

Unit 3: Histograms



SUMMARY OF VIDEO

Many people are afraid of getting hit by lightning. And while getting hit by lightning is against the odds, it is not against all odds. Hundreds of people are struck by lightning every year in the U.S. What's more, fires started by lightning strikes cause hundreds of millions of dollars of property damage. Meteorologist Raul Lopez and his associates began collecting detailed data on lightning strikes back in the 1980s and soon were overwhelmed by the vast amount of data. In one year, they collected three-quarters of a million flashes in a small area of Colorado. They decided to focus on when lightning strikes occurred. The data on the times of the first lightning strike needed to be organized, summarized, and displayed graphically. One of the statistical tools that Raul Lopez turned to was the graphic display called a histogram. For example, data on the percent of first lightning flashes for each hour of the day is displayed in the histogram in Figure 3.1.

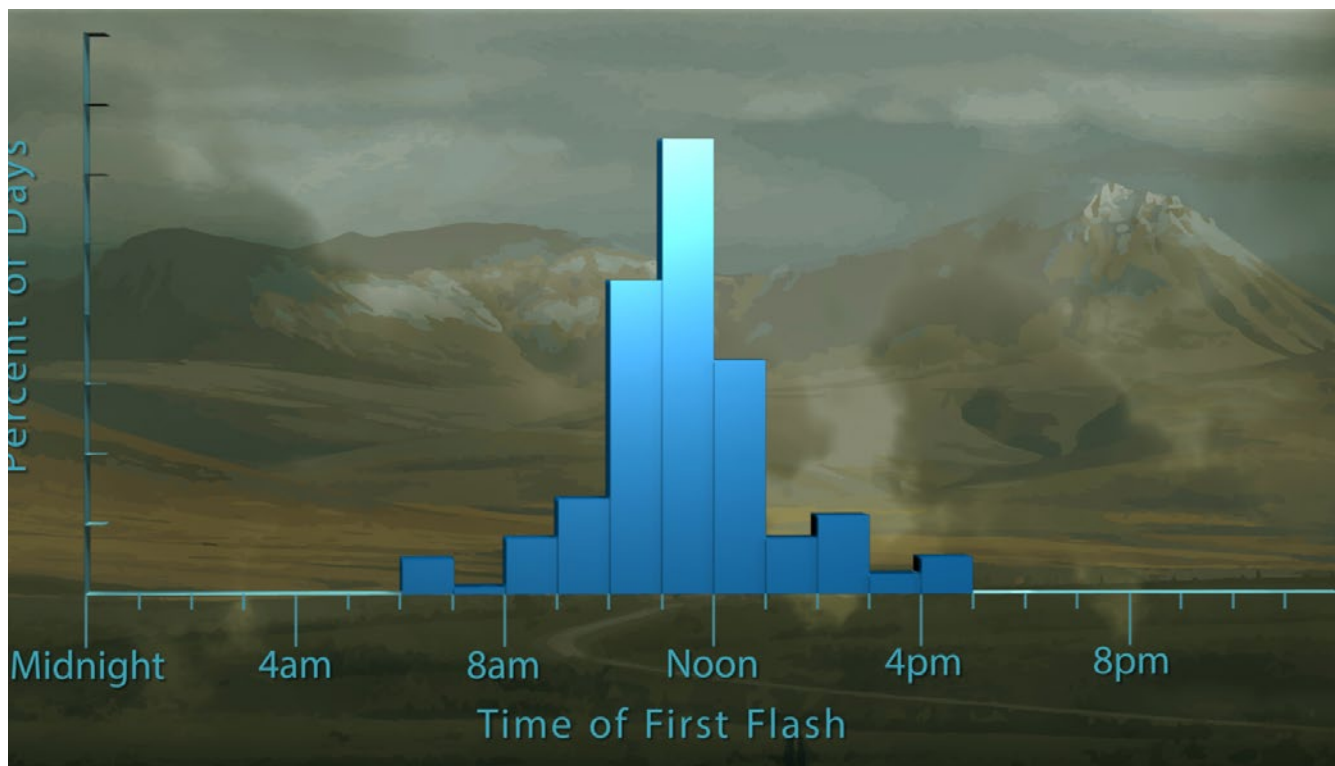


Figure 3.1. Histogram of the time of the first lightning strike.

Before the histogram could be constructed, each day was broken into hours (horizontal axis), the number of first flashes in each hour was counted, and then the counts were converted to percentages (vertical axis). So, in this histogram, each bar represents one hour, and its height is the percentage of days in which the first lightning flash fell in that hour. This histogram has two very striking features. First, it is roughly symmetric about the tallest bar, which represents the percentage of first flashes between 11 a.m. and noon. The second rather surprising feature is how tightly the time of first strike clusters around the center bar, with a range from 10 a.m. to 1 p.m. accounting for most of the days' first strikes. And there are no first strikes at night. This pattern helped explain how lightning storms form in this area. This region is mountainous and winds from the eastern plains carry warm moist air. When the wind hits the mountains it is forced upward where it meets and mixes with colder air higher in the atmosphere forming clouds. And this turns out to be a regular daily occurrence during the Colorado summer.

Lopez and his colleagues next looked at the time of day when the maximum number of lightning flashes occurred. (See Figure 3.2.) They found a similar pattern, with a peak showing that most flashes occur between 4 p.m. and 5 p.m.

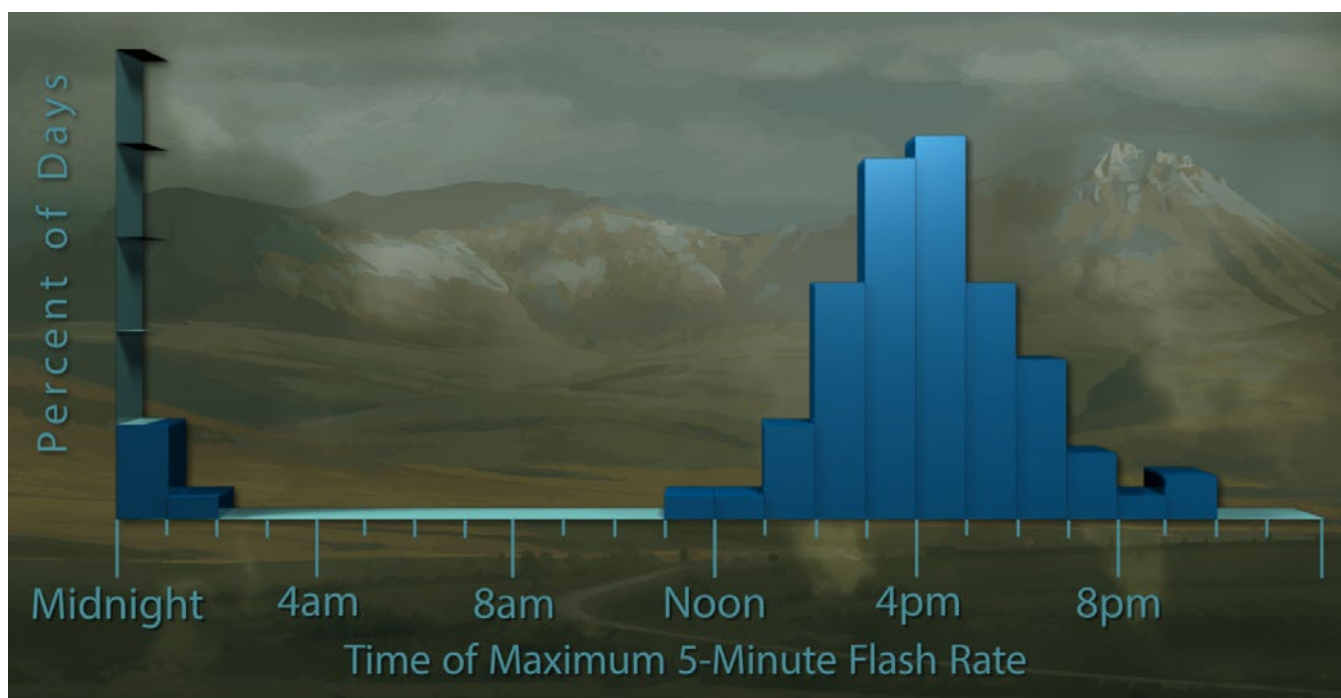


Figure 3.2. Histogram of the time of maximum flash rate.

But there is one big difference from the first flash histogram in Figure 3.1. On a few days the maximum was in the early hours of the morning. Data points like these, which stand out from the overall pattern of the distribution, are called outliers. Outliers are often the most intriguing features of a histogram. Outliers should always be investigated and, if possible, explained.

The explanation that Lopez and his colleagues came up with was that they occur on days when larger weather systems, specifically very strong winds from fast moving weather fronts, overpower the local effect.

Data collection on Colorado lightning has continued since the pioneering work of Raul Lopez and his colleagues. Figure 3.3 shows a histogram produced from more recent data showing the number of people injured or killed by lightning strikes in the last 30 years. It shows the same clustering pattern as Raul Lopez's histograms, but interestingly, the peak time for getting struck by lightning is around 2 p.m., about midway between the peaks of the first strike and maximum activity histograms.

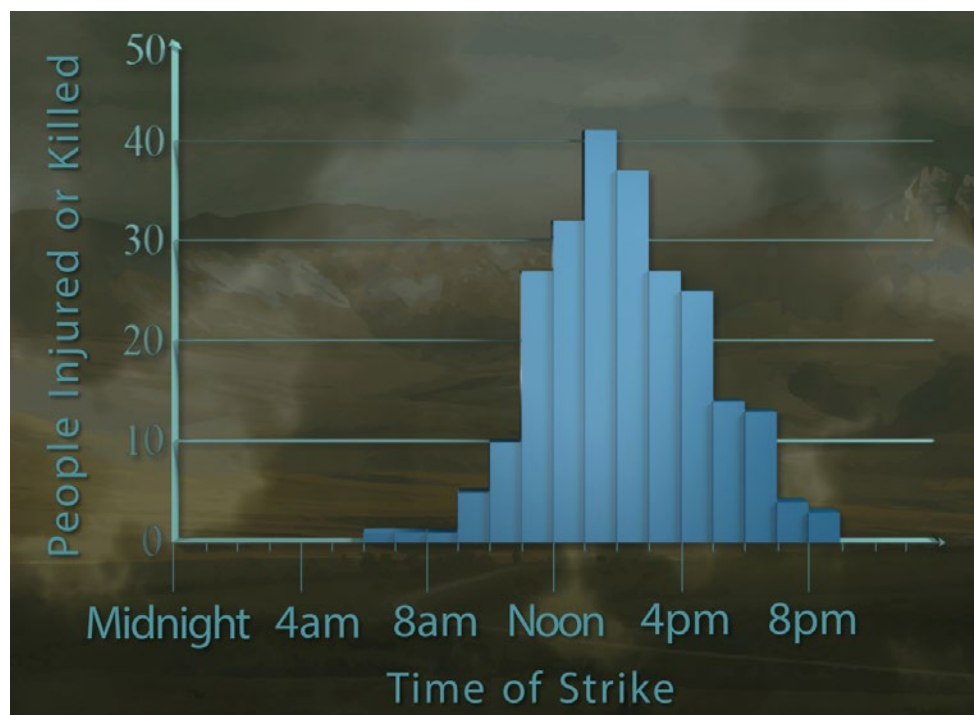


Figure 3.3. Histogram of time when people were struck by lightning.

When constructing histograms it is very important to choose the best class size – that is, the choice of the interval widths for the horizontal axis. Lopez chose one hour for his data, and it works well. But suppose we turn our attention to a different context, the weekday traffic density on a portion of the Massachusetts Turnpike. First, we look at a histogram with class intervals of three hours. (See Figure 3.4.)

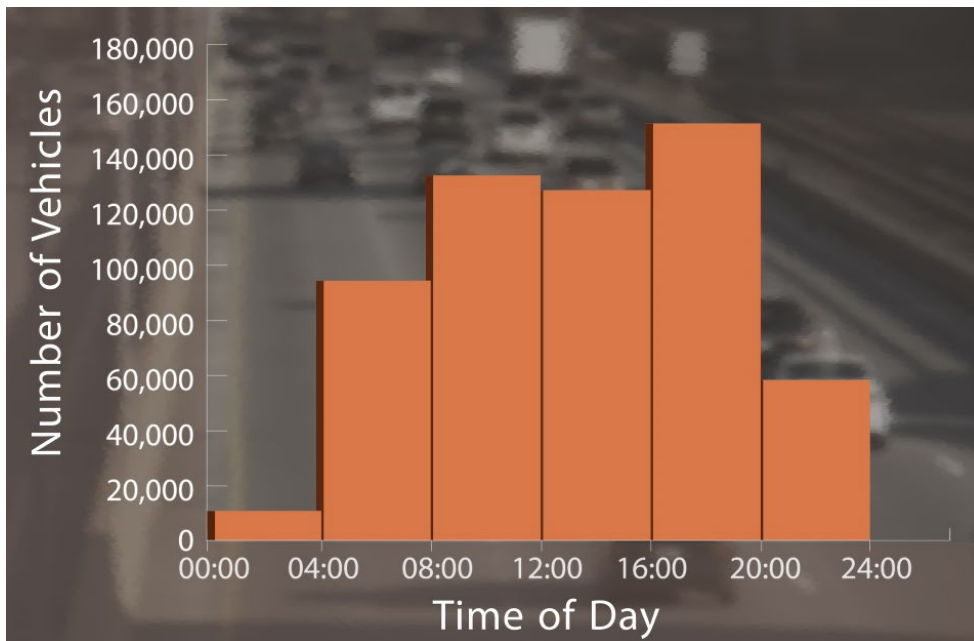


Figure 3.4. Histogram of traffic density in three-hour intervals.

The histogram in Figure 3.4 is not terribly informative. Next, we changed the interval width to one hour, which was better. However, using one-half hour widths as shown in Figure 3.5 is even better. Now, the increased traffic density during morning rush hour and evening rush hour is clearly visible in the pattern of two peaks.

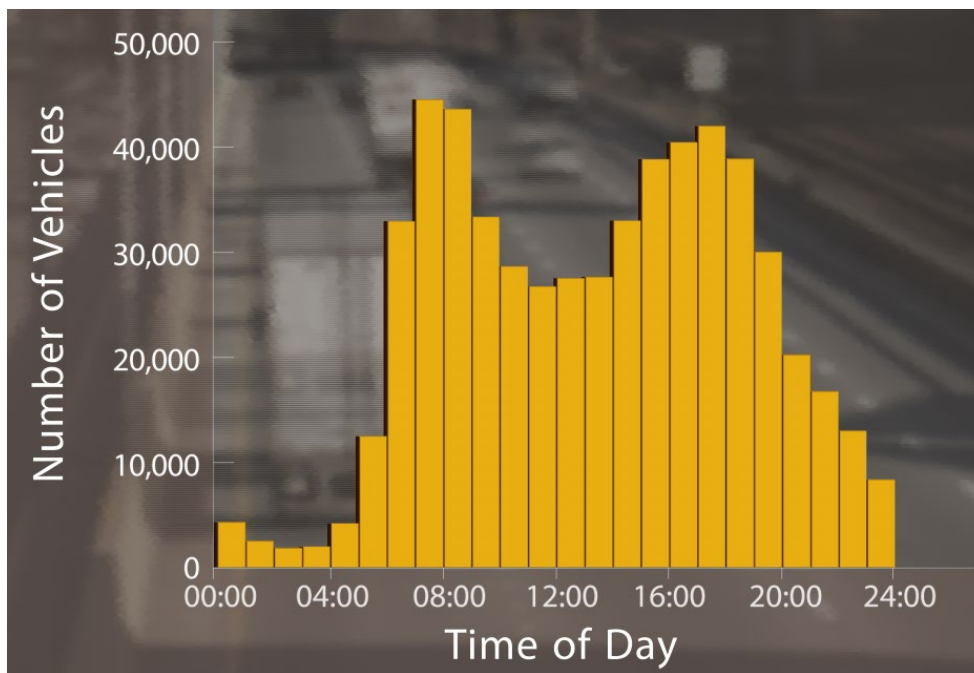


Figure 3.5. Histogram of traffic density in half-hour intervals.

But what if we went even finer-grained and used 5-minute intervals? Take a look at Figure 3.6. Now the peaks begin disappearing again back into the numbers and the histogram becomes less informative.

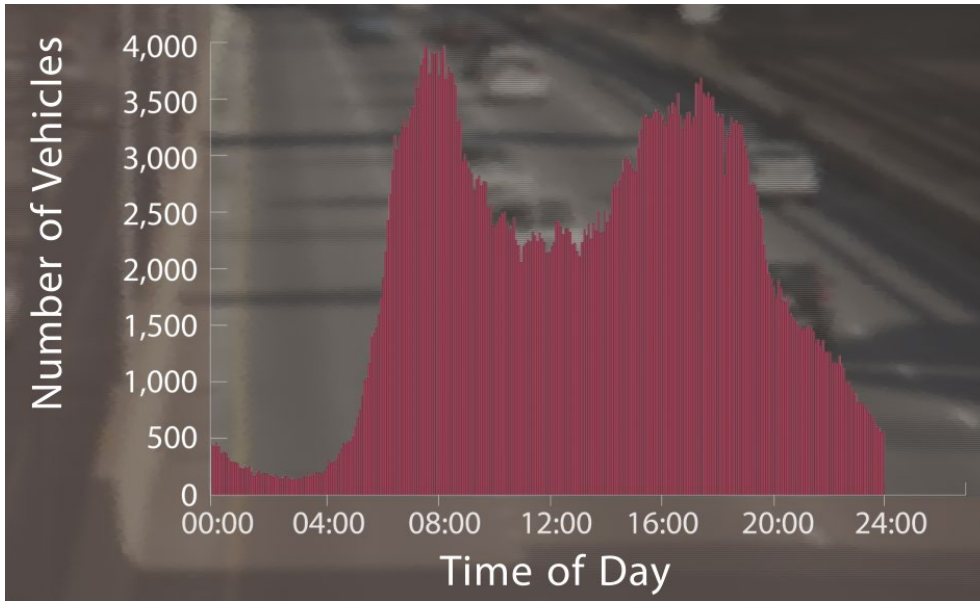


Figure 3.6. Histogram of traffic density in 5-minute intervals.

So, we have seen how histograms can literally show at a glance the essence of a whole lot of numbers. Here is one last example. Figure 3.7 shows a histogram of the weekly wages of workers in the U.S. in the year 1992.

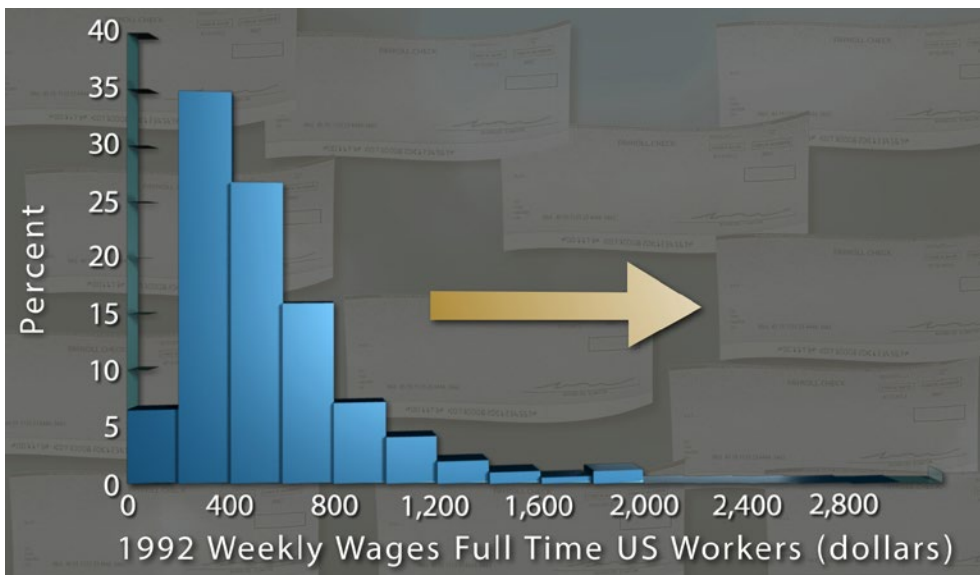


Figure 3.7. Histogram of weekly wages (1992).

Notice how strikingly it is skewed, with most people earning around \$450 per week. As you go out to what is called the tail of the distribution (to the right), the salaries get bigger, but the

percent of people earning those salaries gets smaller. Statisticians say a distribution like this is skewed to the right, because the right side of the histogram extends much further out than the left side. Now look at the histogram in Figure 3.8 of the same variable, weekly wages, but for the year 2011.



Figure 3.8. Histogram of weekly wages (2011).

Now, the skew has become much more pronounced, and the tail has grown much longer. Suddenly our little discourse on histograms could become highly political!

STUDENT LEARNING OBJECTIVES

- A. Understand that the distribution of a variable consists of what values the variable takes and how often. (This is a repeat of an objective from Unit 2, Stemplots.)

- B. Be able to construct a histogram to display the distribution of a variable for moderate amounts of data (say, data sets with fewer than 200 observations).

- C. Understand that class intervals should be of equal width; choose appropriate class widths to effectively reveal informative patterns in the data.

- D. Understand that the vertical axis of the histogram may be scaled for frequency, proportion, or percentage. The choice of vertical scaling for any data set does not affect the important features revealed by a histogram.

- E. Be able to describe a graphical display of data by first describing the overall pattern and then deviations from that pattern. Describe the shape of the overall pattern and identify any gaps in data and potential outliers.

- F. Recognize rough symmetry and clear skewness in the overall pattern of a distribution.

CONTENT OVERVIEW

Rows and rows of data provide little information. For example, below are thickness measurements, in millimeters, from a sample of 25 polished wafers used in the manufacture of microchips. Notice that it is difficult to extract much information from staring at these numbers. The numbers need to be organized, summarized, and displayed graphically in order to unlock the information they contain.

0.402	0.496	0.533	0.387	0.384
0.528	0.411	0.367	0.462	0.499
0.539	0.546	0.425	0.457	0.586
0.558	0.588	0.425	0.437	0.479
0.427	0.485	0.443	0.441	0.658

A **frequency distribution** is one method of organizing and summarizing data in a table. The basic idea behind a frequency distribution is to set up categories (class intervals), classify data values into the categories, and then determine the frequency with which data values are placed into each category. The steps below outline the process of making a frequency distribution table.

Creating a frequency distribution table

Step 1: Identify an interval that is wide enough to contain all the data.

Step 2: Subdivide the interval identified in Step 1 into class intervals of equal width. The class intervals will serve as the categories.

Step 3: Set up a table with three columns for the following: class interval, tally, and frequency. (The tally column can be removed in the final table.)

Step 4: To complete the table, determine the frequency with which data values fall into each class interval.

Convention: Any data value that falls on a class interval boundary is placed in the class interval to the right. If the data value is a maximum, it is generally put in the interval that contains the maximum at its right endpoint.

Now, we apply Steps 1 – 4 to make a frequency distribution table for the thickness measurements.

Step 1: In this case the smallest data value is 0.367 mm and the largest is 0.698 mm. We choose the interval from 0.3 mm to 0.7 mm, which contains all the thickness measurements.

Step 2: The total width of the interval from 0.3 to 0.7 is 0.4. Dividing this interval into eight class intervals works out nicely – each class interval will have width 0.05.

Step 3: We have set up Table 3.1 to have three columns, which we have labeled Thickness, Tally, and Frequency. We have entered the endpoints of the eight class intervals into the Thickness column.

Thickness (mm)	Tally	Frequency
0.30 – 0.35		
0.35 – 0.40		
0.40 – 0.45		
0.45 – 0.50		
0.50 – 0.55		
0.55 – 0.60		
0.60 – 0.65		
0.65 – 0.70		

Table 3.1: Setting up a frequency distribution table.

Step 4: The easiest way to determine the frequencies is to draw a tally line for each data value that falls into a particular class interval. When drawing tally lines, keep the following in mind:

- As you draw tally lines, instead of drawing a fifth tally line, cross out the previous four.
- If a data value falls on the boundary of a class interval, record it in the interval with the larger values.

Once a tally line has been drawn for each data value, count the number of tally lines corresponding to each class interval and record that number in the frequency column as shown in Table 3.2.

Thickness (mm)	Tally	Frequency
0.30 – 0.35		0
0.35 – 0.40		3
0.40 – 0.45		8
0.45 – 0.50		6
0.50 – 0.55		4
0.55 – 0.60		3
0.60 – 0.65		0
0.65 – 0.70		1

Table 3.2: A completed frequency distribution table.

The frequency distribution in Table 3.2 reveals more information about the data than a quick look at the 25 numbers. For example, from the frequency distribution, we learn that more measurements fall in the interval 0.40 – 0.45 than in any of the other class intervals. Also, we learn there is a gap in the data – no data values fall between 0.60 and 0.65.

Although a frequency distribution table is a useful tool for extracting information from data, a histogram can often convey the same information more effectively. Next, we outline how to construct a histogram from a frequency distribution.

Creating a histogram from a frequency distribution

Step 1: Draw a set of axes. On the horizontal axis, mark the boundaries of the class intervals. On the vertical axis, set up a scale appropriate for the frequencies. (Later this scale can be changed to proportion or percent.)

Step 2: Label the horizontal axis with the name of the variable being measured and the units.

Step 3: Over each interval, draw a rectangle with the interval as its base. The height of the rectangle should match the frequency of data contained in that interval.

Next, we apply Steps 1 – 3 for creating a histogram to the frequency distribution in Table 3.2. Figure 3.9 shows the results.

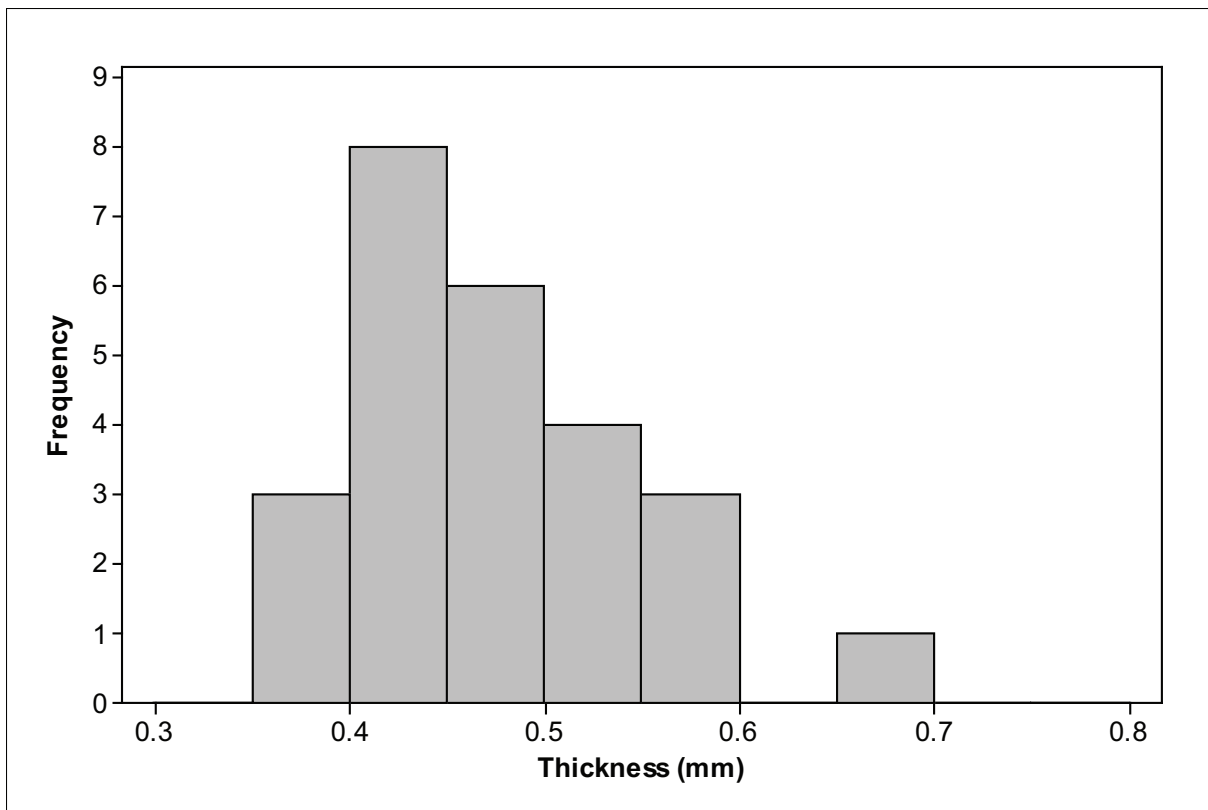


Figure 3.9: Histogram representing frequency distribution in Table 3.2.

Particularly if you are comparing histograms from samples with a different number of data values, it is useful to replace the frequency scale on the vertical axis with the proportion or percent.

Calculating Proportions and Percents

- To calculate a proportion, divide the frequency by the sample size.
- To convert a proportion into a percent, multiply the proportion by 100%.

In describing a histogram, we first look for the overall pattern of the distribution. In sizing up the overall pattern, look for the following:

- center and spread;
- one peak or several (unimodal or multimodal);
- a regular shape, such as symmetric or skewed.

In the case of the histogram in Figure 3.9, the overall pattern is single-peaked (or unimodal) and skewed to the right. Next, we look for any striking deviations from that pattern. An important kind of deviation from an overall pattern is an outlier, an individual observation

that lies clearly outside the overall pattern. Once identified, outliers should be investigated. Sometimes they are errors in the data and sometimes they have interesting stories related to the data. For Figure 3.9, there is a gap between 0.6 and 0.65 and there is one data value between 0.65 and 0.70, which might be an outlier.

KEY TERMS

A **frequency distribution** provides a means of organizing and summarizing data by classifying data values into class intervals and recording the number of data that fall into each class interval.

A **histogram** is a graphical representation of a frequency distribution. Bars are drawn over each class interval on a number line. The areas of the bars are proportional to the frequencies with which data fall into the class intervals.

The shape of a unimodal distribution of a quantitative variable may be **symmetric** (right side close to a mirror image of left side) or skewed to the right or left. A distribution is **skewed to the right** if the right tail of the distribution is longer than the left and is **skewed to the left** if the left tail of the distribution is longer than the right.

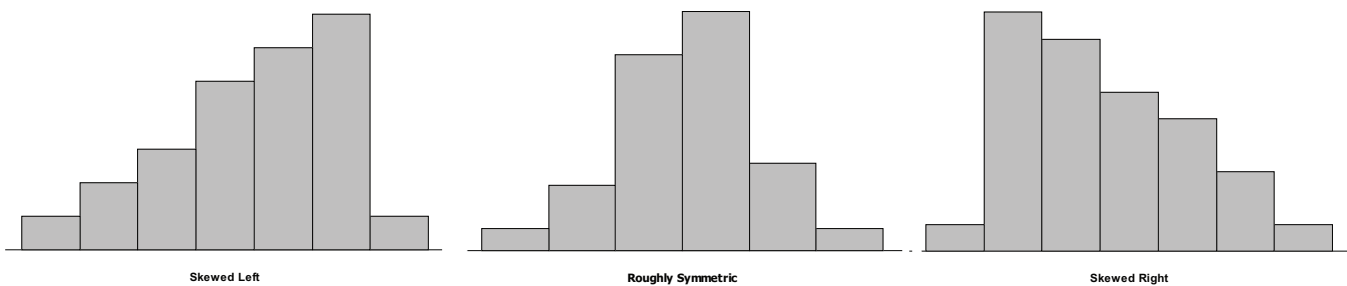


Figure 3.10. Shapes of histograms.

THE VIDEO

Take out a piece of paper and be ready to write down answers to these questions as you watch the video.

1. The video opens by describing a study of lightning strikes in Colorado. What variable does the first histogram display?
2. In this lightning histogram, what does the horizontal scale represent? What does the vertical scale represent?
3. Was the overall shape of this histogram symmetric, skewed, or neither?
4. Why were a few values in the second lightning histogram called outliers?
5. When you choose the classes for a histogram, what property must the classes have if the histogram is to be correct?
6. What happens to a histogram if you use too many classes? What happens if you use too few?

UNIT ACTIVITY:

WAFER THICKNESS

What do automobiles, singing Barbie dolls, cell phones and computers have in common? To a worker in the semiconductor industry, the answer is obvious – they all use microchips, tiny electronic circuits etched on chips of silicon (or some other semiconductor material).

Manufacturing microchips is a complex process. It begins with cylinders of silicon, called ingots, which are 6 to 16 inches in diameter. The ingots are sliced into thin wafers, which are then polished. (See Figure 3.11.) The polished wafers are imprinted with microscopic patterns of circuits, which are etched out with acids and replaced with conductors (such as aluminum or copper). Once completed, the wafers are cut into individual chips. (See Figure 3.12.)

In order to remain competitive in a global market, American companies must process microchips correctly and repeatedly with almost perfect consistency. The only way to accomplish this is to measure and control all of the highly complex processes used to manufacture microchips. These companies rely on statistical techniques to ensure quality control at critical points in the processing. It is simply too costly to wait until the end and then reject defective chips.

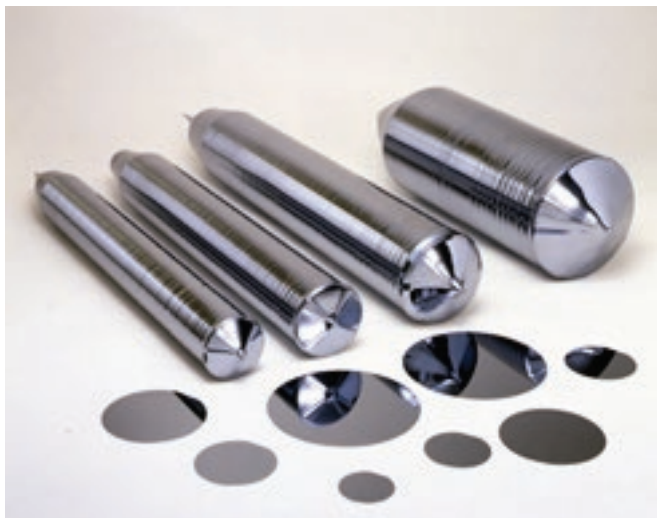


Figure 3.11: Silicon ingots and polished wafers.

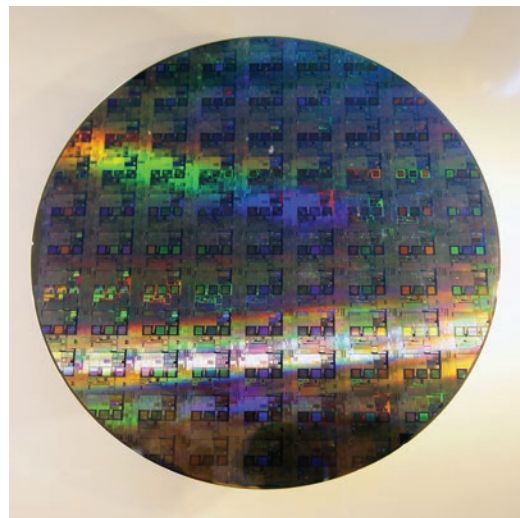


Figure 3.12: The grid pattern shows individual microchips on a wafer. (Credit: Peeldien)

One critical stage in the manufacture of microchips is the grinding and polishing processes used to produce polished wafers. The wafers need to be consistent in thickness, not warped or bowed, and free of surface imperfections. The focus in this activity will be on adjusting controls in order to produce polished wafers that are consistently close to 0.5 mm in thickness.

The Wafer Thickness tool found in the Interactive Tools menu allows you to set three controls that adjust the grinding and polishing processes. Each control has three levels. After setting the controls, you can take a sample of polished wafers and measure their thicknesses.

1. Set all three controls to 1. Select a sample size of 10 and select Real Time mode. Then press the “Collect Sample Data” button. Watch as the sample wafers are measured. A graphic display (called a histogram) is formed in real time as data become available.

a. Describe what happens to the graphic display each time a new data value is added to the table. In other words, how is the histogram constructed?

b. Describe the shape or features of the histogram. Here are some questions to consider when describing the shape of the data:

- Is the histogram roughly symmetric about some center?
- Is there one interval that contains more data than other intervals?
- Are there any gaps between bars? In other words, are there intervals that do not contain data?
- In what interval did the smallest data value fall? In what interval did the largest data value fall?
- If you had to summarize the location (or “center”) of these data with one number, what number would you choose? How did you choose this number?
- Do you think the controls are properly set to produce wafers of consistent 0.5 mm thickness?

2. a. Would another sample of 10 wafers manufactured under the same control settings as your first sample behave exactly as the first sample? To find out leave all settings as they were in Question 1 and click the “Collect Sample Data” button.

b. Answer Question 1b for the new sample.

3. Leave all settings as they are. Click the “Jump To Results” button. The sample size is now set at 25. Collect two more samples by clicking the “Collect Sample Data” button twice. Then click the “Compare to Previous” button. What characteristics do the two histograms have in common? How do the two histograms differ?

The manufacturer wants wafers that are 0.5 mm thick. However, it is not possible to grind and polish wafers so that every wafer has a thickness of exactly 0.5 mm. There will always be some variability in thickness. Hence, the problem is to determine the control settings that produce wafers that are consistently close to 0.5 mm in thickness. For the remainder of this activity, use the “Jump To Results” mode for data collection and select 50 for the sample size. (You could use sample size 25, but you may find that using the larger sample size gives better results.)

4. Your first task is to determine how each control affects the thickness of a sample of wafers. In other words, you should answer the following questions:

- How do the settings of Control 1 affect wafer thickness?
- How do the settings of Control 2 affect wafer thickness?
- How do the settings of Control 3 affect wafer thickness?

a. You will need to be systematic in how you change the controls so that you can determine how each control affects wafer thickness. Describe the strategy you will use to collect data that will allow you to answer the question about the controls. (You may need to collect more than one sample from each set of control settings before you are able to see changes in the data.)

b. Carry out the strategy you have outlined in (a). Describe what affect Controls 1, 2, and 3 have on wafer thickness. Print (or draw) some histograms that support your conclusions.

5. What control settings would you recommend in order to produce wafers that are consistently close to 0.5 mm in thickness?

Explain why you chose the settings that you did.

EXERCISES

Table 3.3 is needed for Exercises 1 – 3.

State	Total	65 and older	Percent 65 and older	State	Total	65 and older	Percent 65 and older
Alabama	4,780	658	13.80%	Montana	989	147	14.90%
Alaska	710	55	7.70%	Nebraska	1,826	247	13.50%
Arizona	6,392	882	13.80%	Nevada	2,701	324	12.00%
Arkansas	2,916	420	14.40%	New Hampshire	1,316	178	13.50%
California	37,254	4,247	11.40%	New Jersey	8,792	1,186	13.50%
Colorado	5,029	550	10.90%	New Mexico	2,059	272	13.20%
Connecticut	3,574	507	14.20%	New York	19,378	2,618	13.50%
Delaware	898	129	14.40%	North Carolina	9,535	1,234	12.90%
District of Columbia	602	69	11.50%	North Dakota	673	97	14.40%
Florida	18,801	3,260	17.30%	Ohio	11,537	1,622	14.10%
Georgia	9,688	1,032	10.70%	Oklahoma	3,751	507	13.50%
Hawaii	1,360	195	14.30%	Oregon	3,831	534	13.90%
Idaho	1,568	195	12.40%	Pennsylvania	12,702	1,959	15.40%
Illinois	12,831	1,609	12.50%	Rhode Island	1,053	152	14.40%
Indiana	6,484	841	13.00%	South Carolina	4,625	362	7.80%
Iowa	3,046	453	14.90%	South Dakota	814	117	14.40%
Kansas	2,853	376	13.20%	Tennessee	6,346	853	13.40%
Kentucky	4,339	578	13.30%	Texas	25,146	2,602	10.30%
Louisiana	4,533	558	12.30%	Utah	2,764	249	9.00%
Maine	1,328	211	15.90%	Vermont	626	91	14.50%
Maryland	5,774	708	12.30%	Virginia	8,001	977	12.20%
Massachusetts	6,548	903	13.80%	Washington	6,725	828	12.30%
Michigan	9,884	1,362	13.80%	West Virginia	1,853	297	16.00%
Minnesota	5,304	683	12.90%	Wisconsin	5,687	777	13.70%
Mississippi	2,967	380	12.80%	Wyoming	564	70	12.40%
Missouri	5,989	838	14.00%				

Table 3.3. Count (in Thousands) of people over 65 by State and the District of Columbia in 2010.

1. How many people in your state are at least 65 years old? The answer varies from state to state. Table 3.3 gives the data for all 50 states and the District of Columbia for the year 2010.
 - a. Make a histogram for these data. Use class intervals of width 500,000.
 - b. Darken the bar in which your state's data value would fall. Does your state tend to have more or fewer residents 65 and older than the other states, or would you say that your state is close to typical?
 - c. Describe the overall shape of the distribution of age 65 and older. Identify any gaps in the distribution and potential outliers.
 - d. Redraw the histogram this time using class intervals of 1,000 thousand. What information is now hidden using this size of class intervals?

2. You would expect highly populated states to have higher numbers of residents over 65 than less populated states. But would the percentage of people 65 and over still be higher?
- Make a histogram of the percentage of people over 65 in each state. Choose interval widths of 1%. Darken the bar in which your state's percentage would fall. Does your state tend to have a higher or lower percentage of residents 65 and older than the other states, or would you say that your state is close to typical?
 - Describe the overall shape of the distribution of percentages. Then identify any gaps in the distribution and potential outliers.
3. Finally, we consider the total population of the states.
- Make a histogram of the total population of the states. Choose a class interval width that shows key features of the distribution.
 - Write a brief description of the most important features of the distribution of total number of state residents. Is the distribution roughly symmetric, clearly skewed, or neither? What states are unusual in their population sizes?
4. In a laboratory experiment, students were asked to estimate the breaking strength of wooden stakes. The dimensions of the stakes, measured in inches, were $8 \times 1.5 \times 1.5$. From the experiment students found the load in pounds needed to break the stakes in a sample of 20 stakes. The class data, measurements of the breaking strength in hundreds of pounds, appear below.

166	161	115	120	159
165	155	151	163	160
156	164	118	152	168
144	166	164	161	160

- Even though the wooden stakes were nearly identical, did the breaking strengths vary? Explain.
- Make a histogram of these data. Use class intervals of width 5.
- Which class interval(s) contained the most data?

d. Modify your histogram in (b) so that the scale on the vertical axis is the percent of the stakes whose breaking strength is in each class interval. How does the shape of your modified histogram compare to your histogram in (b)?

e. Write a short paragraph describing key features of the distribution of breaking strengths.

REVIEW QUESTIONS

Table 3.4, needed for questions 1 and 2, consists of a list of the top 100 major league baseball players ranked according to career batting average. (Notice that because of ties for the 100th place, there are actually 104 players on this list.)

The table contains the following information for each player: number of career years, the last year for which data were collected, career batting average, and career number of home runs.

(See full table on next page...)

First	Last	Career Years	Last Career Year	Career Batting Avg.	Career Home Runs	First	Last	Career Years	Last Career Year	Career Batting Avg.	Career Home Runs
Cap	Anson	22	1897	0.331	97	Willie	Keeler	19	1910	0.341	33
Luke	Appling	20	1950	0.310	45	Joe	Kelly	17	1908	0.317	65
Earl	Averill	13	1941	0.318	238	Chuck	Klein	17	1944	0.320	300
Ginger	Beaumont	12	1910	0.311	39	Nap	Lajoie	21	1916	0.339	83
Wade	Boggs	18	1999	0.328	118	Henry	Larkin	10	1893	0.310	53
Jim	Bottomley	16	1937	0.310	219	Freddie	Lindstrom	13	1936	0.311	103
Dan	Brouthers	19	1904	0.342	106	Denny	Lyons	13	1897	0.318	62
Pete	Browing	13	1894	0.349	46	Heinie	Manush	17	1939	0.330	110
Jessie	Burkett	16	1905	0.338	75	Edgar	Martinez	18	2004	0.312	309
Miguel	Cabrera	10	2012	0.318	312	Joe	Mauer	9	2012	0.322	93
Rod	Carew	19	1985	0.328	92	Barney	McCosky	11	1953	0.312	24
Fred	Clarke	21	1915	0.312	67	John	McGraw	16	1906	0.334	13
Roberto	Clemente	18	1972	0.317	240	Joe	Medwik	17	1948	0.324	205
Ty	Cobb	24	1929	0.366	117	Irish	Meusel	11	1927	0.310	106
Mickey	Cochrane	13	1937	0.320	119	Bing	Miller	16	1936	0.311	116
Eddie	Collins	25	1930	0.333	47	Dale	Mitchell	11	1956	0.312	41
Earle	Combs	12	1935	0.325	58	Johnny	Mize	15	1953	0.312	359
Roger	Connor	18	1897	0.317	138	Stan	Musial	22	1963	0.331	475
Kiki	Cuyler	18	1938	0.321	128	Tip	O'Neill	10	1892	0.334	52
Ed	Delahanty	16	1903	0.346	101	Jim	O'Rourke	9	1904	0.310	80
Bill	Dickey	17	1946	0.313	202	Kirby	Puckett	12	1995	0.318	207
Joe	DiMaggio	13	1951	0.325	361	Albert	Pujols	12	2012	0.325	474
Mike	Donlin	12	1914	0.333	51	Rip	Radcliff	10	1943	0.311	42
Hugh	Duffy	17	1906	0.324	106	Manny	Ramirez	19	2011	0.312	555
Bibb	Falk	12	1931	0.314	69	Sam	Rice	20	1934	0.322	34
Elmer	Flick	13	1910	0.313	48	Jackie	Robinson	10	1956	0.311	137
Bob	Fothergill	12	1933	0.325	36	Edd	Roush	18	1931	0.323	68
Jack	Fournier	15	1927	0.313	136	Babe	Ruth	22	1935	0.342	714
Jimmie	Foxx	20	1945	0.325	534	Joe	Sewell	14	1933	0.312	49
Frankie	Frisch	19	1937	0.316	105	Al	Simmons	15	1944	0.334	307
Nomar	Garciparra	14	2009	0.313	229	George	Sisler	14	1930	0.340	102
Lou	Gehrig	17	1939	0.340	493	Elmer	Smith	14	1901	0.312	37
Charlie	Gehringer	19	1942	0.320	184	Tris	Speaker	22	1928	0.345	117
Goose	Goslin	18	1935	0.316	248	Riggs	Stephenson	14	1934	0.336	63
Hank	Greenberg	13	1947	0.314	331	Ichiro	Suzuki	12	2012	0.322	102
Vladimir	Guerrero	16	2011	0.318	449	Bill	Terry	14	1936	0.341	154
Tony	Gwynn	20	2001	0.338	135	Sam	Thompson	15	1906	0.331	127
Chick	Hafey	13	1937	0.317	164	Mike	Tiernan	13	1899	0.311	106
Billy	Hamilton	14	1901	0.344	40	Cecil	Travis	12	1947	0.314	27
Harry	Heilmann	17	1932	0.342	183	Pie	Traynor	17	1937	0.320	58
Todd	Helton	16	2012	0.320	354	George	Van Haltren	17	1903	0.316	69
Babe	Herman	13	1945	0.324	181	Arky	Vaughan	14	1948	0.318	96
Matt	Holliday	9	2012	0.314	227	Bobby	Veach	14	1925	0.310	64
Rogers	Hornsby	23	1937	0.359	301	Honus	Wagner	21	1917	0.327	101
Baby Doll	Jackson	11	1927	0.311	83	Larry	Walker	17	2005	0.313	383
Joe	Jackson	13	1920	0.356	54	Paul	Waner	20	1945	0.333	113
Hughie	Jennings	17	1918	0.311	18	Lloyd	Waner	18	1945	0.316	27
Derek	Jeter	11	2005	0.313	254	Zack	Wheat	19	1927	0.317	132
Willie	Keeler	19	1910	0.341	33	Ted	Williams	19	1960	0.344	521
Joe	Kelly	17	1908	0.317	65	Ken	Williams	14	1929	0.319	196
Chuck	Klein	17	1944	0.320	300	Taffy	Wright	9	1949	0.311	38
Nap	Lajoie	21	1916	0.339	83	Ross	Youngs	10	1926	0.322	42

Table 3.4: Top 100 career batting averages in baseball (at the end of the 2012 season).

1. Make two histograms for the career home run data. For the first histogram, use class intervals of size 100 and in the second, use class intervals of size 50. Describe the overall shape of the data based on each of your histograms. Also identify any potential outliers. Explain what new information you can obtain from the second histogram that was not visible in the first.

2. a. Make two histograms for Career Years. Use the following class intervals:

Histogram 1: 9 – 14, 14 – 19, 19 – 24, and 24 – 29.

Histogram 2: 9 – 11, 11 – 13, 13 – 15, 15 – 17, 17 – 19, 19 – 21, 21 – 23, 23 – 25.

b. Did any of the career years fall on a boundary of a class interval? If so, how did you classify those data values?

c. Describe the overall shape of each of the two histograms. In particular would you describe the shape as symmetric or skewed? Would you characterize the shape as unimodal (one peak), bimodal (two peaks), or multimodal? Did changing the class intervals affect the shape of the distribution?

3. The duration of 40 phone calls (in minutes) for technical support is given below.

12.0	3.3	0.5	48.7	16.7	1.2	14.8	8.2	9.0	5.7
11.5	17.5	3.2	20.8	7.3	8.0	0.2	51.2	3.3	5.2
12.3	24.5	13.3	7.7	13.5	4.3	13.7	10.7	18.8	15.7
3.2	38.7	16.2	23.3	9.7	4.7	6.5	0.5	45.1	5.3

a. Make a copy of Table 3.5 and then complete the frequency distribution table for the call duration data.

(See table on next page...)

Duration (minutes)	Frequency	Percent
0 – 6		
6 – 12		
12 – 18		
18 – 24		
24 – 30		
30 – 36		
36 – 42		
42 – 48		
48 – 54		

Table 3.5. Frequency distribution table for duration of phone calls.

- b. What percentage of phone calls lasted less than 12 minutes?
- c. What percentage of calls lasted a half hour or more?
- d. Represent the frequency distribution with a histogram. Use a percent scale on the vertical axis.
- e. Describe the shape of the distribution. Are there any gaps in the data? Outliers?