Problem set 1. Descriptive Statistics

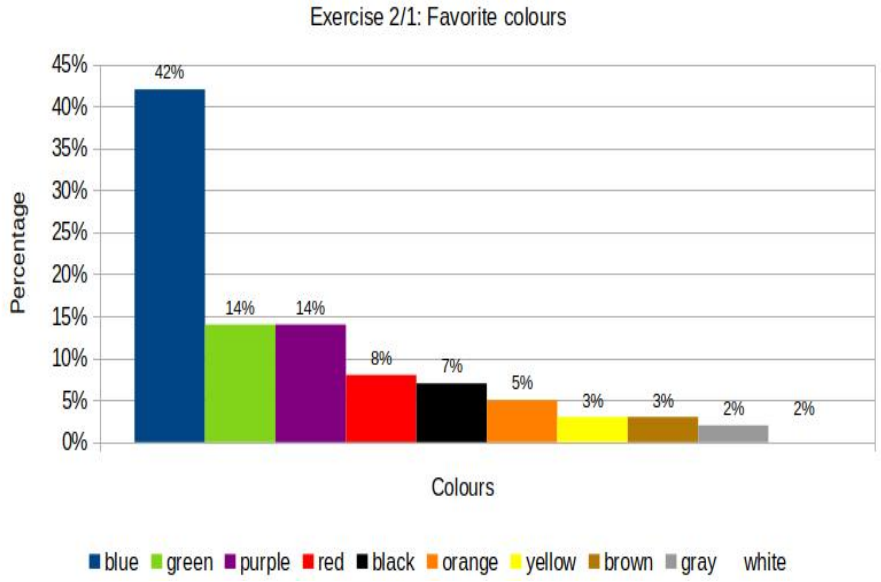
Exercise 1. Survey of students. A survey of students in an introductory statistics class asked the following questions:
(a) age;
(b) do you like to dance? (yes, no);
(c) can you play a musical instrument (not at all, a little, pretty well);
(d) how much did you spend on food last week?
(e) height;
(f) do you like broccoli? (yes, no).
Classify each of these variables as categorical or quantitative and give reasons for your answers.

Odp. 1

a) Quantitative
b) Categorical
c) Categorical
d) Quantitative
e) Quantitative
f) Categorical

Exercise 2. Favorite colors. What is your favorite color? One survey produced the following summary of responses to that question: blue, 42%; green, 14%; purple, 14%; red, 8%; black, 7%; orange, 5%; yellow, 3%; brown, 3%; gray, 2%; and white, 2%. Make a bar graph of the percents and write a short summary of the major features of your graph.

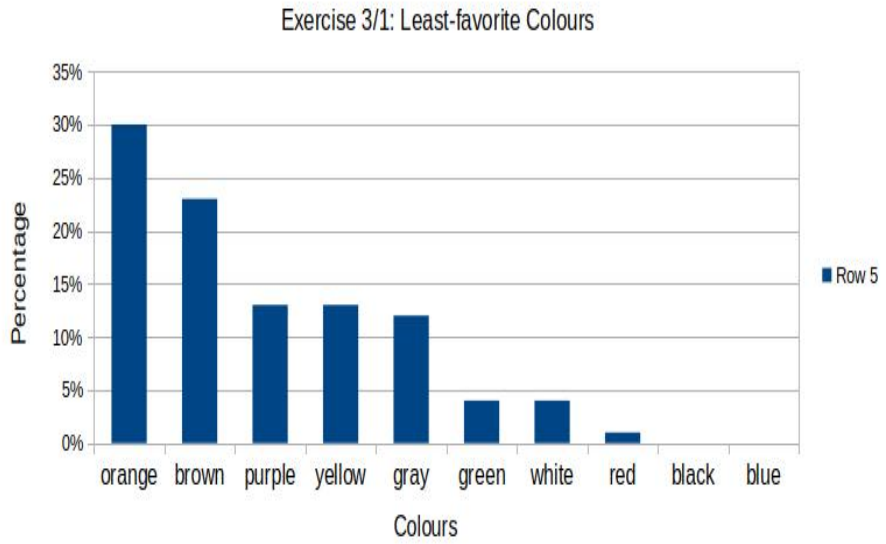
Odp. 2



Blue was by far the most popular choice.
70% of the people in the survey chose 3 out of 10 colours.

Exercise 3. Least-favorite colors. Refer to the previous exercise. The same study also asked people about their least-favorite color. Here are the results: orange, 30%; brown, 23%; purple, 13%; yellow, 13%; gray, 12%; green, 4%; white, 4%; red, 1%; black, 0%; and blue, 0%. Make a bar graph of these percents and write a summary of the results.

Odp. 3



Even if two options are 0%, they must be included in the graph. Orange was the most frequent answer chosen by 30% of the participants, followed by brown at 23% and together, they account for more than half of the least-favorite colour. Black and blue were not disliked by any participant.

Exercise 4. Ages of survey respondents. The survey about color preferences reported the age distribution of the people who responded. Here are the results:

Age group (years)	1-18	19-24	25-35	36-50	51-69	70 and over
Count	10	97	70	36	14	5

(a) Add the counts and compute the percents for each age group.
(b) Make a bar graph of the percents.
(c) Describe the distribution.
(d) Explain why your bar graph is not a histogram.

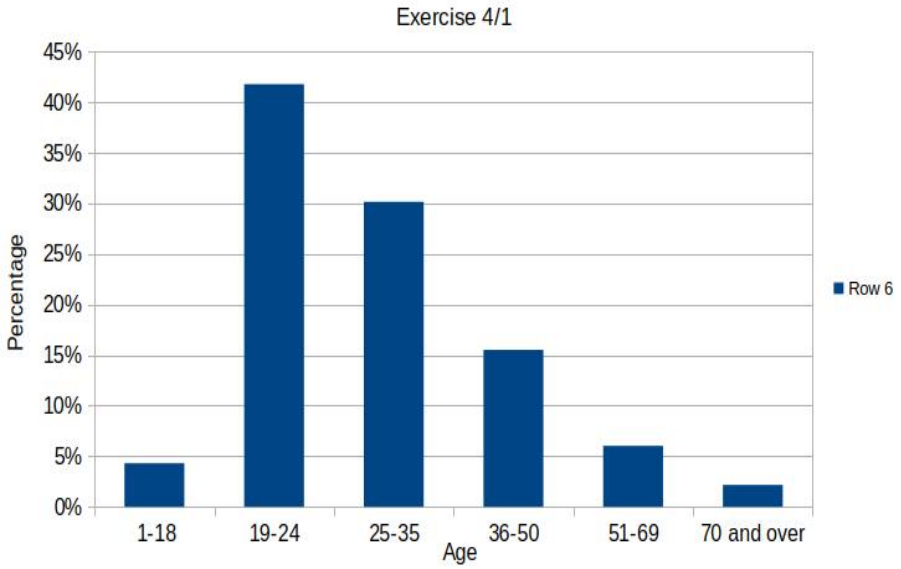
Odp. 4

$$n = \sum_{i=1}^{6} x_i = 232$$

$$Percentage = \frac{x_i}{n} \times 100\%$$

Age group (years)	1-18	19-24	25-35	36-50	51-69	70 and over
Count	10	97	70	36	14	5
Percentage	4.31%	41.81%	30.17%	15.52%	6.03%	2.16%

b) Age has been sorted into groups, and it is considered here categorical data



c) The survey was more focused on younger ages (i.e., 19-35, $\approx 72\%$ of the total survey)

d) A histogram is used for quantitative data where the bars represent continuous intervals in numbers, where the order matters. In a bar graph, we can sort the data according to our preference, and we use it because we have sorted the age groups and the data is categorical.

Exercise 5. Garbage. The formal name for garbage is "municipal solid waste". The table below gives a breakdown of the materials that made up American municipal solid waste.

Material	Weight (million tons)	Percent of total
Food scraps	25.9	11.2
Glass	12.8	5.5
Metals	18.0	7.8
Paper, paperboard	86.7	10.7
Plastics	24.7	37.4
Rubber, leather, textiles	15.8	6.8
Wood	12.7	5.5
Yard trimmings	27.7	11.9
Other	7.5	3.2

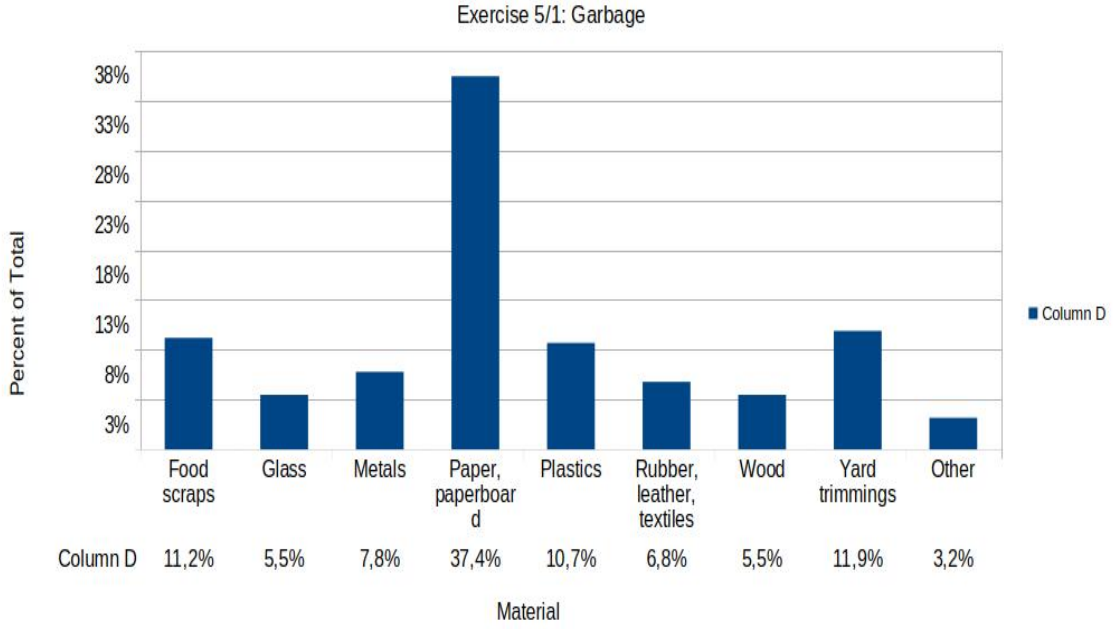
(a) Add the weights for the nine materials given, including "Other". Each entry, including the total, is separately rounded to the nearest tenth. So the sum and the total may differ slightly because of roundoff error.

(b) Make a bar graph of the percents. The graph gives a clearer picture of the main contributors to garbage if you order the bars from tallest to shortest.

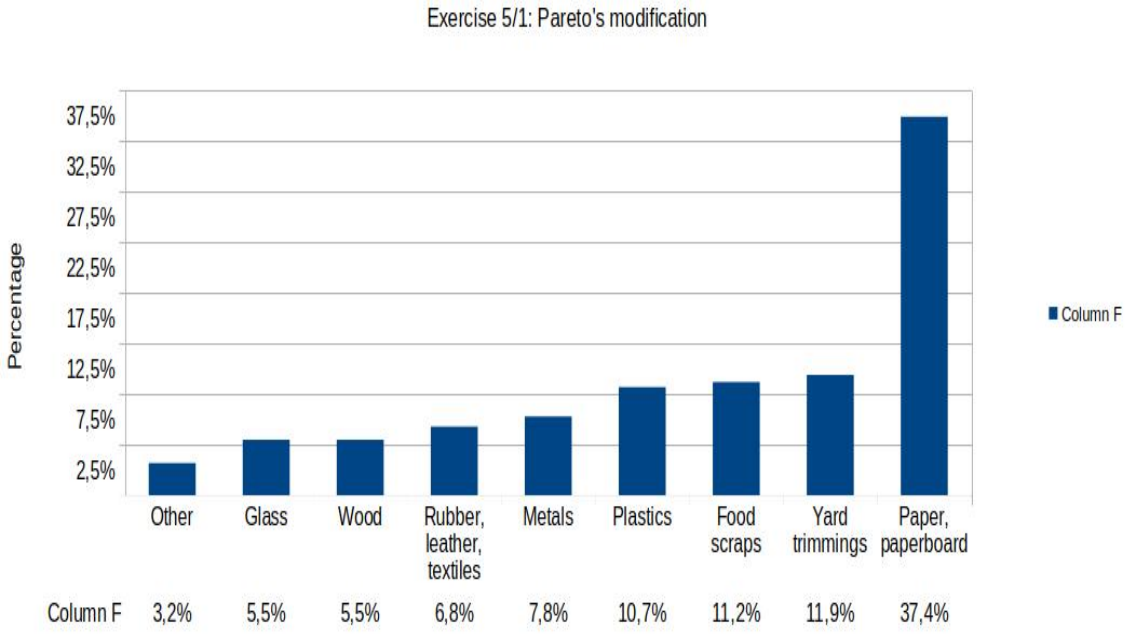
(c) If you use software, also make a pie chart of the percents. Comparing the two graphs, notice that it is easier to see the small differences among "Food scraps", "Plastics", and "Yard trimmings" in the bar graph.

Odp. 5

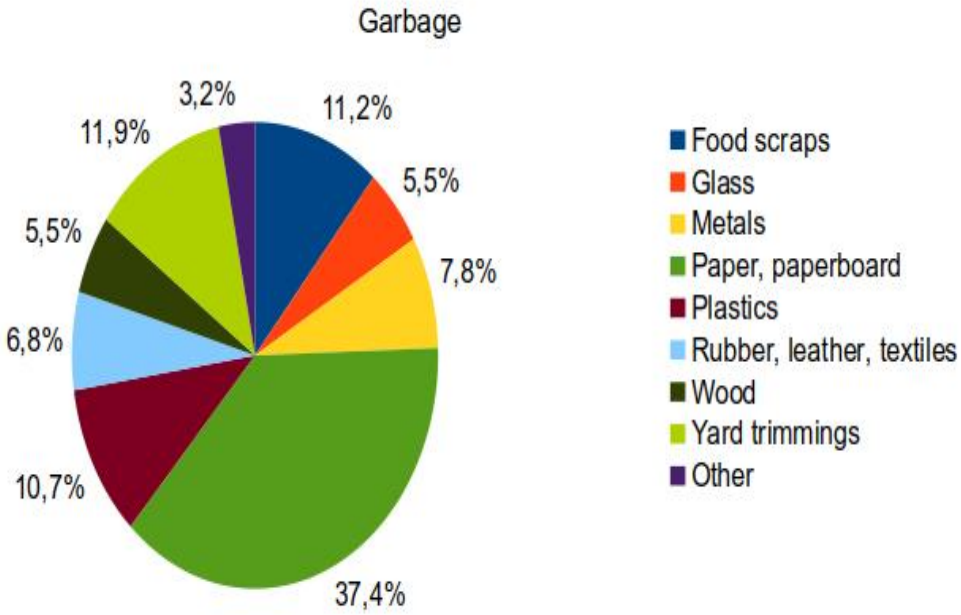
a) Sum: Million Tons: 244, Total percentage: 105.5%. The sum of the total percentage is more than 100 because of the rounding error



b) After reordering:



c) Pie chart

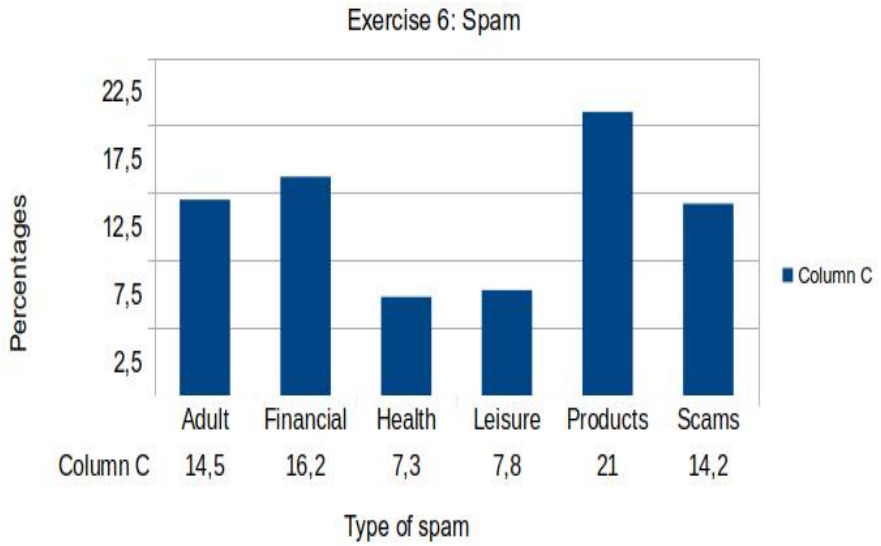


Exercise 6. Spam. Email spam is the curse of the Internet. Here is a compilation of the most common types of spam:

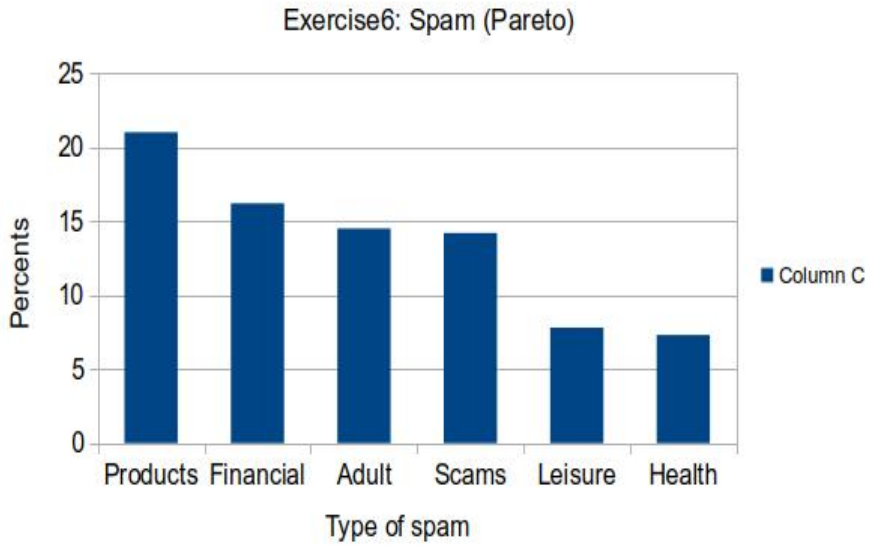
Type of spam	Percent
Adult	14.5
Financial	16.2
Health	7.3
Leisure	7.8
Products	21.0
Scams	14.2

Make two bar graphs of these percents, one with bars ordered as in the table (alphabetical) and the other with bars in order from tallest to shortest. Comparisons are easier if you order the bars by height. A bar graph ordered from tallest to shortest bar is sometimes called a Pareto chart, after the Italian economist who recommended this procedure.

Odp. 6
Alphabetical order:



Tallest to shortest order:

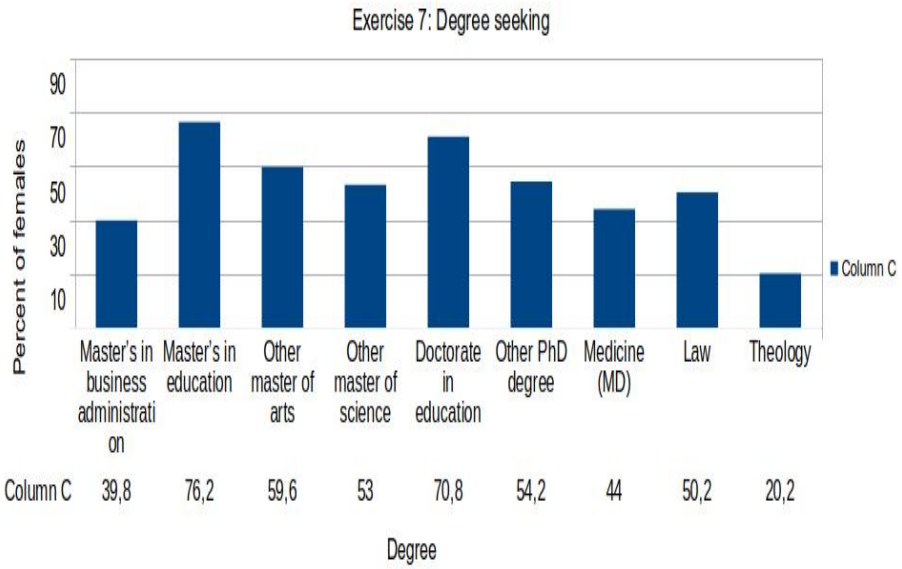


Exercise 7. Women seeking graduate and professional degrees. The table on the next page gives the percents of women among students seeking various graduate and professional degrees:

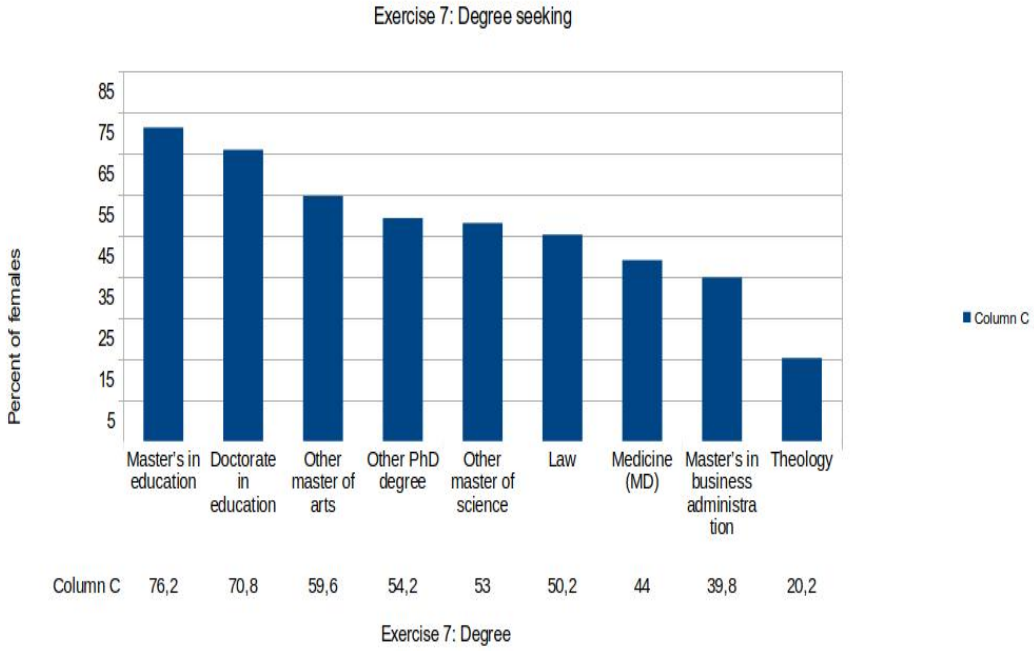
Degree	Percent of females
Master's in business administration	39.8
Master's in education	76.2
Other master of arts	59.6
Other master of science	53.0
Doctorate in education	70.8
Other PhD degree	54.2
Medicine (MD)	44.0
Law	50.2
Theology	20.2

(a) Explain clearly why we cannot use a pie chart to display these data.
(b) Make a bar graph of the data. (Comparisons are easier if you order the bars by height.)

Odp. 7



a) The percentage refers to the value within the specific variable (i.e., the degree). The categories are independent, and they do not sum to 100%. Plotting a pie chart would be very misleading and a bar graph is the ideal choice to compare the percentages across different degrees.
b) Bar graph sorted by highest to lowest value:



Exercise 8. An aging population. The population of the United States is aging, though less rapidly than in other developed countries. Here is a stemplot of the percents of residents aged 65 and over in the 50 states, according to the 2000 census. The stems are whole percents and the leaves are tenths of a percent.

5	7
6	
7	
8	5
9	6 7 9
10	6
11	0 2 2 3 3 6 7 7
12	0 0 1 1 1 1 3 4 4 5 7 8 9
13	0 0 0 1 2 2 3 3 3 4 5 5 6 8
14	0 3 4 5 7 9
15	3 6
16	
17	6

(a) There are two outliers: Alaska has the lowest percent of older residents, and Florida has the highest. What are the percents for these two states?

(b) Ignoring Alaska and Florida, describe the shape, center, and spread of this distribution.

Odp. 8 a) According to the stemplot, Alaska: 5.7%, Florida: 17.6%
b) Ignoring the two outliers from above, the distribution is symmetric, centered around $\approx 13\%$ (more values are located there) with a spread of $8.5\% - 15.6\%$.

Exercise 9. Outliers in percent of older residents. The stemplot in the previous Exercise displays the distribution of the percents of residents aged 65 and over in the 50 states. Stemplots help you find the five-number summary because they arrange the observations in increasing order.
(a) Give the five-number summary of this distribution.
(b) Does the $1.5 \times$ IQR rule identify Alaska and Florida as suspected outliers? Does it also flag any other states?

Odp. 9
Stemplot:

5	7
6	
7	
8	5
9	6 7 9
10	6
11	0 2 2 3 3 6 7 7
12	0 0 1 1 1 1 3 4 4 5 7 8 9
13	0 0 0 1 2 2 3 3 3 4 5 5 6 8
14	0 3 4 5 7 9
15	3 6
16	
17	6

a) Five-number summary
Minimum: 5.7
Q1: 11.7
Median: 12.75
Q3: 13.48
Maximum: 17.6

b) IQR = Q3 – Q1=13.48 – 11.7 = 1.78 (IQR : Inter Quartile Range)
IQR Rule: An outlier is defined when a value is lower than Q1 –

1.5×IQR or higher than Q3 + 1.5×IQR,
1.5 × IQR = 2.67
Q1 – 2.67 = 9.03 and Q3 + 2.67 = 16.15
Therefore, Alaska, with 5.7% and Florida, with 17.6%, which are less than 9.03% and more than 16.15%, are clearly identified as outliers by the 1.5×IQR rule. The only other state that is identified is the one with 8.5% which is also lower than 9.03%.

Exercise 10. Diabetes and glucose. People with diabetes must monitor and control their blood glucose level. The goal is to maintain "fasting plasma glucose" between about 90 and 130 milligrams per deciliter (mg/dl). Here are the fasting plasma glucose levels for 18 diabetics enrolled in a diabetes control class, five months after the end of the class:

141	158	112	153	134	95	96	78	148
172	200	271	103	172	359	145	147	255

Make a stemplot of these data and describe the main features of the distribution. (You will want to trim and also split stems.) Are there outliers? How well is the group as a whole achieving the goal for controlling glucose levels?

Odp. 10

7	8
9	5 6
10	3
11	2
13	4
14	1 5 7 8
15	3 8
17	2 2
20	0
25	5
27	1
35	9

The distribution is right-skewed, centered around 140-150 mg/dl. Outlier: 359 mg/dl. The goal of the glucose level between 90-130 mg/dl is only satisfied by four people, so the group is not doing well.

Exercise 11. Tornado damage. The states differ greatly in the kinds of severe weather that afflict them. Table 1.5

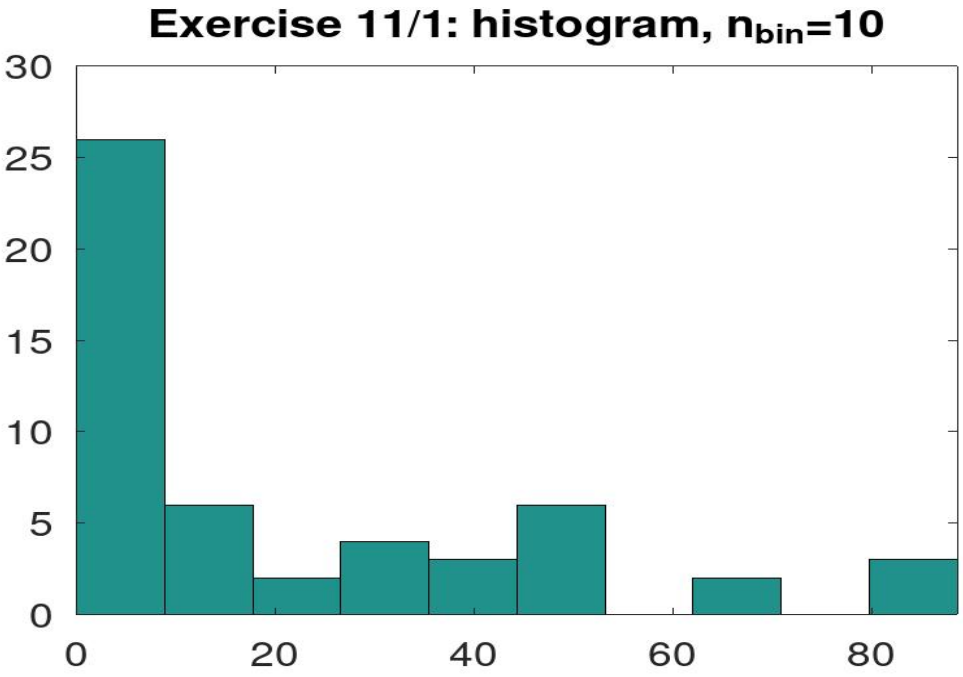
State	Damage ($ millions)	State	Damage ($ millions)	State	Damage ($ millions)
Alabama	51.88	Louisiana	27.75	Ohio	44.36
Alaska	0.00	Maine	0.53	Oklahoma	81.94
Arizona	3.47	Maryland	2.33	Oregon	5.52
Arkansas	40.96	Massachusetts	4.42	Pennsylvania	17.11
California	3.68	Michigan	29.88	Puerto Rico	0.05
Colorado	4.62	Minnesota	84.84	Rhode Island	0.09
Connecticut	2.26	Mississippi	43.62	South Carolina	17.19
Delaware	0.27	Missouri	68.93	South Dakota	10.64
Florida	37.32	Montana	2.27	Tennessee	23.47
Georgia	51.68	Nebraska	30.26	Texas	88.60
Hawaii	0.34	Nevada	0.10	Utah	3.57
Idaho	0.26	New Hampshire	0.66	Vermont	0.24
Illinois	62.94	New Jersey	2.94	Virginia	7.42
Indiana	53.13	New Mexico	1.49	Washington	2.37
Iowa	49.51	New York	15.73	West Virginia	2.14
Kansas	49.28	North Carolina	14.90	Wisconsin	31.33
Kentucky	24.84	North Dakota	14.69	Wyoming	1.78

shows the average property damage caused by tornadoes per year over the period from 1950 to 1999 in each of the 50 states and Puerto Rico. (To adjust for the changing buying power of the dollar over time, all damages were restated in 1999 dollars.)
(a) What are the top five states for tornado damage? The bottom five?
(b) Make a histogram of the data, by hand or using software, with classes "0 $\leq$damage$<$ 10", "10 $\leq$ damage $<$ 20", and so on. Describe the shape, center, and spread of the distribution. Which states may be outliers? (To understand the outliers, note that most tornadoes in largely rural states such as Kansas cause little property damage. Damage to crops is not counted as property damage.)
(c) If you are using software, also display the "default" histogram that your software makes when you give it no instructions. How does this compare with your graph in (b)?

Odp. 11

a) After sorting the table from highest to lowest damage: Top five states: Texas, Minnesota, Oklahoma, Missouri, Illinois Bottom five states: Alaska, Puerto Rico, Rhode Island, Nevada, Vermont

b) Make sure you put the bin width at 10



Strongly right skewed, centered between 15-20 million \$, spread: very wide range. Texas, Minnesota, and Oklahoma are very clear outliers.

c) Default graph:

The first graph with intervals of 10 paints a clearer picture of our data and helps us to draw conclusions better and allow us a fairer comparison.

Exercise 12. Carbon dioxide from burning fuels. Burning fuels in power plants or motor vehicles emits carbon dioxide (CO_2), which contributes to global warming. The table below displays CO_2 emissions per person from countries with population at least 20 million.

Carbon dioxide emissions (metric tons per person)

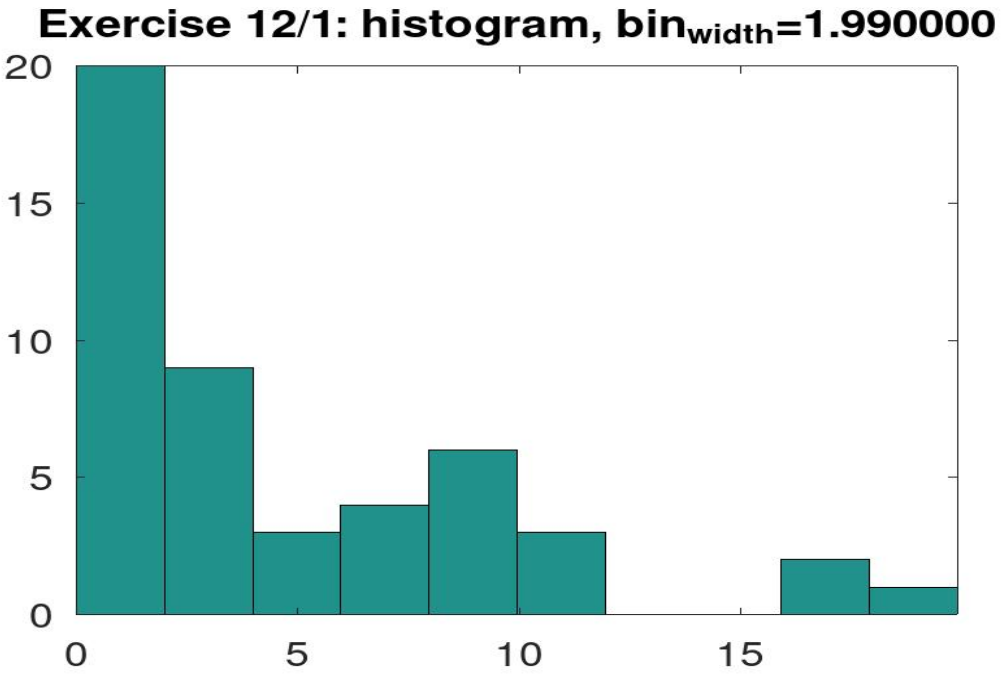
Country	CO_2	Country	CO_2
Algeria	2.3	Mexico	3.7
Argentina	3.9	Morocco	1.0
Australia	17.0	Myanmar	0.2
Bangladesh	0.2	Nepal	0.1
Brazil	1.8	Nigeria	0.3
Canada	16.0	Pakistan	0.7
China	2.5	Peru	0.8
Columbia	1.4	Tanzania	0.1
Congo	0.0	Philippines	0.9
Egypt	1.7	Poland	8.0
Ethiopia	0.0	Romania	3.9
France	6.1	Russia	10.2
Germany	10.0	Saudi Arabia	11.0
Ghana	0.2	South Africa	8.1
India	0.9	Spain	6.8
Indonesia	1.2	Sudan	0.2
Iran	3.8	Thailand	2.5
Iraq	3.6	Turkey	2.8
Italy	7.3	Ukraine	7.6
Japan	9.1	United Kingdom	9.0
Kenya	0.3	United States	19.9
Korea, North	9.7	Uzbekistan	4.8
Korea, South	8.8	Venezuela	5.1
Malaysia	4.6	Vietnam	0.5

(a) Why do you think we choose to measure emissions per person rather than total CO_2 emissions for each country?
(b) Display the data of the table in a graph (histogram). Describe the shape, center, and spread of the distribution. Which countries are outliers?

Odp. 12

a) The statistical analysis will lead to inaccuracies due to countries with higher populations will always have the highest CO2 emissions

b) Set the bin width at 2



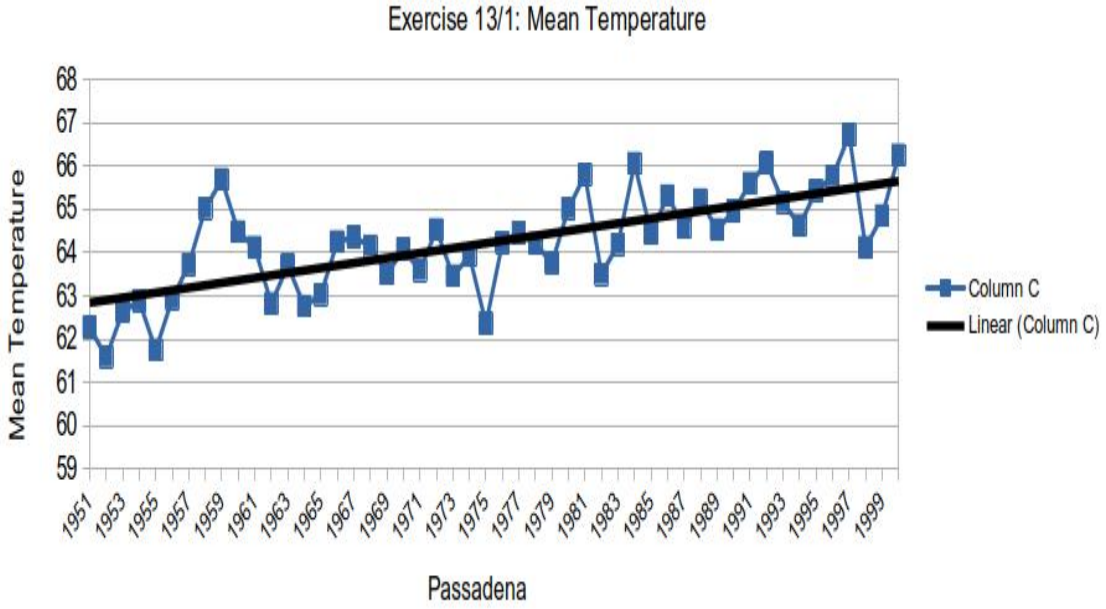
Strongly right skewed, centered around 2 tons, wide spread with $0-19.9$ tons. Outliers: Australia, Canada, and USA.

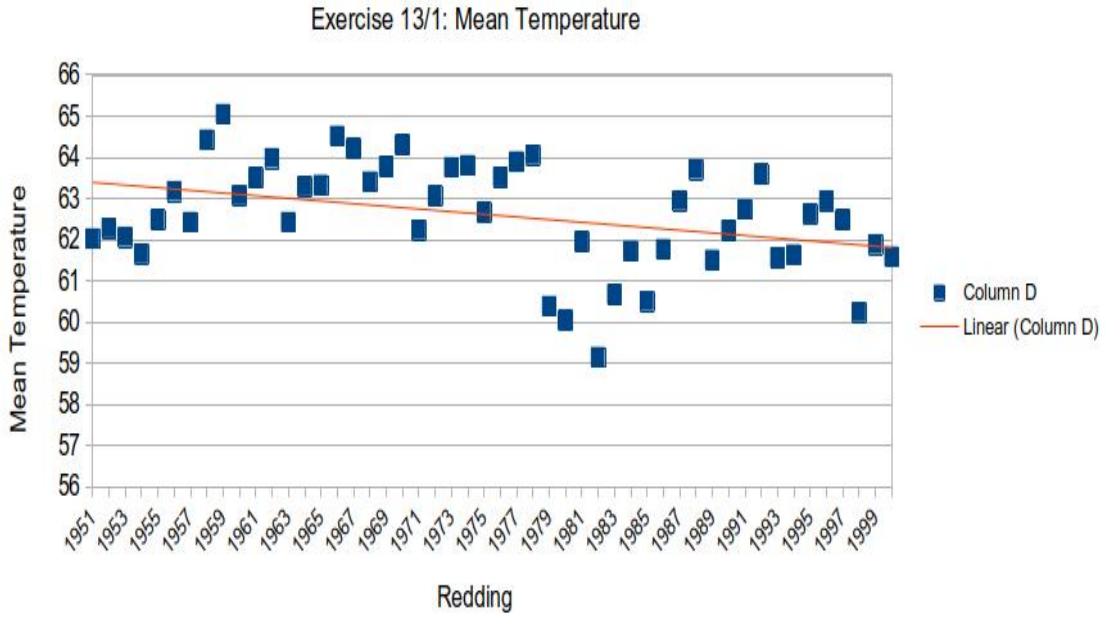
Exercise 13. California temperatures. The table

	Mean Temperature			Mean Temperature	
Year	Pasadena	Redding	Year	Pasadena	Redding
1951	62.27	62.02	1976	64.23	63.51
1952	61.59	62.27	1977	64.47	63.89
1953	62.64	62.06	1978	64.21	64.05
1954	62.88	61.65	1979	63.76	60.38
1955	61.75	62.48	1980	65.02	60.04
1956	62.93	63.17	1981	65.80	61.95
1957	63.72	62.42	1982	63.50	59.14
1958	65.02	64.42	1983	64.19	60.66
1959	65.69	65.04	1984	66.06	61.72
1960	64.48	63.07	1985	64.44	60.50
1961	64.12	63.50	1986	65.31	61.76
1962	62.82	63.97	1987	64.58	62.94
1963	63.71	62.42	1988	65.22	63.70
1964	62.76	63.29	1989	64.53	61.50
1965	63.03	63.32	1990	64.96	62.22
1966	64.25	64.51	1991	65.60	62.73
1967	64.36	64.21	1992	66.07	63.59
1968	64.15	63.40	1993	65.16	61.55
1969	63.51	63.77	1994	64.63	61.63
1970	64.08	64.30	1995	65.43	62.62
1971	63.59	62.23	1996	65.76	62.93
1972	64.53	63.06	1997	66.72	62.48
1973	63.46	63.75	1998	64.12	60.23
1974	63.93	63.80	1999	64.85	61.88
1975	62.36	62.66	2000	66.25	61.58

contains data on the mean annual temperatures (degrees Fahrenheit) for the years 1951 to 2000 at two locations in California: Pasadena and Redding. Make time plots of both time series and compare their main features. You can see why discussions of climate change often bring disagreement

Odp. 13 We make separate timeplots of the two cities and then we plot the trendline.





Both timeplots show random fluctuations of the temperature over the years, but after plotting the trendline, we see an upward trend for Pasadena and a downward trend for Redding, mainly due to dropped temperature in the

mid-1980s. The downward trend in the Redding graph is why discussions of climate change can bring disagreement.

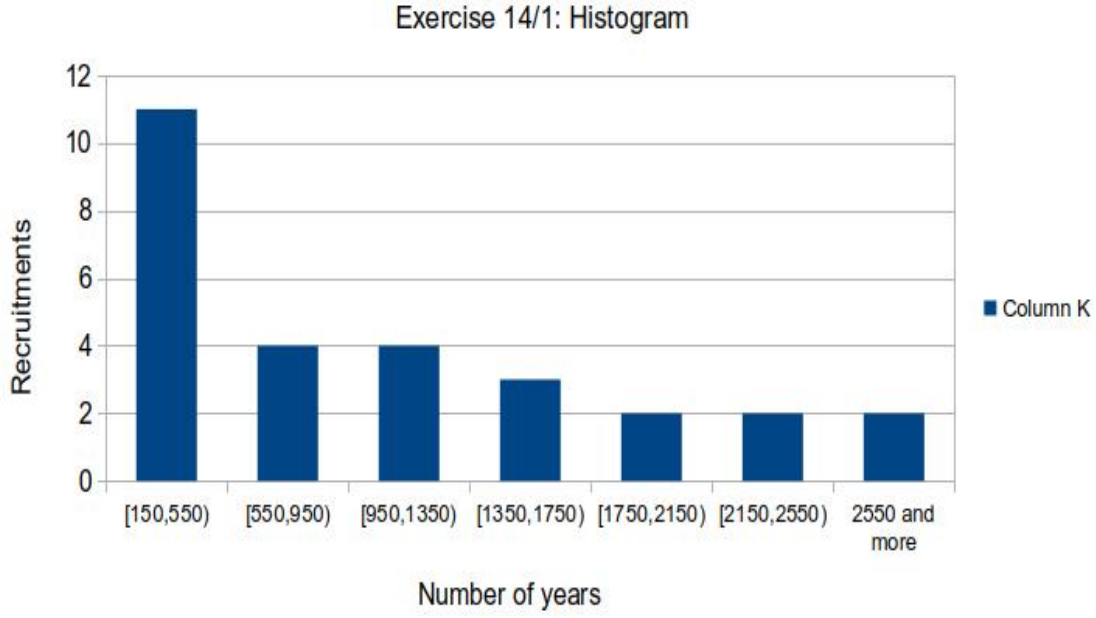
Exercise 14. Fish in the Bering Sea. "Recruitment", the addition of new members to a fish population, is an important measure of the health of ocean ecosystems. Here are data on the recruitment of rock sole in the Bering Sea between 1973 and 2000:

Year	Recruitment (millions)	Year	Recruitment (millions)
1973	173	1987	4700
1974	234	1988	1702
1975	616	1989	1119
1976	344	1990	2407
1977	515	1991	1049
1978	576	1992	505
1979	727	1993	998
1980	1411	1994	505
1981	1431	1995	304
1982	1250	1996	425
1983	2246	1997	214
1984	1793	1998	385
1985	1793	1999	445
1986	2809	2000	676

(a) Make a graph to display the distribution of rock sole recruitment, then describe the pattern and any striking deviations that you see.
(b) Make a time plot of recruitment and describe its pattern. As is often the case with time series data, a time plot is needed to understand what is happening.

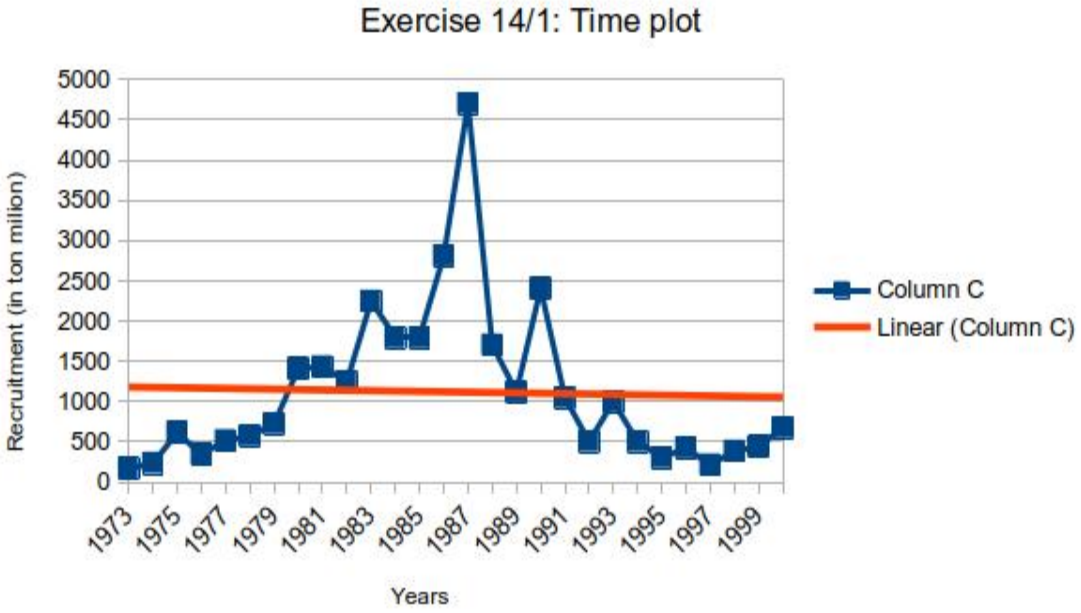
Odp. 14

a) First we plot a histogram



The distribution is right-skewed, centered around 600-800 million with a wide spread of 173-4700 million. There is a very noticeable outlier in 1987.

b) A time plot is more useful to analyze the trend over the years: (Note that the axes are changing)



The trend appears to be a bit downward but not very steady with severe fluctuations over the time period with a big spike in 1987, the outlier as mentioned earlier.Note: Histograms show the distribution, while timeplots reveal the fluctuation over the years.

Exercise 15. Longleaf pine trees. The Wade Tract in Thomas County, Georgia, is an old-growth forest of longleaf pine trees (Pinus palustris) that has survived in a relatively undisturbed state since before the settlement of the area by Europeans. A study collected data about 584 of these trees. One of the variables measured was the diameter at breast height (DBH). This is the diameter of the tree at 4.5 feet and the units are centimeters (cm). Only trees with DBH greater than 1.5 cm were sampled. Here are the diameters of a random sample of 40 of these trees:

10.5	13.3	26.0	18.3	52.2	9.2	26.1	17.6	40.5	31.8
47.2	11.4	2.7	69.3	44.4	16.9	35.7	5.4	44.2	2.2
4.3	7.8	38.1	2.2	11.4	51.5	4.9	39.7	32.6	51.8
43.6	2.3	44.6	31.5	40.3	22.3	43.3	37.5	29.1	27.9

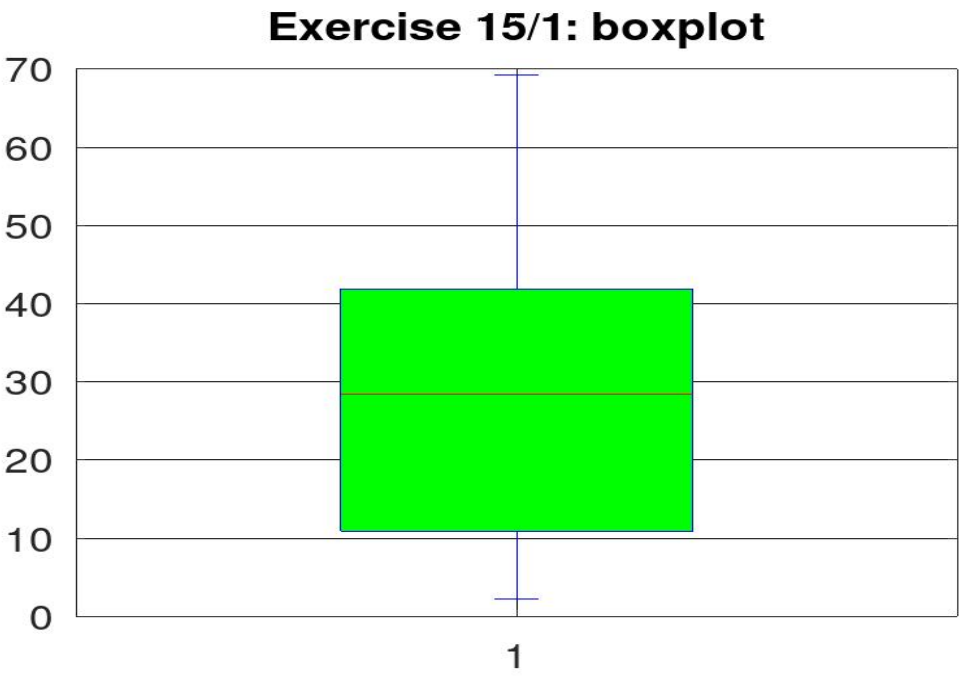
(a) Find the five-number summary for these data.
(b) Make a boxplot.
(c) Make a histogram.

(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

Odp. 15

a) Five-number summary
Minimum: 2.2
Q1: 11.2
Median: 28.5
Q3: 41.2
Maximum: 69.3

b) Boxplot:



c) Histogram:



d) The distribution is right-skewed as evidenced by both plots but more visible in the histogram. The second graph reveals a strong outlier of 69.3 diameter, spread is wide. In this occasion, the histogram is more informative of the dataset.

Exercise 16. Blood proteins in children from Papua New Guinea. C-reactive protein (CRP) is a substance that can be measured in the blood. Values increase substantially within 6 hours of an infection and reach a peak within 24 to 48 hours after. In adults, chronically high values have been linked to an increased risk of cardiovascular disease. In a study of apparently healthy children aged 6 to 60 months in Papua Nw Guinea, CRP was measured in 90 children. The units are milligrams per liter (mg/l). Here are the data from a random sample of 40 of these children:

0.00	3.90	5.64	8.22	0.00	5.62	3.92	6.81	30.61	0.00
73.20	0.00	46.70	0.00	0.00	26.41	22.82	0.00	0.00	3.49
0.00	0.00	4.81	9.57	5.36	0.00	5.66	0.00	59.76	12.38
15.74	0.00	0.00	0.00	0.00	9.37	20.78	7.10	7.89	5.53

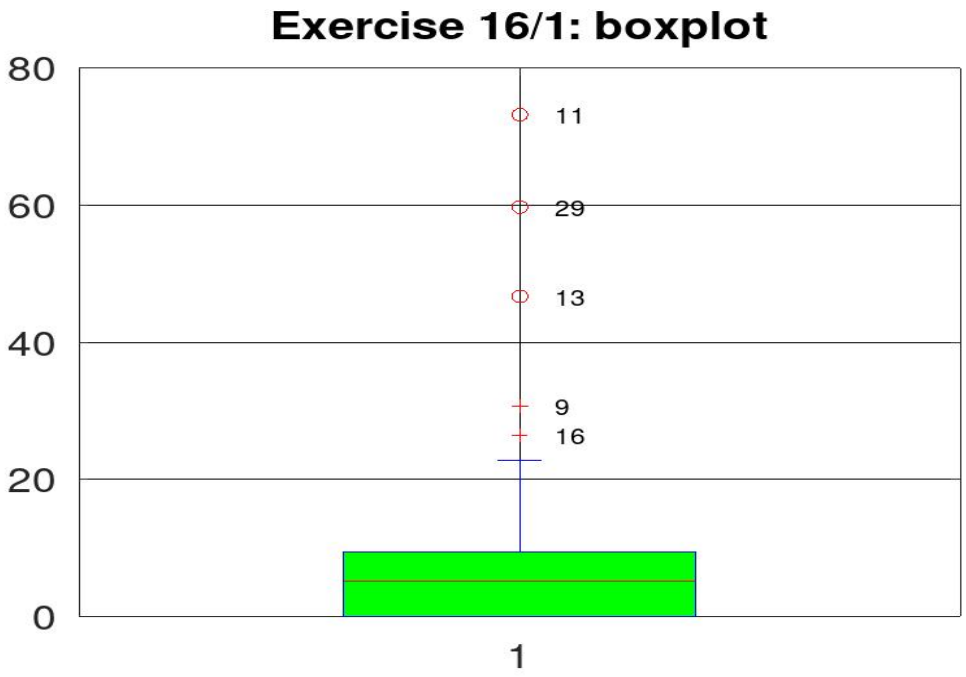
(a) Find the five-number summary for these data.
(b) Make a boxplot.
(c) Make a histogram.

(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

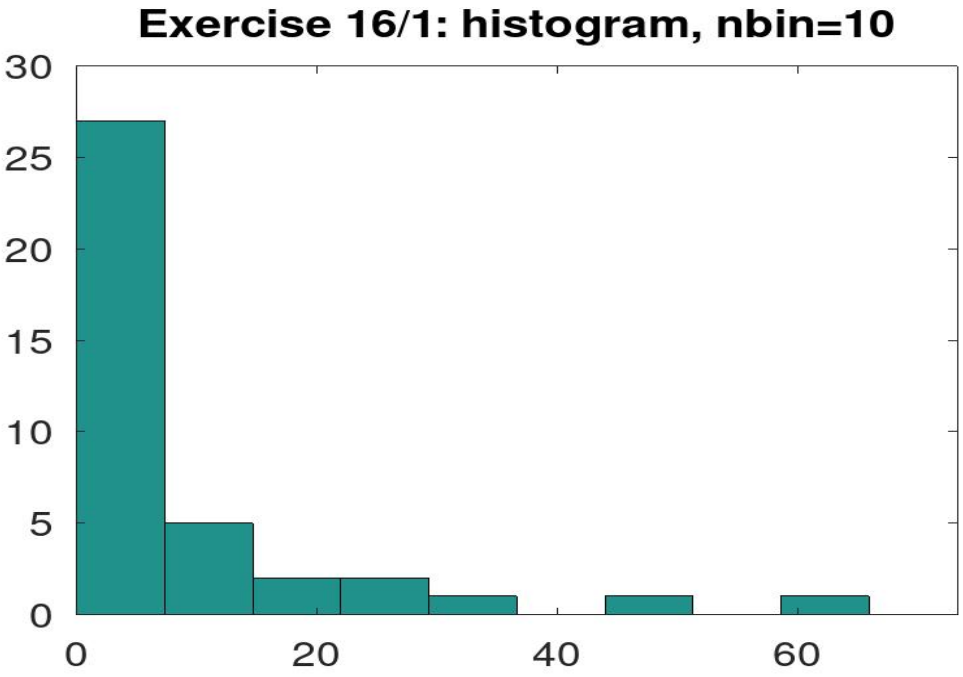
Odp. 16

a) Five-number summary
Minimum: 0
Q1: 0
Median: 5.085
Q3: 9.42
Maximum: 73.20

b) Boxplot



c) Histogram

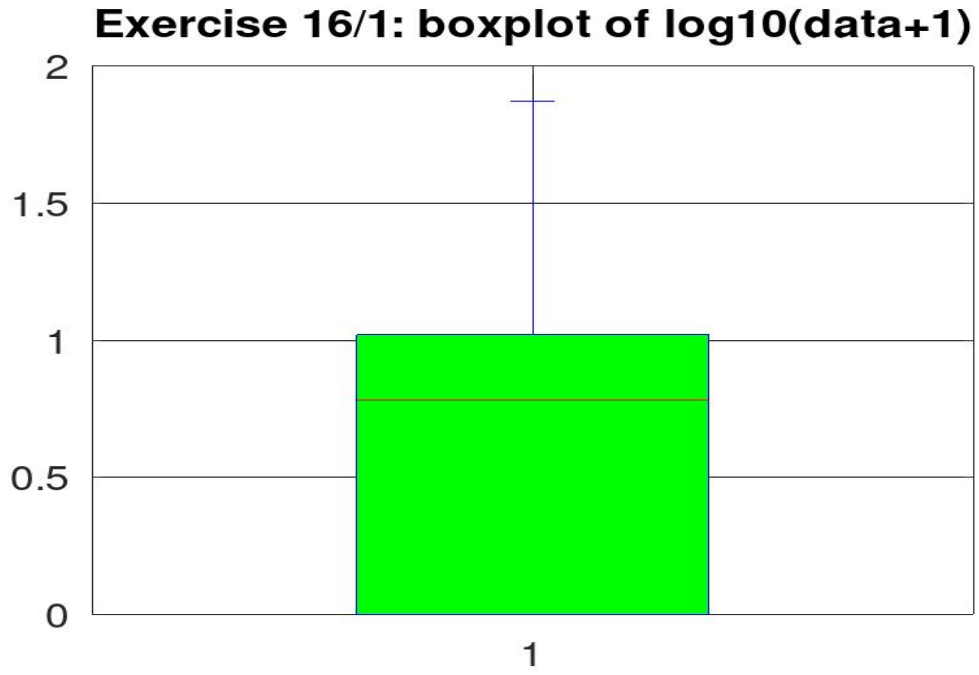


d) The distribution is extremely right-skewed mainly because of a large number of children having 0.00 mg/ml of CRP. Both plots can evidence this, but the histogram conveys the distribution better.

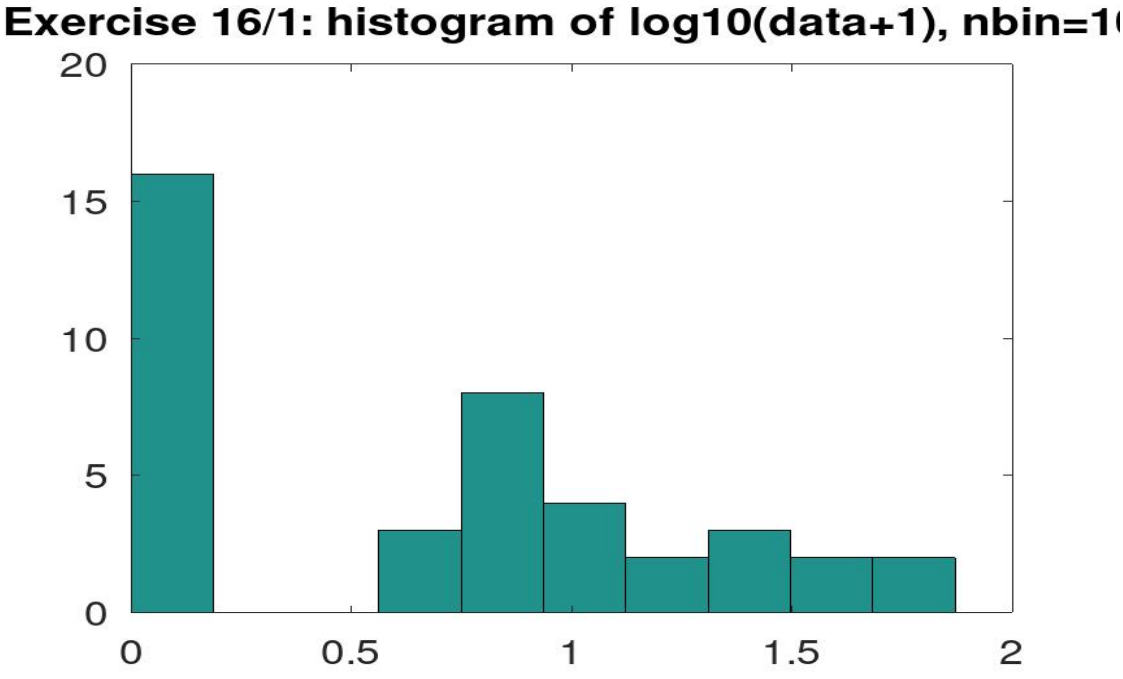
With strongly skewed distributions such as this, we frequently reduce the skewness by taking a log transformation. By adding 1 to all the previous values and then taking the logarithm, we present the new values and plots:

a) Five-number summary
Minimum: 0
Q1: 0
Median: 0.78
Q3: 1.02
Maximum: 1.87

b) Boxplot



c) Histogram



d) The distribution is still right-skewed but not as extreme as in Exercise

15. This is further evidenced by the new five-number summary, with the maximum number being closer to the median, because the range is now smaller. Therefore, the logarithmic transformation can improve the interpretation of data analysis.

Exercise 17. In the Papua New Guinea study, the researchers measured serum retinol. A low value of this variable can be an indicator of vitamin A deficiency. Below are the data on the same sample of 40 children from this study. The units are micromoles per liter (μmol/l).

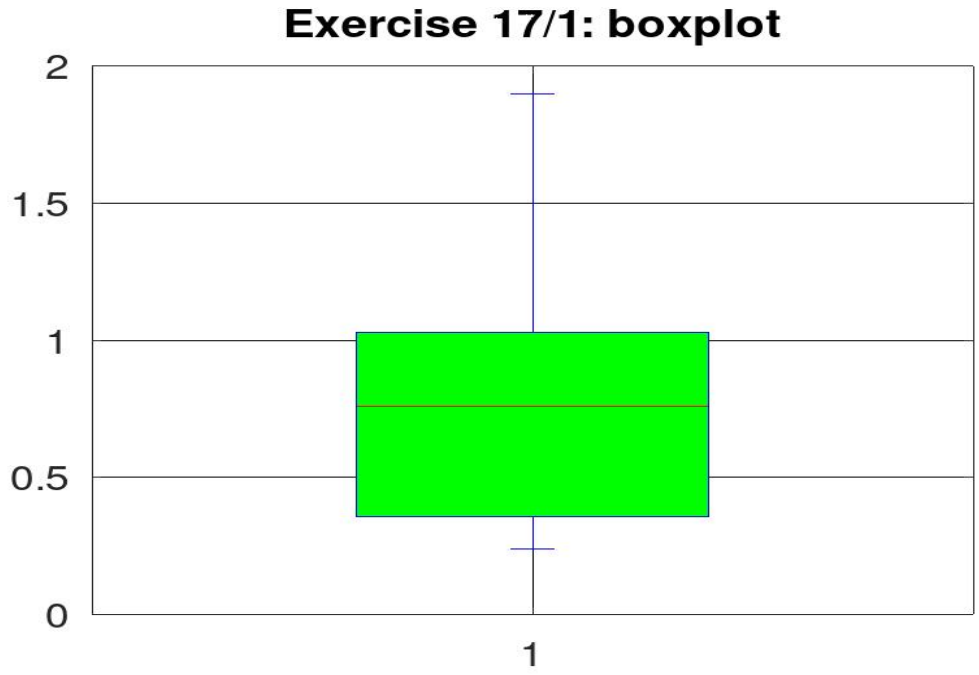
1.15	1.36	0.38	0.34	0.35	0.37	1.17	0.97	0.97	0.67
0.31	0.99	0.52	0.70	0.88	0.36	0.24	1.00	1.13	0.31
1.44	0.35	0.34	1.90	1.19	0.94	0.34	0.35	0.33	0.69
0.69	1.04	0.83	1.11	1.02	0.56	0.82	1.20	0.87	0.41

(a) Find the five-number summary for these data.
(b) Make a boxplot.
(c) Make a histogram.
(d) Write a short summary of the major features of this distribution. Do you prefer the boxplot or the histogram for these data?

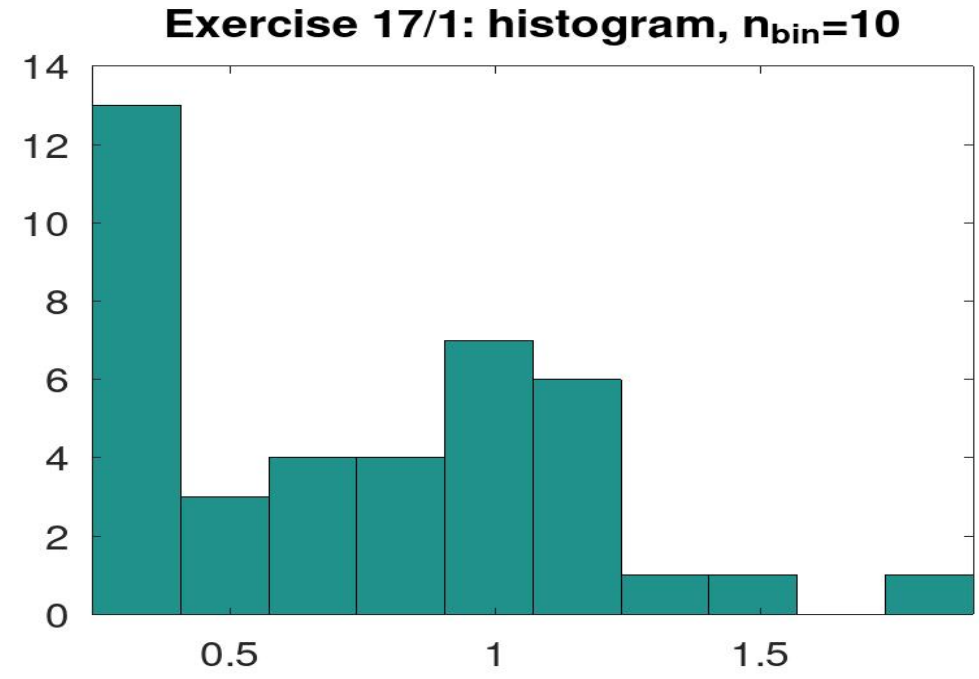
Odp. 17

a) Five-number summary
Minimum: 0.24
Q1: 0.36
Median: 0.76
Q3: 1.03
Maximum: 1.90

b) Boxplot



c) Histogram



d)Overall, the distribution is relatively strong, right-skewed where the

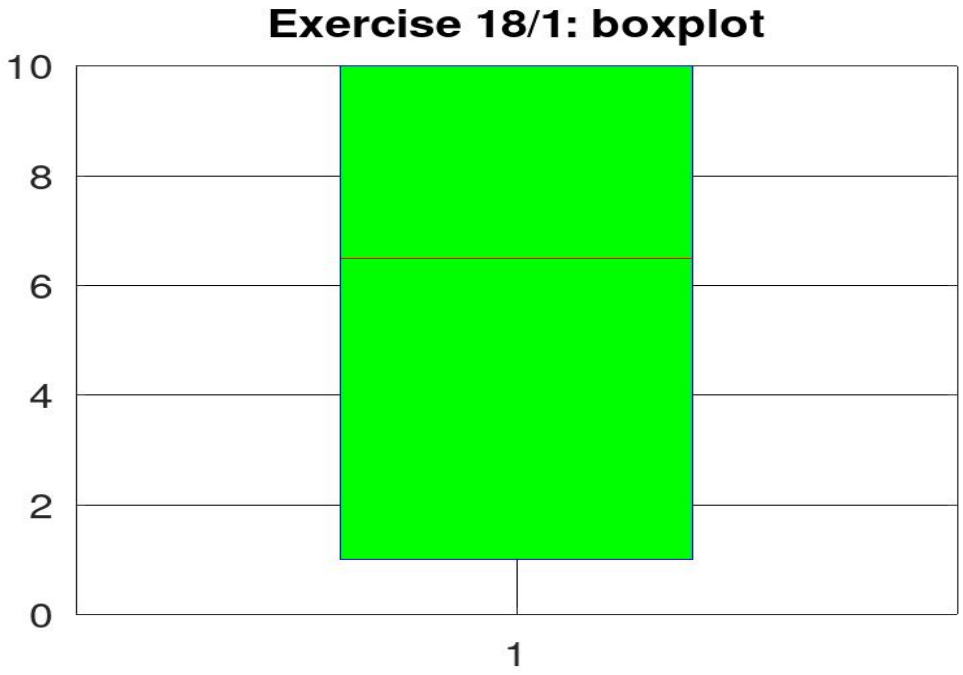
data indicate a predominantly low vitamin A status among the children, with only a few showing normal or elevated levels.

Exercise 18. Luck and puzzle solving. Children in a psychology study were asked to solve some puzzles and were then given feedback on their performance. Then they were asked to rate how luck played a role in determining their scores. This variable was recorded on a 1 to 10 scale with 1 corresponding to very lucky and 10 corresponding to very unlucky. Here are the scores for 60 children:

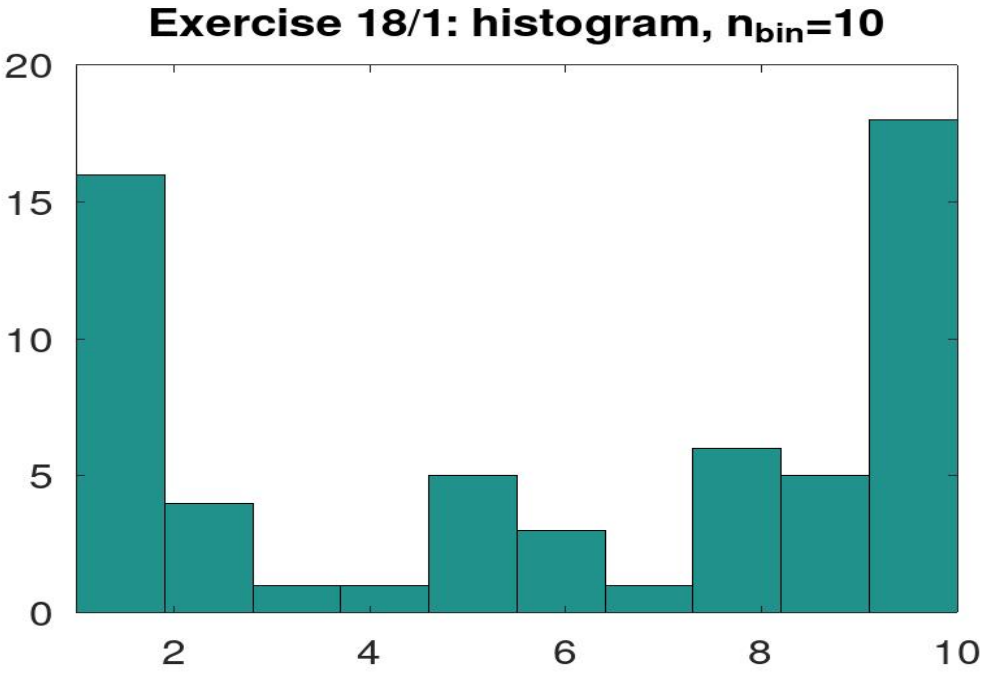
1	10	1	10	1	1	10	5	1	1	8	1	10	2	1
9	5	2	1	8	10	5	9	10	10	9	6	10	1	5
1	9	2	1	7	10	9	5	10	10	10	1	8	1	6
10	1	6	10	10	8	10	3	10	8	1	8	10	4	2

Use numerical and graphical methods to describe these data. Write a short report summarizing your work.

Odp. 18 Five-number summary
Minimum: 1
Q1: 1
Median: 6.5
Q3: 10
Maximum: 10
Mean: 5.9
St. Dev.: 3.77
Boxplot



Histogram



This indicates a U-shaped (bimodal) distribution, with peaks at both ends

and a valley in the middle. The five-number summary suggests a shift towards being more "unlucky" (median: 6.5), but the dataset (and subsequently the histogram) clearly show the polarized distribution with most answers being either 1 or 10.

Exercise 19. College tuition and fees. This figure

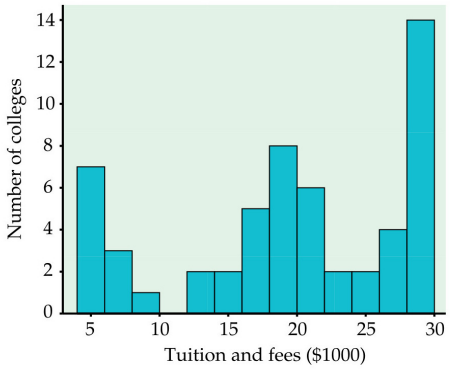


FIGURE 1.16 Histogram of the tuition and fees charged by four-year colleges in Massachusetts, for Exercise 1.27.

is a histogram of the tuition and fees charged by the 56 four-year colleges in the state of Massachusetts. Here are those charges (in dollars), arranged in increasing order:

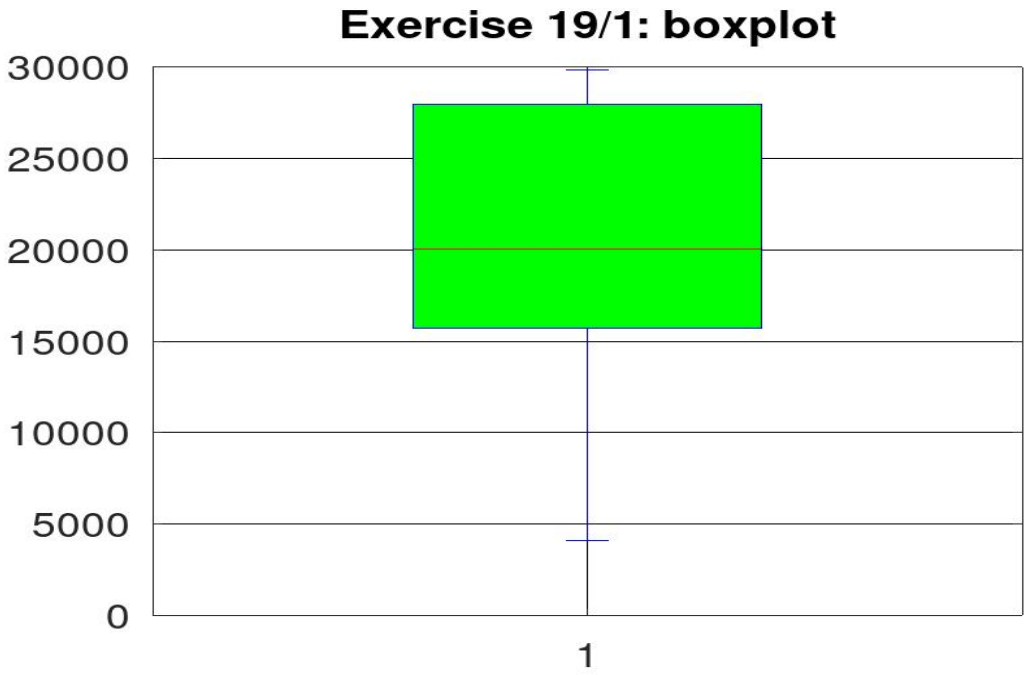
4,123	4,186	4,324	4,342	4,557	4,884	5,397	6,129
6,963	6,972	8,232	13,584	13,612	15,500	15,934	16,230
16,696	16,700	17,044	17,500	18,550	18,750	19,145	19,300
19,410	19,700	19,700	19,910	20,234	20,400	20,640	20,875
21,165	21,302	22,663	23,550	24,324	25,840	26,965	27,522
27,544	27,904	28,011	28,090	28,420	28,420	28,900	28,906
28,950	29,060	29,338	29,392	29,600	29,624	29,630	29,875

Find the five-number summary and make a boxplot. What distinctive feature of the histogram do these summaries miss? Remember that numerical summaries are not a substitute for looking at the data.

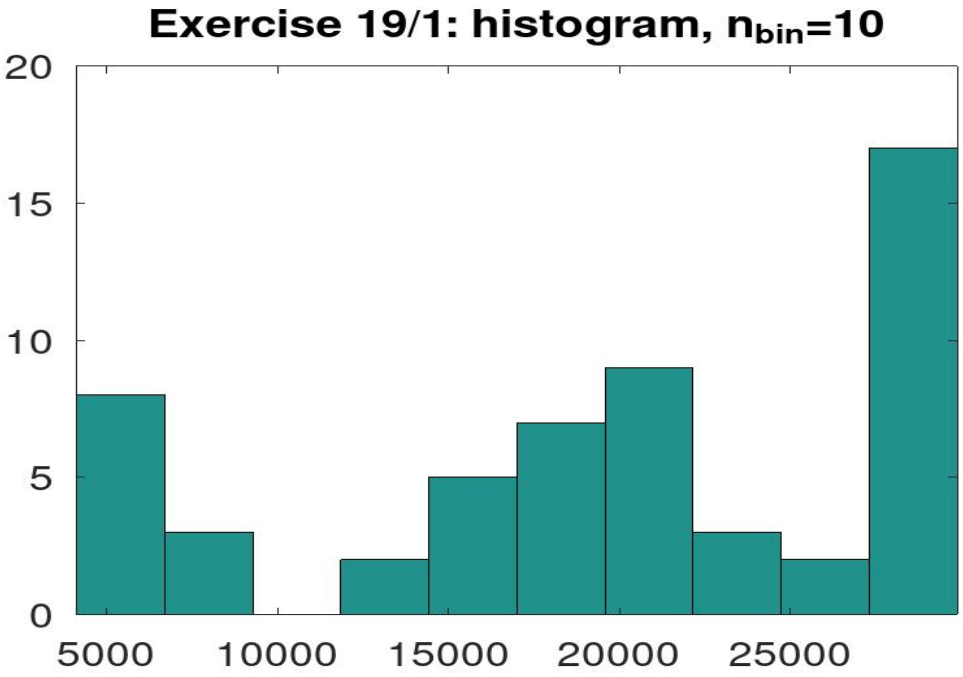
Odp. 19

a) Five-number summary
Minimum: 4 123
Q1: 15 826
Median: 20 072
Q3: 27 931

Maximum: 29 875
Boxplot



Histogram



The boxplot and the five-number summary miss completely the bimodal nature of the distribution as depicted in the histogram, especially on the high end. The boxplot shows a continuous spread of tuition fees being relatively expensive, as opposed to the histogram which depicts three very distinct groups of schools.

Exercise 20. Odp. 20

Exercise 21. Outliers in percent of older residents. The stemplot in Exercise 9 displays the distribution of the percents of residents aged 65 and over in the 50 states. Stemplots help you find the five-number summary because they arrange the observations in increasing order.
(a) Give the five-number summary of this distribution.
(b) Does the 1.5 × IQR rule identify Alaska and Florida as suspected outliers? Does it also flag any other states?

Odp. 21