5.2 Frequency Tables, Histograms, and Frequency Polygons

GOAL

Create frequency tables and graphs from a set of data.

YOU WILL NEED

- calculator
- grid paper

EXPLORE...

- Margaret inherited her grandfather's coin collection. How can she organize a catalogue of the coins to see how many of each type of coin she has?

LEARN ABOUT the Math

Flooding is a regular occurrence in the Red River basin. During the second half of the 20th century, there have been nine notable floods, four of which have been severe, occurring in 1950, 1979, 1996, and 1997. The flood that occurred in 1997 is known as the "flood of the century" in Manitoba and North Dakota.

The following data represents the flow rates of the Red River from 1950 to 1999, as recorded at the Redwood Bridge in Winnipeg, Manitoba.

Maximum Water Flow Rates for the Red River, from 1950 to 1999, Measured at Redwood Bridge*

Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)	Year	Flow Rate (m^3/s)
1950	3058	1960	1965	1970	2280	1980	881	1990	396
1951	1065	1961	481	1971	1526	1981	159	1991	280
1952	1008	1962	1688	1972	1589	1982	1458	1992	1399
1953	357	1963	660	1973	530	1983	1393	1993	946
1954	524	1964	1002	1974	2718	1984	1048	1994	1121
1955	1521	1965	1809	1975	1671	1985	991	1995	1877
1956	1974	1966	2498	1976	1807	1986	1812	1996	3058
1957	654	1967	1727	1977	187	1987	2339	1997	4587
1958	524	1968	510	1978	1750	1988	564	1998	1557
1959	991	1969	2209	1979	3030	1989	1390	1999	2180

National Research Council Canada

(*assumes NO flood protection works in place, for data after 1969 when the floodway was in use)

? How can you approximate the water flow rate that is associated with serious flooding in Winnipeg?

EXAMPLE 1 Creating a frequency distribution

Determine the water flow rate that is associated with serious flooding by creating a **frequency distribution**.

frequency distribution
A set of intervals (table or graph), usually of equal width, into which raw data is organized; each interval is associated with a frequency that indicates the number of measurements in this interval.

Francine's Solution: Creating a frequency distribution table

Highest water flow rate: 4587 m^3/s, in 1997
Lowest water flow rate: 159 m^3/s, in 1981

I decided to organize the data from the table on page 213 into a frequency distribution table because there are too many numbers to order easily. A table would allow me to see, at a glance, the frequency of various flow rates.

Range: $4587 - 159$, or 4428

I determined the range of the data so that I could choose a suitable interval.

$$\frac{4428}{10} = 442.8$$

If the interval width is 500, the intervals will end at 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, and 5000.

Most tables have between 5 and 12 intervals, so I decided to use 10 equally sized intervals to sort the data. Since the range is 4428, I needed an interval width of at least 442.8. I decided to round this value to 500, since 500 will be easier to work with.

Flow Rate (m^3/s)	Tally	Frequency (number of years)
0–500	𝍸 |	6
500–1000	𝍸 𝍸 |	11
1000–1500	𝍸 ||||	9
1500–2000	𝍸 𝍸 ||||	14
2000–2500	𝍸	5
2500–3000	|	1
3000–3500	|||	3
3500–4000		0
4000–4500		0
4500–5000	|	1

Each interval begins just after the first value in the row and includes all the numbers up to and including the last number. The first interval begins at 0.01 and goes up to 500. This ensures that a number like 500 is counted in only one row.

Communication | *Tip*

In frequency tables in this resource, the upper limit of each interval includes that number. For example, in the Flow Rate table to the left, 2000 is included in the interval 1500–2000 and not in the interval 2000–2500.

Flow Rate (m^3/s)	Tally	Frequency (number of years)
0–500	𝍸 |	6
500–1000	𝍸 𝍸 |	11
1000–1500	𝍸 ||||	9
1500–2000	𝍸 𝍸 ||||	14
2000–2500	𝍸	5
2500–3000	|	1
3000–3500	|||	3 floods
3500–4000		0
4000–4500		0
4500–5000	|	1 flood

I could see that 17 of the 50 years had a water flow less than or equal to 1000 m^3/s. There would have been no flooding in those years.

From the data, I knew that the "flood of the century" had a flow rate of 4587 m^3/s. Looking at the last row in my frequency table, I noticed that this was significantly higher than all the other flow rates.

I knew that there were nine floods, and four were severe. Since there were floods in the four years when flow rates were greater than 3000 m^3/s, flow rate and flooding are likely connected. Four of the six flow rates in the 2000–3000 interval would probably have caused floods. The minimum flow rate that results in a flood should be in the 2000–2500 interval.

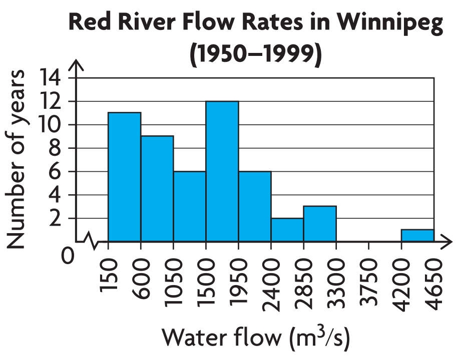
I predict that water flow rates that result in serious flooding are greater than 2000 m^3/s.

I can check my prediction by comparing the years that had flow rates from 2000 to 3000 m^3/s with the historical records of flooding.

Tasha's Solution: Creating a histogram

Flow Rate (m^3/s)	Tally	Frequency (number of years)
150–600	𝍸 𝍸 |	11
600–1050	𝍸 ||||	9
1050–1500	𝍸 |	6
1500–1950	𝍸 𝍸 ||	12
1950–2400	𝍸 |	6
2400–2850	||	2
2850–3300	|||	3
3300–3750		0
3750–4200		0
4200–4650	|	1

I created a frequency table using an interval width of 450 for the Red River water flow data from page 213.



I drew a **histogram** to represent the data in the frequency distribution, since the data is grouped into intervals.

The interval width for my histogram is 450. I labelled the intervals on the horizontal axis. I labelled the frequency on the vertical axis "Number of years."

histogram
The graph of a frequency distribution, in which equal intervals of values are marked on a horizontal axis and the frequencies associated with these intervals are indicated by the areas of the rectangles drawn for these intervals.

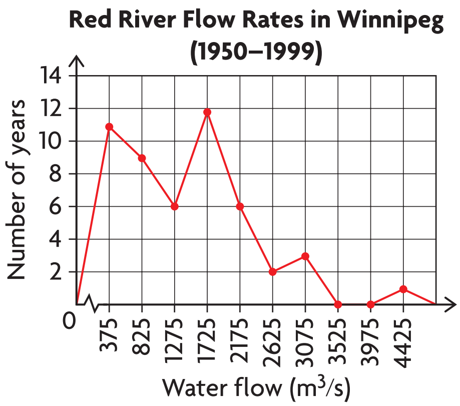
In general, with the exception of the interval 1500–1950, as the maximum flow rate increases, the number of data points in each interval decreases. Low maximum flow rates have been more common than high maximum flow rates.

There were nine floods. Based on my histogram, the flow rate was greater than 1950 m^3/s in only 12 years. These 12 years must include the flood years. I predict that floods occur when the flow rate is greater than 1950 m^3/s.

Monique's Solution: Creating a frequency polygon

Flow Rate (m^3/s)	Midpoint	Frequency (number of years)
150–600	375	11
600–1050	825	9
1050–1500	1275	6
1500–1950	1725	12
1950–2400	2175	6
2400–2850	2625	2
2850–3300	3075	3
3300–3750	3525	0
3750–4200	3975	0
4200–4650	4425	1

I created a frequency table using an interval width of 450 for the Red River water flow data from page 213. I determined the midpoint of each interval by adding the boundaries of each interval and dividing by 2.



Most of the data is in the first four intervals, and the most common water flow is between 1500 and 2000 m^3/s. After this, the frequencies drop off dramatically.

There were six years where the flow rate was around 2625, 3075, or 4425 m^3/s. These must have been flood years. The other three floods should have occurred when the flow rate was around 2175 m^3/s. According to my frequency polygon, there were flows around that midpoint in six years. Assuming that the flow rate in three of those years was 2175 m^3/s or greater, floods should occur when the flow rate is 2175 m^3/s or greater.

Next, I labelled the axes on my graph.

Finally, I plotted the midpoints and joined them to form a **frequency polygon**. I included one interval above the highest value, with a frequency of 0, and connected the first midpoint to 0 to close the polygon.

The graph made it easy to see that the flow rate of 4425 m^3/s was unusual.

frequency polygon
The graph of a frequency distribution, produced by joining the midpoints of the intervals using straight lines.

Reflecting

A. Identify similarities and differences in the frequency distributions created by Francine, Tasha, and Monique.

B. Francine used a different interval width than Tasha and Monique. How did this affect the distribution of the data? Explain.

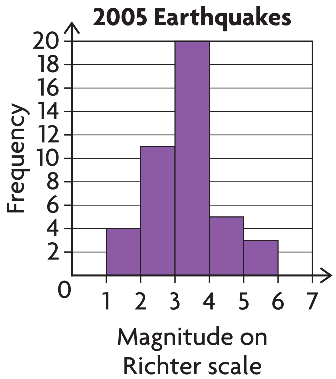
C. Suppose that Francine created a frequency distribution using an interval width of 200. Do you think this interval would make it easier to see which water flow rates result in flooding? Why or why not?

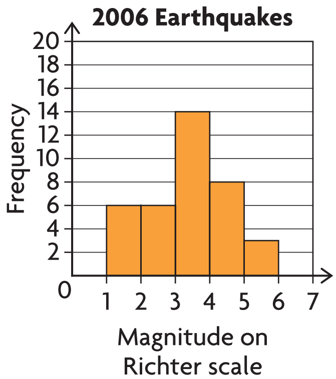
D. Would 2000 be a good interval width for determining critical water flow rates? Explain.

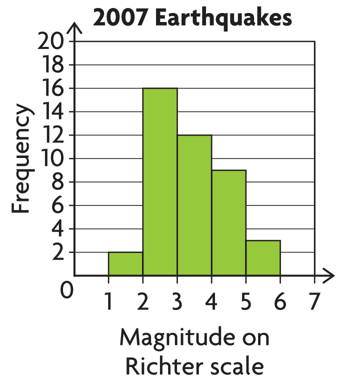
APPLY the Math

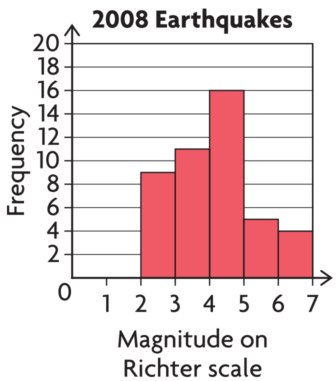
EXAMPLE 2 Comparing data using histograms

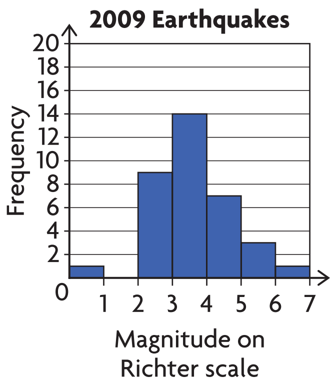
The magnitude of an earthquake is measured using the Richter scale. Examine the histograms for the frequency of earthquake magnitudes in Canada from 2005 to 2009. Which of these years could have had the most damage from earthquakes?











National Research Council Canada

Understanding the Richter Scale*	
Magnitude	**Effects**
less than 3.0	recorded by seismographs; not felt
3.0–3.9	feels like a passing truck; no damage
4.0–4.9	felt by nearly everyone; movement of unstable objects
5.0–5.9	felt by all; considerable damage to weak buildings
6.0–6.9	difficult to stand; partial collapse of ordinary buildings
7.0–7.9	loss of life; destruction of ordinary buildings
more than 7.9	widespread loss of life and destruction

*Every unit increase on the Richter scale represents an earthquake 10 times more powerful. For example, an earthquake measuring 5.6 is 10 times more powerful than an earthquake measuring 4.6.

Bilyana's Solution: Using a frequency table

2005 had the most earthquakes in any one category: 20 earthquakes with a magnitude from 3.0 to 3.9.

At first glance, it seemed that 2005 was the worst year, because it had the highest bar. However, an earthquake of magnitude 3 on the Richter scale does not cause much damage.

Year	2005	2006	2007	2008	2009
Frequency of Earthquakes from 4.0 to 4.9	5	8	9	1	67

I knew that some minor damage will occur if the magnitude of an earthquake is from 4.0 to 4.9 on the Richter scale. I decided to examine the frequency of these earthquakes for 2005 to 2009.

2008 had the greatest number of earthquakes with the potential for minor damage.

Four of the years had three earthquakes with magnitudes from 5.0 to 5.9, while 2008 had five earthquakes with these magnitudes.

Year	Magnitude on Richter Scale			Total
	4.0–4.9	**5.0–5.9**	**6.0–6.9**	
2008	16	5	4	25
2009	7	3	1	11

Only 2008 and 2009 had earthquakes with magnitudes from 6.0 to 6.9, with the potential to cause moderate damage. I decided to examine these years more carefully.

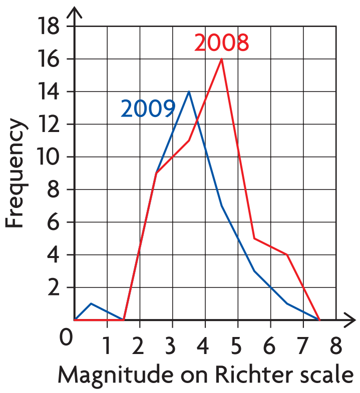
I created a table for earthquakes from 4.0 to 6.9 on the Richter scale for these years.

Therefore, 2008 could have had the most damage from earthquakes.

Miguel's Solution: Comparing frequency polygons

Both 2008 and 2009 had the strongest earthquakes, registering from 6.0 to 6.9 on the Richter scale.

I examined the histograms. There were only two histograms with earthquakes that registered more than 6 on the Richter scale. These years could have had the most damage, because an earthquake registering from 6.0 to 6.9 will result in moderate damage.



I decided to draw frequency polygons, instead of histograms, for these two years. I drew both polygons on the same graph to compare them.

The number of earthquakes in the three highest intervals was greater in 2008 than in 2009, so 2008 could have had the most damage from earthquakes.

I compared the shapes of the frequency polygons.

Your Turn

a) Compare Bilyana's solution with Miguel's solution.

b) What other factors should be considered when determining which year could have had the most damage from earthquakes?

In Summary

Key Ideas

- Large sets of data can be difficult to interpret. Organizing the data into intervals and tabulating the frequency of the data in each interval can make it easier to interpret the data and draw conclusions about how the data is distributed.
- A frequency distribution is a set of intervals and can be displayed as a table, a histogram, or a frequency polygon.

Need to Know

- A frequency distribution should have a minimum of 5 intervals and a maximum of 12 intervals, although any number of intervals is possible. Too many or too few intervals will result in a table or a graph that may not effectively show how the data is distributed.
- The interval width can be determined by dividing the range of the data by the desired number of intervals and then rounding to a suitable interval width.
- The height of each bar in a histogram corresponds to the frequency of the interval it represents.
- Because the individual pieces of data are not visible in a frequency distribution, the minimum and maximum values and the median cannot be determined directly.
- Frequency polygons serve the same purpose as histograms. However, they are especially helpful for comparing multiple sets of data because they can be graphed on top of each other.

CHECK Your Understanding

1. The numbers of earthquakes in the world during two five-year periods are shown in the frequency table.

Magnitude	Years	
	2000–04	2005–09
0.1–0.9	253	5
1.0–1.9	6 957	133
2.0–2.9	28 391	19 120
3.0–3.9	33 717	43 701
4.0–4.9	43 890	58 100
5.0–5.9	6 487	8 948
6.0–6.9	675	770
7.0–7.9	70	61
8.0–9.9	5	8

a) Draw a frequency polygon for the numbers of earthquakes during each five-year period on the same graph.

b) Use your graph to compare the earthquakes in the world during the two five-year periods.

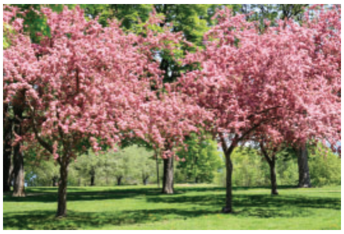
2. Emmanuella walks her golden retriever regularly. She kept track of the lengths of her walks for one month and grouped the data in a frequency table.

a) The first walk was 15 min long. In which interval did she place this piece of data?

b) Draw a frequency polygon to represent the data in the table. Describe how the data is distributed.

Length of Walk (min)	Frequency
5–10	1
10–15	3
15–20	7
20–25	10
25–30	6
30–35	11
35–40	8
40–45	5
45–50	4
50–55	2
55–60	3

PRACTISING



3. A cherry orchard has 30 trees with these heights, given in inches.

78	70	83	79	74	81	80	65	66	76
85	82	74	63	75	76	86	80	72	72
80	69	71	80	77	81	75	75	64	87

   a) Make a frequency table with six intervals to organize the heights.
   b) Construct a histogram of the data.
   c) Which range of heights occurs most frequently? Which occurs least frequently?

4. The amounts withdrawn from an ATM, in dollars, are recorded for a single Wednesday.

20	120	50	70	60	80	140	120	80	160
80	60	110	100	100	80	180	160	40	100
50	80	200	140	160	60	40	80	60	140
100	140	160	200	140	20	80	20	100	70
40	20	120	40	140	100	40	50	180	60

   a) What interval width will give a good representation of how the data is distributed?
   b) Sort the amounts in a frequency distribution table.
   c) Construct a histogram to represent the table in part b).
   d) Describe how the data is distributed.

5. The final scores for the 30 women who competed in the women's figure skating competition at the Vancouver 2010 Olympics are shown. Canadian Joannie Rochette captured the bronze medal.

78.50	63.76	61.02	53.16	50.74	43.84
73.78	63.02	59.22	52.96	49.74	43.80
71.36	62.14	57.46	52.16	49.04	41.94
64.76	61.92	57.16	51.74	49.02	40.64
64.64	61.36	56.70	50.80	46.10	36.10

   a) Make a frequency table to organize the scores.
   b) Draw a histogram of the data.
   c) Does your histogram help you see the range of scores that corresponded to a top-five placement? Explain.

Joannie Rochette trains in St-Leonard, Quebec.

6. **a)** On the same graph, draw frequency polygons to show the populations of males and females in Canada for the year 2009.

b) Examine your graph. Describe any differences you notice in the populations of the two sexes.

Population by Gender and Age Group in 2009		
Age Group	**Male (%)**	**Female (%)**
0–4	5.6	5.3
5–9	5.5	5.1
10–14	6.0	5.7
15–19	6.9	6.5
20–24	7.1	6.6
25–29	7.1	6.8
30–34	6.8	6.6
35–39	6.9	6.7
40–44	7.5	7.2
45–49	8.4	8.2
50–54	7.7	7.6
55–59	6.5	6.6
60–64	5.5	5.7
65–69	4.1	4.3
70–74	3.0	3.4
75–79	2.4	2.9
80–84	1.6	2.4
85–89	0.9	1.6
90+	0.3	0.9
Total	99.8%	100.1%

Statistics Canada

7. The following frequency table shows the number of production errors in vehicles coming off an assembly line during the first, second, third, and fourth hour of the day shift.

Number of Errors		1	2	3	4	5	6	7	8
Frequency	**1st Hour**	1	71	38	5	4	3	1	1
	2nd Hour	2	61	06	5	4	2	0	0
	3rd Hour	3	01	05	3	1	1	0	0
	4th Hour	3	58	3	2	1	0	0	0

a) Draw all the frequency polygons for the data on the same graph.

b) What conclusions can you make, based on your graph?

8. Holly and Jason have 14-week training programs to prepare them to run a marathon. On different days during their programs, they run different distances. Holly plans to run the half marathon (21.1 km). Jason plans to run the full marathon (42.2 km). The distances that they run on various training days are shown below.

a) Construct a frequency distribution table for each training program. Explain the size of interval that you choose.

b) Use your frequency distribution table to graph a frequency polygon for each runner on the same graph.

c) Compare the two training programs.

	Holly's Program (km)					Jason's Program (km)					
Week	**Tues.**	**Wed.**	**Thurs.**	**Fri.**	**Sun.**	**Mon.**	**Tues.**	**Wed.**	**Thurs.**	**Fri.**	**Sun.**
1	8	5	8	5	10	off	13	5	13	6	16
2	8	5	10	5	10	off	13	6	13	6	19
3	10	5	13	5	13	off	13	6	16	6	22
4	10	5	13	6	13	5	13	6	16	6	26
5	10	6	13	6	13	6	16	6	16	6	19
6	13	6	13	6	16	6	16	6	16	6	29
7	13	6	13	6	16	6	16	6	19	6	29
8	13	6	13	6	13	6	16	8	19	8	29
9	13	6	16	6	16	6	16	8	19	10	32
10	13	8	13	8	13	8	16	8	16	13	35
11	13	8	16	8	19	6	16	8	16	8	22
12	16	6	16	6	22	10	16	10	19	6	35
13	10	5	13	5	10	5	13	5	10	5	16
14	8	off	8	3	race day	5	10	5	8	3	race day

Serious Injuries in Car Accidents		
Age Group	**Driver (%)**	**Passenger (%)**
0–4	0	5.2
5–14	0.4	10.8
15–24	24.6	33.5
25–34	20.6	12.5
35–44	18.9	9.7
45–54	15.7	8.1
55–64	9.5	6.7
65+	9.9	9.5
not stated	0.4	3.9
Total	100	99.9

9. Examine the given data for injuries, taken from the Canadian Motor Vehicle Traffic Collision Statistics. On the same graph, draw two frequency polygons to illustrate the percent of serious injuries for drivers and for passengers, by age group. Compare the data distributions. What conclusions can you make?

10. For a histogram to display the distribution of data accurately, intervals of equal width must be used. Explain why, using examples.

Closing

11. a) What are the advantages of grouping raw data into intervals?
b) How does a histogram differ from a frequency polygon? When would using frequency polygons be better than using histograms?

Extending

12. This table gives the Aboriginal population in 151 metropolitan areas across Canada, based on statistics from the 2006 census. In 2006, the total Aboriginal population was 1 172 785.
a) Use this table to estimate the mean population. Explain what you did.
b) Use this table to estimate the median population. Explain what you did.
c) Using the actual data, the value for the median population is 1700 and the value for the mean is 4275. How do your estimates compare with these values? Explain any discrepancies.

Aboriginal Populations in Cities, 2006

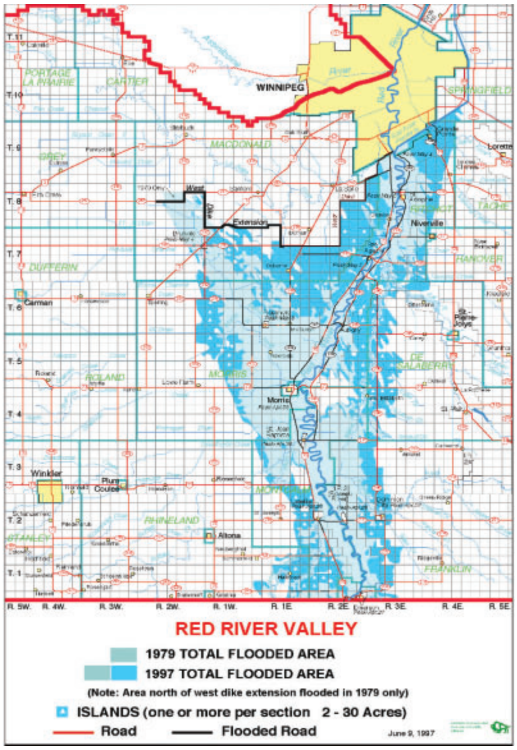
Population	Frequency
0–5 000	122
5 000–10 000	16
10 000–15 000	4
15 000–20 000	2
20 000–25 000	2
25 000–30 000	2
30 000–35 000	0
35 000–40 000	0
40 000–45 000	1
45 000–50 000	0
50 000–55 000	1
55 000–60 000	0
60 000–65 000	0
65 000–70 000	1

Statistics Canada

History Connection

Duff's Ditch

Winnipeg's Red River Floodway was constructed between 1962 and 1968 to protect the city from severe flooding. Affectionately known as Duff's Ditch (after Premier Duff Roblin, who insisted that the project go forward), the floodway is a 47 km channel that diverts water around the city to allow river levels in Winnipeg to remain below flood level. The gates to the floodway are opened whenever the city is threatened by flooding.



A. Flooding starts to happen in Winnipeg when the water level is 5.5 m. This is equivalent to a flow rate of about 1470 m^3/s. When the level gets close to 4.6 m, sandbagging is needed in low-lying areas and the floodway is opened. How many times between its first use in 1969 and 1999 was the floodway opened? (Use the data tables on page 213.)

B. Do you think that high flow rates are becoming the norm rather than the exception? Refer to the data tables on page 213, and do research to find more recent data. Justify your answer.