people who are HIV-positive not to take the oral vaccine. The second group would likely take a placebo.

**b.** This would have been a between-groups experiment because the people who are HIV-positive would have been in only one group: either vaccine or no vaccine.

**c.** This limits the researchers' ability to draw causal conclusions because the participants who received the vaccine may have been different in some way from those who did not receive the vaccine. There may have been a confounding variable that led to these findings. For example, those who received the vaccine might have had better access to health care and better sanitary conditions to begin with, making them less likely to contract cholera regardless of the vaccine's effectiveness.

**d.** The researchers might not have used random assignment because it would have meant recruiting participants, likely immunizing half, then following up with all of them. The researchers likely did not want to deny the vaccine to people who were HIV-positive because they might have contracted cholera and died without it.

**1-44** **a.** Ability level, graduate level (high school versus university), and race

**b.** Wages

**c.** 12,000 men and women in the United States who were between 14 and 22 years old in 1979

**d.** All high school and college graduate men and women in the United States

**e.** Participants were studied over time to measure change during that period.

**f.** Age could be a confounding variable, as those who are older will have more exposure to the various areas measured via the AFQT, in addition to the education they received at the college level.

**g.** Ability could be operationalized by having managers rate each participant's ability to perform his or her job. Another way ability could be operationalized is via high school and college grades or a standardized ability test.

**1-45** **a.** A "good charity" is operationally defined as one that spends more of its money for the cause it is supporting and less for fundraising or administration.

**b.** The rating is a scale variable, as it has a meaningful zero point, has equal distance between intervals, and is continuous.

**c.** The tier is an ordinal variable, as it involves ranking the organizations into categories (1st, 2nd, 3rd, 4th, or 5th tier) and it is discrete.

**d.** The type of charity is a nominal variable, as it uses names or categories to classify the values (e.g., health and medical needs) and it is discrete.

**e.** Measuring finances is more objective and easier to measure than some of the criteria mentioned by Ord, such as importance of the problem and competency and honesty.

**f.** Charity Navigator's ratings are more likely to be reliable than GiveWell's ratings because they are based on an objective measure. It is more likely that different assessors would come up with the same rating for Charity Navigator than for GiveWell.

**g.** GiveWell's ratings are likely to be more valid than Charity Navigator's, provided that they can attain some level of reliability. GiveWell's more comprehensive rating system incorporates a better-rounded assessment of a charity.

**h.** This would be a correlational study because donation funds, the independent variable, would not be randomly assigned based on country but measured as they naturally occur.

**i.** This would be an experiment because the levels of donation funds, the independent variable, are randomly assigned to different regions to determine the effect on death rate.

# Chapter 2

**2-1** Raw scores are the original data, to which nothing has been done.

**2-2** To create a frequency table: (1) Determine the highest and lowest scores. (2) Create two columns; label the first with the variable name and label the second "Frequency." (3) List the full range of values that encompasses all the scores in the data set, from lowest to highest, even those for which the frequency is 0. (4) Count the number of scores at each value, and write those numbers in the frequency column.

**2-3** A frequency table is a visual depiction of data that shows how often each value occurred; that is, it shows how many scores are at each value. Values are listed in one column, and the numbers of individuals with scores at that value are listed in the second column. A grouped frequency table is a visual depiction of data that reports the frequency within each given interval, rather than the frequency for each specific value.

**2-4** Statisticians might use *interval* to describe a type of variable. Interval variables have numbers as their values, and the distance (or interval) between numbers is assumed to be equal. Statisticians might also use *interval* to refer to the range of values to be used in a grouped frequency table, histogram, or polygon.

**2-5** Bar graphs typically provide scores for nominal data, whereas histograms typically provide frequencies for scale data. Also, the categories in bar graphs do not need to be arranged in a particular order and the bars should not touch, whereas the intervals in histograms are arranged in a meaningful order (lowest to highest) and the bars should touch each other.

**2-6** The $x$-axis is typically labeled with the name of the variable of interest. The $y$-axis is typically labeled "Frequency."

**2-7** A histogram looks like a bar graph but is usually used to depict scale data, with the values (or midpoints of intervals) of the variable on the $x$-axis and the frequencies on the $y$-axis. A frequency polygon is a line graph, with the $x$-axis representing values (or midpoints of intervals) and the $y$-axis representing frequencies; a dot is placed at the frequency for each value (or midpoint), and the points are connected.

**2-8** Visual displays of data often help us see patterns that are not obvious when we examine a long list of numbers. They help us organize the data in meaningful ways.

**2-9** In everyday conversation, you might use the word *distribution* in a number of different contexts, from the distribution of food to a marketing distribution. A statistician would use *distribution* only to describe the way that a set of scores, such as a set of grades, is distributed. A statistician is looking at the overall pattern of the data—what the shape is, where the data tend to cluster, and how they trail off.

**2-10** A normal distribution is a specific frequency distribution that is a bell-shaped, symmetric, unimodal curve.

**2-11** With positively skewed data, the distribution's tail extends to the right, in a positive direction, and with negatively skewed data, the distribution's tail extends to the left, in a negative direction.

**2-12** A floor effect occurs when there are no scores below a certain value; a floor effect leads to a positively skewed distribution because the lower part of the distribution is constrained.

**2-13** A ceiling effect occurs when there are no scores above a certain value; a ceiling effect leads to a negatively skewed distribution because the upper part of the distribution is constrained.

**2-14** A stem-and-leaf plot retains information about every unique data point in a set, whereas a histogram does not. Additionally, it is easy to create side-by-side stem-and-leaf plots for different groups to compare their distributions. Such a side-by-side comparison of groups is not as easy to do with histograms.

**2-15** A stem-and-leaf plot is much like a histogram in that it conveys how often different values in a data set occur. Also, when a stem-and-leaf plot is turned on its side, it has the same shape as a histogram of the same data set.

**2-16** 4.98% and 2.27%

**2-17** 17.95% and 40.67%

**2-18** 3.69% and 18.11% are scale variables, both as counts and as percentages.

**2-19** 0.10% and 96.77%

**2-20** 1,889.00, 2.65, and 0.08

**2-21** 0.04, 198.22, and 17.89

**2-22** **a.** The full range is the maximum (27) minus the minimum (0), plus 1, which equals 28.

**b.** Five

**c.** The intervals would be 0–4, 5–9, 10–14, 15–19, 20–24, and 25–29.

**2-23** The full range of data is 68 minus 2, plus 1, or 67. The range (67) divided by the desired seven intervals gives us an interval size of 9.57, or 10 when rounded. The seven intervals are: 0–9, 10–19, 20–29, 30–39, 40–49, 50–59, and 60–69.

**2-24** 37.5, 52.5, and 67.5

**2-25** Four countries had at least 30 volcanoes.

**2-26** Fourteen countries had 1 volcano and 12 countries had 2 volcanoes. So, 26 countries had one or two volcanoes.

**2-27** Serial killers would create positive skew, adding high numbers of murders to the data that are clustered around 1.

**2-28** People convicted of murder are assumed to have killed at least one person, so observations below one are not seen, which creates a floor effect.

**2-29** **a.** For the college population, the range of ages extends farther to the right (with a larger number of years) than to the left, creating positive skew.

**b.** The fact that youthful prodigies have limited access to college creates a sort of floor effect that makes low scores less possible.

**2-30** **a.** Assuming that most people go for the maximum number of friends, for the range of Facebook friends, the number of friends extends farther to the left (with fewer number of friends) than to the right, creating a negative skew.

**b.** The fact that Facebook cuts off or limits the number of friends to 5000 means there is a ceiling effect that makes higher scores impossible.

**2-31** **a.** The lines for each stem number need to be combined so that the new stem-and-leaf plot looks like this (the leaves for both 3s, 2s, 1s, and 0s need to be combined):

3 001333444455568888

2 003445778889

1 03688

0 15

**b.** This stem-and-leaf plot depicts a negatively skewed distribution.

**2-32** **a.** The stem-and-leaf plot is depicted below:

4 00055

3 00005555555

2 00555

**b.** This stem-and-leaf plot depicts a symmetric distribution.

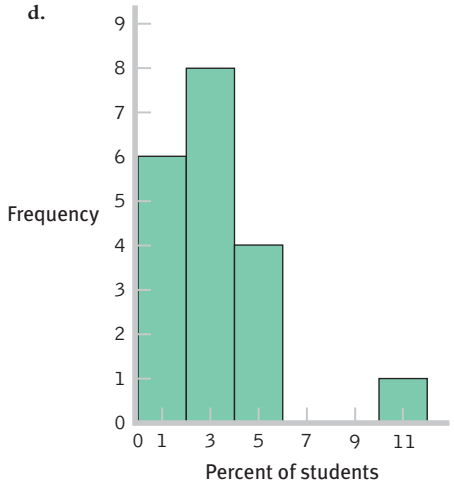
**2-33** **a.**

| Percentage | Frequency | Percentage |
|---|---|---|
| 10 | 1 | 5.26 |
| 9 | 0 | 0.00 |
| 8 | 0 | 0.00 |
| 7 | 0 | 0.00 |
| 6 | 0 | 0.00 |
| 5 | 2 | 10.53 |
| 4 | 2 | 10.53 |
| 3 | 4 | 21.05 |
| 2 | 4 | 21.05 |
| 1 | 5 | 26.32 |
| 0 | 1 | 5.26 |

**b.** In 10.53% of these schools, exactly 4% of the students reported that they wrote between 5 and 10 twenty-page papers that year.

**c.** This is not a random sample. It includes schools that chose to participate in this survey and opted to have their results made public.

**d.**




**e.** One

**f.** The data are clustered around 1% to 4%, with a high outlier, 10%.

**2-34** **a.**

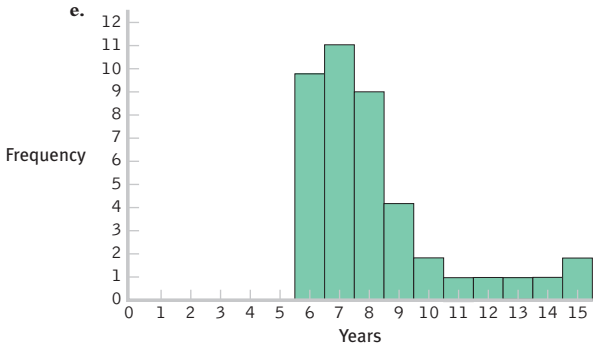
| Years to Complete | Frequency |
|---|---|
| 15 | 2 |
| 14 | 1 |
| 13 | 1 |
| 12 | 1 |
| 11 | 1 |
| 10 | 2 |
| 9 | 4 |
| 8 | 9 |
| 7 | 11 |
| 6 | 10 |

**b.** 30

**c.** A grouped frequency table is not necessary here. These data are relatively easy to interpret in the frequency table. Grouped frequency tables are useful when the list of data is long and difficult to interpret.

**d.** These data are clustered around 6 to 8 years, with a long tail of data out to a greater number of years to complete. These data show positive skew.

**e.**



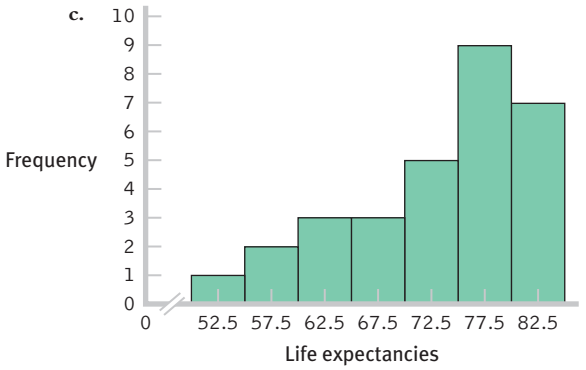

**f.** Eight

**2-35** **a.**

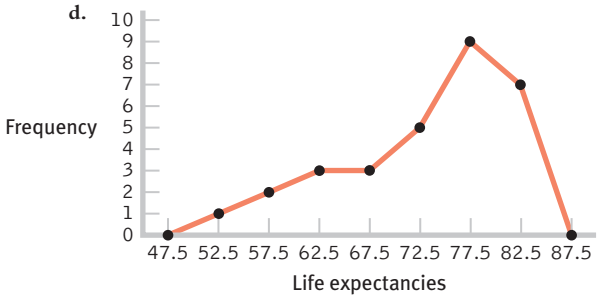
| Interval | Frequency |
|---|---|
| 80–84 | 7 |
| 75–79 | 9 |
| 70–74 | 5 |
| 65–69 | 3 |
| 60–64 | 3 |
| 55–59 | 2 |
| 50–54 | 1 |

**b.** There are many possible answers. One research hypothesis might be that the economic status of different regions of the world predicts life expectancies.

**c.**




**d.**

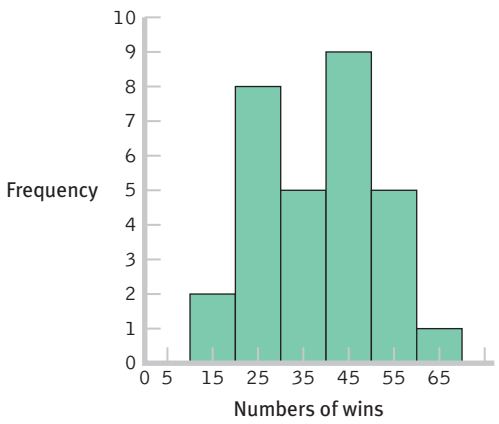



**e.** The data presented here demonstrate a negatively skewed distribution. There are more countries that have a life expectancy on the higher end—longer than 70 years—than there are in the rest of the distribution.

**2-36** **a.**

| Interval | Frequency |
|---|---|
| 60–69 | 1 |
| 50–59 | 5 |
| 40–49 | 9 |
| 30–39 | 5 |
| 20–29 | 8 |
| 10–19 | 2 |

**b.**




**c.** The summary will differ for each student but should include the following information: The data appear to be roughly symmetric.

**d.** With so few data points, it is easy to view patterns in the data without grouping the data into intervals.

**2-37** **a.** Extroversion scores are most likely to have a normal distribution. Most people would fall toward the middle, with some people having higher levels and some having lower levels.

**b.** The distribution of finishing times for a marathon is likely to be positively skewed. The floor is the fastest possible time, a little over 2 hours; however, some runners take as long as 6 hours or more. Unfortunately for the very, very slow but unbelievably dedicated runners, many marathons shut down the finish line 6 hours after the start of the race.

**c.** The distribution of numbers of meals eaten in a dining hall in a semester on a three-meal-a-day plan is likely to be negatively skewed. The ceiling is three times per day, multiplied by the number of days; most people who choose to pay for the full plan would eat many of these meals. A few would hardly ever eat in the dining hall, pulling the tail in a negative direction.

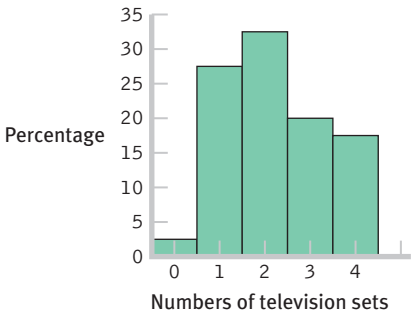
**2-38** **a.** You would present individual data values because the few categories of eye color would result in a readable list. A frequency table would be most appropriate.

**b.** You would present grouped data because it is possible for each person to use a different number of minutes and such a long list would be unreadable. A grouped frequency table, histogram, or frequency polygon would be most appropriate.

**c.** You would present grouped data because time to complete carried out to seconds would produce too many unique numbers to organize meaningfully without groupings. A grouped frequency table, histogram, or frequency polygon would be most appropriate.

**d.** You would present individual data values because number of siblings tends to take on limited values. A frequency table, histogram, or frequency polygon would be most appropriate.

**2-39**




**2-40** **a.** This is a grouped frequency table because each row includes an interval of frequencies rather than a single frequency.

**b.** The intervals are not all the same size. The researchers likely created the table this way because it helps tell a story. If they had created intervals of, say, 500,000 each, we would not gain information about how many surnames are very infrequent—those toward the bottom of this table.

**c.** Based on the data in this table, the distribution is positively skewed. The data trail off in the positive direction with only a few extremely common last names.

**d.** There is a floor effect. A name cannot occur fewer than one time, so the data are bunched up around the low-frequency surnames.

**2-41** **a.** A frequency polygon based on these data is likely to be negatively skewed. The scale is 1–10 and most films are rated above the midpoint. Very few are as low as *Gunday*.

**b.** There is more likely to be a ceiling effect. With most films earning high ratings, it seems that the limiting factor is the top score of 10. No film earned the lowest possible score of 1, and few were as low as *Gunday's* 1.4. So, there doesn't seem to be a floor effect of 1.

**c.** IMDb ratings don't seem to be a good way to operationalize movie quality. Audience ratings may be based on something other than how good the film is. In this case, many of those who rated *Gunday* based their scores on politics rather than on the qualities of the film itself. Another way to operationalize movie quality is a rating based on critics' reviews, such as the system used by rottentomatoes.com. This site provides an average rating from critics, based on published reviews, in addition to one by movie audiences. Critics are unlikely to rate a movie simply based on politics.

**2-42** The stem-and-leaf plot is depicted below:

3 3
2 0026
1 89
0 2

**2-43** The stem-and-leaf plot is depicted below:

6 0
5 01367
4 002234489
3 13779
2 33567999
1 89

**2-44** **a.** A histogram of grouped frequencies

**b.** Approximately 32

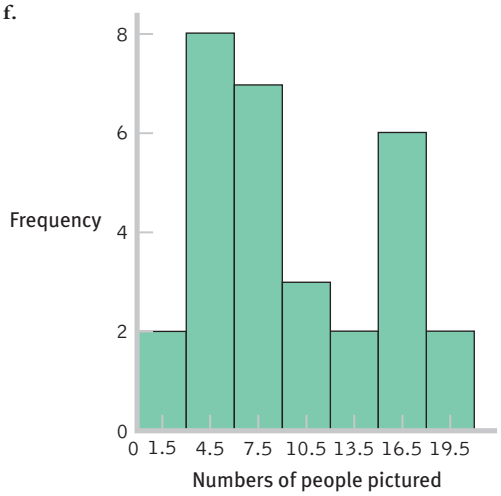
**c.** Approximately 27

**d.** Answers will vary, but two questions we might ask are (1) How close is the person to those photographed? and (2) What might account for the two peaks in these data?

**e.**

| Interval | Frequency |
|---|---|
| 18–20 | 2 |
| 15–17 | 6 |
| 12–14 | 2 |
| 9–11 | 3 |
| 6–8 | 7 |
| 3–5 | 8 |
| 0–2 | 2 |

**f.**




**g.** The data have two high points around 3–9 and 15–18. We can see that the data are asymmetric to the right, creating positive skew.

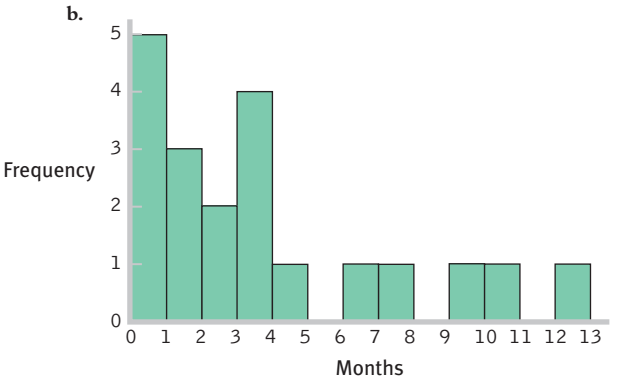
**h.** The stem-and-leaf plot is depicted below:
2 0
1 01235566778
0 013333345567777889

**i.** These data reflect a floor effect because most of the observations are clustered on the lower end of the distribution between 0 and 9. This floor effect is likely caused by the fact that people cannot have fewer than 0 pictures of others.

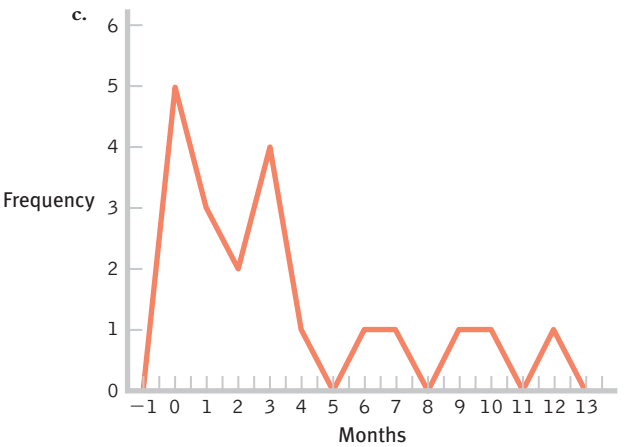
**2-45** **a.**

| Months | Frequency | Percentage |
|---|---|---|
| 12 | 1 | 5 |
| 11 | 0 | 0 |
| 10 | 1 | 5 |
| 9 | 1 | 5 |
| 8 | 0 | 0 |
| 7 | 1 | 5 |
| 6 | 1 | 5 |
| 5 | 0 | 0 |
| 4 | 1 | 5 |
| 3 | 4 | 20 |
| 2 | 2 | 10 |
| 1 | 3 | 15 |
| 0 | 5 | 25 |

**b.**



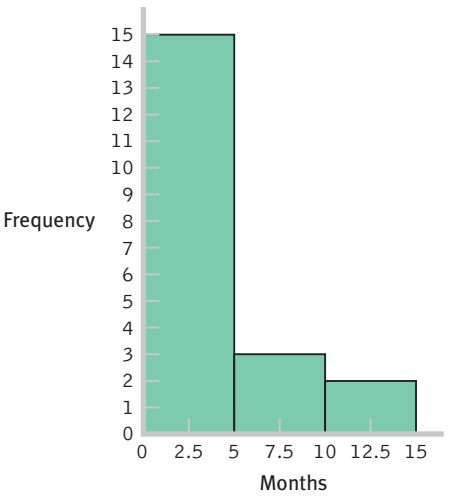

**c.**



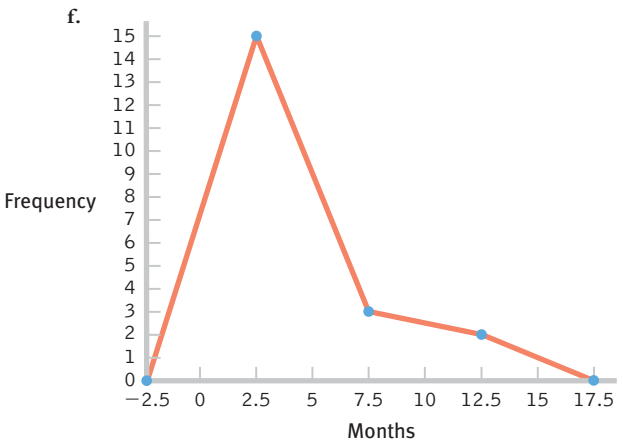
**d.**

| Interval | Frequency |
|---|---|
| 10–14 months | 2 |
| 5–9 months | 3 |
| 0–4 months | 15 |

**e.**




**f.**




**g.** These data are centered around the 3-month period, with positive skew extending the data out to the 12-month period.

**h.** The bulk of the data would need to be shifted from the 3-month period to approximately 12 months, so the women who have breast-fed for 3 months so far might be the focus of attention. Perhaps early contact at the hospital and at follow-up visits after birth would help encourage mothers to breast-feed, and to breast-feed longer. One could also consider studying the women who create the positive skew to learn what unique characteristics or knowledge they have that influenced their behavior.

**2-46** **a.** The column for faculty shows a high point from 0–7 friends.

**b.** The column for students shows two high points around 4–11 and 16–23, with some high outliers creating positive skew.

**c.** The independent variable would be status, with two levels (faculty, student).

**d.** The dependent variable would be number of friends.

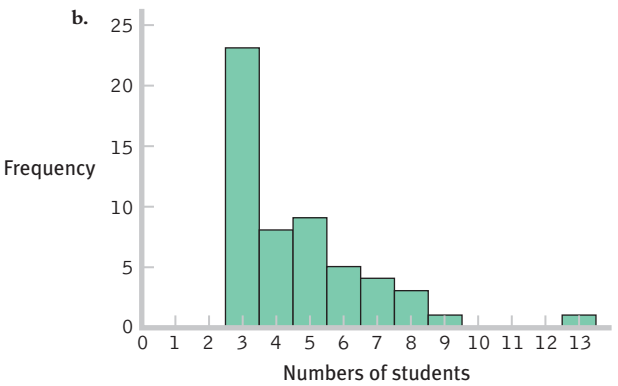
**e.** A confounding variable could be age, as faculty are older than students and tend to be less involved in social activities or situations where making friends is common.

**f.** The dependent variable was operationalized as the number of people who appear in photographs on display in dorm rooms and offices across campus. There are several additional ways these data could be operationalized. One way would be to record the number of Facebook friends each person has. Another way would be to count the number of friends each person reports interacting with on a regular basis. This latter method of measuring number of friends is more likely to reveal the quality of friendship via the amount of interaction.

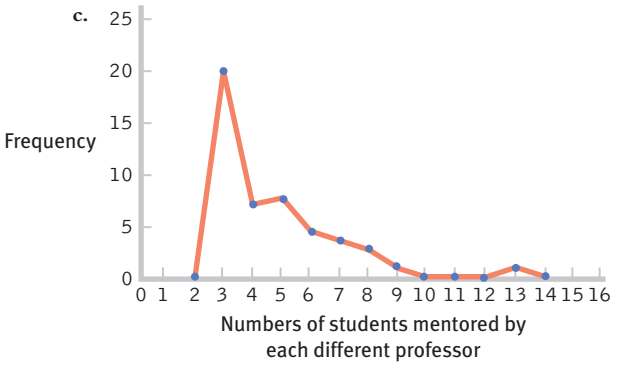
**2-47** **a.**

| Former Students Now in Top Jobs | Frequency | Percentage |
|---|---|---|
| 13 | 1 | 1.85 |
| 12 | 0 | 0.00 |
| 11 | 0 | 0.00 |
| 10 | 0 | 0.00 |
| 9 | 1 | 1.85 |
| 8 | 3 | 5.56 |
| 7 | 4 | 7.41 |
| 6 | 5 | 9.26 |
| 5 | 9 | 16.67 |
| 4 | 8 | 14.81 |
| 3 | 23 | 42.59 |

**b.**



**c.**




**d.** This distribution is positively skewed.

**e.** The researchers operationalized the variable of mentoring success as numbers of students placed into top professorial positions. There are many other ways this variable could have been operationalized. For example, the researchers might have counted numbers of student publications while in graduate school or might have asked graduates to rate their satisfaction with their graduate mentoring experiences.

**f.** The students might have attained their positions as professors because of the prestige of their advisor, not because of his mentoring.

**g.** There are many possible answers to this question. For example, the attainment of a top professorial position might be predicted by the prestige of the institution, the number of publications while in graduate school, or the graduate student's academic ability.

# Chapter 3

**3-1** The five techniques for misleading with graphs are the biased scale lie, the sneaky sample lie, the interpolation lie, the extrapolation lie, and the inaccurate values lie.

**3-2** (1) Organize the data by participant; each participant will have two scores, one on each scale variable. (2) Label the horizontal *x*-axis with the name of the independent variable and its possible values, starting with 0 if practical. (3) Label the vertical *y*-axis with the name of the dependent variable and its possible values, starting with 0 if practical. (4) Make a mark on the graph above each study participant's score on the *x*-axis and across from his or her score on the *y*-axis.

**3-3** Each dot on a scatterplot represents an individual's scores on two scale variables. It falls above the individual's score on the *x*-axis and across from the individual's score on the *y*-axis.

**3-4** A linear relation between variables means that the relation between variables is best described by a straight line.

**3-5** With scale data, a scatterplot allows for a helpful visual analysis of the relation between two variables. If the data points appear to fall approximately along a straight line, the variables may have a linear relation. If the data form a line that changes direction along its path, the variables may have a nonlinear relation. If the data points show no particular relation, it is possible that the two variables are not related.

**3-6** A line graph is used to illustrate the relation between two scale variables. One type of line graph is based on a scatterplot and allows us to construct a line of best fit that represents the predicted *y* scores for each *x* value. A second type of line graph allows us to visualize changes in the values on the *y*-axis over time. A time plot, or time series plot, is a specific type of line graph. It is a graph that plots a scale variable on the *y*-axis as it changes over an increment of time (e.g., a second, a day, a century) recorded on the *x*-axis.

**3-7** A bar graph is a visual depiction of data in which the independent variable is nominal or ordinal and the dependent variable is scale. Each bar typically represents the mean value of the dependent variable for each category. A Pareto chart is a specific type of bar graph in which the categories along the *x*-axis are ordered from highest bar on the left to lowest bar on the right.

**3-8** Bar graphs typically depict summary statistics, such as frequencies or averages, for several different levels of one or more nominal or ordinal independent variables. Histograms typically depict frequencies for different values of one scale variable. Bars represent counts or percentages for different values of a scale variable or for different intervals of that scale variable.

**3-9** A pictorial graph is a visual depiction of data typically used for a nominal independent variable with very few levels (categories) and a scale dependent variable. Each level uses a picture or symbol to represent its value on the scale dependent variable. A pie chart is a graph in the shape of a circle, with a slice for every level. The size of each slice represents the proportion (or percentage) of each category. In most cases, a bar graph is preferable to a pictorial graph or a pie chart.

**3-10** Bar graphs are straightforward presentations of data, whereas the elements of pictorial graphs and pie charts can often distract from the data being presented. Also, mistakes in presentation style are much more common in pictorial graphs and pie charts than in bar graphs.

**3-11** The independent variable typically goes on the horizontal *x*-axis and the dependent variable goes on the vertical *y*-axis.

**3-12** Whenever possible, graph axes should start at 0, although sometimes this is not practical. For example, when the data do not contain low values (and including 0 would minimize the depiction of the actual data), cut marks should be used to indicate axes that do not start at 0.

**3-13** Moiré vibrations are any visual patterns that create a distracting impression of vibration and movement. A grid is a background pattern, almost like graph paper, on which the data representations, such as bars, are superimposed. Ducks are features of the data that have been dressed up to be something other than merely data.

**3-14** Geographic information systems are particularly powerful for analyzing demographic patterns or demographic differences in a variable. Knowing how several variables change over geographic regions could lead researchers to detect important relations among variables.

**3-15** Like a traditional scatterplot, the locations of the points on the bubble graph simultaneously represent the values that a single

CHAPTER TWO

# Frequency Distributions

## NOTE TO INSTRUCTORS

In this chapter, instructors should emphasize the importance of visually representing data. The chapter describes the different ways of organizing data in terms of a frequency distribution, as well as the various shapes of distributions. Students often forget the importance of visually representing the data that they work with, so it would be useful to show students how valuable visual representation can be by demonstrating how frequency distributions can be used to aid in getting a quick "snapshot" of data that are collected.

## OUTLINE OF RESOURCES

**I. Frequency Distributions**
- Discussion Question 2-1
- Discussion Question 2-2
- Discussion Question 2-3
- Discussion Question 2-4
- Classroom Activity 2-1: Comparing Visualization Methods
- LaunchPad Statistical Applets: One-Variable Statistical Calculator

**II. Shapes of Distributions**
- Discussion Question 2-5
- Classroom Activity 2-2: Exploring Shapes of Distribution

**III. Next Steps: Stem-and-Leaf Plot**
- LaunchPad Video Resources
- Additional Reading
- Online Resources

**IV. Handouts**
- Handout 2-1: Exploring Shapes of Distribution

## CHAPTER GUIDE

**I. Frequency Distributions**

**1.** When we organize data that are composed of **raw scores,** or data that have not yet been analyzed, it is useful to look at the distribution of scores. The distribution allows us to examine the pattern of our data.

2. We organize our raw scores into a **frequency distribution,** which describes the pattern of a set of numbers by displaying a count or proportion for each possible value of a variable.
3. The best and simplest way to arrange data is to use a **frequency table,** which visually displays the data so that we can see how often each value occurs.
4. To create a frequency table, we determine the range of our scores. Then, we create two columns. In the first column, add the highest value to the top of the column and the lowest value to the bottom. In the second column, mark the number of times each of these values has occurred in our data set.
5. Sometimes it is better to use a **grouped frequency table,** which displays the frequencies for an interval rather than a specific value. A grouped frequency table is a better choice than a frequency table when the data are composed of continuous interval variables, cover a huge range, or are very large.

### > Discussion Question 2-1

***What is the difference between a frequency table and a grouped frequency table? When would you want to use one type rather than the other?***

Your students' answers should include the following:

- A frequency table reports every value in a given data set, whereas a grouped frequency table reports intervals or ranges of values.
- A frequency table is used to depict data showing how often certain values occurred and how many scores were at each value. A grouped frequency table is used when the values are:
  - **a.** Vast in number (such as when reporting hundreds of values)
  - **b.** Several decimal places
  - **c.** Both vast in number and several decimal places long

6. To create a grouped frequency table, find the highest and lowest scores in the distribution. If the highest and lowest values are decimals, round down. Subtract the lowest score from the highest score and add one. Next, determine the number of intervals and best interval size. List the intervals from lowest to highest in a column. Then, in the other column, count the number of values in each interval.

### > Discussion Question 2-2

***What steps are involved in creating a frequency table? A grouped frequency table?***

Your students' answers should include the following:

- To create a frequency table:
  - **a.** Examine the data.
  - **b.** Create two columns; in the first column record the values, putting highest at the top and lowest at the bottom.
  - **c.** Tally the occurrence of each value.
  - **d.** Record the tallies in the second column.

- To create a grouped frequency table:
  - **a.** Find the highest and lowest scores.
  - **b.** Use the full range of data, but round scores down to whole numbers.
  - **c.** Determine the number of intervals and best interval size.
  - **d.** Determine which number will be the bottom of the lowest interval.
  - **e.** List the intervals from highest to lowest and then count the numbers of scores in each.

**7.** Another way to organize the data is to use a **histogram.** Histograms typically depict just one variable, usually based on scale data, with the values of the variable on the *x*-axis and the frequencies on the *y*-axis. Each bar represents the frequencies for each value or interval.

**8.** To create a histogram, start with a frequency table. Draw your *x*- and *y*-axis and label them with your variable of interest. Draw a bar for each value, centering the bar over that value on the *x*-axis. The bar should be as high as the frequency for that value.

**9.** Histograms can also be created from a grouped frequency table. Instead of values, the midpoints of the intervals are listed on the *x*-axis. The remaining steps are the same as those used when constructing a histogram from a frequency table.

**10.** **Frequency polygons** are another way of visually representing data using a line graph, where the *x*-axis represents the value (or interval midpoint) and the *y*-axis represents the frequency. Frequency polygons are similar to histograms except that dots are used instead of bars and a line is used to connect the dots.

### > Discussion Question 2-3

***What is the difference between a histogram and a frequency polygon?***

Your students' answers should include the following:

- A histogram looks like a bar graph and often depicts interval data, with the values of the variables represented on the *x*-axis and the frequencies represented on the *y*-axis.
- A frequency polygon is a line graph depicting interval data. It also represents values on the *x*-axis and frequencies on the *y*-axis.

### > Discussion Question 2-4

***What steps are involved in creating a histogram? A frequency polygon?***

Your students' answers should include the following:

- To create a histogram:
  - **a.** Determine the midpoint for each interval, if needed.
  - **b.** Draw and label the *x*-axis and the *y*-axis of a graph.
  - **c.** Draw a bar for each value.
- To create a frequency polygon:
  - **a.** Determine the midpoint for each interval, if needed.
  - **b.** Draw and label the *x*-axis and the *y*-axis.
  - **c.** Mark a dot above each value and connect the dots with a line.

**d.** Add hypothetical values at both ends of the *x*-axis and mark dots for the frequency of 0 for each value to create a grounded shape rather than a floating line.

**Classroom Activity 2-1**

**Comparing Visualization Methods**

In this exercise, you will work with one set of data and compare how the different visualization methods capture the data. You could collect any sort of information from your students, perhaps using the data that you collected earlier to demonstrate different measurement scales, (e.g., height, year in school, age, etc.) and then demonstrate how frequency tables, grouped frequency tables, histograms, and so forth might be used to visualize the data.

**LaunchPad Statistical Applets**

**One-Variable Statistical Calculator**

This applet calculates standard numerical statistics (e.g., mean, standard deviation, quartiles) and shows graphical displays (a histogram and a stem-and-leaf plot) of one-variable data sets. You can choose to view data sets from the textbook or enter your own set of data.

## II. Shapes of Distributions

1. A **normal distribution** refers to a bell-shaped, symmetrical, and unimodal frequency distribution.
2. We could also have a **skewed distribution.** Skewed distributions are distributions where one of the tails of the distribution is pulled away from the center.
3. When the tail of our distribution extends to the right, we say that our data are **positively skewed.** We typically observe positively skewed data when there is a **floor effect**—when a variable is prevented from taking values below a certain point.
4. Data can also be **negatively skewed,** meaning that the tail of our distribution extends to the left. We may observe negatively skewed data in the case of a **ceiling effect**—when a variable is prevented from taking values above a certain point.

### > Discussion Question 2-5

***What is skewness? What is the difference between the two different types of skewness?***

Your students' answers should include the following:

- Skewness is the amount that a tail of a distribution is pulled away from the center.
  - **a.** Positively skewed data: The tail of the distribution extends to the right
  - **b.** Negatively skewed data: The tail of the distribution extends to the left

## Classroom Activity 2-2
## Exploring Shapes of Distribution

In this exercise, students will generate examples of two variables.

- Have the students predict whether the variables will be positively skewed or negatively skewed.
- Students can then develop questionnaires in groups to measure these two variables.
- Have them hand out versions of their questionnaires in class to see if they were correct in their predictions.

See Handout 2-1.

**III. Next Steps: Stem-and-Leaf Plot**

**1.** The **stem-and-leaf plot** is a graph that displays all the data points of a single variable both numerically and visually. It displays the same information as a histogram—just in a different way and with more detail.

**2.** To create a stem-and-leaf plot, first create the stem by writing down the first digit for each number of your data from highest to lowest. The leaves consist of the last digit for each score and are added in ascending order.

### LaunchPad Video Resources

Snapshots: Data and Distributions
StatClips Examples: Exploratory Pictures for Quantitative Data, Example A
StatClips Examples: Summaries and Pictures for Categorical Data, Examples A and B
StatClips: Summaries and Pictures for Categorical Data

### Additional Reading

Moore, Thomas L., Ed. (2001). *Teaching statistics: Resources for undergraduates.* Mathematical Association of America.

This book is an instructor's manual for teaching undergraduate statistics that advocates a hands-on approach.

### Online Resources

The following Web site provides a wealth of information on statistics: http://www.math.yorku.ca/SCS/StatResource.html.

For information on good and bad visual graphic presentations, see: http://www.datavis.ca/gallery/index.php.

## HANDOUT 2-1: EXPLORING SHAPES OF DISTRIBUTION

Directions: Answer the following questions regarding skewness.

1. What two variables do you think would generate data from your class that would be positively skewed or negatively skewed? Why do you think the data would follow this pattern?

2. Develop short questionnaires to measure these two variables.

3. Distribute these questionnaires to your classmates and draw histograms of the data collected. Were you correct in any of your predictions? Why or why not?