in the book.

Variables

by Heidi Ziemer

Prerequisites

- none

Learning Objectives

1. Define and distinguish between independent and dependent variables
2. Define and distinguish between discrete and continuous variables
3. Define and distinguish between qualitative and quantitative variables

Independent and dependent variables

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts (as opposed to constants such as π that do not vary). When conducting research, experimenters often manipulate variables. For example, an experimenter might compare the effectiveness of four types of antidepressants. In this case, the variable is "type of antidepressant." When a variable is manipulated by an experimenter, it is called an independent variable. The experiment seeks to determine the effect of the independent variable on relief from depression. In this example, relief from depression is called a dependent variable. In general, the independent variable is manipulated by the experimenter and its effects on the dependent variable are measured.

Example #1: Can blueberries slow down aging? A study indicates that antioxidants found in blueberries may slow down the process of aging. In this study, 19-month-old rats (equivalent to 60-year-old humans) were fed either their standard diet or a diet supplemented by either blueberry, strawberry, or spinach powder. After eight weeks, the rats were given memory and motor skills tests. Although all supplemented rats showed improvement, those supplemented with blueberry powder showed the most notable improvement.

1. What is the independent variable? (dietary supplement: none, blueberry, strawberry, and spinach)

2. What are the dependent variables? (memory test and motor skills test)

Example #2: Does beta-carotene protect against cancer? Beta-carotene supplements have been thought to protect against cancer. However, a study published in the Journal of the National Cancer Institute suggests this is false. The study was conducted with 39,000 women aged 45 and up. These women were randomly assigned to receive a beta-carotene supplement or a placebo, and their health was studied over their lifetime. Cancer rates for women taking the beta-carotene supplement did not differ systematically from the cancer rates of those women taking the placebo.

1. What is the independent variable? (supplements: beta-carotene or placebo)

2. What is the dependent variable? (occurrence of cancer)

Example #3: How bright is right? An automobile manufacturer wants to know how bright brake lights should be in order to minimize the time required for the driver of a following car to realize that the car in front is stopping and to hit the brakes.

1. What is the independent variable? (brightness of brake lights)

2. What is the dependent variable? (time to hit brakes)

Levels of an Independent Variable

If an experiment compares an experimental treatment with a control treatment, then the independent variable (type of treatment) has two levels: experimental and control. If an experiment were comparing five types of diets, then the independent variable (type of diet) would have 5 levels. In general, the number of levels of an independent variable is the number of experimental conditions.

Qualitative and Quantitative Variables

An important distinction between variables is between qualitative variables and quantitative variables. Qualitative variables are those that express a qualitative attribute such as hair color, eye color, religion, favorite movie, gender, and so on. The values of a qualitative variable do not imply a numerical ordering. Values of the variable “religion” differ qualitatively; no ordering of religions is implied. Qualitative variables are sometimes referred to as categorical variables. Quantitative variables are those variables that are measured in terms of numbers. Some examples of quantitative variables are height, weight, and shoe size.

In the study on the effect of diet discussed previously, the independent variable was type of supplement: none, strawberry, blueberry, and spinach. The variable “type of supplement” is a qualitative variable; there is nothing quantitative about it. In contrast, the dependent variable “memory test” is a quantitative variable since memory performance was measured on a quantitative scale (number correct).

Discrete and Continuous Variables

Variables such as number of children in a household are called discrete variables since the possible scores are discrete points on the scale. For example, a household could have three children or six children, but not 4.53 children. Other variables such as “time to respond to a question” are continuous variables since the scale is continuous and not made up of discrete steps. The response time could be 1.64 seconds, or it could be 1.64237123922121 seconds. Of course, the practicalities of measurement preclude most measured variables from being truly continuous.

Percentiles

by David Lane

Prerequisites
- none

Learning Objectives
1. Define percentiles
2. Use three formulas for computing percentiles

A test score in and of itself is usually difficult to interpret. For example, if you learned that your score on a measure of shyness was 35 out of a possible 50, you would have little idea how shy you are compared to other people. More relevant is the percentage of people with lower shyness scores than yours. This percentage is called a percentile. If 65% of the scores were below yours, then your score would be the 65th percentile.

Two Simple Definitions of Percentile

There is no universally accepted definition of a percentile. Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this "Definition 1." The 65th percentile can also be defined as the smallest score that is greater than or equal to 65% of the scores. This we will call "Definition 2." Unfortunately, these two definitions can lead to dramatically different results, especially when there is relatively little data. Moreover, neither of these definitions is explicit about how to handle rounding. For instance, what rank is required to be higher than 65% of the scores when the total number of scores is 50? This is tricky because 65% of 50 is 32.5. How do we find the lowest number that is higher than 32.5% of the scores? A third way to compute percentiles (presented below) is a weighted average of the percentiles computed according to the first two definitions. This third definition handles rounding more gracefully than the other two and has the advantage that it allows the median to be defined conveniently as the 50th percentile.

A Third Definition

Unless otherwise specified, when we refer to “percentile,” we will be referring to this third definition of percentiles. Let's begin with an example. Consider the 25th percentile for the 8 numbers in Table 1. Notice the numbers are given ranks ranging from 1 for the lowest number to 8 for the highest number.

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

The first step is to compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = \frac{P}{100} \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = \frac{25}{100} \times (8 + 1) = \frac{9}{4} = 2.25$$

If R is an integer, the Pth percentile is be the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

1. Define IR as the integer portion of R (the number to the left of the decimal point). For this example, IR = 2.
2. Define FR as the fractional portion of R. For this example, FR = 0.25.

3. Find the scores with Rank I_R and with Rank $I_R + 1$. For this example, this means the score with Rank 2 and the score with Rank 3. The scores are 5 and 7.
4. Interpolate by multiplying the difference between the scores by F_R and add the result to the lower score. For these data, this is (0.25)(7 - 5) + 5 = 5.5.

Therefore, the 25th percentile is 5.5. If we had used the first definition (the smallest score greater than 25% of the scores), the 25th percentile would have been 7. If we had used the second definition (the smallest score greater than or equal to 25% of the scores), the 25th percentile would have been 5.

For a second example, consider the 20 quiz scores shown in Table 2.

Table 2. 20 Quiz Scores.

Score	Rank
4	1
4	2
5	3
5	4
5	5
5	6
6	7
6	8
6	9
7	10
7	11
7	12
8	13
8	14
9	15
9	16
9	17
10	18
10	19
10	20

We will compute the 25th and the 85th percentiles. For the 25th,

$$R = \frac{25}{100} \times (20 + 1) = \frac{21}{4} = 5.25$$

IR = 5 and FR = 0.25.

Since the score with a rank of IR (which is 5) and the score with a rank of IR + 1 (which is 6) are both equal to 5, the 25th percentile is 5. In terms of the formula:

25th percentile = (.25) x (5 - 5) + 5 = 5.

For the 85th percentile,

$$R = \frac{85}{100} \times (20 + 1) = 17.85$$

IR = 17 and FR = 0.85

Caution: FR does not generally equal the percentile to be computed as it does here.

The score with a rank of 17 is 9 and the score with a rank of 18 is 10. Therefore, the 85th percentile is:

(0.85)(10 - 9) + 9 = 9.85

Consider the 50th percentile of the numbers 2, 3, 5, 9.

$$R = \frac{50}{100} \times (4 + 1) = 2.5$$

IR = 2 and FR = 0.5.

The score with a rank of IR is 3 and the score with a rank of IR + 1 is 5. Therefore, the 50th percentile is:

(0.5)(5 - 3) + 3 = 4.

Finally, consider the 50th percentile of the numbers 2, 3, 5, 9, 11.

$$R = \frac{50}{100} \times (5 + 1) = 3$$

IR = 3 and FR = 0.

Whenever FR = 0, you simply find the number with rank IR. In this case, the third number is equal to 5, so the 50th percentile is 5. You will also get the right answer if you apply the general formula:

50th percentile = (0.00) (9 - 5) + 5 = 5.

Distributions

by David M. Lane and Heidi Ziemer

Prerequisites

- Chapter 1: Variables

Learning Objectives

1. Define "distribution"
2. Interpret a frequency distribution
3. Distinguish between a frequency distribution and a probability distribution
4. Construct a grouped frequency distribution for a continuous variable
5. Identify the skew of a distribution
6. Identify bimodal, leptokurtic, and platykurtic distributions

Distributions of Discrete Variables

I recently purchased a bag of Plain M&M's. The M&M's were in six different colors. A quick count showed that there were 55 M&M's: 17 brown, 18 red, 7 yellow, 7 green, 2 blue, and 4 orange. These counts are shown below in Table 1.

Table 1. Frequencies in the Bag of M&M's

Color	Frequency
Brown	17
Red	18
Yellow	7
Green	7
Blue	2
Orange	4

This table is called a frequency table and it describes the distribution of M&M color frequencies. Not surprisingly, this kind of distribution is called a frequency distribution. Often a frequency distribution is shown graphically as in Figure 1.

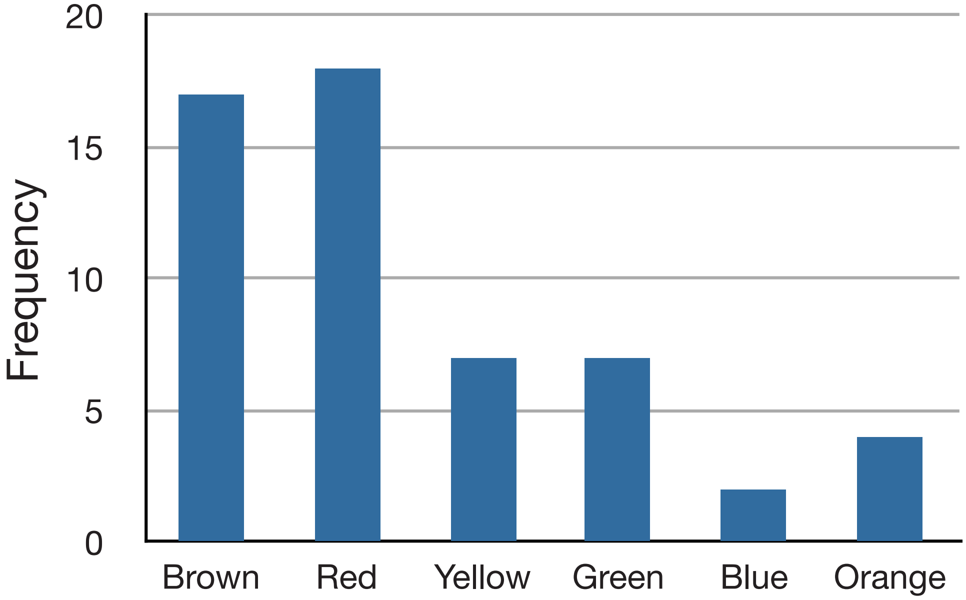


Figure 1. Distribution of 55 M&M's.

The distribution shown in Figure 1 concerns just my one bag of M&M's. You might be wondering about the distribution of colors for all M&M's. The manufacturer of M&M's provides some information about this matter, but they do not tell us exactly how many M&M's of each color they have ever produced. Instead, they report proportions rather than frequencies. Figure 2 shows these proportions. Since every M&M is one of the six familiar colors, the six proportions shown in the figure add to one. We call Figure 2 a probability distribution because if you choose an M&M at random, the probability of getting, say, a brown M&M is equal to the proportion of M&M's that are brown (0.30).

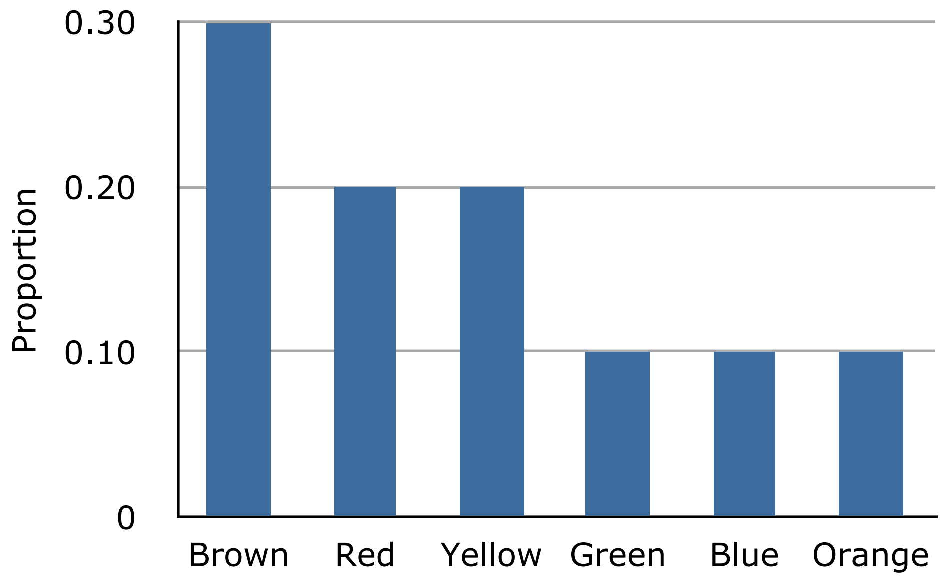


Figure 2. Distribution of all M&M's.

Notice that the distributions in Figures 1 and 2 are not identical. Figure 1 portrays the distribution in a sample of 55 M&M's. Figure 2 shows the proportions for 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; others will be further away.

Continuous Variables

The variable "color of M&M" used in this example is a discrete variable, and its distribution is also called discrete. Let us now extend the concept of a distribution to continuous variables.

The data shown in Table 2 are the times it took one of us (DL) to move the cursor over a small 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 accurately (to many decimal places), no two response times would be expected to be the same. Measuring time in milliseconds (thousandths of a second) is often precise enough to approximate a continuous variable in psychology. As you can see in Table 2, measuring DL's responses this way produced times no two of which were the same. As a result, a frequency distribution would be uninformative: it would consist of the 20 times in the experiment, each with a frequency of 1.

Table 2. Response Times

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

The solution to this problem is to create a grouped frequency distribution. In a grouped frequency distribution, scores falling within various ranges are tabulated. Table 3 shows a grouped frequency distribution for these 20 times.

Table 3. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Grouped frequency distributions can be portrayed graphically. Figure 3 shows a graphical representation of the frequency distribution in Table 3. This kind of graph is called a histogram. Chapter 2 contains an entire section devoted to histograms.

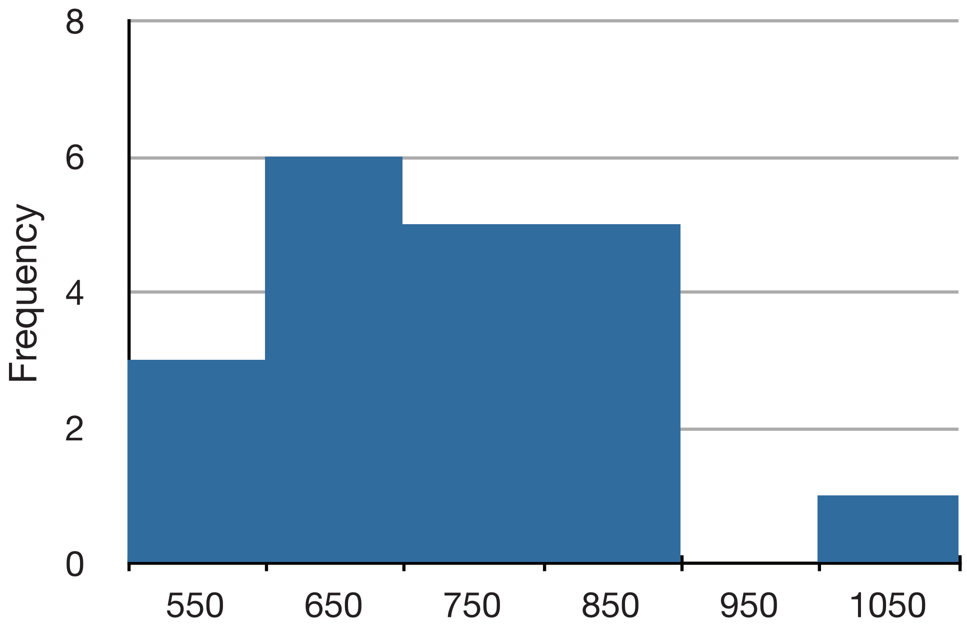


Figure 3. A histogram of the grouped frequency distribution shown in Table 3. The labels on the X-axis are the middle values of the range they represent.

Probability Densities

The histogram in Figure 3 portrays just DL's 20 times in the one experiment he performed. To represent the probability associated with an arbitrary movement (which can take any positive amount of time), we must represent all these potential times at once. For this purpose, we plot the distribution for the continuous variable of time. Distributions for continuous variables are called continuous distributions. They also carry the fancier name probability density. Some probability densities have particular importance in statistics. A very important one is shaped like a bell, and called the normal distribution. Many naturally-occurring phenomena can be approximated surprisingly well by this distribution. It will serve to illustrate some features of all continuous distributions.

An example of a normal distribution is shown in Figure 4. Do you see the “bell”? The normal distribution doesn't represent a real bell, however, since the left and right tips extend indefinitely (we can't draw them any further so they look like they've stopped in our diagram). The Y-axis in the normal distribution represents the “density of probability.” Intuitively, it shows the chance of obtaining values near corresponding points on the X-axis. In Figure 4, for example, the probability of an observation with value near 40 is about half of the probability of an observation with value near 50. (For more information, see Chapter 7.)

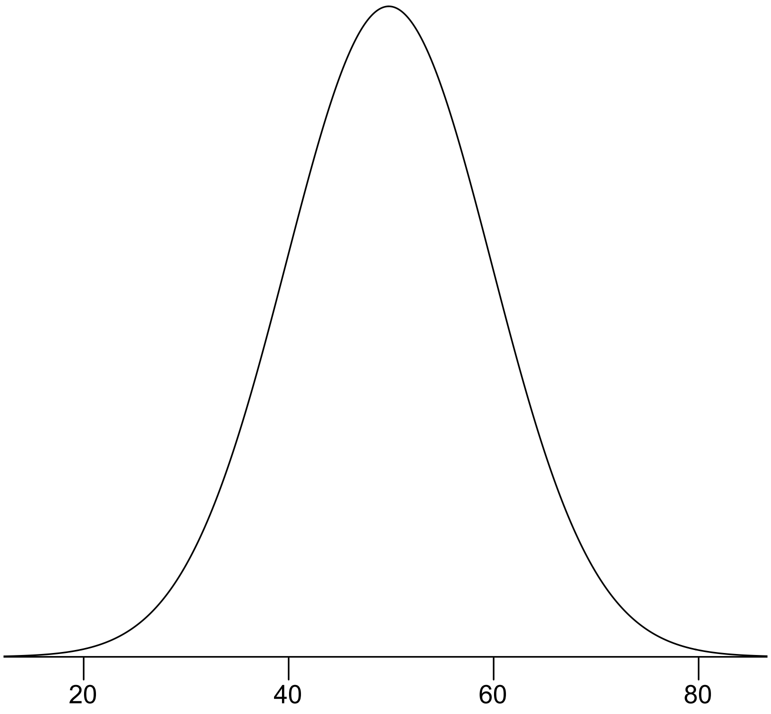


Figure 4. A normal distribution.

Although this text does not discuss the concept of probability density in detail, you should keep the following ideas in mind about the curve that describes a continuous distribution (like the normal distribution). First, the area under the curve equals 1. Second, the probability of any exact value of X is 0. Finally, the area under the curve and bounded between two given points on the X-axis is the probability that a number chosen at random will fall between the two points. Let us illustrate with DL's hand movements. First, the probability that his movement takes some amount of time is one! (We exclude the possibility of him never finishing his gesture.) Second, the probability that his movement takes exactly 598.956432342346576 milliseconds is essentially zero. (We can make the probability as close as we like to zero by making the time measurement more and more precise.) Finally, suppose that the probability of DL's movement taking between 600 and 700 milliseconds is one tenth. Then the continuous distribution for DL's possible times would have a shape that places 10% of the area below the curve in the region bounded by 600 and 700 on the X-axis.

Shapes of Distributions

Distributions have different shapes; they don't all look like the normal distribution in Figure 4. For example, the normal probability density is higher in the middle compared to its two tails. Other distributions need not have this feature. There is even variation among the distributions that we call “normal.” For example, some normal distributions are more spread out than the one shown in Figure 4 (their tails begin to hit the X-axis further from the middle of the curve --for example, at 10 and 90 if drawn in place of Figure 4). Others are less spread out (their tails might approach the X-axis at 30 and 70). More information on the normal distribution can be found in a later chapter completely devoted to them.

The distribution shown in Figure 4 is symmetric; if you folded it in the middle, the two sides would match perfectly. Figure 5 shows the discrete distribution of scores on a psychology test. This distribution is not symmetric: the tail in the positive direction extends further than the tail in the negative direction. A distribution with the longer tail extending in the positive direction is said to have a positive skew. It is also described as “skewed to the right.”

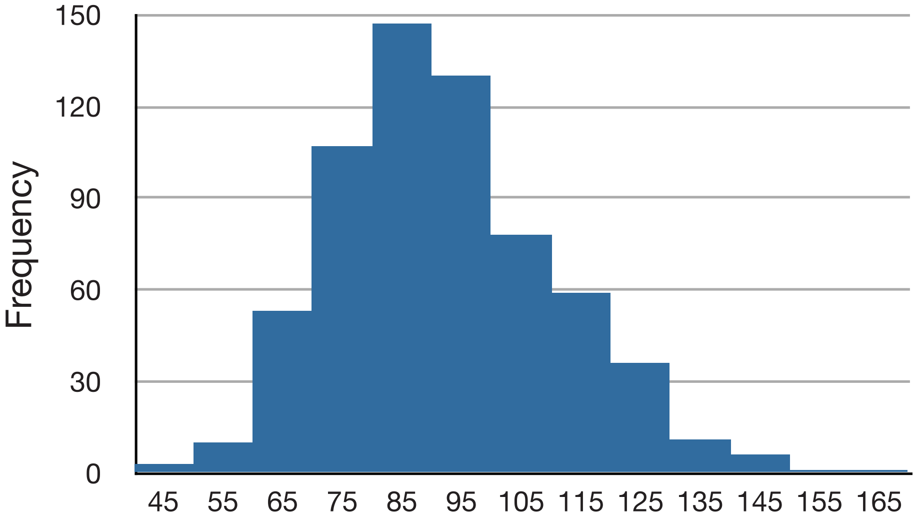


Figure 5. A distribution with a positive skew.

Figure 6 shows the salaries of major league baseball players in 1974 (in thousands of dollars). This distribution has an extreme positive skew.

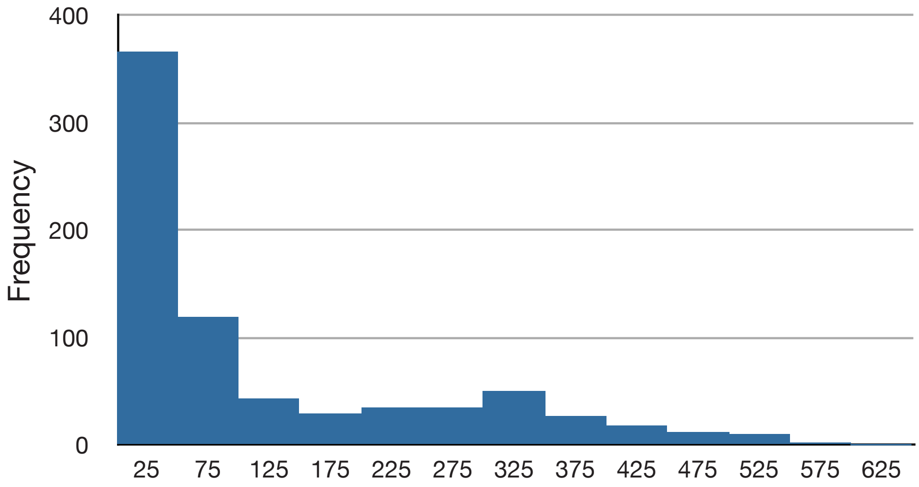


Figure 6. A distribution with a very large positive skew.

A continuous distribution with a positive skew is shown in Figure 7.

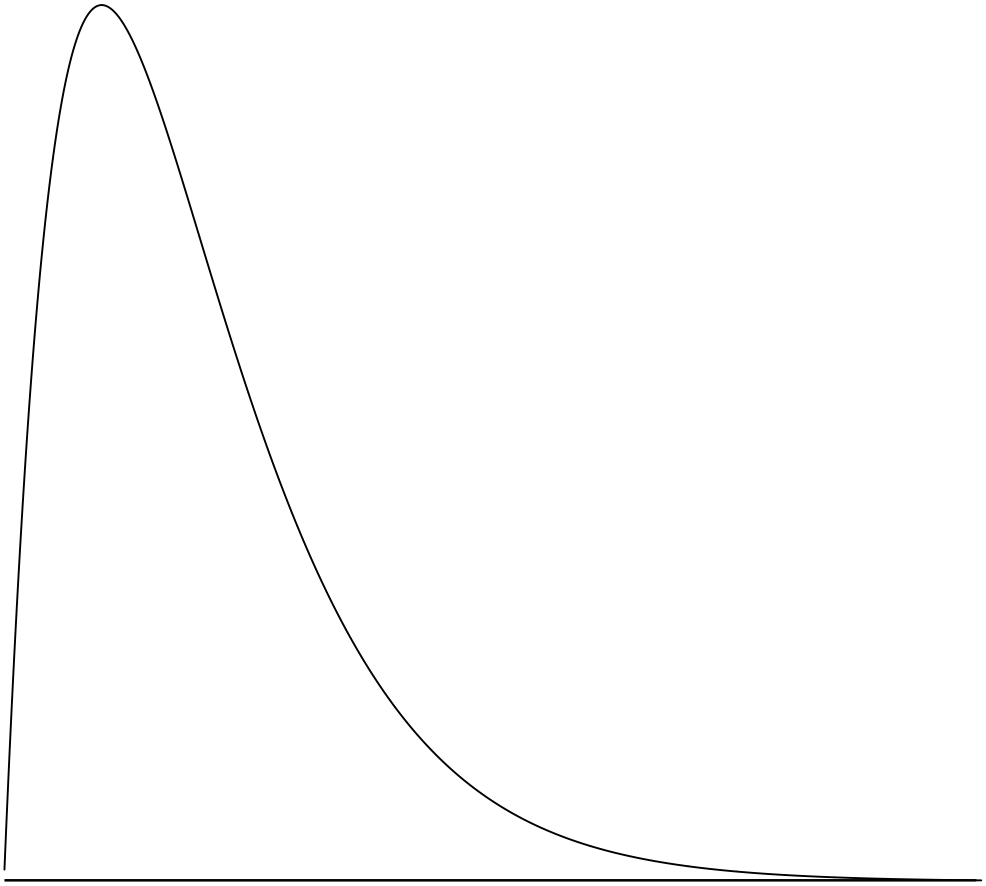


Figure 7. A continuous distribution with a positive skew.

Although less common, some distributions have a negative skew. Figure 8 shows the scores on a 20-point problem on a statistics exam. Since the tail of the distribution extends to the left, this distribution is skewed to the left.

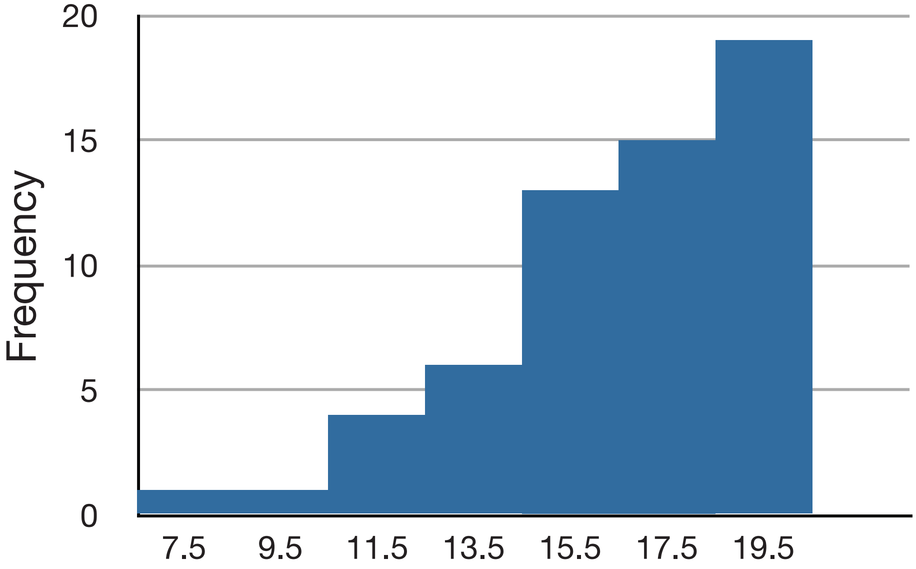


Figure 8. A distribution with negative skew. This histogram shows the frequencies of various scores on a 20-point question on a statistics test.

A continuous distribution with a negative skew is shown in Figure 9.

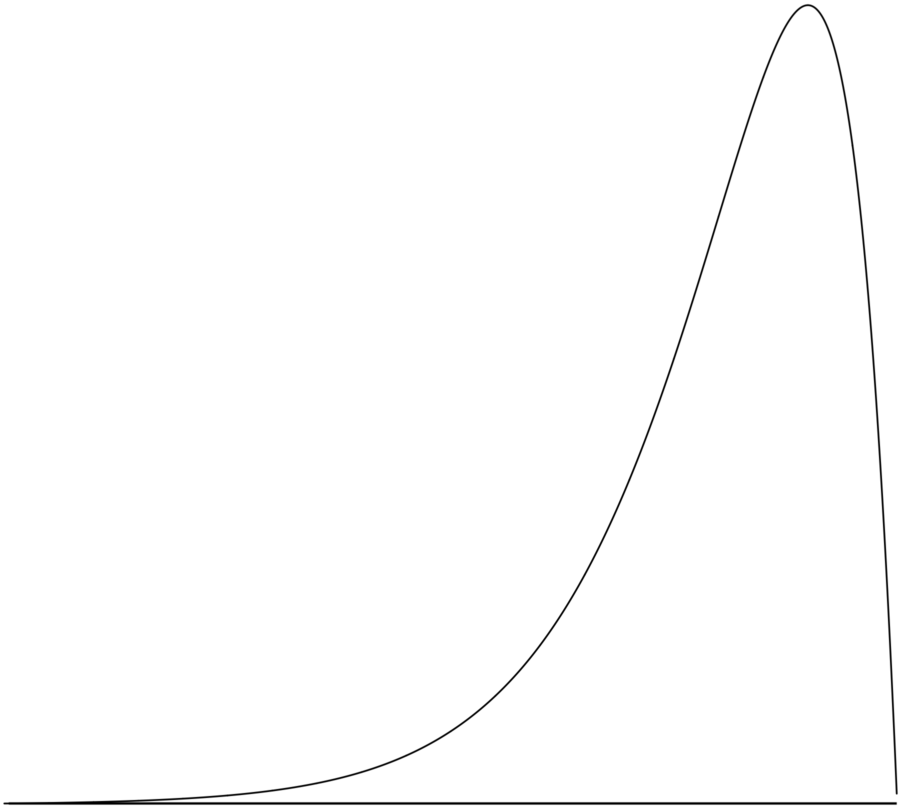


Figure 9. A continuous distribution with a negative skew.

The distributions shown so far all have one distinct high point or peak. The distribution in Figure 10 has two distinct peaks. A distribution with two peaks is called a bimodal distribution.

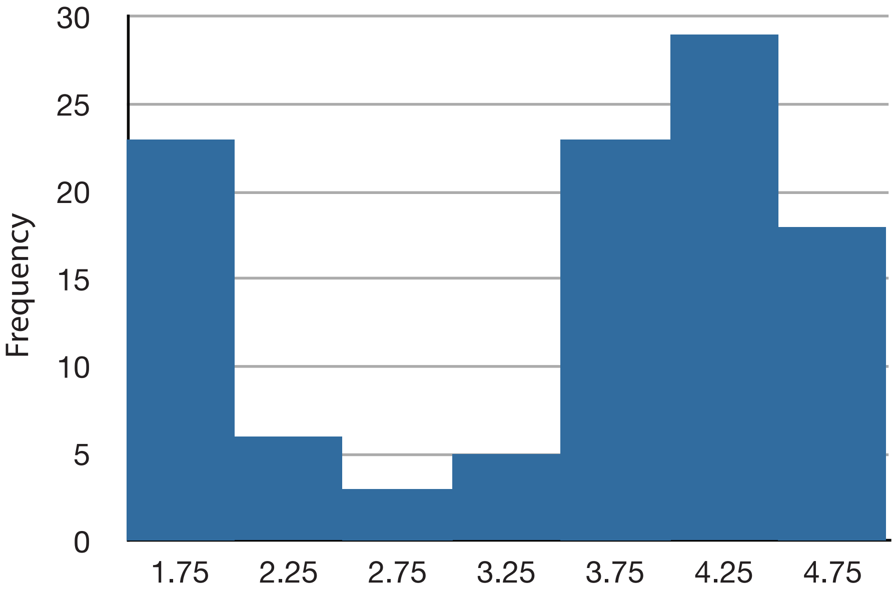
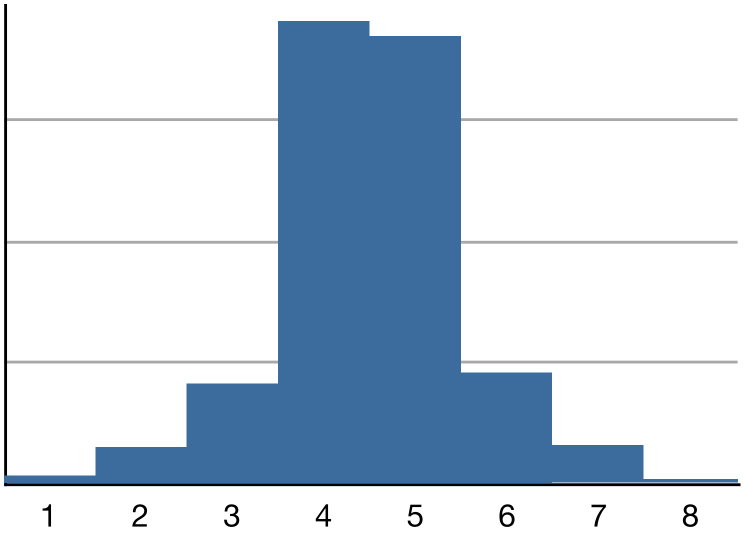


Figure 10. Frequencies of times between eruptions of the Old Faithful geyser. Notice the two distinct peaks: one at 1.75 and the other at 4.25.

Distributions also differ from each other in terms of how large or “fat” their tails are. Figure 11 shows two distributions that differ in this respect. The upper distribution has relatively more scores in its tails; its shape is called leptokurtic. The lower distribution has relatively fewer scores in its tails; its shape is called platykurtic.



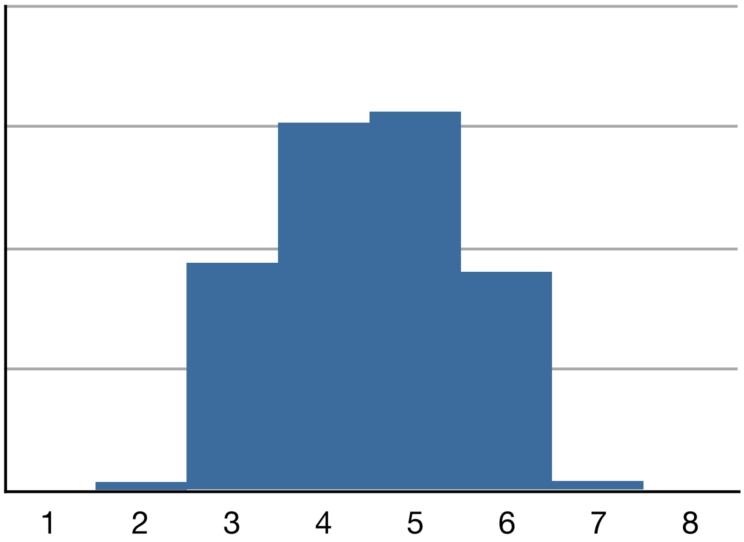


Figure 11. Distributions differing in kurtosis. The top distribution has long tails. It is called "leptokurtic." The bottom distribution has short tails. It is called "platykurtic."

Summation Notation

by David M. Lane

Prerequisites

- None

Learning Objectives

1. Use summation notation to express the sum of all numbers
2. Use summation notation to express the sum of a subset of numbers
3. Use summation notation to express the sum of squares

Many statistical formulas involve summing numbers. Fortunately there is a convenient notation for expressing summation. This section covers the basics of this summation notation.

Let's say we have a variable X that represents the weights (in grams) of 4 grapes. The data are shown in Table 1.

Table 1. Weights of 4 grapes.

Grape	X
1	4.6
2	5.1
3	4.9
4	4.4

We label Grape 1's weight X_1, Grape 2's weight X_2, etc. The following formula means to sum up the weights of the four grapes:

$$\sum_{i=1}^{4} X_i$$

The Greek letter Σ indicates summation. The “i = 1” at the bottom indicates that the summation is to start with X_1 and the 4 at the top indicates that the summation will end with X_4. The “X_i” indicates that X is the variable to be summed as i goes from 1 to 4. Therefore,

$$\sum_{i=1}^{4} X_i = X_1 + X_2 + X_3 + X_4 = 4.6 + 5.1 + 4.9 + 4.4 = 19$$

The symbol

$$\sum_{i=1}^{3} X_i$$

indicates that only the first 3 scores are to be summed. The index variable i goes from 1 to 3.

When all the scores of a variable (such as X) are to be summed, it is often convenient to use the following abbreviated notation:

$$\sum X$$

Thus, when no values of i are shown, it means to sum all the values of X.

Many formulas involve squaring numbers before they are summed. This is indicated as

$$\sum X^2 = 4.6^2 + 5.1^2 + 4.9^2 + 4.4^2$$

$$= 21.16 + 26.01 + 24.01 + 19.36 = 90.54.$$

Notice that:

$$\left(\sum X\right)^2 \neq \sum X^2$$

because the expression on the left means to sum up all the values of X and then square the sum ($19^2 = 361$), whereas the expression on the right means to square the numbers and then sum the squares (90.54, as shown).

Some formulas involve the sum of cross products. Table 2 shows the data for variables X and Y. The cross products (XY) are shown in the third column. The sum of the cross products is 3 + 4 + 21 = 28.

Table 2. Cross Products.

X	Y	XY
1	3	3
2	2	4
3	7	21

In summation notation, this is written as:

$$\sum XY = 28$$

Exercises

Relevant exercises:
3, 10, 11, 12

Prerequisites

- All material presented in Chapter: "Introduction"

1. A teacher wishes to know whether the males in his/her class have more conservative attitudes than the females. A questionnaire is distributed assessing attitudes and the males and the females are compared. Is this an example of descriptive or inferential statistics?

2. A cognitive psychologist is interested in comparing two ways of presenting stimuli on sub- sequent memory. Twelve subjects are presented with each method and a memory test is given. What would be the roles of descriptive and inferential statistics in the analysis of these data?

3. If you are told only that you scored in the 80th percentile, do you know from that description exactly how it was calculated? Explain.

4. A study is conducted to determine whether people learn better with spaced or massed practice. Subjects volunteer from an introductory psychology class. At the beginning of the semester 12 subjects volunteer and are assigned to the massed-practice condition. At the end of the semester 12 subjects volunteer and are assigned to the spaced-practice condition. This experiment involves two kinds of non-random sampling: (1) Subjects are not randomly sampled from some specified population and (2) subjects are not randomly assigned to conditions. Which of the problems relates to the generality of the results? Which of the problems relates to the validity of the results? Which problem is more serious?

5. Give an example of an independent and a dependent variable.

6. Categorize the following variables as being qualitative or quantitative:
   Rating of the quality of a movie on a 7-point scale
   Age
   Country you were born in
   Favorite Color
   Time to respond to a question

7. Specify the level of measurement used for the items in Question 6.

8. Which of the following are linear transformations?
Converting from meters to kilometers
Squaring each side to find the area
Converting from ounces to pounds
Taking the square root of each person's height.
Multiplying all numbers by 2 and then adding 5
Converting temperature from Fahrenheit to Centigrade

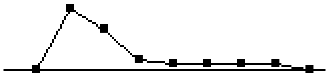
9. The formula for finding each student's test grade (g) from his or her raw score (s) on a test is as follows: g = 16 + 3s

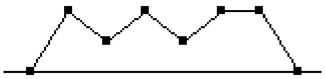
Is this a linear transformation?

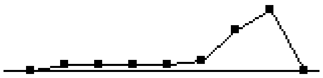
If a student got a raw score of 20, what is his test grade?

10. For the numbers 1, 2, 4, 16, compute the following:
ΣX
ΣX^2
$(\Sigma X)^2$

11. Which of the frequency polygons has a large positive skew? Which has a large negative skew?

A.


B.


C.


12. What is more likely to have a skewed distribution: time to solve an anagram problem (where the letters of a word or phrase are rearranged into another

word or phrase like “dear” and “read” or “funeral” and “real fun”) or scores on a vocabulary test?

Questions from Case Studies

Angry Moods (AM) case study

13. (AM) Which variables are the participant variables? (They act as independent variables in this study.)

14. (AM) What are the dependent variables?

15. (AM) Is Anger-Out a quantitative or qualitative variable?

Teacher Ratings (TR) case study

16. (TR) What is the independent variable in this study?

ADHD Treatment (AT) case study

17. (AT) What is the independent variable of this experiment? How many levels does it have?

18. (AT) What is the dependent variable? On what scale (nominal, ordinal, interval, ratio) was it measured?

2. Graphing Distributions

Graphing data is the first and often most important step in data analysis. In this day of computers, researchers all too often see only the results of complex computer analyses without ever taking a close look at the data themselves. This is all the more unfortunate because computers can create many types of graphs quickly and easily.

This chapter covers some classic types of graphs such bar charts that were invented by William Playfair in the 18th century as well as graphs such as box plots invented by John Tukey in the 20th century.

by David M. Lane

Prerequisites

- Chapter 1: Variables

Learning Objectives

1. Create a frequency table
2. Determine when pie charts are valuable and when they are not
3. Create and interpret bar charts
4. Identify common graphical mistakes

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owner, a previous Windows owner, or a new computer purchaser.

This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous Macintosh users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" of 0.17 = 85/500.

Table 1. Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Pie Charts

The pie chart in Figure 1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

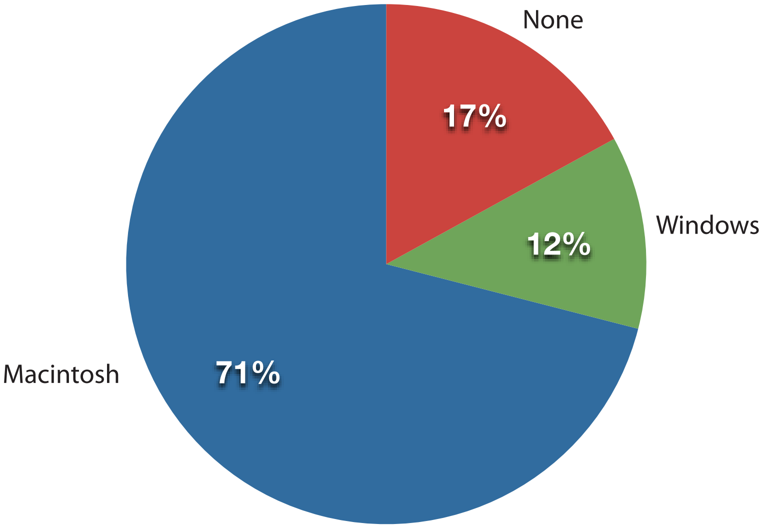


Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.

Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted "The only worse design than a pie chart is several of them."

Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have occurred since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3) instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 2. Frequencies are shown on the Y-axis and the type of computer previously owned is shown on the X-axis. Typically, the Y-axis shows the number of observations in each category rather than the percentage of observations in each category as is typical in pie charts.

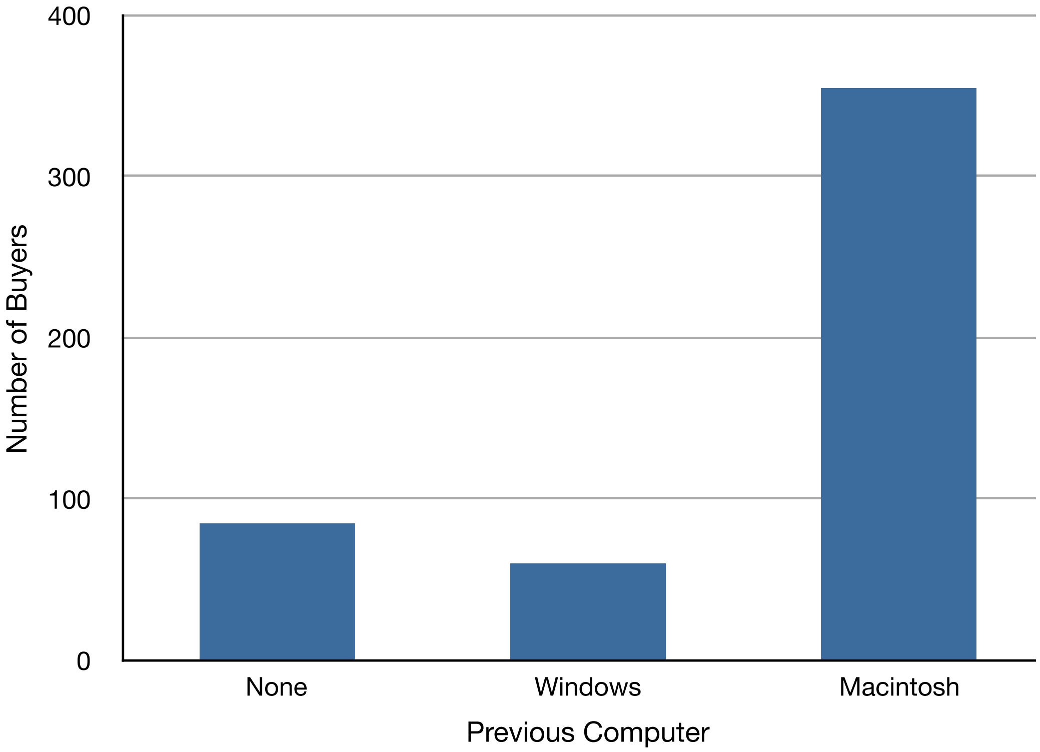


Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 3 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday in the spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

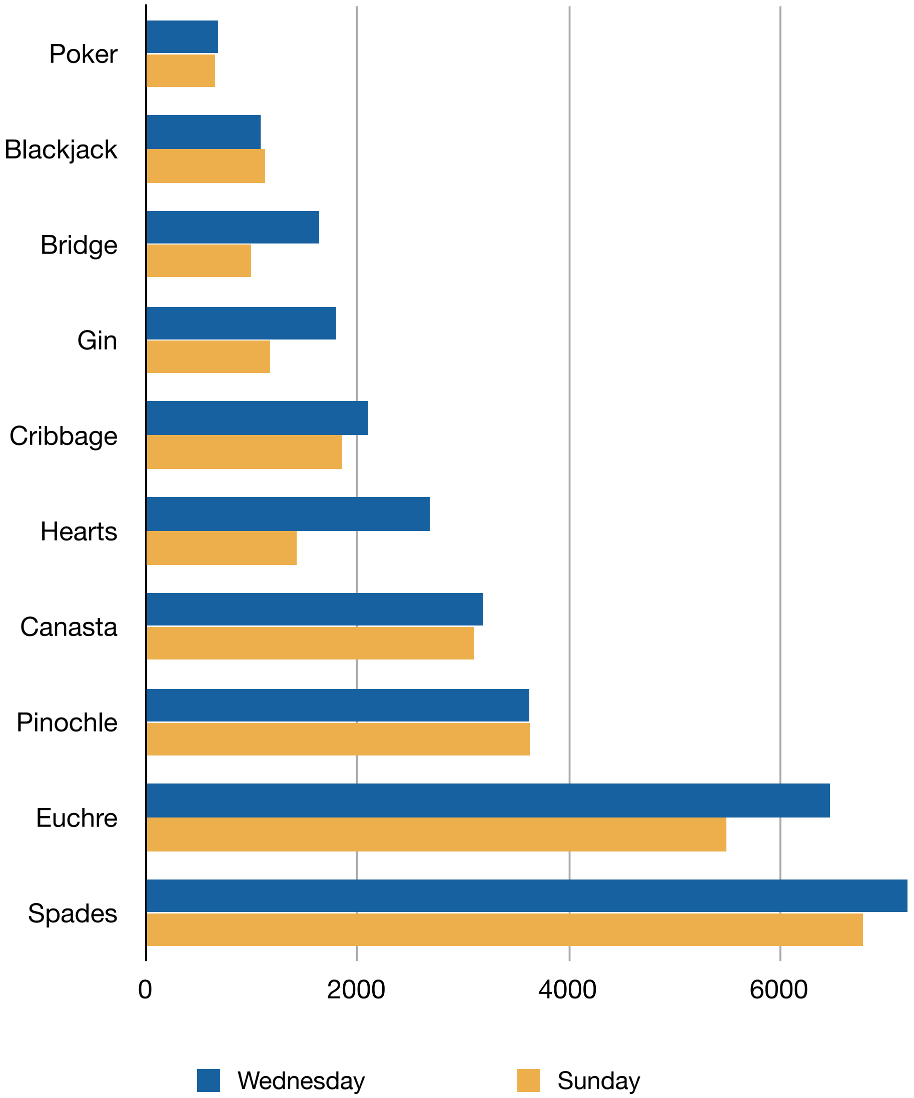


Figure 3. A bar chart of the number of people playing different card games on Sunday and Wednesday.

The bars in Figure 3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in this chapter.

Some graphical mistakes to avoid

Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts such as the one shown in Figure 4 are usually not as effective as their two-dimensional counterparts.

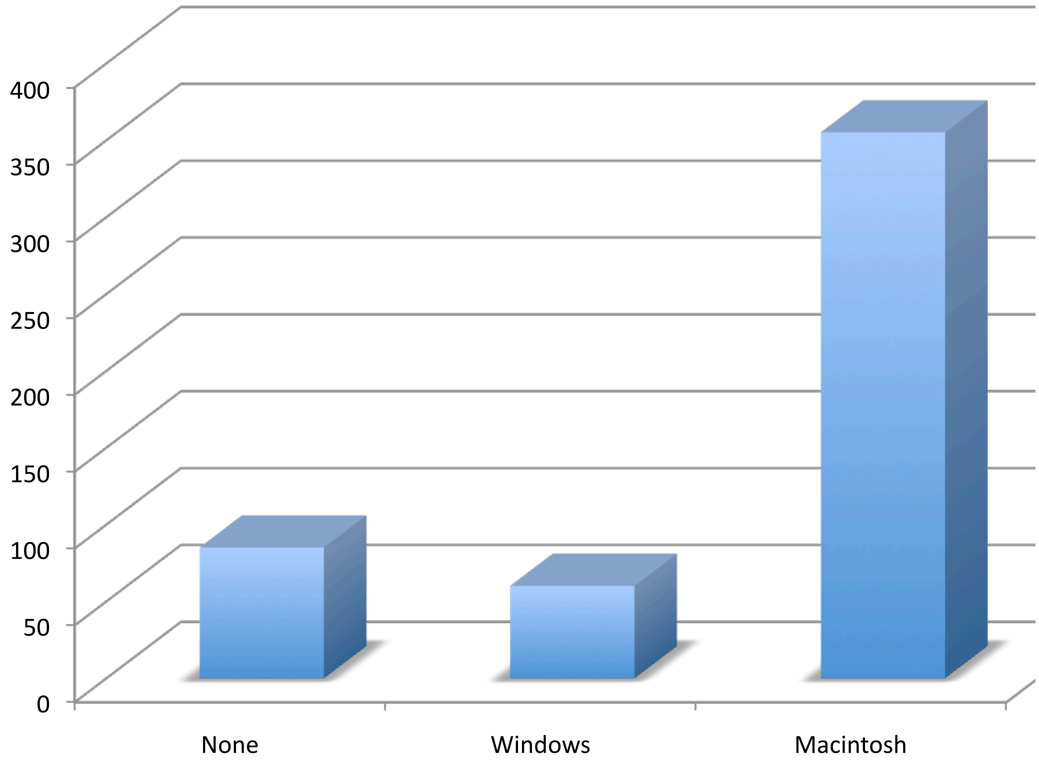


Figure 4. A three-dimensional version of Figure 2.

Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 5 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 5 is misleading because the viewer's attention will be captured by areas. The areas can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 5 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 5 instead of Figure 2! Edward Tufte coined the term “lie factor” to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

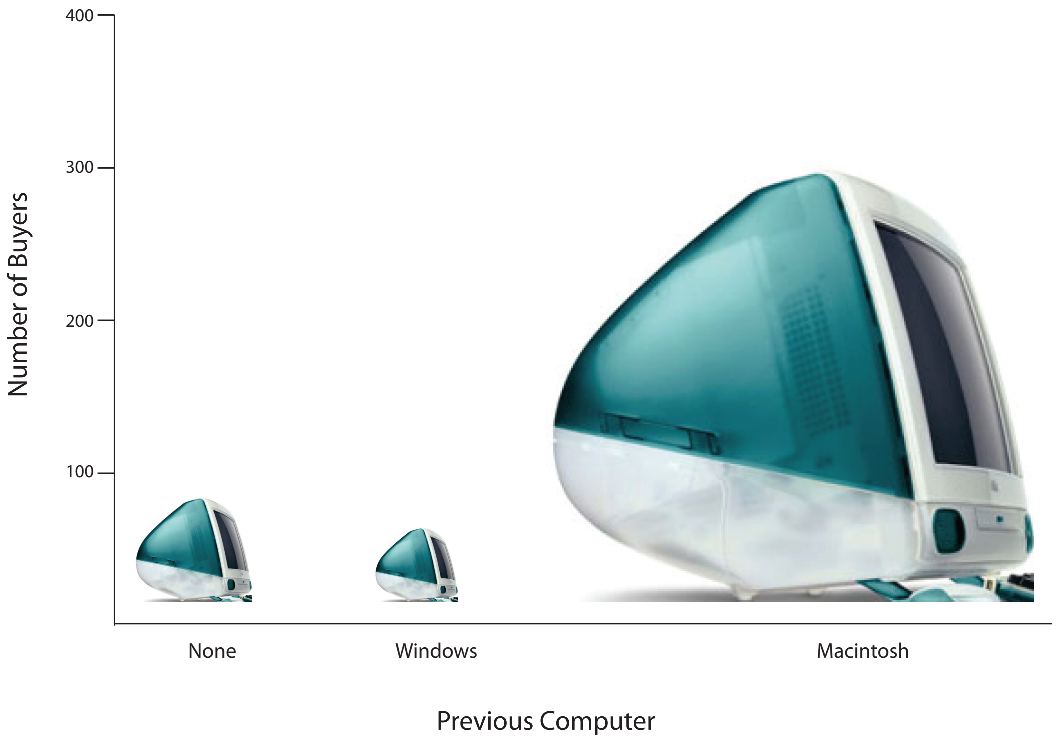


Figure 5. A redrawing of Figure 2 with a lie factor greater than 8.

Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, but not always, this number should be zero. Figure 6 shows the iMac data with a baseline of 50. Once again, the differences in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

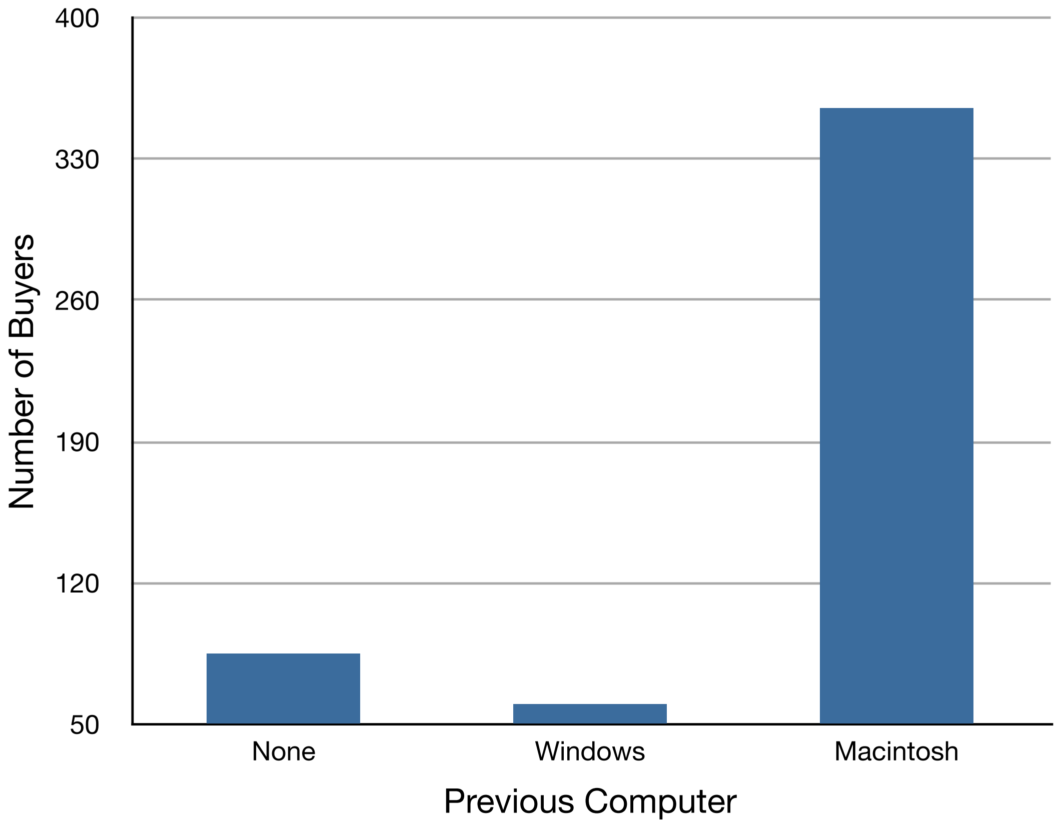


Figure 6. A redrawing of Figure 2 with a baseline of 50.

Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 7 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 7 is that it gives the false impression that the games are naturally ordered in a numerical way when, in fact, they are ordered alphabetically.

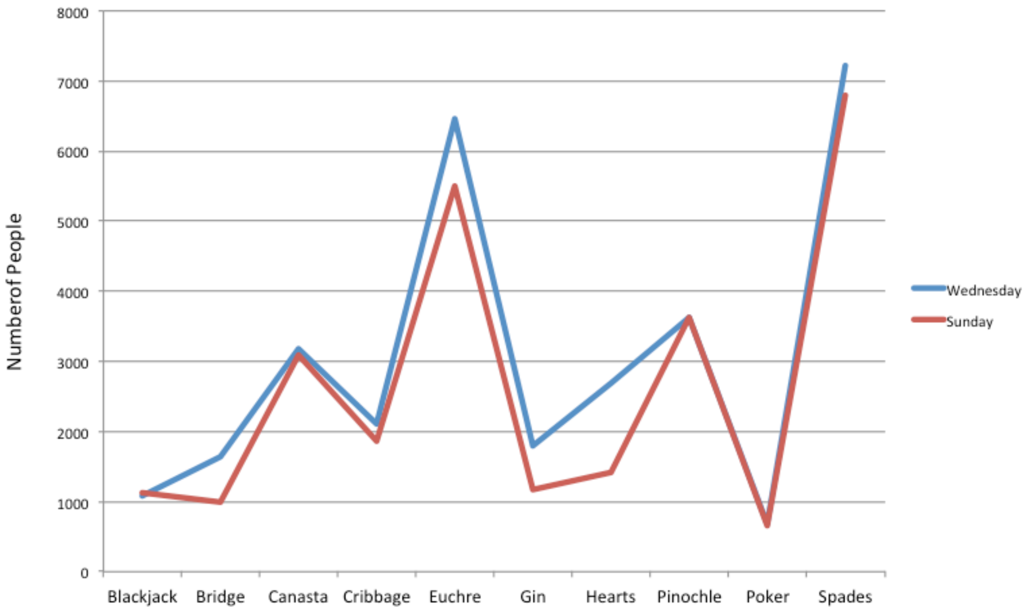


Figure 7. A line graph used inappropriately to depict the number of people playing different card games on Sunday and Wednesday.

Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.

Graphing Quantitative Variables

1. Stem and Leaf Displays
2. Histograms
3. ~~Frequency Polygons~~
4. Box Plots
5. Bar Charts
6. ~~Line Graphs~~
7. ~~Dot Plots~~

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: (1) stem and leaf displays, (2) histograms, (3) frequency polygons, (4) box plots, (5) bar charts, (6) line graphs, (7) dot plots, and (8) scatter plots (discussed in a different chapter). Some graph types such as stem and leaf displays 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.

Stem and Leaf Displays

by David M. Lane

Prerequisites

- Chapter 1: Distributions

Learning Objectives

1. Create and interpret basic stem and leaf displays
2. Create and interpret back-to-back stem and leaf displays
3. Judge whether a stem and leaf display is appropriate for a given data set

A stem and leaf display is a graphical method of displaying data. It is particularly useful when your data are not too numerous. In this section, we will explain how to construct and interpret this kind of graph.

As usual, we will start with an example. Consider Table 1 that shows the number of touchdown passes (TD passes) thrown by each of the 31 teams in the National Football League in the 2000 season.

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28,
28, 23, 22, 22, 22, 21,
21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6

A stem and leaf display of the data is shown in Figure 1. The left portion of Figure 1 contains the stems. They are the numbers 3, 2, 1, and 0, arranged as a column to the left of the bars. Think of these numbers as 10's digits. A stem of 3, for example, can be used to represent the 10's digit in any of the numbers from 30 to 39. The numbers to the right of the bar are leaves, and they represent the 1's digits. Every leaf in the graph therefore stands for the result of adding the leaf to 10 times its stem.

```
3|2337
2|001112223889
1|2244456888899
0|69
```

Figure 1. Stem and leaf display of the number of touchdown passes.

To make this clear, let us examine Figure 1 more closely. In the top row, the four leaves to the right of stem 3 are 2, 3, 3, and 7. Combined with the stem, these leaves represent the numbers 32, 33, 33, and 37, which are the numbers of TD passes for the first four teams in Table 1. The next row has a stem of 2 and 12 leaves. Together, they represent 12 data points, namely, two occurrences of 20 TD passes, three occurrences of 21 TD passes, three occurrences of 22 TD passes, one occurrence of 23 TD passes, two occurrences of 28 TD passes, and one occurrence of 29 TD passes. We leave it to you to figure out what the third row represents. The fourth row has a stem of 0 and two leaves. It stands for the last two entries in Table 1, namely 9 TD passes and 6 TD passes. (The latter two numbers may be thought of as 09 and 06.)

One purpose of a stem and leaf display is to clarify the shape of the distribution. You can see many facts about TD passes more easily in Figure 1 than in Table 1. For example, by looking at the stems and the shape of the plot, you can tell that most of the teams had between 10 and 29 passing TD's, with a few having more and a few having less. The precise numbers of TD passes can be determined by examining the leaves.

We can make our figure even more revealing by splitting each stem into two parts. Figure 2 shows how to do this. The top row is reserved for numbers from 35 to 39 and holds only the 37 TD passes made by the first team in Table 1. The second row is reserved for the numbers from 30 to 34 and holds the 32, 33, and 33 TD passes made by the next three teams in the table. You can see for yourself what the other rows represent.

```
3|7
3|233
2|889
2|001112223
1|56888899
1|22444
0|69
```

Figure 2. Stem and leaf display with the stems split in two.

Figure 2 is more revealing than Figure 1 because the latter figure lumps too many values into a single row. Whether you should split stems in a display depends on the exact form of your data. If rows get too long with single stems, you might try splitting them into two or more parts.

There is a variation of stem and leaf displays that is useful for comparing distributions. The two distributions are placed back to back along a common column of stems. The result is a "back-to-back stem and leaf display." Figure 3 shows such a graph. It compares the numbers of TD passes in the 1998 and 2000 seasons. The stems are in the middle, the leaves to the left are for the 1998 data, and the leaves to the right are for the 2000 data. For example, the second-to-last row shows that in 1998 there were teams with 11, 12, and 13 TD passes, and in 2000 there were two teams with 12 and three teams with 14 TD passes.

11	4	
	3	7
332	3	233
8865	2	889
44331110	2	001112223
987776665	1	56888899
321	1	22444
7	0	69

Figure 3. Back-to-back stem and leaf display. The left side shows the 1998 TD data and the right side shows the 2000 TD data.

Figure 3 helps us see that the two seasons were similar, but that only in 1998 did any teams throw more than 40 TD passes.

There are two things about the football data that make them easy to graph with stems and leaves. First, the data are limited to whole numbers that can be represented with a one-digit stem and a one-digit leaf. Second, all the numbers are positive. If the data include numbers with three or more digits, or contain decimals, they can be rounded to two-digit accuracy. Negative values are also easily handled. Let us look at another example.

Table 2 shows data from the case study Weapons and Aggression. Each value is the mean difference over a series of trials between the times it took an experimental subject to name aggressive words (like "punch") under two conditions. In one condition, the words were preceded by a non-weapon word such as “bug.” In the second condition, the same words were preceded by a weapon word such as “gun” or “knife.” The issue addressed by the experiment was whether a preceding weapon word would speed up (or prime) pronunciation of the aggressive word compared to a non-weapon priming word. A positive difference implies greater priming of the aggressive word by the weapon word. Negative differences imply that the priming by the weapon word was less than for a neutral word.

Table 2. The effects of priming (thousandths of a second).

43.2, 42.9, 35.6, 25.6, 25.4, 23.6, 20.5, 19.9, 14.4, 12.7, 11.3,
10.2, 10.0, 9.1, 7.5, 5.4, 4.7, 3.8, 2.1, 1.2, -0.2, -6.3, -6.7,
-8.8, -10.4, -10.5, -14.9, -14.9, -15.0, -18.5, -27.4

You see that the numbers range from 43.2 to -27.4. The first value indicates that one subject was 43.2 milliseconds faster pronouncing aggressive words when they were preceded by weapon words than when preceded by neutral words. The value -27.4 indicates that another subject was 27.4 milliseconds slower pronouncing aggressive words when they were preceded by weapon words.

The data are displayed with stems and leaves in Figure 4. Since stem and leaf displays can only portray two whole digits (one for the stem and one for the leaf) the numbers are first rounded. Thus, the value 43.2 is rounded to 43 and represented with a stem of 4 and a leaf of 3. Similarly, 42.9 is rounded to 43. To represent negative numbers, we simply use negative stems. For example, the bottom row of the figure represents the number –27. The second-to-last row represents the numbers -10, -10, -15, etc. Once again, we have rounded the original values from Table 2.

```
 4|33
 3|6
 2|00456
 1|00134
 0|1245589
-0|0679
-1|005559
-2|7
```

Figure 4. Stem and leaf display with negative numbers and rounding.

Observe that the figure contains a row headed by “0” and another headed by “-0.” The stem of 0 is for numbers between 0 and 9, whereas the stem of -0 is for numbers between 0 and -9. For example, the fifth row of the table holds the numbers 1, 2, 4, 5, 5, 8, 9 and the sixth row holds 0, -6, -7, and -9. Values that are exactly 0 before rounding should be split as evenly as possible between the “0” and “-0” rows. In Table 2, none of the values are 0 before rounding. The “0” that appears in the “-0” row comes from the original value of -0.2 in the table.

Although stem and leaf displays are unwieldy for large data sets, they are often useful for data sets with up to 200 observations. Figure 5 portrays the distribution of populations of 185 US cities in 1998. To be included, a city had to have between 100,000 and 500,000 residents.

```
4|899
4|6
4|4455
4|333
4|01
3|99
3|677777
3|55
3|223
3|111
2|8899
2|666667
2|444455
2|22333
2|000000
1|88888888888999999999999
1|666666777777
1|444444444444555555555555
1|2222222222222222222333333333
1|000000000000000111111111111111111111111111
```

Figure 5. Stem and leaf display of populations of 185 US cities with populations between 100,000 and 500,000 in 1988.

Since a stem and leaf plot shows only two-place accuracy, we had to round the numbers to the nearest 10,000. For example the largest number (493,559) was rounded to 490,000 and then plotted with a stem of 4 and a leaf of 9. The fourth highest number (463,201) was rounded to 460,000 and plotted with a stem of 4 and a leaf of 6. Thus, the stems represent units of 100,000 and the leaves represent units of 10,000. Notice that each stem value is split into five parts: 0-1, 2-3, 4-5, 6-7, and 8-9.

Whether your data can be suitably represented by a stem and leaf display depends on whether they can be rounded without loss of important information. Also, their extreme values must fit into two successive digits, as the data in Figure 5 fit into the 10,000 and 100,000 places (for leaves and stems, respectively). Deciding what kind of graph is best suited to displaying your data thus requires good judgment. Statistics is not just recipes!

Histograms

by David M. Lane

Prerequisites

- Chapter 1: Distributions
- Chapter 2: Graphing Qualitative Data

Learning Objectives

1. Create a grouped frequency distribution
2. Create a histogram based on a grouped frequency distribution
3. Determine an appropriate bin width

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 1.

Table 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 1.

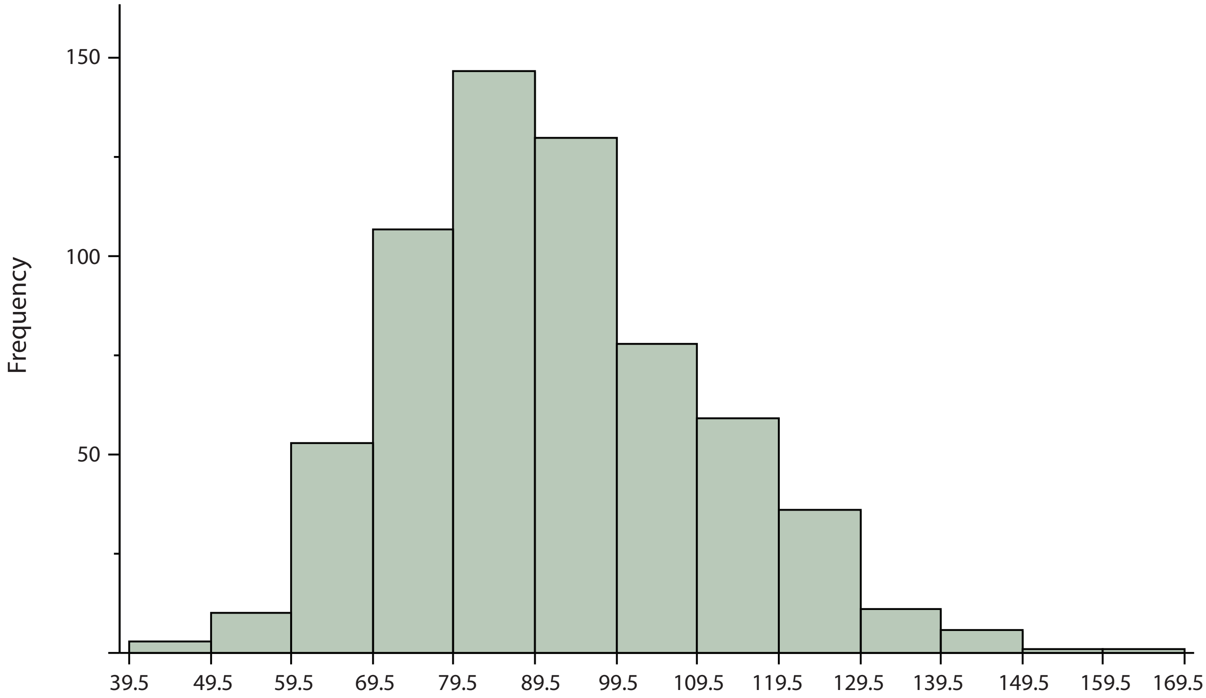


Figure 1. Histogram of scores on a psychology test.

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. There are some “rules of thumb” that can help you choose an appropriate width. (But keep in mind that none of the rules is perfect.) Sturges’ rule is to set the number of intervals as close as possible to 1 + $\text{Log}_2(N)$, where $\text{Log}_2(N)$ is the base 2 log of the number of observations. The formula can also be written as 1 + 3.3 $\text{Log}_{10}(N)$ where $\text{Log}_{10}(N)$ is the log base 10 of the number of observations. According to Sturges’ rule, 1000 observations would be graphed with 11 class intervals since 10 is the closest integer to Log$_2$(1000). We prefer the Rice rule, which is to set the number of intervals to twice the cube root of the number of observations. In the case of 1000 observations, the Rice rule yields 20 intervals instead of the 11 recommended by Sturges' rule. For the psychology test example used above, Sturges' rule recommends 10 intervals while the Rice rule recommends 17. In the end, we compromised and chose 13 intervals for Figure 1 to create a histogram that seemed clearest. 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.

To provide experience in constructing histograms, we have developed an interactive demonstration (external link; Java required). The demonstration reveals the consequences of different choices of bin width and of lower boundary for the first interval.

Box Plots

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 2: Histograms
- Chapter 2: Frequency Polygons

Learning Objectives

1. Define basic terms including hinges, H-spread, step, adjacent value, outside value, and far out value
2. Create a box plot
3. Create parallel box plots
4. Determine whether a box plot is appropriate for a given data set

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section we present another important graph, called a box plot. Box plots are useful for identifying outliers and for comparing distributions. We will explain box plots with the help of data from an in-class experiment. Students in Introductory Statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible. Their times (in seconds) were recorded. We'll compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender. Such a display is said to involve parallel box plots.

There are several steps in constructing a box plot. The first relies on the 25th, 50th, and 75th percentiles in the distribution of scores. Figure 1 shows how these three statistics are used. For each gender we draw a box extending from the 25th percentile to the 75th percentile. The 50th percentile is drawn inside the box. Therefore, the bottom of each box is the 25th percentile, the top is the 75th percentile, and the line in the middle is the 50th percentile.

The data for the women in our sample are shown in Table 1.

Table 1. Women's times.

14	17	18	19	20	21	29
15	17	18	19	20	22	
16	17	18	19	20	23	
16	17	18	20	20	24	
17	18	18	20	21	24	

For these data, the 25th percentile is 17, the 50th percentile is 19, and the 75th percentile is 20. For the men (whose data are not shown), the 25th percentile is 19, the 50th percentile is 22.5, and the 75th percentile is 25.5.

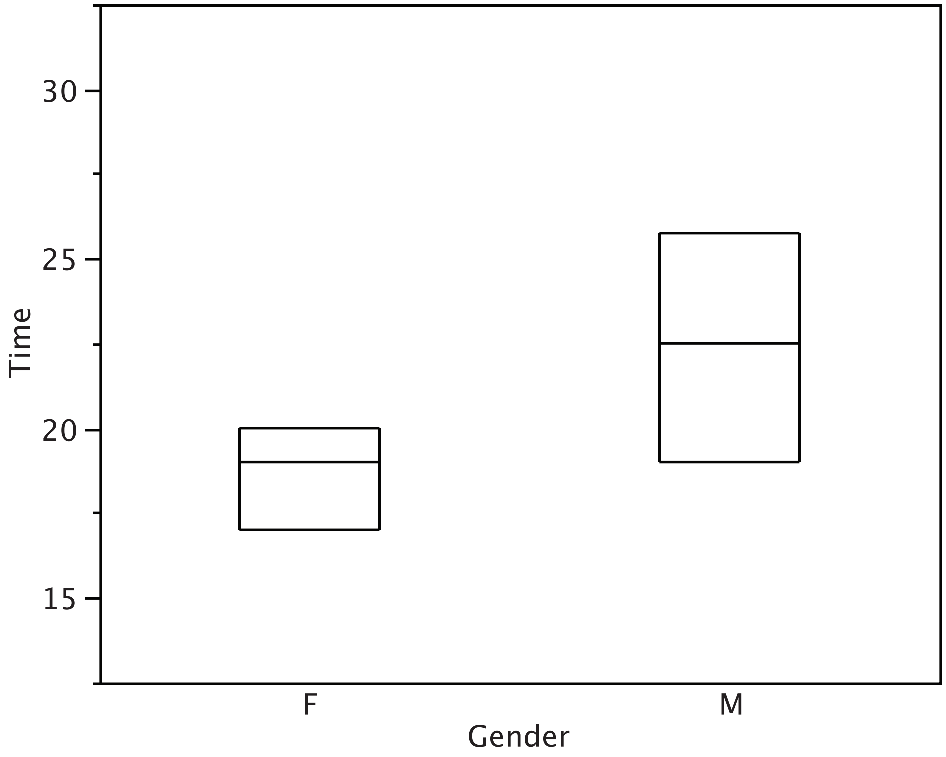


Figure 1. The first step in creating box plots.

Before proceeding, the terminology in Table 2 is helpful.

Table 2. Box plot terms and values for women's times.

Name	Formula	Value
Upper Hinge	75th Percentile	20
Lower Hinge	25th Percentile	17

H-Spread	Upper Hinge - Lower Hinge	3
Step	1.5 x H-Spread	4.5
Upper Inner Fence	Upper Hinge + 1 Step	24.5
Lower Inner Fence	Lower Hinge - 1 Step	12.5
Upper Outer Fence	Upper Hinge + 2 Steps	29
Lower Outer Fence	Lower Hinge - 2 Steps	8
Upper Adjacent	Largest value below Upper Inner Fence	24
Lower Adjacent	Smallest value above Lower Inner Fence	14
Outside Value	A value beyond an Inner Fence but not beyond an Outer Fence	29
Far Out Value	A value beyond an Outer Fence	None

Continuing with the box plots, we put “whiskers” above and below each box to give additional information about the spread of data. Whiskers are vertical lines that end in a horizontal stroke. Whiskers are drawn from the upper and lower hinges to the upper and lower adjacent values (24 and 14 for the women's data).

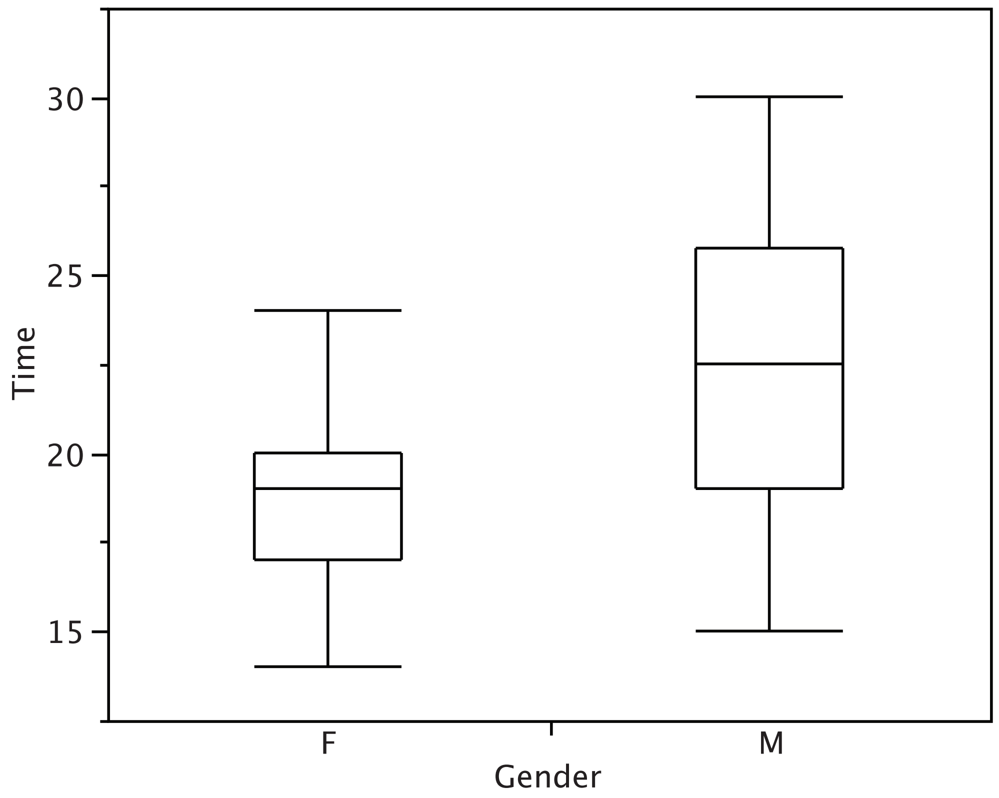


Figure 2. The box plots with the whiskers drawn.

Although we don't draw whiskers all the way to outside or far out values, we still wish to represent them in our box plots. This is achieved by adding additional marks beyond the whiskers. Specifically, outside values are indicated by small “o's” and far out values are indicated by asterisks (*). In our data, there are no far-out values and just one outside value. This outside value of 29 is for the women and is shown in Figure 3.

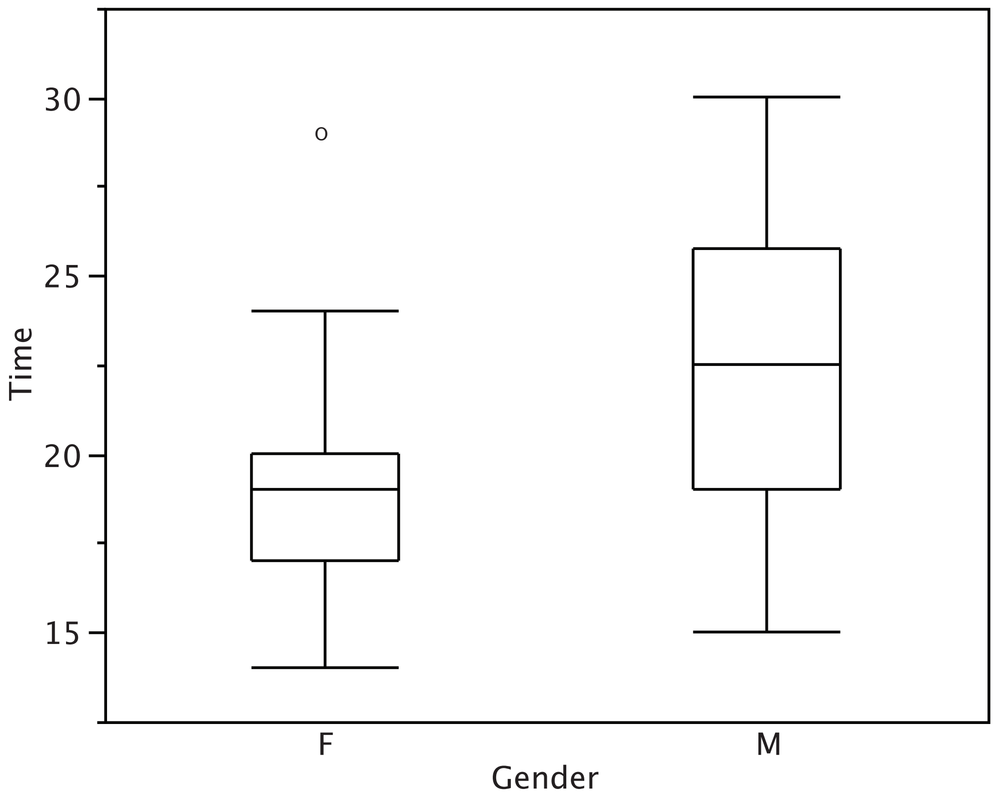


Figure 3. The box plots with the outside value shown.

There is one more mark to include in box plots (although sometimes it is omitted). We indicate the mean score for a group by inserting a plus sign. Figure 4 shows the result of adding means to our box plots.

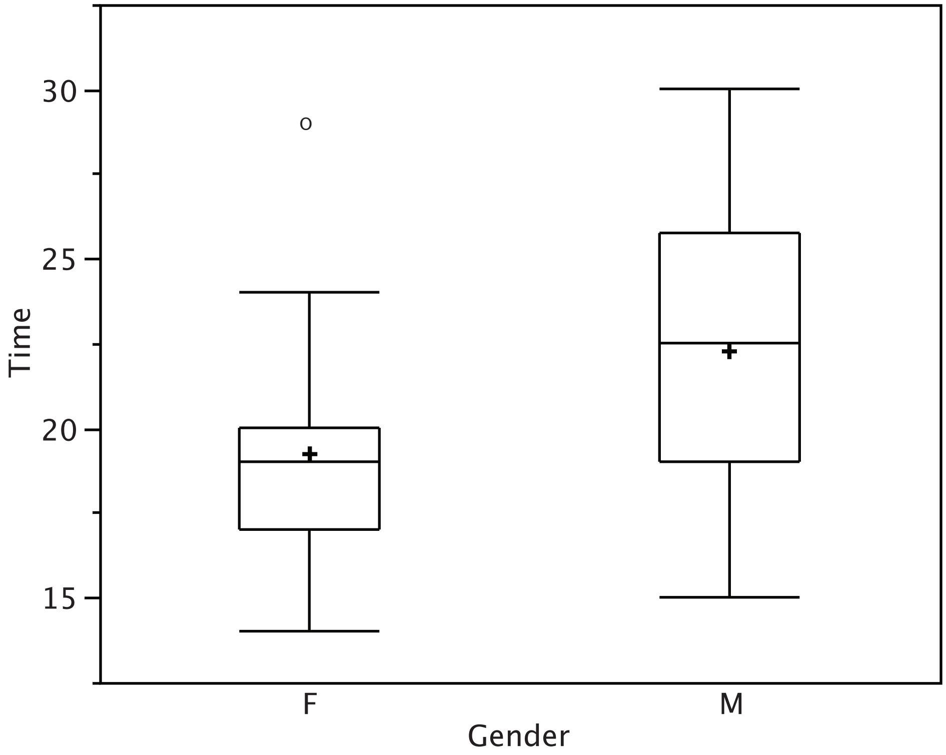


Figure 4. The completed box plots.

Figure 4 provides a revealing summary of the data. Since half the scores in a distribution are between the hinges (recall that the hinges are the 25th and 75th percentiles), we see that half the women's times are between 17 and 20 seconds whereas half the men's times are between 19 and 25.5 seconds. We also see that women generally named the colors faster than the men did, although one woman was slower than almost all of the men. Figure 5 shows the box plot for the women's data with detailed labels.

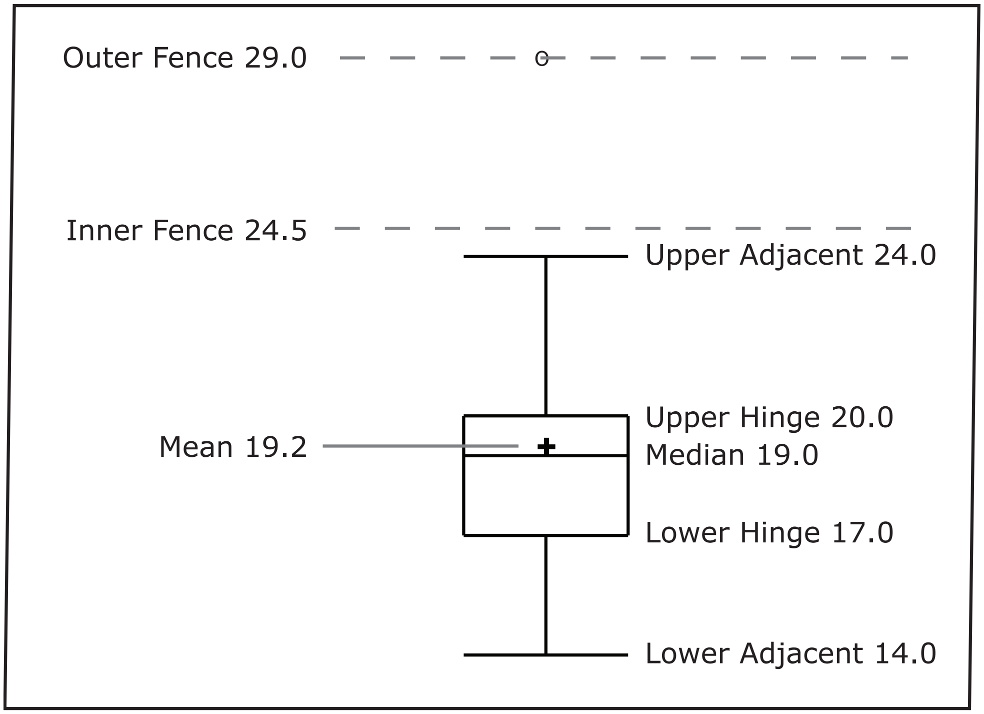


Figure 5. The box plots for the women's data with detailed labels.

Box plots provide basic information about a distribution. For example, a distribution with a positive skew would have a longer whisker in the positive direction than in the negative direction. A larger mean than median would also indicate a positive skew. Box plots are good at portraying extreme values and are especially good at showing differences between distributions. However, many of the details of a distribution are not revealed in a box plot and to examine these details one should use create a histogram and/or a stem and leaf display.

Variations on box plots

Statistical analysis programs may offer options on how box plots are created. For example, the box plots in Figure 6 are constructed from our data but differ from the previous box plots in several ways.

1. It does not mark outliers.
2. The means are indicated by green lines rather than plus signs.
3. The mean of all scores is indicated by a gray line.
4. Individual scores are represented by dots. Since the scores have been rounded to the nearest second, any given dot might represent more than one score.

5. The box for the women is wider than the box for the men because the widths of the boxes are proportional to the number of subjects of each gender (31 women and 16 men).

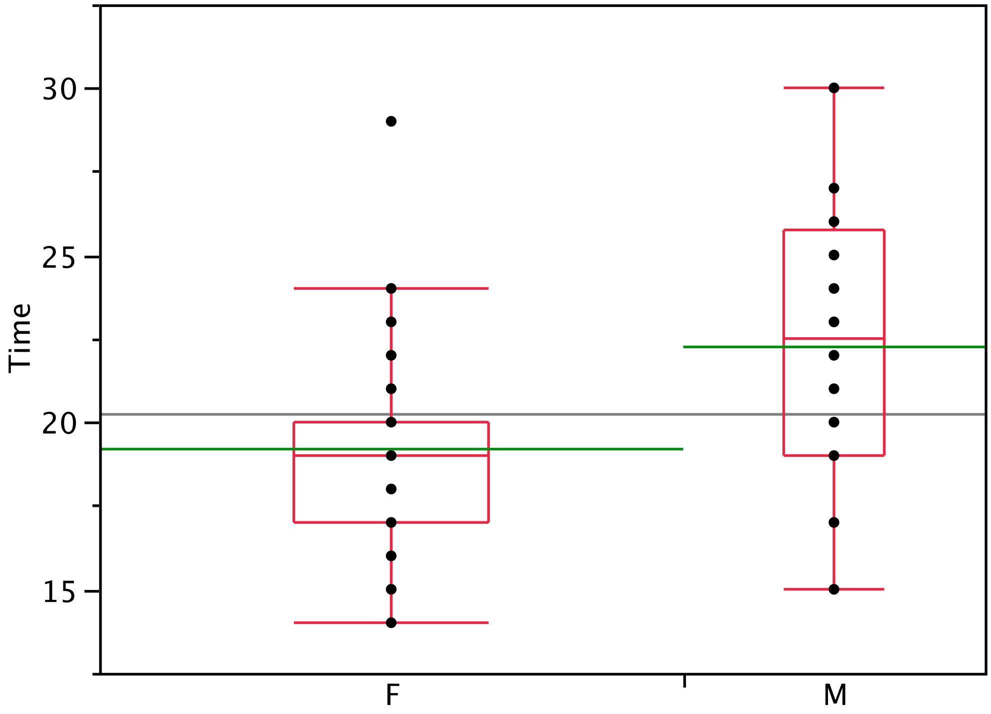


Figure 6. Box plots showing the individual scores and the means.

Each dot in Figure 6 represents a group of subjects with the same score (rounded to the nearest second). An alternative graphing technique is to jitter the points. This means spreading out different dots at the same horizontal position, one dot for each subject. The exact horizontal position of a dot is determined randomly (under the constraint that different dots don't overlap exactly). Spreading out the dots helps you to see multiple occurrences of a given score. However, depending on the dot size and the screen resolution, some points may be obscured even if the points are jittered. Figure 7 shows what jittering looks like.

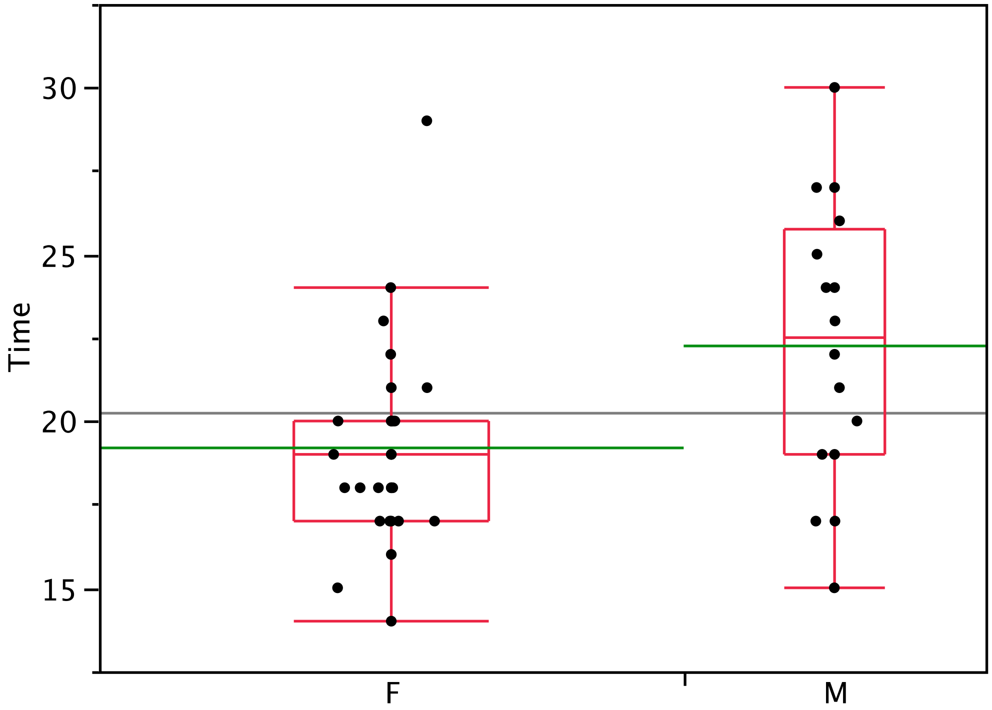


Figure 7. Box plots with the individual scores jittered.

Different styles of box plots are best for different situations, and there are no firm rules for which to use. When exploring your data, you should try several ways of visualizing them. Which graphs you include in your report should depend on how well different graphs reveal the aspects of the data you consider most important.

Bar Charts

by David M. Lane

Prerequisites

- Chapter 2: Graphing Qualitative Variables

Learning Objectives

1. Create and interpret bar charts
2. Judge whether a bar chart or another graph such as a box plot would be more appropriate

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 1 shows how many purchasers of iMac computers were previous Macintosh users, previous Windows users, and new computer purchasers.

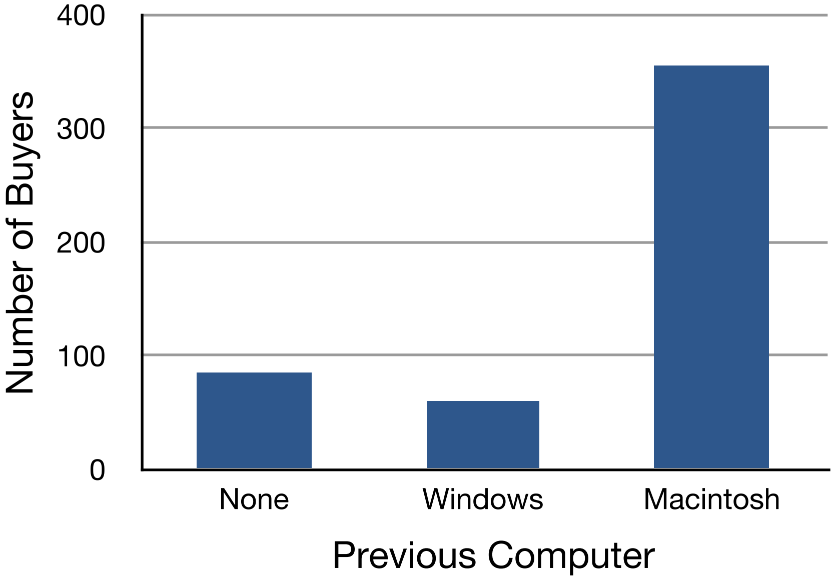


Figure 1. iMac buyers as a function of previous computer ownership.

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 2 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*.

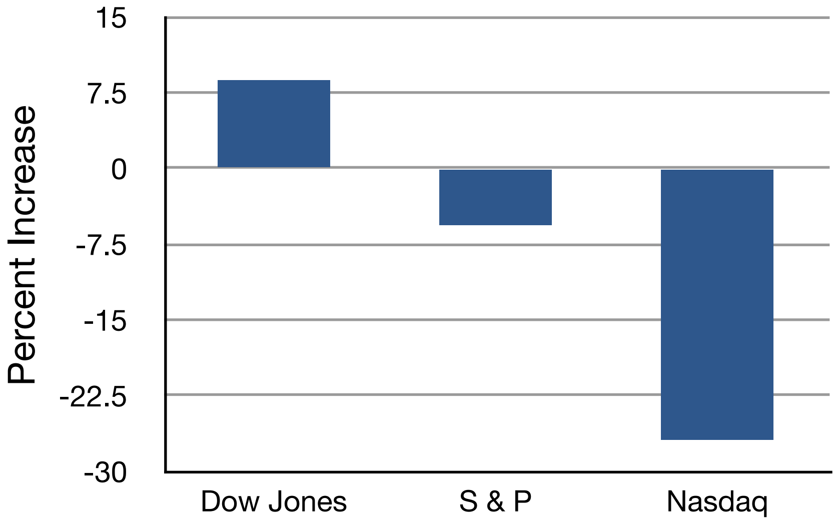


Figure 2. Percent increase in three stock indexes from May 24th 2000 to May 24th 2001.

Bar charts are particularly effective for showing change over time. Figure 3, 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.

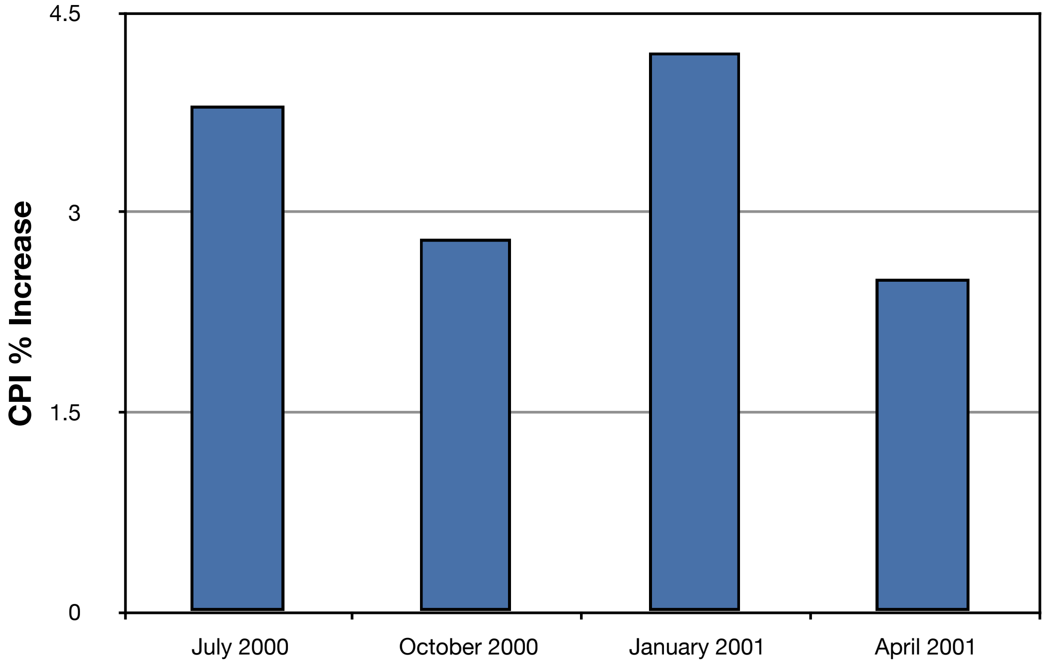


Figure 3. Percent change in the CPI over time. Each bar represents percent increase for the three months ending at the date indicated.

Bar charts are often used to compare the means of different experimental conditions. Figure 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.

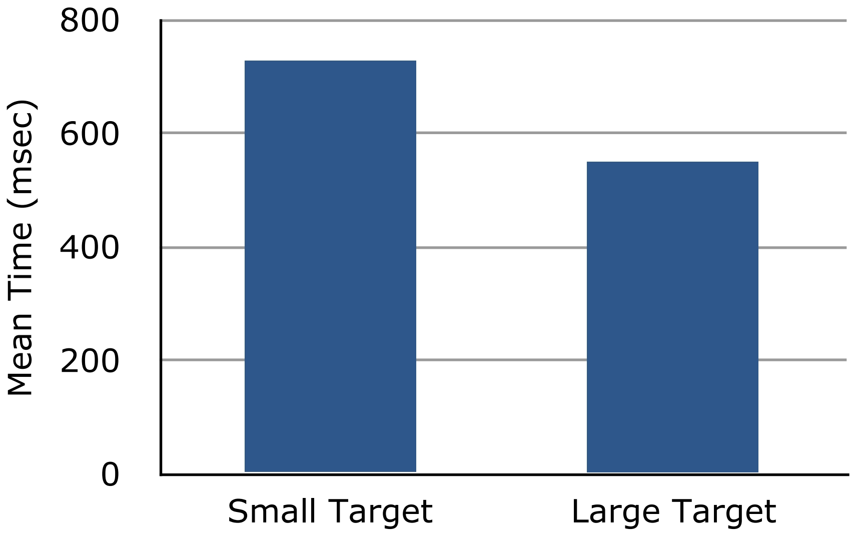


Figure 4. Bar chart showing the means for the two conditions.

Although bar charts can display means, we do not recommend them for this purpose. Box plots should be used instead since they provide more information than bar charts without taking up more space. For example, a box plot of the cursor-movement data is shown in Figure 5. You can see that Figure 5 reveals more about the distribution of movement times than does Figure 4.

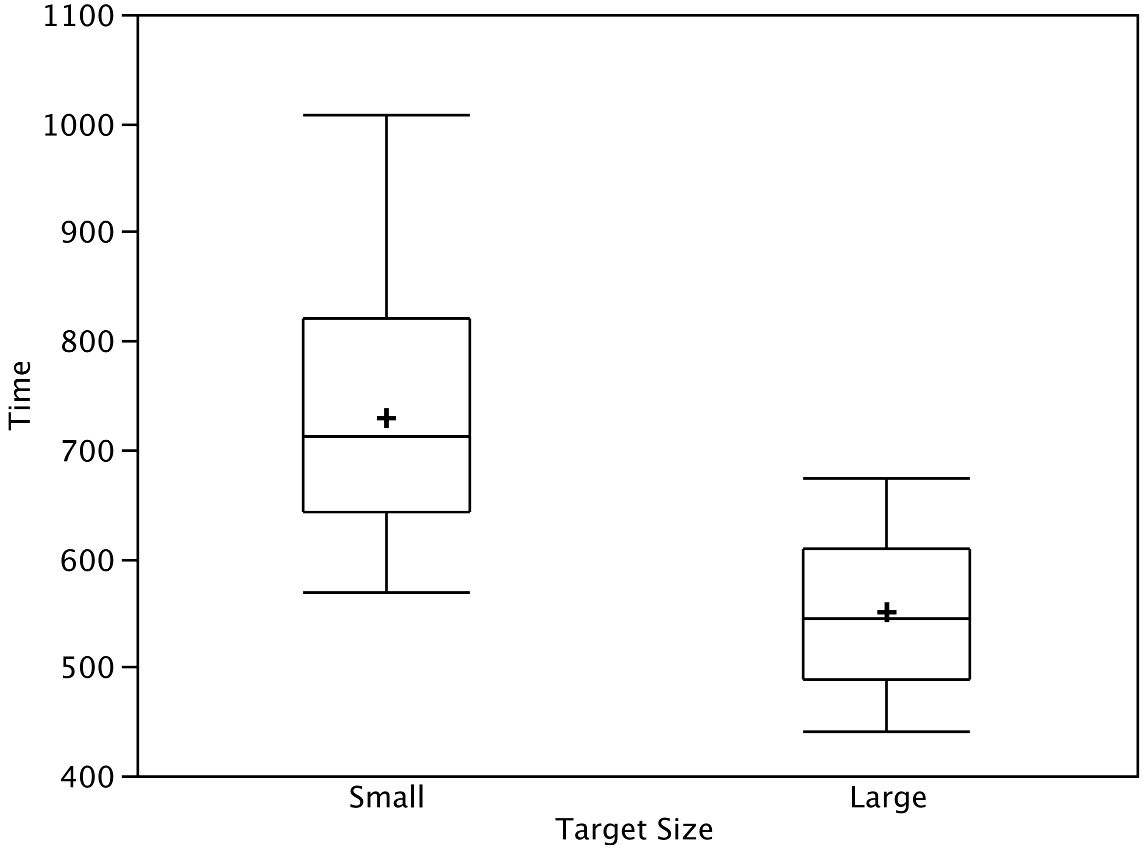


Figure 5. Box plots of times to move the cursor to the small and large targets.

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.

Exercises

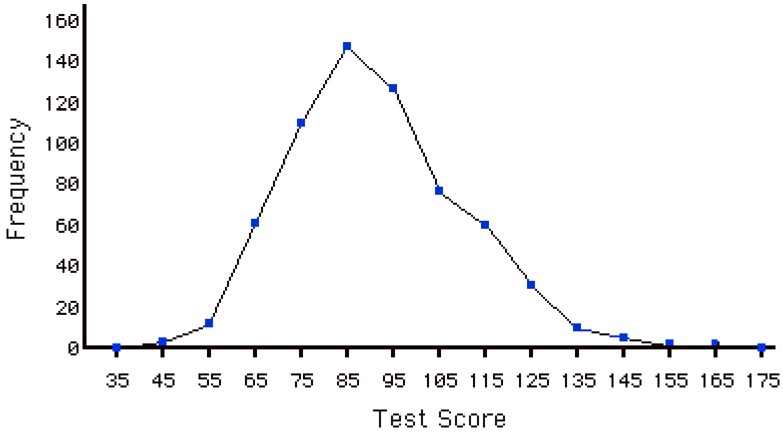
Relevant exercises:
1, 3, 4, 5, 7, 8, 9, 27

Prerequisites

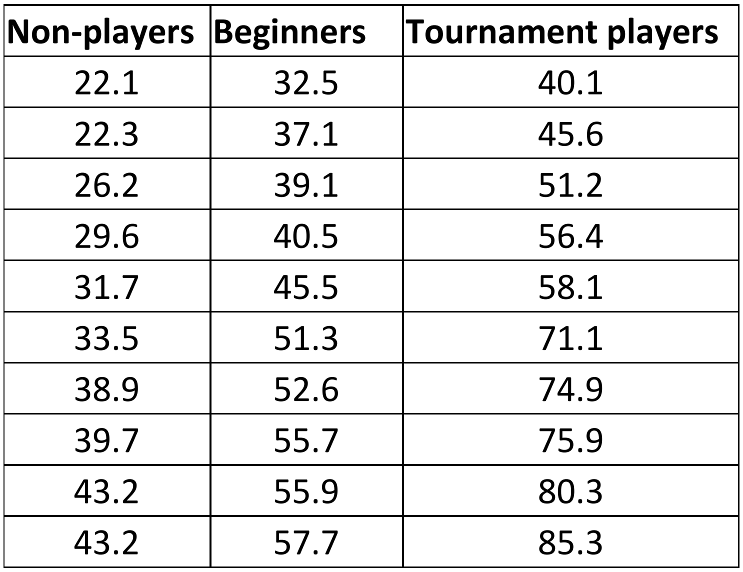
- All material presented in the Graphing Distributions chapter

1. Name some ways to graph quantitative variables and some ways to graph qualitative variables.

2. Based on the frequency polygon displayed below, the most common test grade was around what score? Explain.



3. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?



Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

4. You have to decide between displaying your data with a histogram or with a stem and leaf display. What factor(s) would affect your choice?

5. In a box plot, what percent of the scores are between the lower and upper hinges?

6. A student has decided to display the results of his project on the number of hours people in various countries slept per night. He compared the sleeping patterns of people from the US, Brazil, France, Turkey, China, Egypt, Canada, Norway, and Spain. He was planning on using a line graph to display this data. Is a line graph appropriate? What might be a better choice for a graph?

7. For the data from the 1977 Stat. and Biom. 200 class for eye color, construct:

   a. pie graph

   b. horizontal bar graph

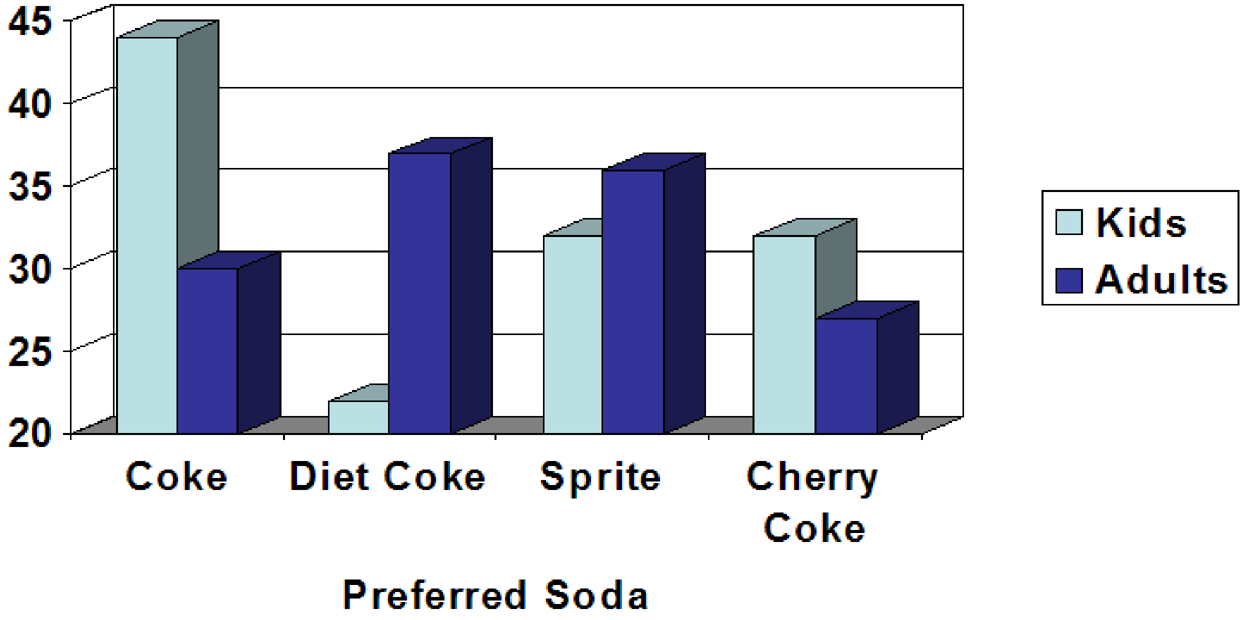
   c. vertical bar graph

   d. a frequency table with the relative frequency of each eye color

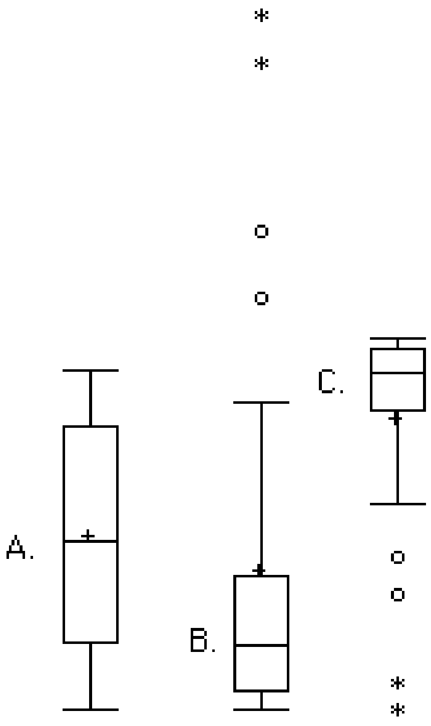
Eye Color	Number of students
Brown	11
Blue	10
Green	4
Gray	1

(Question submitted by J. Warren, UNH)

8. A graph appears below showing the number of adults and children who prefer each type of soda. There were 130 adults and kids surveyed. Discuss some ways in which the graph below could be improved.



9. Which of the box plots on the graph has a large positive skew? Which has a large negative skew?



Question from Case Studies

Angry Moods (AM) case study

10. (AM) Is there a difference in how much males and females use aggressive behavior to improve an angry mood? For the “Anger-Out” scores:

    a. Create parallel box plots.

    b. Create a back to back stem and leaf displays (You may have trouble finding a computer to do this so you may have to do it by hand. Use a fixed-width font such as Courier.)

11. (AM) Create parallel box plots for the Anger-In scores by sports participation.

12. (AM) Plot a histogram of the distribution of the Control-Out scores.

13. (AM) Create a bar graph comparing the mean Control-In score for the athletes and the non- athletes. What would be a better way to display this data?

14. (AM) Plot parallel box plots of the Anger Expression Index by sports participation. Does it look like there are any outliers? Which group reported expressing more anger?

Flatulence (F) case study

15. (F) Plot a histogram of the variable "per day."

16. (F) Create parallel box plots of "how long" as a function gender. Why is the 25th percentile not showing? What can you say about the results?

17. (F) Create a stem and leaf plot of the variable "how long." What can you say about the shape of the distribution?

Physicians' Reactions (PR) case study

18. (PR) Create box plots comparing the time expected to be spent with the average-weight and overweight patients.

19. (PR) Plot histograms of the time spent with the average-weight and overweight patients.

20. (PR) To which group does the patient with the highest expected time belong?

Smiles and Leniency (SL) case study

21. (SL) Create parallel box plots for the four conditions.

22. (SL) Create back to back stem and leaf displays for the false smile and neutral conditions. (It may be hard to find a computer program to do this for you, so be prepared to do it by hand).

ADHD Treatment (AT) case study

23. (AT) Create a line graph of the data. Do certain dosages appear to be more effective than others?

24. (AT) Create a stem and leaf plot of the number of correct responses of the participants after taking the placebo (d0 variable). What can you say about the shape of the distribution?

25. (AT) Create box plots for the four conditions. You may have to rearrange the data to get a computer program to create the box plots.

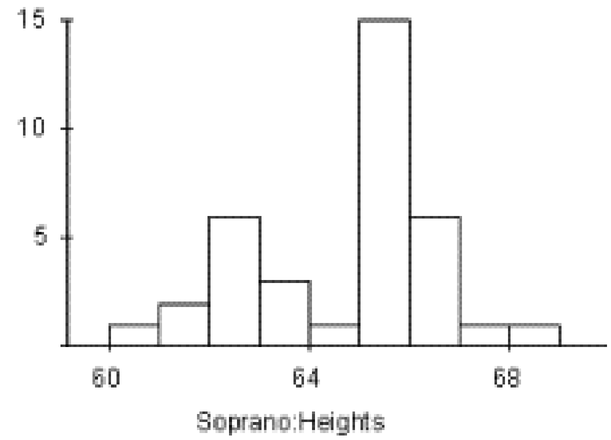
SAT and College GPA (SG) case study

26. (SG)Create histograms and stem and leaf displays of both high-school grade point average and university grade point average. In what way(s) do the distributions differ?

27. The April 10th issue of the Journal of the American Medical Association reports a study on the effects of anti-depressants. The study involved 340 subjects who were being treated for major depression. The subjects were randomly assigned to receive one of three treatments: St. John's wort (an herb), Zoloft (Pfizer's cousin of Lilly's Prozac) or placebo for an 8-week period. The following are the mean scores (approximately) for the three groups of subjects over the eight-week experiment. The first column is the baseline. Lower scores mean less depression. Create a graph to display these means.

Placebo	22.5	19.1	17.9	17.1	16.2	15.1	12.1	12.3
Wort	23.0	20.2	18.2	18.0	16.5	16.1	14.2	13.0
Zoloft	22.4	19.2	16.6	15.5	14.2	13.1	11.8	10.5

28. For the graph below, of heights of singers in a large chorus. What word starting with the letter "B" best describes the distribution?



29. Pretend you are constructing a histogram for describing the distribution of salaries for individuals who are 40 years or older, but are not yet retired. (a) What is on the Y-axis? Explain. (b) What is on the X-axis? Explain. (c) What would be the probable shape of the salary distribution? Explain why.

3. Summarizing Distributions

Descriptive statistics often involves using a few numbers to summarize a distribution. One important aspect of a distribution is where its center is located. Measures of central tendency are discussed first. A second aspect of a distribution is how spread out it is. In other words, how much the numbers in the distribution vary from one another. The second section describes measures of variability. Distributions can differ in shape. Some distributions are symmetric whereas others have long tails in just one direction. The third section describes measures of the shape of distributions. The final two sections concern (1) how transformations affect measures summarizing distributions and (2) the variance sum law, an important relationship involving a measure of variability.

What is Central Tendency?

by David M. Lane and Heidi Ziemer

Prerequisites

- Chapter 1: Distributions
- Chapter 2: Stem and Leaf Displays

Learning Objectives

1. Identify situations in which knowing the center of a distribution would be valuable
2. Give three different ways the center of a distribution can be defined
3. Describe how the balance is different for symmetric distributions than it is for asymmetric distributions.

What is “central tendency,” and why do we want to know the central tendency of a group of scores? Let us first try to answer these questions intuitively. Then we will proceed to a more formal discussion.

Imagine this situation: You are in a class with just four other students, and the five of you took a 5-point pop quiz. Today your instructor is walking around the room, handing back the quizzes. She stops at your desk and hands you your paper. Written in bold black ink on the front is “3/5.” How do you react? Are you happy with your score of 3 or disappointed? How do you decide? You might calculate your percentage correct, realize it is 60%, and be appalled. But it is more likely that when deciding how to react to your performance, you will want additional information. What additional information would you like?

If you are like most students, you will immediately ask your neighbors, “Whad'ja get?” and then ask the instructor, “How did the class do?” In other words, the additional information you want is how your quiz score compares to other students' scores. You therefore understand the importance of comparing your score to the class distribution of scores. Should your score of 3 turn out to be among the higher scores, then you'll be pleased after all. On the other hand, if 3 is among the lower scores in the class, you won't be quite so happy.

This idea of comparing individual scores to a distribution of scores is fundamental to statistics. So let's explore it further, using the same example (the pop quiz you took with your four classmates). Three possible outcomes are shown in Table 1. They are labeled “Dataset A,” “Dataset B,” and “Dataset C.” Which of the three datasets would make you happiest? In other words, in comparing your score with your fellow students' scores, in which dataset would your score of 3 be the most impressive?

In Dataset A, everyone's score is 3. This puts your score at the exact center of the distribution. You can draw satisfaction from the fact that you did as well as everyone else. But of course it cuts both ways: everyone else did just as well as you.

Table 1. Three possible datasets for the 5-point make-up quiz.

Student	Dataset A	Dataset B	Dataset C
You	3	3	3
John's	3	4	2
Maria's	3	4	2
Shareecia's	3	4	2
Luther's	3	5	1

Now consider the possibility that the scores are described as in Dataset B. This is a depressing outcome even though your score is no different than the one in Dataset A. The problem is that the other four students had higher grades, putting yours below the **center of the distribution**.

Finally, let's look at Dataset C. This is more like it! All of your classmates score lower than you so your score is above the center of the distribution.

Now let's change the example in order to develop more insight into the center of a distribution. Figure 1 shows the results of an experiment on memory for chess positions. Subjects were shown a chess position and then asked to reconstruct it on an empty chess board. The number of pieces correctly placed was recorded. This was repeated for two more chess positions. The scores represent the total number of chess pieces correctly placed for the three chess positions. The maximum possible score was 89.

	8	05
	7	156
	6	233
	5	168
330	4	06
9420	3	
622	2	

Figure 1. Back-to-back stem and leaf display. The left side shows the memory scores of the non-players. The right side shows the scores of the tournament players.

Two groups are compared. On the left are people who don't play chess. On the right are people who play a great deal (tournament players). It is clear that the location of the center of the distribution for the non-players is much lower than the center of the distribution for the tournament players.

We're sure you get the idea now about the center of a distribution. It is time to move beyond intuition. We need a formal definition of the center of a distribution. In fact, we'll offer you three definitions! This is not just generosity on our part. There turn out to be (at least) three different ways of thinking about the center of a distribution, all of them useful in various contexts. In the remainder of this section we attempt to communicate the idea behind each concept. In the succeeding sections we will give statistical measures for these concepts of central tendency.

Definitions of Center

Now we explain the three different ways of defining the center of a distribution. All three are called measures of central tendency.

Balance Scale

One definition of central tendency is the point at which the distribution is in balance. Figure 2 shows the distribution of the five numbers 2, 3, 4, 9, 16 placed upon a balance scale. If each number weighs one pound, and is placed at its position along the number line, then it would be possible to balance them by placing a fulcrum at 6.8.

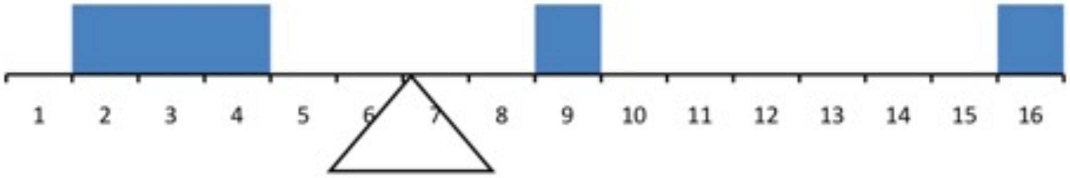


Figure 2. A balance scale.

For another example, consider the distribution shown in Figure 3. It is balanced by placing the fulcrum in the geometric middle.

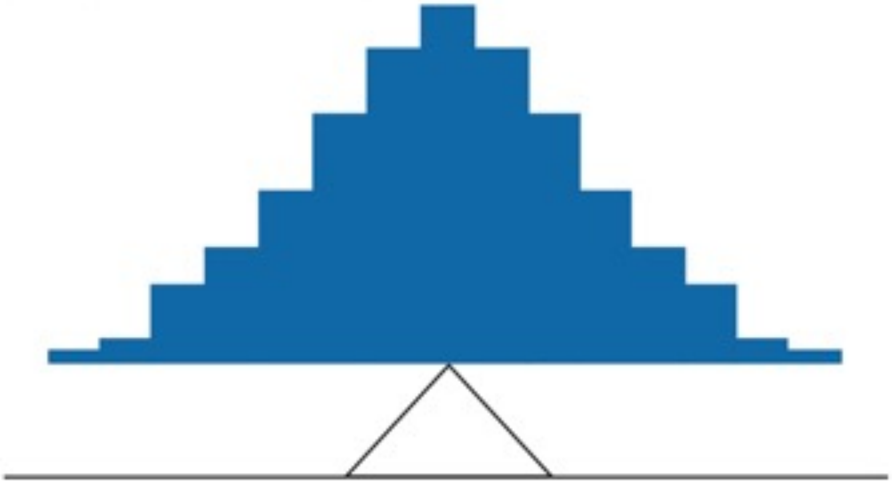


Figure 3. A distribution balanced on the tip of a triangle.

Figure 4 illustrates that the same distribution can't be balanced by placing the fulcrum to the left of center.

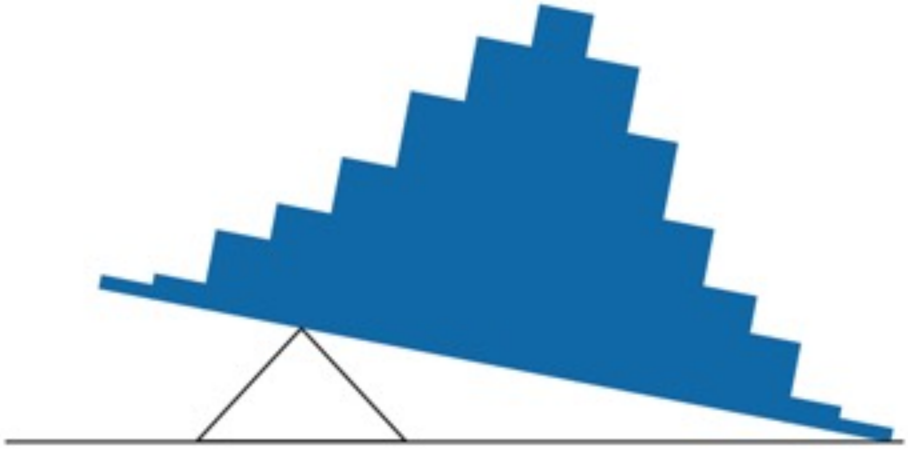


Figure 4. The distribution is not balanced.

Figure 5 shows an asymmetric distribution. To balance it, we cannot put the fulcrum halfway between the lowest and highest values (as we did in Figure 3). Placing the fulcrum at the "half way" point would cause it to tip towards the left.

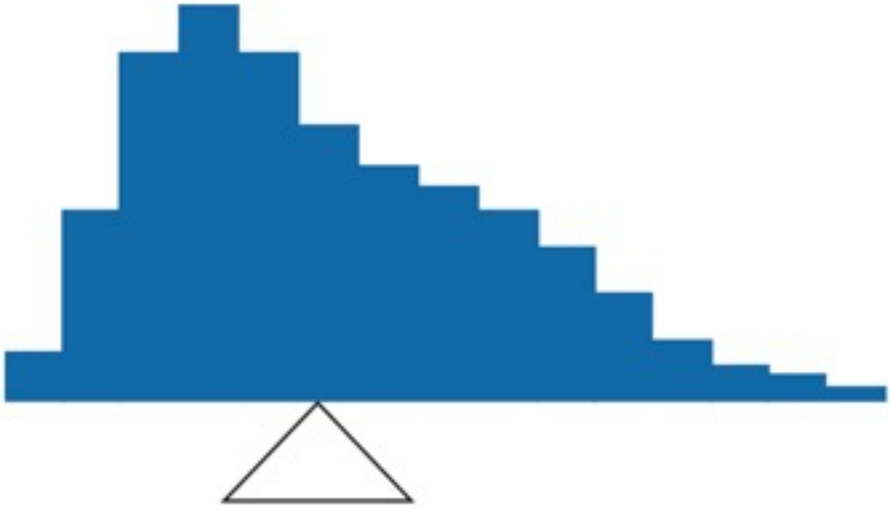


Figure 5. An asymmetric distribution balanced on the tip of a triangle.

The balance point defines one sense of a distribution's center.

Smallest Absolute Deviation

Another way to define the center of a distribution is based on the concept of the sum of the absolute deviations (differences). Consider the distribution made up of the five numbers 2, 3, 4, 9, 16. Let's see how far the distribution is from 10 (picking a number arbitrarily). Table 2 shows the sum of the absolute deviations of these numbers from the number 10.

Table 2. An example of the sum of absolute deviations

Values	Absolute Deviations from 10
2	8
3	7
4	6
9	1
16	6
Sum	28

The first row of the table shows that the absolute value of the difference between 2 and 10 is 8; the second row shows that the absolute difference between 3 and 10 is 7, and similarly for the other rows. When we add up the five absolute deviations, we get 28. So, the sum of the absolute deviations from 10 is 28. Likewise, the sum of the absolute deviations from 5 equals 3 + 2 + 1 + 4 + 11 = 21. So, the sum of the absolute deviations from 5 is smaller than the sum of the absolute deviations from 10. In this sense, 5 is closer, overall, to the other numbers than is 10.

We are now in a position to define a second measure of central tendency, this time in terms of absolute deviations. Specifically, according to our second definition, the center of a distribution is the number for which the sum of the absolute deviations is smallest. As we just saw, the sum of the absolute deviations from 10 is 28 and the sum of the absolute deviations from 5 is 21. Is there a value for which the sum of the absolute deviations is even smaller than 21? Yes. For these data, there is a value for which the sum of absolute deviations is only 20. See if you can find it.

Smallest Squared Deviation

We shall discuss one more way to define the center of a distribution. It is based on the concept of the sum of squared deviations (differences). Again, consider the distribution of the five numbers 2, 3, 4, 9, 16. Table 3 shows the sum of the squared deviations of these numbers from the number 10.

Table 3. An example of the sum of squared deviations.

Values	Squared Deviations from 10
2	64
3	49
4	36
9	1
16	36
Sum	186

The first row in the table shows that the squared value of the difference between 2 and 10 is 64; the second row shows that the squared difference between 3 and 10 is 49, and so forth. When we add up all these squared deviations, we get 186. Changing the target from 10 to 5, we calculate the sum of the squared deviations from 5 as 9 + 4 + 1 + 16 + 121 = 151. So, the sum of the squared deviations from 5 is smaller than the sum of the squared deviations from 10. Is there a value for which the sum of the squared deviations is even smaller than 151? Yes, it is possible to reach 134.8. Can you find the target number for which the sum of squared deviations is 134.8?

The target that minimizes the sum of squared deviations provides another useful definition of central tendency (the last one to be discussed in this section). It can be challenging to find the value that minimizes this sum.

Measures of Central Tendency

by David M. Lane

Prerequisites
- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Central Tendency

Learning Objectives
1. Compute mean
2. Compute median
3. Compute mode

In the previous section we saw that there are several ways to define central tendency. This section defines the three most common measures of central tendency: the mean, the median, and the mode. The relationships among these measures of central tendency and the definitions given in the previous section will probably not be obvious to you.

This section gives only the basic definitions of the mean, median and mode. A further discussion of the relative merits and proper applications of these statistics is presented in a later section.

Arithmetic Mean

The arithmetic mean is the most common measure of central tendency. It is simply the sum of the numbers divided by the number of numbers. The symbol "μ" is used for the mean of a population. The symbol "M" is used for the mean of a sample. The formula for μ is shown below:

$$\mu = \frac{\sum X}{N}$$

where ΣX is the sum of all the numbers in the population and N is the number of numbers in the population.

The formula for M is essentially identical:

$$M = \frac{\sum X}{N}$$

where ΣX is the sum of all the numbers in the sample and N is the number of numbers in the sample.

As an example, the mean of the numbers 1, 2, 3, 6, 8 is 20/5 = 4 regardless of whether the numbers constitute the entire population or just a sample from the population.

Table 1 shows the number of touchdown (TD) passes thrown by each of the 31 teams in the National Football League in the 2000 season. The mean number of touchdown passes thrown is 20.4516 as shown below.

$$\mu = \frac{\sum X}{N} = \frac{634}{31} = 20.4516$$

Table 1. Number of touchdown passes.

37, 33, 33, 32, 29, 28,
28, 23, 22, 22, 22, 21,
21, 21, 20, 20, 19, 19,
18, 18, 18, 18, 16, 15,
14, 14, 14, 12, 12, 9, 6

Although the arithmetic mean is not the only “mean” (there is also a geometric mean), it is by far the most commonly used. Therefore, if the term “mean” is used without specifying whether it is the arithmetic mean, the geometric mean, or some other mean, it is assumed to refer to the arithmetic mean.

Median

The median is also a frequently used measure of central tendency. The median is the midpoint of a distribution: the same number of scores is above the median as below it. For the data in Table 1, there are 31 scores. The 16th highest score (which equals 20) is the median because there are 15 scores below the 16th score and 15 scores above the 16th score. The median can also be thought of as the 50th percentile.

Computation of the Median

When there is an odd number of numbers, the median is simply the middle number. For example, the median of 2, 4, and 7 is 4. When there is an even number of numbers, the median is the mean of the two middle numbers. Thus, the median of the numbers 2, 4, 7, 12 is:

$$\frac{(4+7)}{2} = 5.5$$

When there are numbers with the same values, then the formula for the third definition of the 50th percentile should be used.

Mode

The mode is the most frequently occurring value. For the data in Table 1, the mode is 18 since more teams (4) had 18 touchdown passes than any other number of touchdown passes. With continuous data, such as response time measured to many decimals, the frequency of each value is one since no two scores will be exactly the same (see discussion of continuous variables). Therefore the mode of continuous data is normally computed from a grouped frequency distribution. Table 2 shows a grouped frequency distribution for the target response time data. Since the interval with the highest frequency is 600-700, the mode is the middle of that interval (650).

Table 2. Grouped frequency distribution

Range	Frequency
500-600	3
600-700	6
700-800	5
800-900	5
900-1000	0
1000-1100	1

Median and Mean

by David M. Lane

Prerequisites

- Chapter 3: What is Central Tendency
- Chapter 3: Measures of Central Tendency

Learning Objectives

1. State when the mean and median are the same
2. State whether it is the mean or median that minimizes the mean absolute deviation
3. State whether it is the mean or median that minimizes the mean squared deviation
4. State whether it is the mean or median that is the balance point on a balance scale

In the section "What is central tendency," we saw that the center of a distribution could be defined three ways: (1) the point on which a distribution would balance, (2) the value whose average absolute deviation from all the other values is minimized, and (3) the value whose squared difference from all the other values is minimized. The mean is the point on which a distribution would balance, the median is the value that minimizes the sum of absolute deviations, and the mean is the value that minimizes the sum of the squared deviations.

Table 1 shows the absolute and squared deviations of the numbers 2, 3, 4, 9, and 16 from their median of 4 and their mean of 6.8. You can see that the sum of absolute deviations from the median (20) is smaller than the sum of absolute deviations from the mean (22.8). On the other hand, the sum of squared deviations from the median (174) is larger than the sum of squared deviations from the mean (134.8).

Table 1. Absolute and squared deviations from the median of 4 and the mean of 6.8.

Value	Absolute Deviation from Median	Absolute Deviation from Mean	Squared Deviation from Median	Squared Deviation from Mean
2	2	4.8	4	23.04
3	1	3.8	1	14.44
4	0	2.8	0	7.84
9	5	2.2	25	4.84
16	12	9.2	144	84.64
Total	20	22.8	174	134.8

Figure 1 shows that the distribution balances at the mean of 6.8 and not at the median of 4. The relative advantages and disadvantages of the mean and median are discussed in the section "Comparing Measures" later in this chapter.

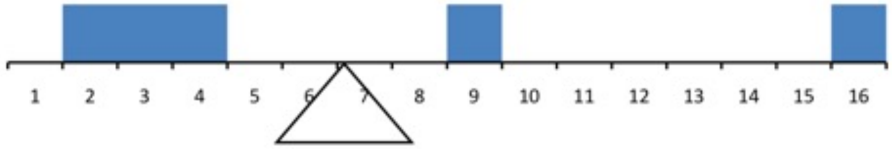


Figure 1. The distribution balances at the mean of 6.8 and not at the median of 4.0.

When a distribution is symmetric, then the mean and the median are the same. Consider the following distribution: 1, 3, 4, 5, 6, 7, 9. The mean and median are both 5. The mean, median, and mode are identical in the bell-shaped normal distribution.

Measures of Variability

by David M. Lane

Prerequisites

- Chapter 1: Percentiles
- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency

Learning Objectives

1. Determine the relative variability of two distributions
2. Compute the range
3. Compute the inter-quartile range
4. Compute the variance in the population
5. Estimate the variance from a sample
6. Compute the standard deviation from the variance

What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider graphs in Figure 1. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.

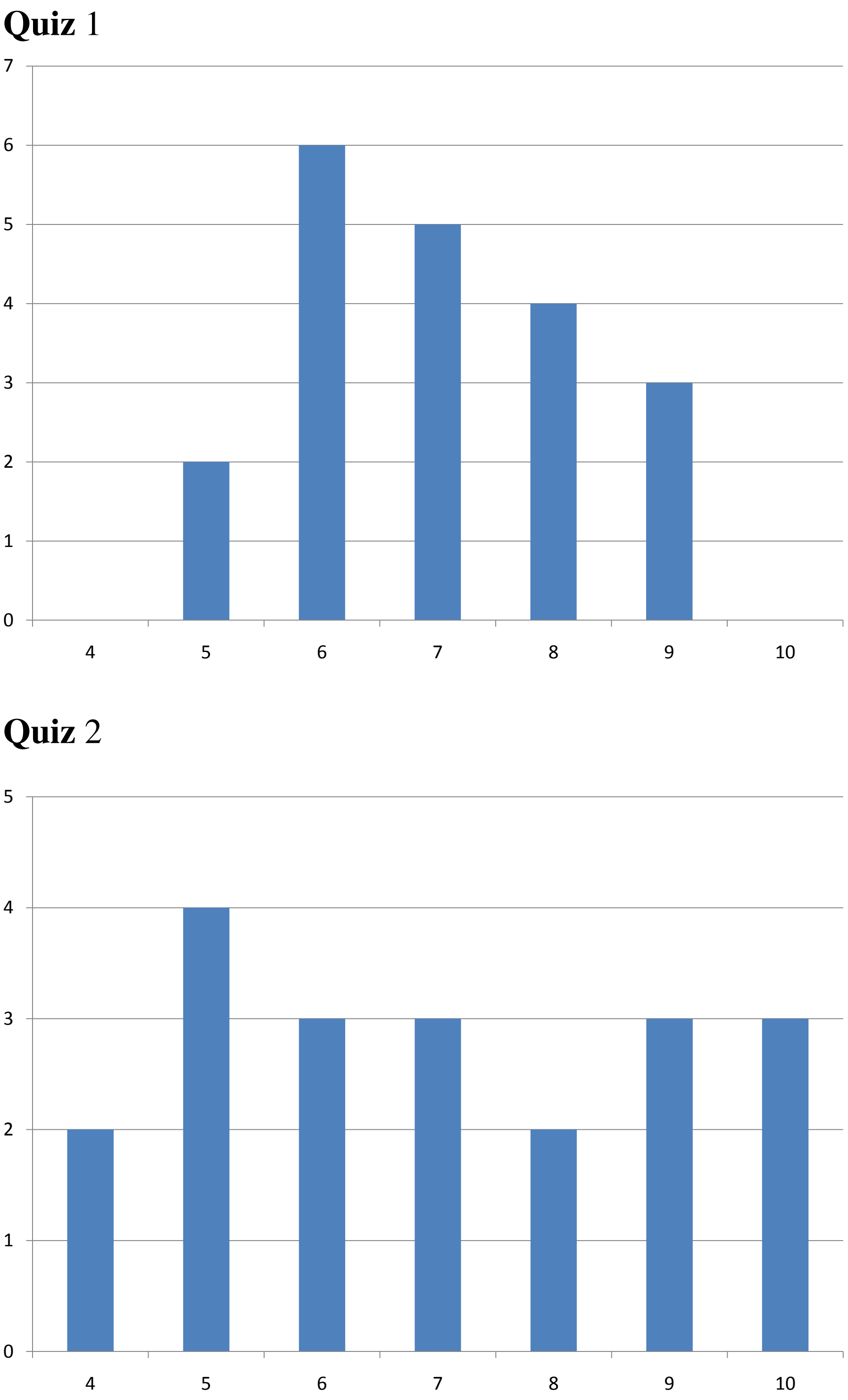


Figure 1. Bar charts of two quizzes.

The terms variability, spread, and dispersion are synonyms, and refer to how spread out a distribution is. Just as in the section on central tendency where we discussed measures of the center of a distribution of scores, in this chapter we will discuss measures of the variability of a distribution. There are four frequently used measures of variability: range, interquartile range, variance, and standard deviation. In the next few paragraphs, we will look at each of these four measures of variability in more detail.

Range

The range is the simplest measure of variability to calculate, and one you have probably encountered many times in your life. The range is simply the highest score minus the lowest score. Let's take a few examples. What is the range of the following group of numbers: 10, 2, 5, 6, 7, 3, 4? Well, the highest number is 10, and the lowest number is 2, so 10 - 2 = 8. The range is 8. Let's take another example. Here's a dataset with 10 numbers: 99, 45, 23, 67, 45, 91, 82, 78, 62, 51. What is the range? The highest number is 99 and the lowest number is 23, so 99 - 23 equals 76; the range is 76. Now consider the two quizzes shown in Figure 1. On Quiz 1, the lowest score is 5 and the highest score is 9. Therefore, the range is 4. The range on Quiz 2 was larger: the lowest score was 4 and the highest score was 10. Therefore the range is 6.

Interquartile Range

The interquartile range (IQR) is the range of the middle 50% of the scores in a distribution. It is computed as follows:

```
IQR = 75th percentile - 25th percentile
```

For Quiz 1, the 75th percentile is 8 and the 25th percentile is 6. The interquartile range is therefore 2. For Quiz 2, which has greater spread, the 75th percentile is 9, the 25th percentile is 5, and the interquartile range is 4. Recall that in the discussion of box plots, the 75th percentile was called the upper hinge and the 25th percentile was called the lower hinge. Using this terminology, the interquartile range is referred to as the H-spread.

A related measure of variability is called the semi-interquartile range. The semi-interquartile range is defined simply as the interquartile range divided by 2. If a distribution is symmetric, the median plus or minus the semi-interquartile range contains half the scores in the distribution.

Variance

Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean. The data from Quiz 1 are shown in Table 1. The mean score is 7.0. Therefore, the column "Deviation from Mean" contains the score minus 7. The column "Squared Deviation" is simply the previous column squared.

Table 1. Calculation of Variance for Quiz 1 scores.

Scores	Deviation from Mean	Squared Deviation
9	2	4
9	2	4
9	2	4
8	1	1
8	1	1
8	1	1
8	1	1
7	0	0
7	0	0
7	0	0
7	0	0
7	0	0
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
6	-1	1
5	-2	4
5	-2	4
Means		
7	0	1.5

One thing that is important to notice is that the mean deviation from the mean is 0. This will always be the case. The mean of the squared deviations is 1.5. Therefore, the variance is 1.5. Analogous calculations with Quiz 2 show that its variance is 6.7. The formula for the variance is:

$$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$$

where σ^2 is the variance, μ is the mean, and N is the number of numbers. For Quiz 1, $\mu = 7$ and $N = 20$.

If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

where s^2 is the estimate of the variance and M is the sample mean. Note that M is the mean of a sample taken from a population with a mean of μ. Since, in practice, the variance is usually computed in a sample, this formula is most often used.

Let's take a concrete example. Assume the scores 1, 2, 4, and 5 were sampled from a larger population. To estimate the variance in the population you would compute s^2 as follows:

$$M = \frac{1+2+4+5}{4} = \frac{12}{4} = 3$$

$$s^2 = \frac{(1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2}{4-1} = \frac{4+1+1+4}{3} = \frac{10}{3} = 3.333$$

There are alternate formulas that can be easier to use if you are doing your calculations with a hand calculator:

$$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$$

and

$$s^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1}$$

For this example,

$$\left(\sum X\right)^2 = \frac{(1+2+4+5)^2}{4} = \frac{144}{4} = 36$$

$$\sigma^2 = \frac{(46-36)}{4} = 2.5$$

$$s^2 = \frac{(46-36)}{3} = 3.333$$

as with the other formula.

Standard Deviation

The standard deviation is simply the square root of the variance. This makes the standard deviations of the two quiz distributions 1.225 and 2.588. The standard deviation is an especially useful measure of variability when the distribution is normal or approximately normal (see Chapter 7) because the proportion of the distribution within a given number of standard deviations from the mean can be calculated. For example, 68% of the distribution is within one standard deviation of the mean and approximately 95% of the distribution is within two standard deviations of the mean. Therefore, if you had a normal distribution with a mean of 50 and a standard deviation of 10, then 68% of the distribution would be between 50 - 10 = 40 and 50 +10 =60. Similarly, about 95% of the distribution would be between 50 - 2 x 10 = 30 and 50 + 2 x 10 = 70. The symbol for the population standard deviation is σ; the symbol for an estimate computed in a sample is s. Figure 2 shows two normal distributions. The red distribution has a mean of 40 and a standard deviation of 5; the blue distribution has a mean of 60 and a standard deviation of 10. For the red distribution, 68% of the distribution is between 45 and 55; for the blue distribution, 68% is between 50 and 70.

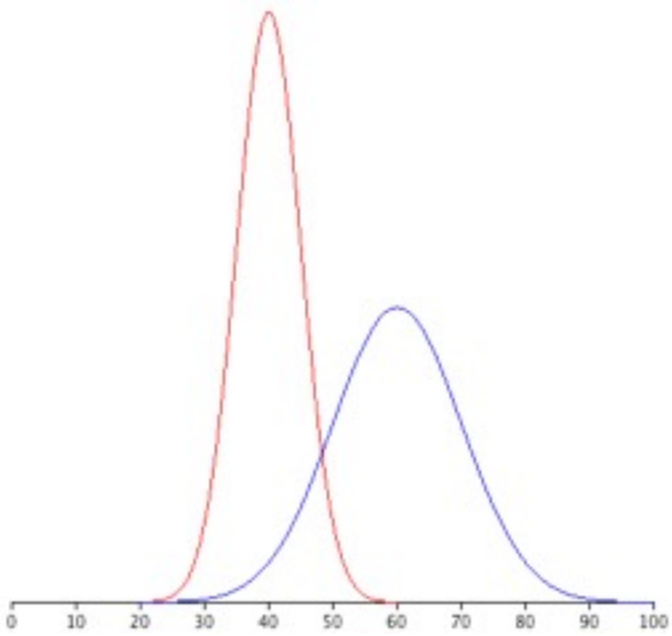


Figure 2. Normal distributions with standard deviations of 5 and 10.

Shapes of Distributions

by David M. Lane

Prerequisites

- Chapter 1: Distributions
- Chapter 3: Measures of Central Tendency
- Chapter 3: Variability

Learning Objectives

1. Compute skew using two different formulas
2. Compute kurtosis

We saw in the section on distributions in Chapter 1 that shapes of distributions can differ in skew and/or kurtosis. This section presents numerical indexes of these two measures of shape.

Skew

Figure 1 shows a distribution with a very large positive skew. Recall that distributions with positive skew have tails that extend to the right.

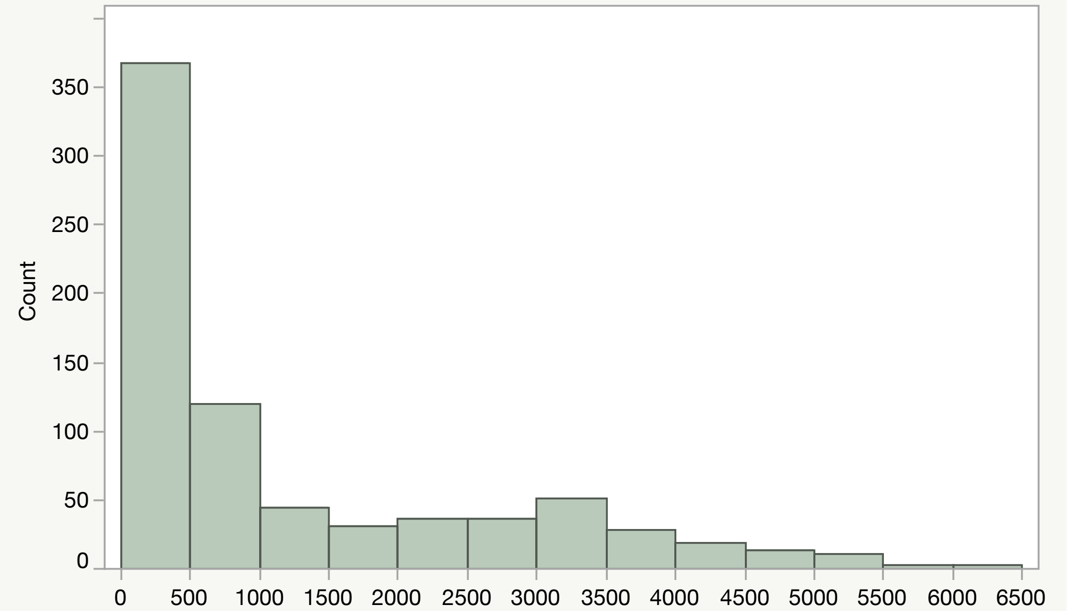


Figure 1. A distribution with a very large positive skew. This histogram shows the salaries of major league baseball players (in thousands of dollars).

Distributions with positive skew normally have larger means than medians. The mean and median of the baseball salaries shown in Figure 1 are $1,183,417 and $500,000 respectively. Thus, for this highly-skewed distribution, the mean is more than twice as high as the median. The relationship between skew and the relative size of the mean and median lead the statistician Pearson to propose the following simple and convenient numerical index of skew:

$$\frac{3(Mean - Median)}{\sigma}$$

The standard deviation of the baseball salaries is 1,390,922. Therefore, Pearson's measure of skew for this distribution is 3(1,183,417 - 500,000)/1,390,922 = 1.47.

Just as there are several measures of central tendency, there is more than one measure of skew. Although Pearson's measure is a good one, the following measure is more commonly used. It is sometimes referred to as the third moment about the mean.

$$\sum \frac{(X - \mu)^3}{\sigma^3}$$

Kurtosis

The following measure of kurtosis is similar to the definition of skew. The value “3” is subtracted to define “no kurtosis” as the kurtosis of a normal distribution. Otherwise, a normal distribution would have a kurtosis of 3.

$$\sum \frac{(X - \mu)^4}{\sigma^4} - 3$$

Relevant exercises:
1, 2, 3, 4, 5, 6 (leave out trimmed mean), 7, 9, 10, 11, 12, 13, 14, 16, 19

Exercises

Prerequisites

- All material presented in the Summarizing Distributions chapter

1. Make up a dataset of 12 numbers with a positive skew. Use a statistical program to compute the skew. Is the mean larger than the median as it usually is for distributions with a positive skew? What is the value for skew?

2. Repeat Problem 1 only this time make the dataset have a negative skew.

3. Make up three data sets with 5 numbers each that have:

   (a) the same mean but different standard deviations.

   (b) the same mean but different medians.

   (c) the same median but different means.

4. Find the mean and median for the following three variables:

A	B	C
8	4	6
5	4	2
7	6	3
1	3	4
3	4	1

5. A sample of 30 distance scores measured in yards has a mean of 10, a variance of 9, and a standard deviation of 3 (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation? (b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?

6. You recorded the time in seconds it took for 8 participants to solve a puzzle. These times appear below. However, when the data was entered into the statistical program, the score that was supposed to be 22.1 was entered as 21.2.

You had calculated the following measures of central tendency: the mean, the median, and the mean trimmed 25%. Which of these measures of central tendency will change when you correct the recording error?

Time (seconds)
15.2
18.8
19.3
19.7
20.2
21.8
22.1
29.4

7. For the test scores in question #6, which measures of variability (range, standard deviation, variance) would be changed if the 22.1 data point had been erroneously recorded as 21.2?

8. You know the minimum, the maximum, and the 25th, 50th, and 75th percentiles of a distribution. Which of the following measures of central tendency or variability can you determine?

   mean, median, mode, trimean, geometric mean, range, interquartile range, variance, standard deviation

9. For the numbers 1, 3, 4, 6, and 12:

   Find the value (v) for which $\Sigma(X\text{-}v)^2$ is minimized.

   Find the value (v) for which $\Sigma|x\text{-}v|$ is minimized.

10. Your younger brother comes home one day after taking a science test. He says that some- one at school told him that “60% of the students in the class scored above the median test grade.” What is wrong with this statement? What if he had said “60% of the students scored below the mean?”

11. An experiment compared the ability of three groups of participants to remember briefly- presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess

positions. Compare the performance of each group. Consider spread as well as central tendency.

Non-players	Beginners	Tournament players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

12. True/False: A bimodal distribution has two modes and two medians.

13. True/False: The best way to describe a skewed distribution is to report the mean.

14. True/False: When plotted on the same graph, a distribution with a mean of 50 and a standard deviation of 10 will look more spread out than will a distribution with a mean of 60 and a standard deviation of 5.

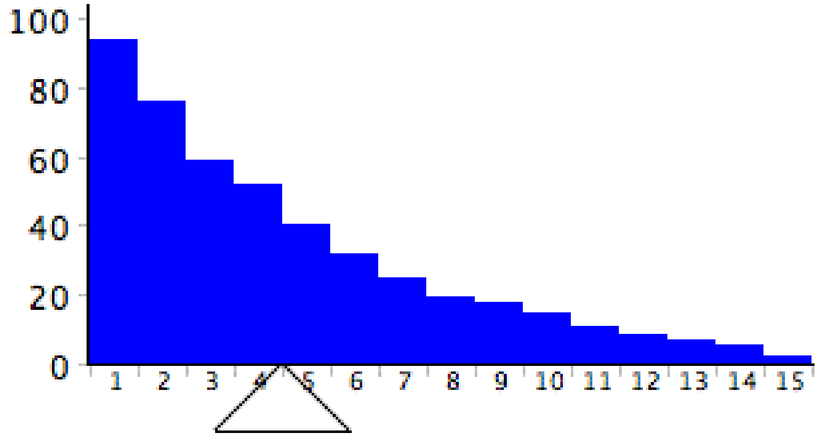
15. Compare the mean, median, trimean in terms of their sensitivity to extreme scores.

16. If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?

17. A set of numbers is transformed by taking the log base 10 of each number. The mean of the transformed data is 1.65. What is the geometric mean of the untransformed data?

18. Which measure of central tendency is most often used for returns on investment?

19. The histogram is in balance on the fulcrum. What are the mean, median, and mode of the distribution (approximate where necessary)?



Questions from Case Studies

Angry Moods (AM) case study

20. (AM) Does Anger-Out have a positive skew, a negative skew, or no skew?

21. (AM) What is the range of the Anger-In scores? What is the interquartile range?

22. (AM) What is the overall mean Control-Out score? What is the mean Control-Out score for the athletes? What is the mean Control-Out score for the non-athletes?

23. (AM) What is the variance of the Control-In scores for the athletes? What is the variance of the Control-In scores for the non-athletes?

Flatulence (F) case study

24. (F) Based on a histogram of the variable “perday”, do you think the mean or median of this variable is larger? Calculate the mean and median to see if you are right.

Stroop (S) case study

25.(S) Compute the mean for “words”.

26. (S#2) Compute the mean and standard deviation for “colors”.

Physicians’ Reactions (PR) case study

27.(PR) What is the mean expected time spent for the average-weight patients? What is the mean expected time spent for the overweight patients?

28.(PR) What is the difference in means between the groups? By approximately how many standard deviations do the means differ?

Smiles and Leniency (SL) case study

29.(SL) Find the mean, median, standard deviation, and interquartile range for the leniency scores of each of the four groups.

ADHD Treatment (AT) case study

30.(AT) What is the mean number of correct responses of the participants after taking the placebo (0 mg/kg)?

31.(AT) What are the standard deviation and the interquartile range of the d0 condition?