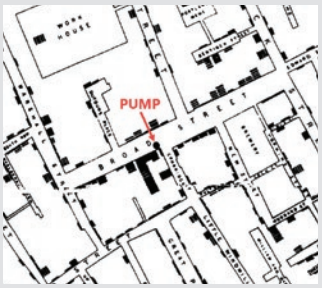


Absence of Evidence is Not Evidence of Absence

During the London cholera outbreaks of the mid-1800s, thousands of people died within a relatively short period. At the time, the prevailing theory regarding how cholera was spread was called the miasma theory. It stated that the disease was spread through "bad air" that emanated from rotting organic matter. However, Dr. John Snow suspected that unsanitary water from the River Thames was the true culprit. Unfortunately, germ theory had not been developed yet, so Dr. Snow didn't fully understand how the alternative transmission method worked. In 1854, Dr. Snow utilized sampling and data visualization to illustrate that most of the cholera outbreaks happening at the time were occurring in houses that were close to the water pump on Broad Street. Still, the skeptics endured. However, even though his examination of the water was absent of evidence for harmful microbes, that does not mean that the microbes themselves were absent. Over a decade later, Louis Pasteur would officially propose germ theory, vindicating the work of Dr. Snow.

Table 3.3.6 - Percentage of Obese Adults by US County

FIPS Code	Percentage of Obese Adults, 2016 Normalized Values
1001	30.5
1003	26.6
1005	37.3
1007	34.3
1009	30.4
. . .	

Using our new data, we generate the following map.

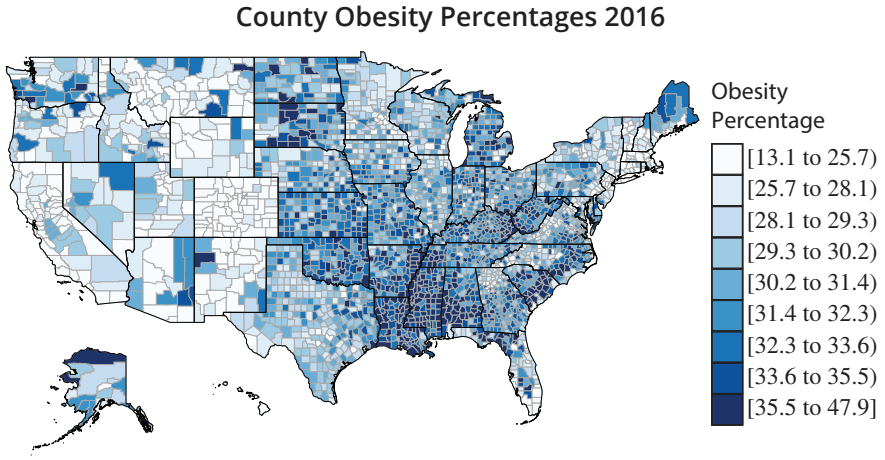


Figure 3.3.14

There are over 3000 counties in the U.S. The incredible aspect of these graphs is you can literally see the 3000+ data points and their geographic distribution. If we use Figure 3.3.13 to come to a conclusion about the geographic distribution of obesity in the U.S., we would have assumed that the Southwest and Northeast regions are the areas in the country that struggle with obesity. Comparing Figure 3.3.14 to Figure 3.3.13 produces an entirely new set of conclusions. Once we normalize the obesity variable by making it a ratio of the total county population, Figure 3.3.14 suggests that it is actually the Southeast and the Midwest that have the greatest struggle with obesity. In many instances, the data you start with will need to undergo some form of transformation to make it valuable for your intended purposes or goals.

3.3 Exercises

Basic Concepts

1. What is the main characteristic of data that a histogram reveals?
2. Describe the type of data that could be usefully described with a histogram.
3. True or false: A frequency distribution contains all of the information needed to construct a histogram.
4. List the important features to look for when studying a histogram.

5. Explain why the stem-and-leaf plot is sometimes called a "hybrid graphical method."
6. Identify the advantages of a stem-and-leaf plot.
7. Consider the following data value: 39. What would be the stem and the leaf for this value if we identified the stem as the tens digit? What would be the stem and the leaf if we identified the stem as the hundreds digit?
8. When constructing a stem-and-leaf plot, how do you determine which part to make the stem and which part to make the leaf?
9. What is an ordered array?
10. What are some advantages of the ordered array?
11. Dot plots are most useful for what types of data?
12. What are some advantages of using a dot plot?
13. How can the most frequently occurring value be identified by studying a dot plot?
14. Why is it important to plot time series data?
15. The time variable is always graphed on which axis?
16. What is a choropleth map?

Exercises

17. The weights (in pounds) for the players on an NFL football roster are shown below.

153	183	185	189	190	190	195	196	198	200	202	204
205	207	209	212	213	215	220	220	225	225	228	233
235	236	238	243	244	245	246	247	248	254	254	255
255	256	257	265	268	280	298	302	305	308	310	311

   a. Construct a frequency distribution for the weights of players on the roster.

   b. Construct a histogram for the weights of players on the roster.

18. Using the OECD Better Life 2022 data set from the web resource, create a frequency distribution and histogram for the variable *Voter Turnout* and use them to answer the following questions. For the histogram and frequency distribution, use 40% for the minimum value and use class widths of 10%.

   a. What is the level of measurement of the variable?

   b. Construct a relative frequency distribution for voter turnout.

   c. How many of the countries have a voter turnout of 70% or greater?

   d. What percent of the countries have less than a 50% voter turnout?

   e. What percent of the countries have a voter turnout of 80% or greater?

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022**

19. A chemist is interested in knowing the amount of alcohol contained in American-brewed beers. To study this, the chemist uses data containing information about several different kinds of American-brewed beers, and evaluates the alcohol by volume for each. Using the Beers and Breweries data set from the web resource, perform the following:

   a. Construct a frequency distribution for the alcohol by volume (ABV) variable. Use 0.001 for the minimum value and 0.130 for the maximum value. Use bin widths of 0.010.

Data

stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > Beers and Breweries**

b. Construct a relative frequency distribution for the ABV. Round the relative frequencies to four decimal places.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the alcohol by volume in American-brewed beers which you were able to ascertain by examining the distributions and the histogram.

20. Some final heat times (in seconds) for the boys 100-meter races at the 2023 Georgia state track meet are listed below.[17]

10.72	10.75	10.78	10.79	10.80	10.83	10.86	10.94
10.97	10.97	10.99	11.00	11.00	11.02	11.03	11.04
11.05	11.05	11.08	11.09	11.11	11.14	11.17	11.22
11.23	11.24	11.25	11.27	11.40	11.40	11.46	11.60

a. Construct a frequency distribution for the 100-meter race times.

b. Construct a relative frequency distribution for the 100-meter race times.

c. Construct a histogram of the relative frequency distribution.

d. Comment on any information about the 100-meter race times which you were able to ascertain by examining the distributions and the histogram.

21. To attend a friend's wedding this summer, you'll fly from the Hartsfield-Jackson Airport in Atlanta, GA to Denver, CO. The following data shows the cost (in dollars) of several flights that are compatible with your travel plans.

209	198	188	198	227	246	256	220
198	205	246	198	227	199	198	194
198	188	198	209	188	198	231	205

a. What level of measurement does the data possess?

b. Construct a stem-and-leaf display for the data using the tens digits as the stems.

c. Comment on the shape of the distribution.

22. The following data shows the credit scores of recent lease applicants at an apartment complex.

645	756	668	590	713	647	811	725	806
675	632	740	583	689	739	791	826	670
782	654	619	672	689	702	717		

a. Construct a stem-and-leaf display for the data using the hundreds digits as the stems.

b. Construct a histogram using the classes 500-599, 600-699, 700-799, 800-899.

c. Comment on the shape of the distribution. What do you notice about the shapes of the stem-and-leaf display and the histogram of the credit score data?

d. Applicants with a credit score above 700 can remit a reduced security deposit. What percentage of these applicants have a credit score above 700?

e. Individuals with a credit score below 660 are considered "subprime" and will likely not qualify for the lease. What percentage of applicants have a credit score below 660?

23. Daily high temperatures (°F) in June 2022 for two US cities are shown in the side-by-side stem-and-leaf display shown below.

Charleston, SC	Stem	Milwaukee, WI
	6	1 3 5 6 7 9
2	7	4 4 4 5 5 6 6 6 8 9 9
2 2 2 2 2 2 4 4 4 4 4 6 6 6 6 6 8 8 8	8	0 0 0 5 7 8 8
0 0 0 1 1 3 3 3 5 5	9	2 4 4 5 9
0	10	0

a. What level of measurement does the data possess?

b. Based upon the stem-and-leaf display, compare the June temperatures for the two cities. State several observations about the ways that the distributions are similar and different.

c. Suppose that someone does not like extremely warm summer temperatures. What percentage of days had temperatures less than 90 degrees in Charleston? in Milwaukee?

d. What was the most frequent temperature in Charleston?

24. *Fortune* magazine publishes a list of the top 100 best companies to work for. For the top 10 companies on this list, the average annual employee salaries are given in the following table (in thousands of dollars).

Average Salaries (Thousands of Dollars)									
121	122	136	74	118	101	114	61	95	132

a. Construct a stem-and-leaf display for the data using the tens digits as the stems.

b. Comment on any information about the average annual salaries (in thousands of dollars) of the top 10 companies which you were able to ascertain by examining the stem-and-leaf display.

c. Construct an ordered array of the average annual salaries in rank order.

d. Does the ordered array provide any additional insight into the nature of the data?

25. The pH level of drinking water obtained from a well was regularly tested. The recorded data is as follows:

7.1	6.5	6.9	6.6	7.0	7.3	7.1	6.8	6.8	6.5
6.7	6.9	6.8	6.5	6.3	6.6	6.7	7.0	7.2	6.9
6.8	6.6	6.9	6.8	7.1	6.8	6.7	6.5	6.8	7.0

a. Construct a dot plot of the data.

b. Which data value occurs most often?

c. The recommended pH levels for drinking water should fall in the range of 6.5 to 8.5. Is that true of this data?

d. Does this well water tend to be acidic or alkaline? Explain your response.

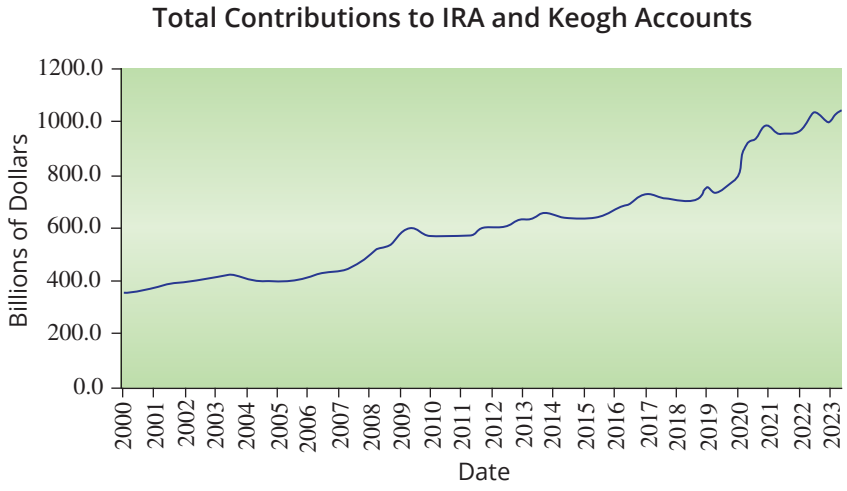
26. Listed in the following table is the number of passing attempts per game by Super Bowl champion Patrick Mahomes in the 2022 NFL regular season.[18]

a. Construct a dot plot of the data.

b. Which data value occurs most often?

Passing Attempts by Patrick Mahomes								
39	35	35	37	43	40	34	68	35
34	42	27	42	41	28	42	26	

27. The following line graph displays the total IRA and Keogh accounts (in billions of dollars) in the U.S., charted from January 2000 to May 2023.[19]

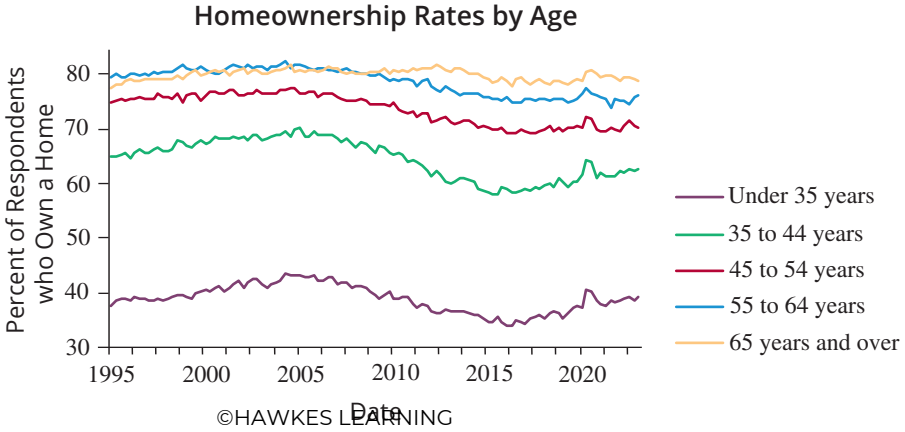


a. What conclusions can you make regarding the total contributed to the accounts?

b. Is the data time series data?

c. If the data is time series data, is the series stationary or nonstationary?

28. The following chart displays the homeownership rate data collected by the United States Census Bureau for January 1995 through January 2023. The percentage of owner-occupied housing units is reported for various age ranges.[20]



a. Examine the graph and discuss the data. What conclusions can you make?

b. If the data is time series data, is it a stationary or nonstationary time series? Explain your reasoning.

29. The unemployment rate is a key economic indicator that measures the percentage of the labor force that is unemployed and actively seeking employment. The unemployment rates for the state of North Carolina from January 2010 to January 2022 are given in the table below.[21,22]

North Carolina Unemployment Rate							
Year	2010	2011	2012	2013	2014	2015	2016
Percent	11.2	10.4	9.7	9.4	6.4	5.7	5.3
Year	2017	2018	2019	2020	2021	2022	
Percent	4.9	4.2	4.0	3.8	5.6	3.6	

a. What is the level of measurement of the unemployment rate data?

b. Construct a time series plot for the data.

c. What conclusions can you make from the plot?

3.4 Analyzing Graphs

Graphs that help us visualize data can either be enlightening, in the sense that they give us insight and understanding of a set of data, or misleading, either intentionally or unintentionally. When you see graphs in the media, you need to be cautious to ensure the data has been accurately represented by the graph. This section will help you analyze graphs for accuracy and appropriate presentation of the given information. Here are a few key ideas to consider when interpreting information displayed in graphical form.

Graph Labeling

Every graph should be properly labeled with an appropriate title that tells you what type of information is being displayed. Also, if the graph has a horizontal and vertical axis, these should be labeled and should include the unit of measurement when necessary for the understanding of the data. For example, in Figure 3.4.1, the title does not provide enough information about the data. Why were those countries chosen? Do they have relatively high or low prison populations compared to the rest of the world? Furthermore, we do not know whether this information is relevant to modern times. Is this data for a specific year? The countries are labeled along the horizontal axis but note that the vertical axis is just labeled *Population*. We have no idea what the values along the vertical axis represent. Is the prisoner population in units of thousands, millions, or billions? In fact, this chart shows the countries with the top ten highest prisoner populations for the year 2021. The unit for the vertical axis should be thousands, which means that the United States had a prison population of approximately 1690 thousand, or 1.690 million, in the year 2021. Without these seemingly small pieces of information, the graph is not very informative. It is also good practice to use the largest possible unit for the scale of an axis, which in this case is correctly chosen to be thousands.