Chapter 3: Displaying and Summarizing Quantitative Data

Section 3.1

1. **Explain**
   a) Bar chart is for categorical data, whereas histogram is for quantitative data. Bar chart uses gaps to separate categories and a gap (or gaps) in histogram indicate that there is a bin (or bins) with no values.
   b) A stem-and-leaf plot is like a histogram but it shows the individual values. It is easier to make by hand for data sets with small number of data.

2. **Stem-and-leaf**
   a) 0 | 8
      1 | 28
      2 | 134447788899
      3 | 11237
      where 1 | 2 = 12
   b) 0 | 8
      1 | 2
      1 | 8
      2 | 13444
      2 | 7788899
      3 | 1123
      3 | 7
      where 1 | 2 = 12
   c) Display in part (a) is a little crowded. Display in part (b) gives a clearer view on the overall shape of the distribution of the data set.

Section 3.2

3. **Incomes in Canada.** Probably it will be unimodal and skewed to the high end. The reason is that the higher the individual income, the less number of individual will be able to achieve it.

4. **Incarcerated.** Overall, it looks unimodal and skewed to the high end with probably some extraordinarily large values (three data points over 100000).

Section 3.3

5. **Blue Jays.** The median number of wins by the Blue Jays since they last won the World Series is 80.5.

6. **Divorce Rate.** The median divorce rate (per 100,000 residents per year) in Canada between 1989 to 2008 was 230.55.

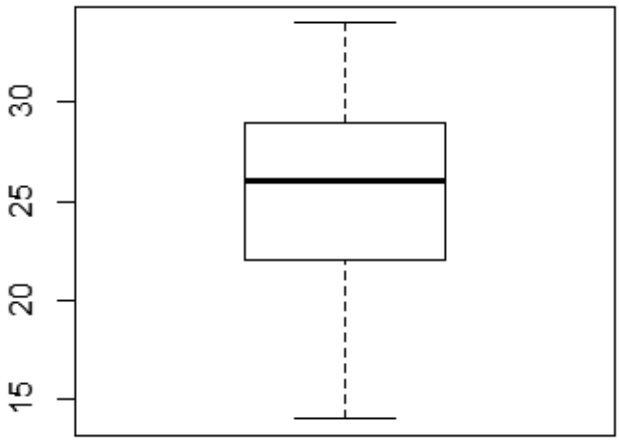
Section 3.4

7. Blue Jays. The IQR is 85 – 74 = 11 wins.

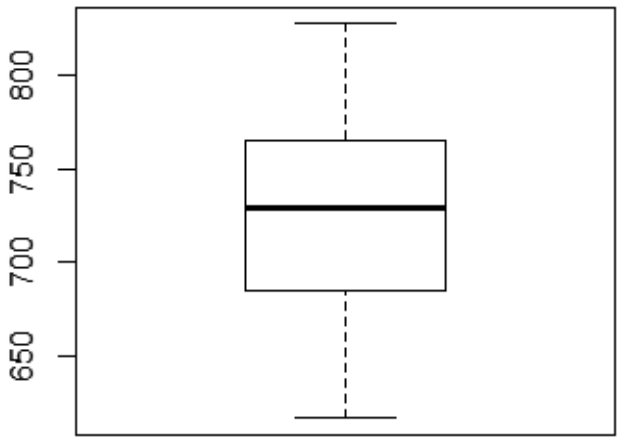
8. Divorce Rate. The IQR is 268.45 – 223.85 = 44.60.

Section 3.5

9. Temperature. The five number summary is: 14, 22, 26, 29, 34.



10. Bicycle Deaths. The five number summary is: 616, 684.5, 729.5, 765, 828.



Section 3.6

11. Adoptions II. The mean number of adoptions is expected to by higher than the median number of adoptions, since the distribution of the number of adoptions is skewed to the right.

12. Test score centers. The median test score will not be affected, but the mean test score will increase by 0.4 points.

Section 3.7

13. Test score spreads. The IQR of the test scores will not be affected, but the standard deviation of the test scores will increase.

14. Fuel economy. If the outlier is removed, the standard deviation will decrease. The IQR will not change substantially. (It may change slightly, since removing one value from the data set may change the quartiles, which would change the IQR.)

Chapter Exercises

15. Histogram Answers will vary.

16. Not a histogram Answers will vary.

17. In the news Answers will vary.

18. In the news II Answers will vary.

19. Thinking about shape

a) The distribution of the number of speeding tickets each student in final year of university has ever had is likely to be unimodal and skewed to the right. Most students will have very few speeding tickets (maybe 0 or 1), but a small percentage of students will likely have comparatively many (3 or more) tickets.

b) The distribution of player's scores at the U.S. Open Golf Tournament would most likely be unimodal and slightly skewed to the right. The best golf players in the game will likely have around the same average score, but some golfers might be off their game and score 15 strokes above the mean. (Remember that high scores are undesirable in the game of golf!)

c) The weights of female babies in a particular hospital over the course of a year will likely have a distribution that is unimodal and symmetric. Most newborns have about the same weight, with some babies weighing more and less than this average. There may be slight skew to the left, since there seems to be a greater likelihood of premature birth (and low birth weight) than post-term birth (and high birth weight).

d) The distribution of the length of the average hair on the heads of students in a large class would likely be bimodal and skewed to the right. The average hair length of the males would be at one mode, and the average hair length of the females would be at the other mode, since women typically have longer hair than men. The distribution would be skewed to the right, since it is not possible to have hair length less than zero, but it is possible to have a variety of lengths of longer hair.

20. More shapes

a) The distribution of the ages of people at a pee-wee hockey game would likely be bimodal and skewed to the right. The average age of the players would be at one mode and the average age of the spectators (probably mostly parents) would be at the other mode. The distribution would be skewed to the right, since it is possible to have a greater variety of ages among the older people, while there is a natural left endpoint to the distribution at zero years of age.
b) The distribution of the number of siblings of people in your class is likely to be unimodal and skewed to the right. Most people would have 0, 1, or 2 siblings, with some people having more siblings.
c) The distribution of pulse rate of college- or university-age males would likely be unimodal and symmetric. Most males' pulse rates would be around the average pulse rate for college- or university-age males, with some males having lower and higher pulse rates.
d) The distribution of the number of times each face of a die shows in 100 tosses would likely be uniform, with around 16 or 17 occurrences of each face (assuming the die had six sides).

21. Sugar in cereals

a) The distribution of the sugar content of breakfast cereals is bimodal, with a cluster of cereals with sugar content around 10% sugar and another cluster of cereals around 45% sugar. The lower cluster shows a bit of skew to the right. Most cereals in the lower cluster have between 0% and 10% sugar. The upper cluster is symmetric, with centre around 45% sugar.
b) There are two different types of breakfast cereals, those for children and those for adults. The children's cereals are likely to have higher sugar contents, to make them taste better (to kids, anyway!). Adult cereals often advertise low sugar content.

22. Chorus heights

a) The distribution of the heights of singers in the chorus is bimodal, with a mode at around 65 inches and another mode around 71 inches. No chorus member has height below 60 inches or above 76 inches.
b) The two modes probably represent the mean heights of the male and female members of the chorus.

23. Test scores

a) The number of students scoring 40 or higher is approximately 16 + 5 + 3 + 1 = 25 (these are only approximate heights of the bars in the histogram after 40) and the percentage = 25/110 = 22.7 percent.

b) The number of students scoring between 25 and 35 is the sum of the heights of the two bars between 25 and 35. This is approximately 21 + 31 = 52, and so the percentage is 52/110 = 47.3 percent.

c) The distribution is symmetric. The centre (the median) is around 32. The scores range from 0–60, but the few scores close to zero may be outliers (with a gap of just five we might not be able to conclude as a clear outlier, but they are somewhat unusual compared to the rest of the scores).

24. Run times The distribution of run times is skewed to the right. The shortest run time was around 28.5 minutes, and the longest run time was around 35.5 minutes. A typical run time was between 30 and 31 minutes, and the majority of run times were between 29 and 32 minutes. It is easier to run slightly slower than usual and end up with a longer run time than it is to run slightly faster than usual and end up with a shorter run time. This could account for the skew to the right seen in the distribution.

25. Election 2011

a) The distribution is right skewed and so it is logical to expect the mean to be greater than the median. The median is the (308+1)/2 = 154.5th value in the ordered data set. This value must be in the fourth interval from the left of the histogram, i.e., the median must be in the interval 0.45–0.55.

b) The distribution is right skewed, and could be bimodal. The median is between 0.45 and 0.55. The values of the percentage of rejected ballots range from 0–2.5 (approx.). The largest value could be an outlier, so might warrant further investigation. The apparent bimodality might also warrant further investigation.

26. Emails

a) The distribution of the number of e-mails sent is skewed to the right, so the mean is larger than the median. There appears to be about 155 observations. The median would be the 78th observation, which falls at 1.

b) The distribution of the number of e-mails received from each student by a professor in a large introductory statistics class during an entire term is skewed to the right, with the number of e-mails ranging from 1 to 21 e-mails. The distribution is centred at about 2 e-mails, with many students only sending 1 e-mail. There is one outlier in the distribution, a student who sent 21 e-mails. The next highest number of e-mails sent was only 8.

27. Summaries

a) The mean price of the electric smoothtop ranges is $1001.50.

b) To find the median and the quartiles, first order the list.
565 750 850 900 1000 1050 1050 1200 1250 1400
The median price of the electric smoothtop ranges is $1025.
Quartile 1 = $850 and Quartile 3 = $1200.

c) The range of the distribution of prices is Max – Min = $1400 – $565 = $835.
The IQR = Q3 – Q1 = $1200 – $850 = $350.

28. Tornadoes 2011

a) The mean number of annual deaths from tornadoes in the United States from 1998 through 2011 is 133.5.

b) The median number of deaths is 60.5. The first quartile is 40 deaths and the third quartile is 125 deaths.

c) The range is 555 – 21 = 534 deaths. The IQR is 125 – 40 = 85 deaths.

29. Mistake

a) As long as the boss's true salary of $200 000 is still above the median, the median will be correct. The mean will be too large, since the total of all the salaries will decrease by $2 000 000 – $200 000 = $1 800 000, once the mistake is corrected.

b) The range will likely be too large. The boss's salary is probably the maximum, and a lower maximum would lead to a smaller range. The IQR will likely be unaffected, since the new maximum has no effect on the quartiles. The standard deviation will be too large, because the $2 000 000 salary will have a large squared deviation from the mean.

30. Sick days The company probably used the mean number of sick days, while the union used the median number. The mean will likely be higher, since it is affected by a probable right skew. Some employees may have many sick days, while most have relatively few.

31. Standard deviation I

a) Set 2 has the greater standard deviation. Both sets have the same mean (6), but set 2 has values that are generally farther away from the mean.
SD(Set 1) = 2.24 SD(Set 2) = 3.16

b) Set 2 has the greater standard deviation. Both sets have the same mean (15), maximum (20), and minimum (10), but 11 and 19 are farther from the mean than 14 and 16.
SD(Set 1) = 3.61 SD(Set 2) = 4.53

c) The standard deviations are the same. Set 2 is simply {Set 1} + 80. Although the measures of centre and position change, the spread is exactly the same.
SD(Set 1) = 4.24 SD(Set 2) = 4.24

32. Standard deviation II

a) Set 2 has the greater standard deviation. Both sets have the same mean (7), maximum (10), and minimum (4), but 6 and 8 are farther from the mean than 7.
SD(Set 1) = 2.12 SD(Set 2) = 2.24

b) The standard deviations are the same. Set 1 is simply Set 2 + 90. Although the measures of centre and position are different, the spread is exactly the same.
SD(Set 1) = 36.06 SD(Set 2) = 36.06

c) Set 2 has the greater standard deviation. The central 4 values of Set 2 are simply the central 4 values of Set 1 +40, but the maximum and minimum of Set 2 are farther away from the mean than the maximum and minimum of Set 1.
Range(Set 1) = 18 and Range(Set 2) = 22.
The Range of Set 2 is greater than the Range of Set 1, and the standard deviation is also larger.
SD(Set 1) = 6.03 SD(Set 2) = 7.24

33. Payroll

a) The mean salary is $\frac{1200+700+6(400)+4(500)}{12} = 525$
The median salary is the middle of the ordered list:
400 400 400 400 400 400 500 500 500 500 700 1200
The median is \$450.

b) Only two employees, the supervisor and the inventory manager, earn more than the mean wage.

c) The median better describes the wage of the typical worker. The mean is affected by the two higher salaries.

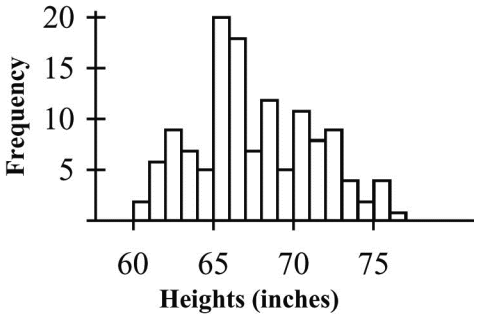
d) The IQR is the better measure of spread for the payroll distribution. The standard deviation and the range are both affected by the two higher salaries.

34. Singers

a) Five-number summary: 60, 65, 66, 70, 76, so the median is 66 inches and the IQR is 70 – 65 = 5 inches.

b) The mean height of the singers is 67.12 inches, and the standard deviation of the heights is 3.79 inches.

c) The histogram of heights of the choir members is shown at the right.



d) The distribution of the heights of the choir members is bimodal (probably due to differences in height of men and women) and skewed slightly to the right. The median is 66 inches. The distribution is fairly spread out, with the middle 50% of the heights falling between 65 and 70 inches. There are no gaps or outliers in the distribution.

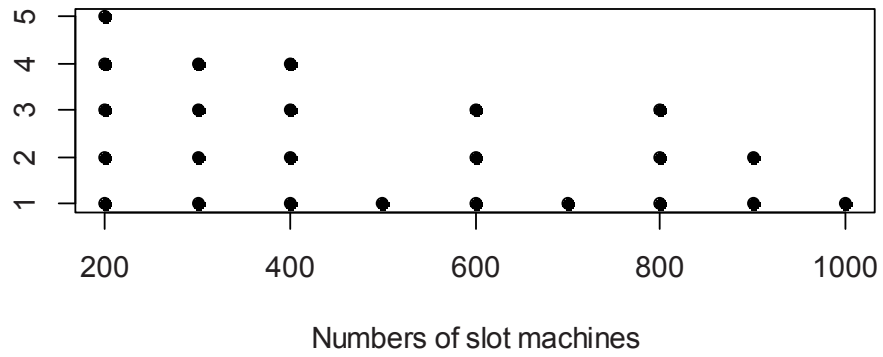
35. Alberta casinos 2013

The stem and leaf plot, a dotplot, and the five-number summary (plus mean) for these data are shown below. The distribution looks slightly right-skewed (mean larger than median). The median number of slot machines is 428. The interquartile range is 731 – 299 = 432.

Stem and leaf plot:

```
 1 | 7
 2 | 0035
 3 | 0033
 4 | 0224
 5 | 3
 6 | 000
 7 | 0678
 8 | 66
 9 |
10 | 0
 (6 | 0 means 600 slot machines)
```

Dotplot:



Descriptive Statistics: numbers of slot machines

Min.	1st Qu.	Median	3rd Qu.	Max.
170	299	428	731	1000

36. He shoots, he scores.

a) The stemplot is shown below.

```
0 | 4
0 | 66
0 | 99
1 | 0
1 | 233
1 | 44
1 | 666
1 | 89
```

2 | 0011
points scored
(2 | 0 means 200 points)

b) The distribution looks slightly left skewed. The scores range from about 40–210 (note that these are truncated values). The median (the average of the 10th and the 11th values) is about 140. There are no outliers.

c) If we consider this as a left-skewed distribution, then the five-number summary is an appropriate summary and is given below.

Descriptive Statistics: points scored

Variable	Minimum	Q1	Median	Q3	Maximum
points scored	48.0	99.5	145.5	189.5	215.0

37. How tall? The histogram shows some low outliers in the distribution of height estimates. These are probably poor estimates and will pull the mean down. The median is likely to give a better estimate of the professor's true height.

38. He shoots, he scores again

a) The distribution of the points scored per season by Wayne Gretzky is slightly skewed to the left and has no outliers. The median is more resistant to the skewness than the mean.

b) The median, or middle of the ordered list, is 145.5 points. This is the average of the 10th value (142) and the 11th value (149).

c) The mean should be slightly lower as the distribution is slightly left-skewed.

39. World Series champs

The distribution of the number of homeruns hit by Joe Carter during the 1983–1998 seasons is skewed to the left, with a typical number of homeruns per season in the high 20s to low 30s. The season in which Joe hit no homeruns looks to be an outlier. With the exception of this no-homerun season, Joe's total number of homeruns per season was between 13 and 35. The median is 27 homeruns.

40. Bird species 2013.

a) The results of the 2013 Laboratory of Ornithology Christmas Bird Count are displayed in the stem and leaf display below.

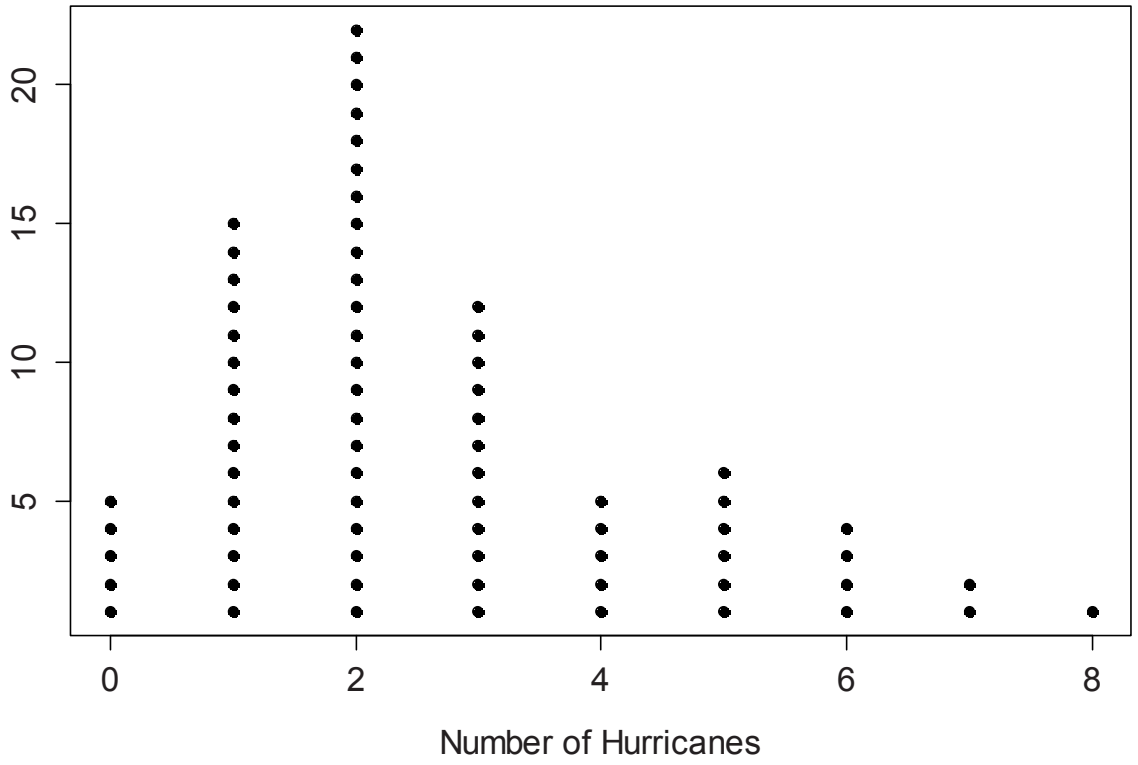
Number of Birds
8 | 2368
9 | 78
10 | 1156
11 | 8
12 | 468
13 | 136
14 |
15 | 0

16 | 6
17 |
18 | 4
Key: 15 | 0 = 150 birds

b) The distribution of the number of birds spotted by participants in the 2013 Laboratory of Ornithology Christmas Bird Count is skewed right, with a median of 112 birds. There are three high potential outliers, with participants spotting 150, 166, and 184 birds. With the exception of these outliers, most participants saw between 82 and 136 birds.

41. Hurricanes 2015.

a) A dotplot of the number of hurricanes each year from 1944 through 2015 is shown below. Each dot represents a year in which there were that many hurricanes.



b) The distribution of the number of hurricanes per year is unimodal and skewed to the right, with centre around 2 hurricanes per year. The number of hurricanes per year ranges from 0 to 8. There are no outliers. There may be a second mode at 5 hurricanes per year, but since there were only 6 years in which 5 hurricanes occurred, this may simply be natural variability.

42. Horsepower

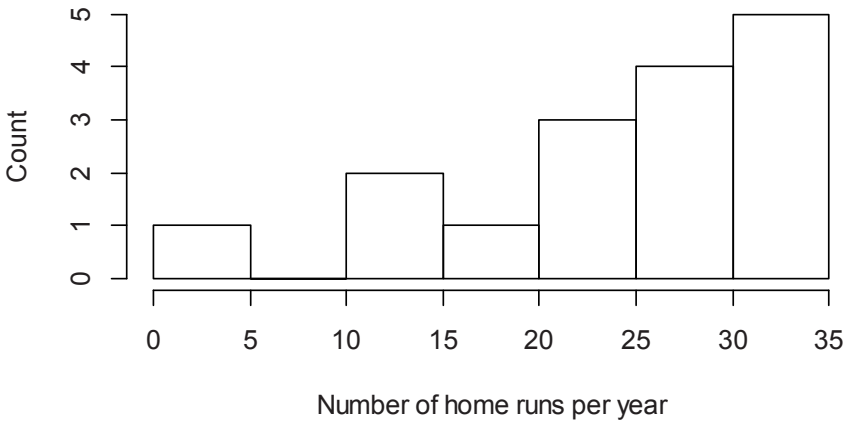
The distribution of horsepower of cars reviewed by *Consumer Reports* is nearly uniform. The lowest horsepower was 65 and the highest was 155. The centre of the distribution was around 100 horsepower.

```
 6 | 55889
 7 | 01158
 8 | 0058
 9 | 00577
10 | 359
11 | 00555
12 | 0559
13 | 0358
14 | 2
15 | 05
( 10|3 means 103 )
```

43. World Series champs again

a) This is not a histogram. The horizontal axis should be the number of homeruns per year, split into bins of a convenient width. The vertical axis should show the frequency; that is, the number of years in which Carter hit a number of home runs within the interval of each bin. The display shown is a bar chart / time series plot hybrid that simply displays the data table visually, in time order. It is of no use in describing the shape, center, spread, or unusual features of the distribution of home runs hit per year by Carter.

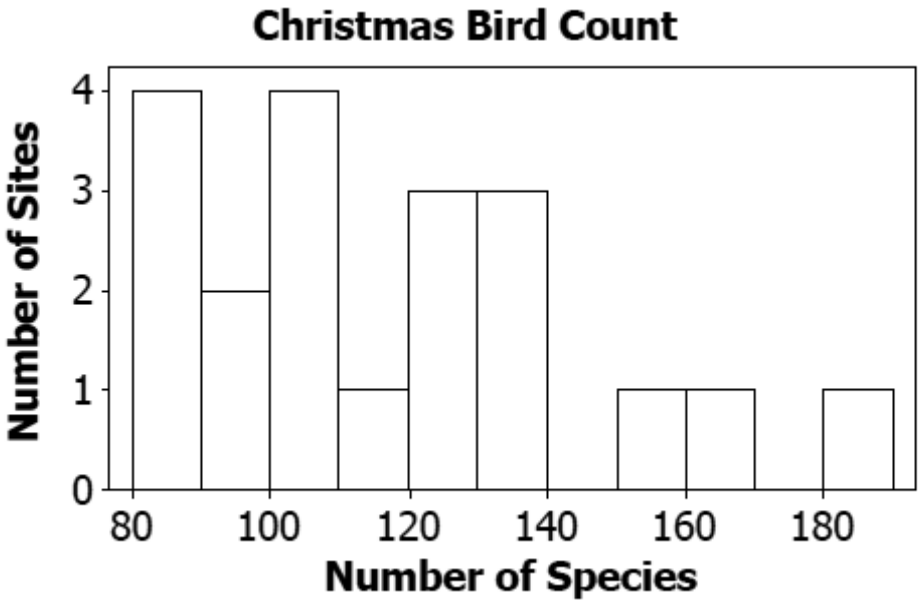
b) The histogram is shown below.



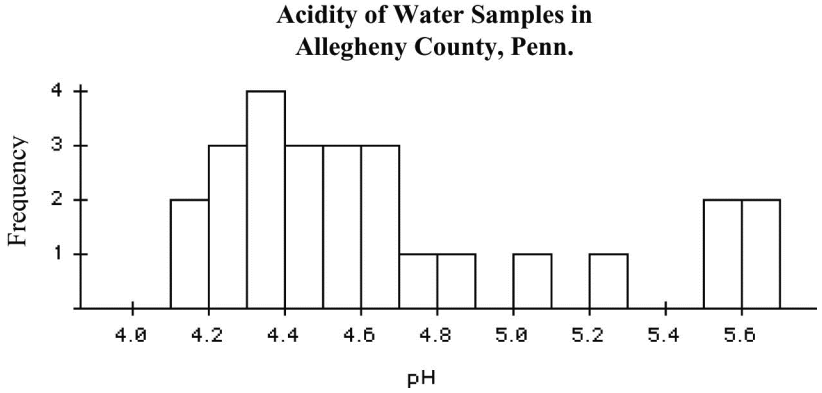
44. Return of the birds 2013.

a) This is not a histogram. The horizontal axis should split the number of counts from each site into bins. The vertical axis should show the number of sites in each bin. The given graph is nothing more than a bar chart, showing the bird count from each site as its own bar. It is of absolutely no use for describing the shape, center, spread, or unusual features of the distribution of bird counts.

b) The histogram is below.



45. Acid rain The distribution of the pH readings of water samples in Allegheny County, Pennsylvania, is bimodal. A roughly uniform cluster is centred on a pH of 4.4. This cluster ranges from pH of 4.1 to 4.9. Another smaller, tightly packed cluster is centred on a pH of 5.6. Two readings in the middle seem to belong to neither cluster.



46. Sip size

a) The distribution of height is most nearly symmetric and shows no outliers. That makes it the best candidate for summarizing with a mean. The distribution of weight is skewed to the right, which would inflate the value of the mean. The distribution of sip size has an outlier, which would inflate the value of the mean.

b) The distribution of sip size shows a high outlier. The standard deviation is sensitive to outliers, so the IQR is a better choice for summarizing spread. It is resistant to outliers.

47. Housing Price Boom.

a) The stem-and-leaf plot is shown below:

–0 | 421
 0 | 000000001222244
 0 | 789

Percent increase (rounded)
(0 | 9 means 9 percent)

b) Minimum = –4.0
The first quartile = median of lower 10 values = average of 5th and 6th values in ordered data = 0.0
Median = middle value = 11th value = 0
The third quartile = median of upper 10 values = average of 5th and 6th values from the end of the ordered data = 3.0
Maximum = 9.0

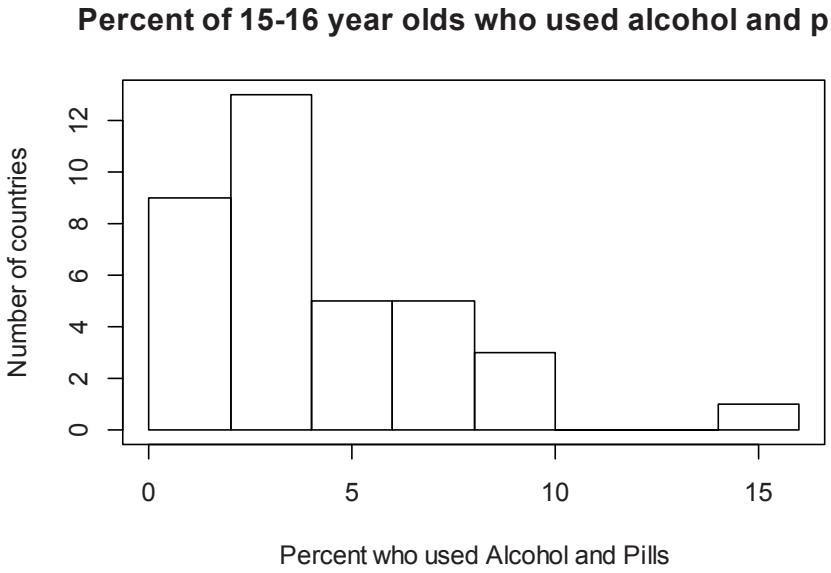
c) The mean of the data is approximately 1.6. The mean is larger than the median because the distribution is slightly right-skewed.

d) The percentage increases in price range from -4% to 9%. The median increase is 0%. The distribution is slightly right-skewed, but it is difficult to tell from the shape of the stem and leaf plot alone.

e) The data sorted by the percentage increase is shown below. The largest increases are in the Greater Toronto Area and Niagara regions of Ontario, and in British Columbia.

Metropolitan.Area	Percentage.increase
Saskatoon	-3.5
Calgary	-1.9
Québec	-0.6
Regina	-0.4
Halifax	-0.3
Charlottetown	-0.3
Edmonton	-0.3
Greater Sudbury and Thunder Bay	-0.1
Saint John, Fredericton and Moncton	0.1
St. John's	0.3
Ottawa-Gatineau	0.4
Montréal	1.4
Windsor	1.8
Kitchener	2.3
Winnipeg	2.3
London	2.4
Hamilton	3.7
Victoria	3.8
St. Catharines-Niagara	6.8
Vancouver	7.7
Toronto and Oshawa	9.0

48. Getting high 2011



The distribution of the percentage of 15–16 year-olds in 36 European countries who have used alcohol together with pills to get high is unimodal and skewed to the right. The country with the highest percentage of 15–16 year olds that have used alcohol and pills together is the Czech Republic (16%), which appears to be an outlier. A typical country might have a percentage of between 2.5% and 6.5%. The median is 4%.

Descriptive Statistics:

Min.	1st Qu.	Median	3rd Qu.	Max.
1.00	2.50	4.00	6.50	16.00

49. Trimmed mean

a) The sum of the 10 values given is 64, so the mean is 64/10 = 6.4
The data arranged in increasing order: 1, 5, 6, 6, 7, 7, 8, 8, 8, 8. The median is 7 (the average of the 5th and the 6th values).
There are 10 values in the data set and 10% of 10 is 1. To calculate the 10% trimmed mean, delete the smallest and the largest value in the sorted data set and calculate the average of the remaining values. The 10% trimmed mean is 55/8 = 6.875.

b) The data of Exercise 34, arranged in increasing order:
1 1 2 2 2 2 2 2 2 3 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 6 7 7 8 8 8 10 10 10 16. There are 36 values in the data set and 5% of 36 is 1.8, or approximately 2. Hence, to calculate the 5% trimmed mean, delete the two smallest and two largest values in the sorted data set and calculate the average of the remaining values. The 5% trimmed mean is 144/32 = 4.5. The mean is 4.78, and the median is 4.0.

c) The data in part (a) is left skewed. As a result, the median is bigger than the trimmed mean, which is bigger than the mean. The data in part (b) is right skewed, so the median is smaller than the trimmed mean, which is smaller than the mean.

50. Raptors 2011

a) The stem and leaf plot for the data is given below:

```
1 | 000000
2 | 0000
3 | 000
4 | 000
5 |
6 |
7 |
8 | 0
9 |
```

(4|0 means 4.0 or just 4)

b) The median is 2.0 (9th value in the ordered data set).
There are 17 values in the data set and 5% of 17 = 1 (approx.). As a result, the 5% trimmed mean is 34/15 = 2.27 (the mean of the 15 values, deleting the first and the last observations in the ordered data set). Ordinary mean is 43/17 = 2.53; median = 2.0; trimmed mean = 2.3; mean = 2.5. The mean is greater than the median because the distribution of length is right skewed. The maximum value (8) is a more extreme observation than the minimum; in fact, the 1.5 × IQR shows this as an outlier. This observation pulls the mean toward it. When you delete it, the mean should decrease.

c) The mode of the distribution is 1.0.

d) They could consider where the games were played, at home or on the road.

51. Final grades. The width of the bars is much too wide to be of much use. The distribution of grades is skewed to the left, but not much more information can be gathered.

52. Final grades revisited.

a) This display has a bar width that is much too narrow. As it is, the histogram is only slightly more useful than a list of scores. It does little to summarize the distribution of final exam scores.

b) The distribution of test scores is skewed to the left, with center at approximately 170 points. There are several low outliers below 100 points, but other than that, the distribution of scores is fairly tightly clustered.

53. Zip codes Even though zip codes are numbers, they are not quantitative in nature. Zip codes are categories. A histogram is not an appropriate display for categorical data. The histogram the Holes-R-Us staff member displayed doesn't take into

account that some 5-digit numbers do not correspond to zip codes or that zip codes falling into the same classes may not even represent similar cities or towns. The summary statistics are meaningless for the same reason. The employee could design a better display by constructing a bar chart that groups together zip codes representing areas with similar demographics and geographic locations.

54. Industry codes

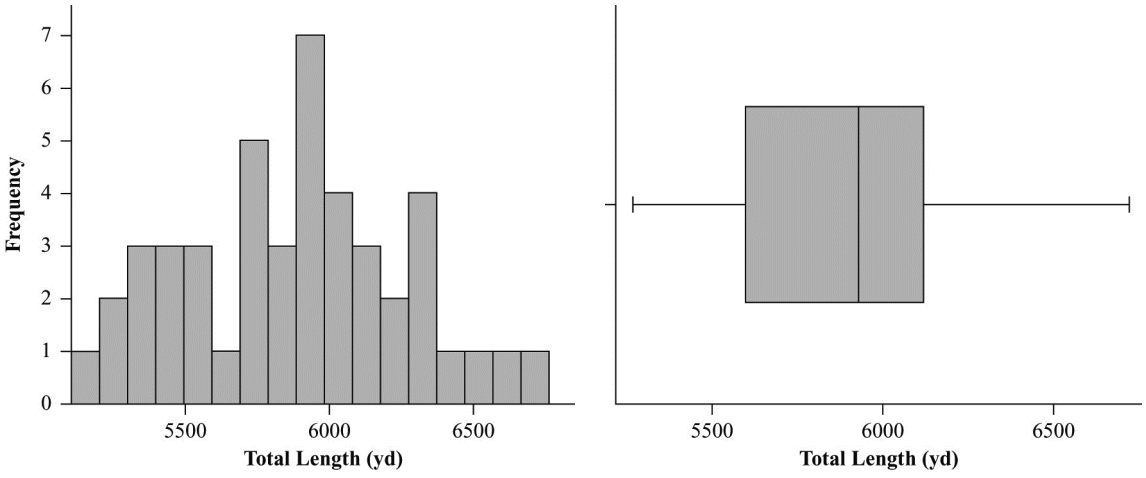
a) First of all, it must be noted that industry codes are categorical, so the use of a histogram as a display is inappropriate. Strangely enough, the investment analyst made even more mistakes! This display is really just a poorly constructed bar chart (of sorts). There are gaps in the display because all of the industry codes are integers and the widths of the bars are all less than 1. With 5 bars between 0 and 3.75, we can find the width to be 0.75. So, for example, there appears to be a gap between 2.25 and 3, simply because there are no integers between 2.25 and 3 (remember, the upper boundary is not included). The gaps aren't really there, but a poor choice of scale makes them appear.

b) This question doesn't really make any sense. "Unimodal" is a vocabulary word specific to describing distributions of quantitative data. As mentioned before, the industry codes are categorical. Saying that the mode is 1 means that most of the companies are code 1, or Financial services. Saying that the median is 6 is completely meaningless because these are unordered categorical data.

c) A histogram can never be used to summarize categorical data. The analyst would be better off displaying the data in a bar chart, with relative heights of the bars representing the number of companies involved in each industry, or a pie chart displaying the percentage of companies involved in each industry.

55. Golf courses

a)



b) In any distribution, 50% of scores lie between Quartile 1 and Quartile 3. In this case, Quartile 1 = 5585.75 yards and Quartile 3 = 6131 yards.

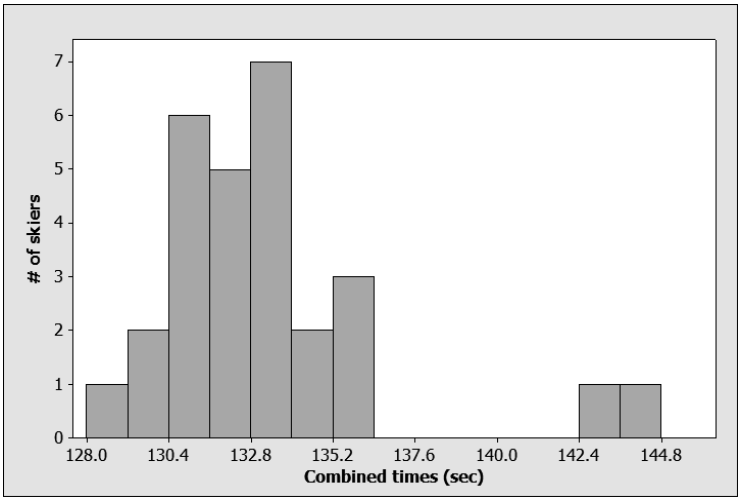
c) The distribution of golf course lengths appears roughly symmetric, so the mean and standard deviation are the preferred measures of centre and spread.

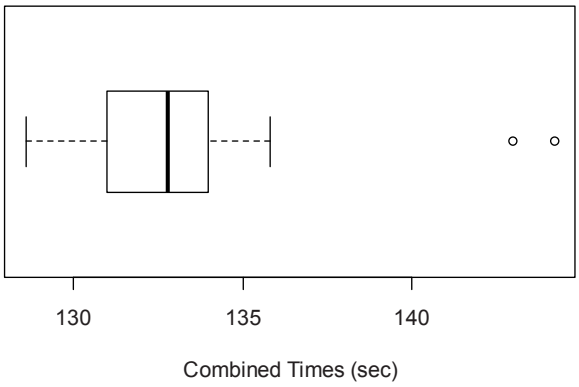
d) The distribution of the lengths of all the golf courses in Vermont is roughly unimodal and symmetric. The mean length of the golf courses is approximately 5900 yards. Vermont has golf courses anywhere from 5185 yards to 6796 yards long. There are no outliers in the distribution.

e) (Mean – Std Dev, Mean + Std Dev) → (5892.91 – 386.59, 5892.91 + 386.59) → (5506.32, 6279.5). There are 28 observations in this range. Or 28/45 = 62% are in this range.
(Mean – 2Std Dev, Mean + 2Std Dev) → (5892.91 – 2(386.59), 5892.91 + 2(386.59)) → (5119.73, 6666.09). There are 44 observations in this range. Or 44/45 = 98% are in this range.
These are reasonably close to what we expect according to the empirical rule for roughly bell-shaped distributions.

56. Women's Olympic alpine 2010

a)





b) In any distribution, 50% of scores lie between Quartile 1 and Quartile 3. In this case, Quartile 1 = 131.0 seconds and Quartile 3 = 134.0 seconds.

c) The distribution of the Women's Combined times at the 2010 Olympics is unimodal and skewed to the right. The median time was 132.8 seconds. Half the times fall between 131.0 seconds and 134.0 seconds. There were two very slow times of 143.0 and 144.2 that were outliers.

57. Math scores 2009

a) The 5-number summary of the national averages is given below:

Descriptive Statistics: Ave Score

Min.	1st Qu.	Median	3rd Qu.	Max.
419	487	496.5	514	546

The IQR, mean, and the standard deviation of the national averages is given below:

Descriptive Statistics: Ave Score

Variable	Mean	StDev	IQR
Avg. Score	495.7	29.66	27

The distribution of average scores appears to be left-skewed. The long left tail pulls the mean down toward the smaller values (in fact, three of these small values are smaller than Q1 – 1.5 × IQR, and are thus suspect outliers). The low outliers attract the mean toward them, whereas the median is resistant to outliers. (Alternatively, one might say that there are several very small scores, and putting them aside, there is some slight right-skewness)

b) Since there are outliers and asymmetry in the data set, the 5-number summary is better than the mean and standard deviation which are not resistant and are more useful for symmetric distributions.

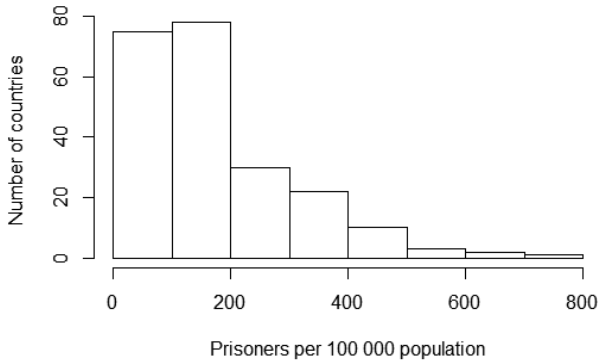
c) Thirty-four countries participated in the program. The highest national average is 546 and the lowest is 419. The median national average is 496.5. The interquartile range is 27. Twenty-five percent (i.e., 9 countries) of the participating countries had a national average 487 or below and at least 25% of the countries had a national average of 514 or above. Canada's national average (which is 527) is in the top 25% of all participating countries, more specifically, the 5th highest of all participating countries. The United States' national average (which is 487) is the 25th highest of all participating countries.

d) Using the empirical rule, the middle 68% of the students have their scores within one standard deviation from the mean, i.e., 527 – 88 = 439 to 527 + 88 = 615.
The middle 95% of the students have their scores within two standard deviations from the mean, i.e., 527 – 2 × 88 = 351 to 527 + 2 × 86 = 703.
Just about all of the students will have their scores within three standard deviations from the mean, i.e., 527 – 3 × 88 = 263 to 527 + 3 × 88 = 791.
Using this information we can fill in the blanks as shown below:

About two-thirds of students scored between 439 and 615.
Only about 5% scored less than 351 or more than 703.
Only a real math genius could have scored above 791.

58. Incarceration rates 2013

a) The histogram of incarceration rates is shown below. A stemplot is also an appropriate display.



0 | 22223333333444444444
0 | 55555555666666666666677777777777888888888999999
1 | 000000000001111111111222222222223333333333444444
1 | 555555555556666667777778899999
2 | 000011122223334444444
2 | 556667788899
3 | 01111111223334
3 | 56778899
4 | 0011234
4 | 6799
5 | 134
5 |
6 | 4
6 | 5
7 | 2

(6|4 means 640 prisoners per 100 000 of nation's population)

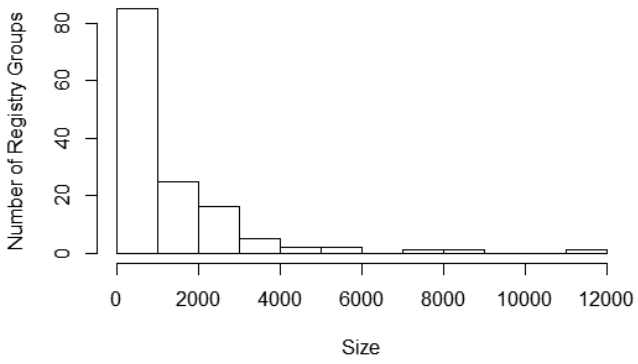
b) For skewed distributions the 5-number summary is more appropriate.

Minimum	Q1	Median	Q3	Maximum
17	76.5	130	238	716

c) The distribution of incarceration rates is right skewed with median 130 and interquartile range 238 – 76.5 = 161.5. The maximum value (716 in US) is an outlier, along with St Kitts and Seychelles. Canada's incarceration rate is 114 and this is below the median but greater than the first quartile.

59. First Nations 2010

a) A histogram (or a stemplot) is an appropriate graphical display. Both these displays are given below. The distribution is right skewed. This means the portion of large registry groups is relatively small. There are some (one or two) outliers (indicated by the gaps in the histogram or stemplot). The median size is 717. The interquartile range is 1220.




```
 0 | 011122222222333333344444444444444
 0 | 555555555555566666666666777777777777788888899
 1 | 000000111122222344
 1 | 56677889999
 2 | 01112233344
 2 | 5567
 3 | 0234
 3 | 8
 4 | 033
 4 |
 5 |
 5 | 57
 6 |
 6 |
 7 | 4
 7 |
 8 | 0
 8 |
 9 |
 9 |
10 |
10 |
11 | 2
  (7 | 4 means 7400)
```

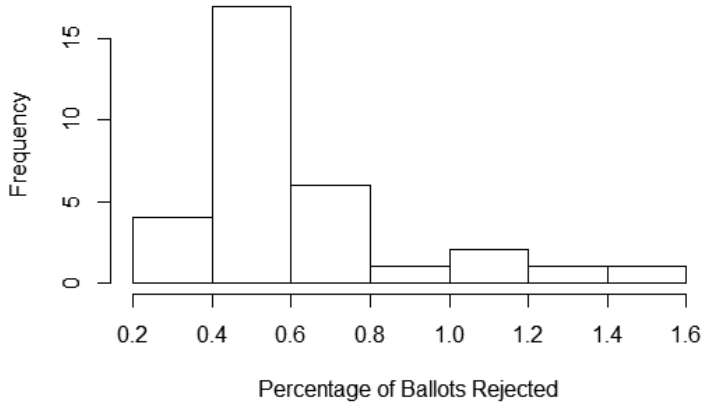
The 5-number summary is:

Minimum	Q1	Median	Q3	Maximum
41	452	717	1672	11202

b) If we calculate the band sizes, most band sizes will be same as the registry group sizes since only Six Nations of the Grand River in Ontario consists of more than one registry group. This very large band, consisting of 13 registry groups, will increase considerably the mean and standard deviation, while having a much smaller effect on the more resistant median and IQR. The histogram will have a bigger gap due to this very large band.

60. Elections 2011 Maritimes

a) A histogram or a stemplot is a suitable display. Both these are shown below.




```
 2 | 000
 4 | 00000000000
 6 | 0000000000
 8 | 0000
10 | 0
12 | 00
14 |
16 | 0
 (2|0 means 0.20 percent)
```

b) The mean is 0.659 and the standard deviation is 0.292.

c) The 5-number summary is given below.

Descriptive Statistics: Percentage of Ballots Rejected

Minimum	Q1	Median	Q3	Maximum
0.3	0.5	0.6	0.75	1.6

d) The mean and median differ here because the distribution of the percentage ballots rejected is skewed.

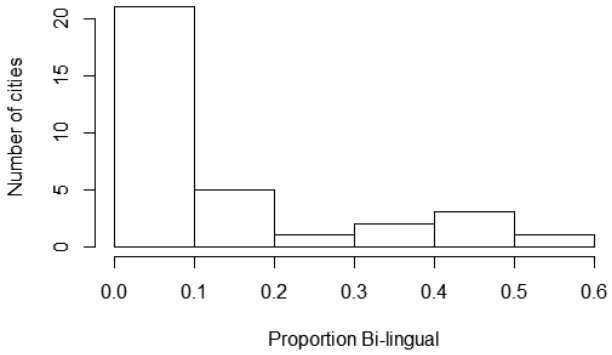
e) For skewed distributions (like the distribution of the percentage ballots rejected in this question) the 5-number summary is better than the mean and the standard deviation. From the histogram (and the stemplot above), we see that there are some possible outliers in the data. Mean and the standard deviation are measures not resistant to outliers. This also favours the use of the 5-number summary to summarize this data set.

f) The mean will increase.
The current median is 0.6%. Both the 0.8 and 1.8 are on the same side of the median. Changing any value to a different value on the same side of the median does not change the median.
The standard deviation will increase.
The IQR will not change. The current third quartile is 0.75 and the value 0.8, which currently falls to the right of Q3 remains to the right in the new ordered data set. Thus Q3 is unchanged, as is the IQR.

g) The distribution is right skewed. This means the proportion of ridings with high percentages of ballots rejected is relatively small. The median is 0.6%. The interquartile range is 0.75% – 0.5% = 0.25%, so about half the ridings have rejection percentages between 0.5% and 0.75%. Madawaska–Restigouche's 1.6% appears to be an outlier.

61. Bi-lingual ni-lingual 2011

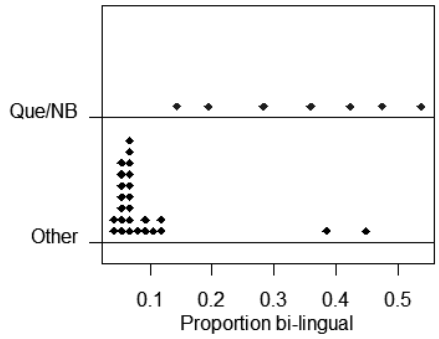
a) The histogram of the proportion of city residents who are bilingual is shown below. The highest proportion of bilinguals is in Montreal with more than 50% bilinguals, and the lowest proportion is in Brantford with less than 5% bilinguals. The median is about 8%. The distribution of the proportion of city residents who are bilingual is right skewed. This means a relatively smaller number of cities with high proportion of bilinguals and more cities with a low proportion of bilinguals. The distribution of proportion of bilinguals appears bimodal it looks like the distribution of a few cities with a large proportion of bilinguals has a distinct shape compared to the other cities. Summary statistics are given below. About 25% of the cities have less than 7% bilinguals.



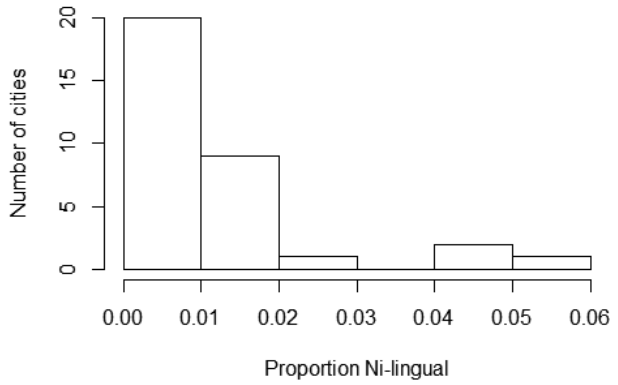
Descriptive Statistics: Proportion Bi-lingual

Mean	StDev	Min.	Q1	Median	Q3	Max.
0.153	0.145	0.042	0.066	0.077	0.173	0.539

b) The dotplot below shows the cities in Quebec and New Brunswick on top. They have relatively high proportions of bilinguals.



c) The histogram of the proportion of city residents who speak neither English nor French is shown below. The highest proportion of residents who speak neither English nor French is in Vancouver with more than 5%, and the lowest proportion is in Saguenay with about 0.03%. The median is about 0.69%. The distribution of the proportion of city residents who speak neither English nor French is right skewed. This means there is a relatively smaller number of cities with a high proportion of residents who speak neither English nor French and more cities with a low proportion. Three cities (Vancouver, Abbotsford, and Toronto, the bars above 0.031 on the histogram) appear to be outliers.

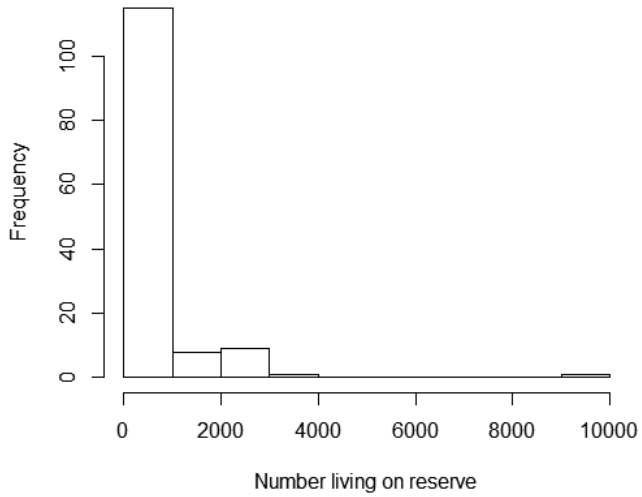


Descriptive Statistics: Proportion Ni-lingual

Mean	StDev	Minimum	Q1	Median	Q3	Maximum
0.0111	0.0135	0.0003	0.0026	0.0069	0.0144	0.0560

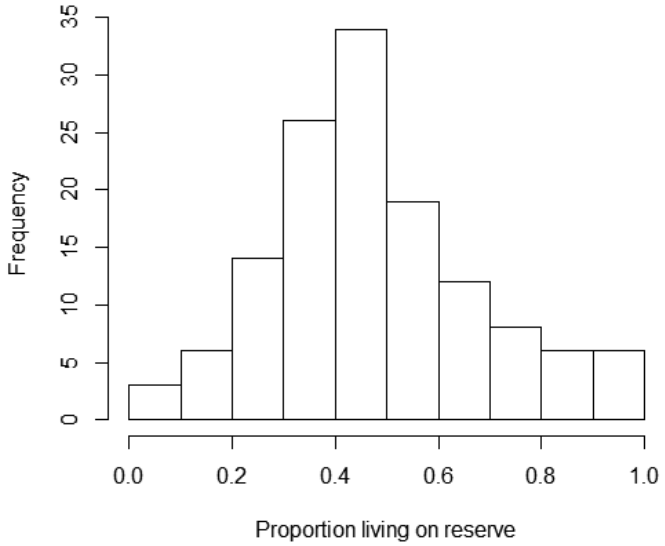
62. First Nations 2010 on reserves

a) The histogram of numbers living on reserve is shown below :



The distribution is very right-skewed, with numbers living on reserve ranging from near 0 to over 3000, plus a possible outlier around 9000, and mode around 500. Band size distribution is very similar in shape, very right-skewed, but with bigger numbers and mode around 1000.

b) The histogram of the proportion living on reserve is shown below:



This distribution is much more symmetric and looks fairly bell-shaped (maybe slightly right-skewed), centred around 0.5, ranging from nearly 0 to nearly 100%. A guess of the mean might be about 0.5, and standard deviation about 0.2 (0.3 to 0.7 catches about 2/3 of the distribution, and 0.1 to 0.9 catches about 95%, as indicated by the empirical rule).

c) mean = 0.4785, sd = 0.2053

d) Mean should exceed median for the right-skewed on reserve counts, but they should be similar for the more symmetric proportions.

e) Likely to be small bands. Around 50% of of the members of a band on average are living on reserve, so if the band size is 10, having 7 or 70% on reserve seems much more likely than a band of size 10 000 having 7000 on reserve.

f) Not the same, unless all band sizes were the same (which they are not), or a mathematical fluke.

63. Elections 2011 again

a) The closeness of the mean and median suggests that the distribution of the percentage of voter turnout in the 2011 election is approximately symmetric.

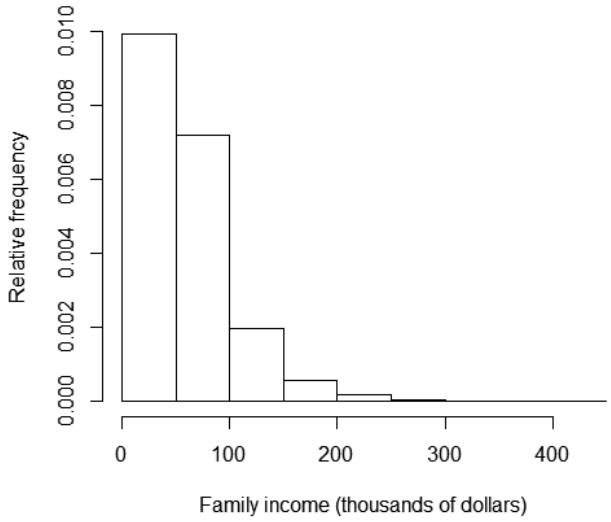
b) The distribution of the percentage of voter turnout in the 2011 election is approximately symmetric, with mean 61.068 and standard deviation 5.772.

c) Mean + 2 Stdev = 61.068 + 2 × 5.772 = 72.612
Mean – 2 Stdev = 61.068 – 2 × 5.772 = 49.524
From the histogram there appears to be about 12.5 observations below 50 and 6.5 observations above 72.5 (assuming uniform distribution within bins). This is about 19/308 = 6.2%. Thus, about 93.8% of the observations are within 2 standard deviations from the mean. This is pretty close to 95% as we would expect according to the empirical rule for bell-shaped distributions. The exact answer is 292/308 or 94.8%.

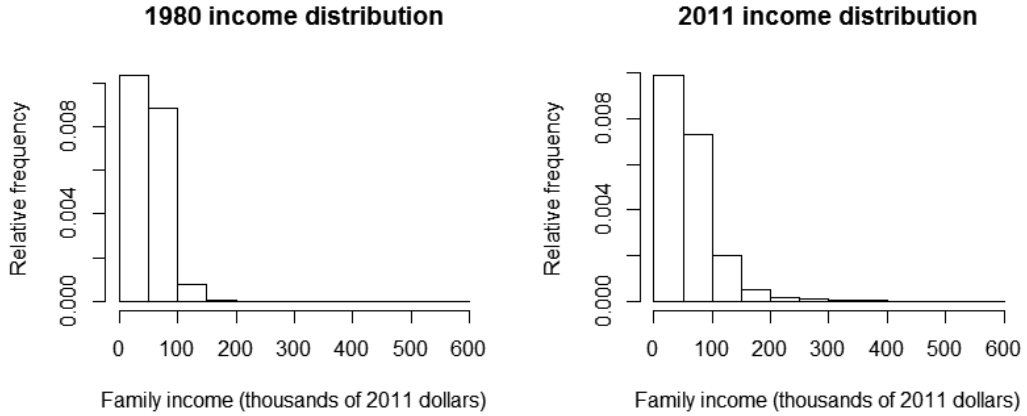
d) The overall percentage won't be exactly the same as the mean of the percentages as the number of eligible voters is not exactly the same in each electoral district. However, since the number of eligible voters in each riding is pretty close, we should expect the overall percentage to be fairly close to 61.07%.

64. Family income 2011

a) This is because the distribution of income is right skewed due to the relatively smaller proportion of very high-income families. Shown below is what a histogram of family income might look like:



b) In 2011, the mean was (63 000 – 50 700)/50 700 = 24.3% higher compared to the median. In 1980, the mean was (54 000 – 49 500)/49 500 = 9.1% higher compared to the median. This shows that the distribution has become more skewed. This means the relatively small proportion of the rich has gotten richer compared to the larger population. Note that the median income has also risen, however, so that the middle income has increased, although not as much as the average income. Shown below are hypothetical histograms of family income for 1980 and 2011:



c) In 2030, the mean is (80 000-60 000)/60 000 = 33.3% higher compared to the median. As in part b), this shows that the distribution has become more skewed compared to 2011. This means the relatively small proportion of the rich has gotten richer compared to the larger population. Note that the median income has also risen, however, so that the middle income has increased, although not as much as the average income.

65. Run times continued

a) The mean will be smaller because it depends on the sum of all of the values. The standard deviation will be smaller because we have truncated the distribution at 32. The range will be smaller for the same reason. The statistics based on quartiles of the data (median, Q1, Q3, IQR) won't change because the distribution was truncated to the right of Q3.

b) The spread of the distribution will be relatively smaller because we have added observations near the middle. Thus, we expect the standard deviation and IQR to be smaller. We should only see a small change to the mean and median because the added observations are near the middle of the distribution. The range will not change.

c) The spread of the distribution will be relatively larger because we have added observations at about 2 minutes left and right of the mean. Thus, we expect the standard deviation and IQR to be bigger. Because the additional observations were added symmetrically to either side of the mean, there should be little change to the mean or median, and the range will remain unchanged.

d) Subtracting one minute from each time shifts the entire distribution to the left by one minute. The mean and median should decrease by one minute. However, there has been no change to the shape of the distribution, so the standard deviation, IQR, and range will all remain the same.

e) The values added at 35 are further from the mean than the values at 29.5, so the mean will increase. The overall spread of the distribution has increased, so the standard deviation will increase. The added observations at 35 minutes lie above Q3, so the IQR will increase. Because the same number of observations was added on either side of the median, its value won't change. The minimum and maximum values have not changed, so the range remains unchanged as well.

f) We have moved some values outside of the Q1 to Q3 interval further away from the centre of the distribution. Q1 and Q3 remain unchanged, so the IQR remains unchanged. Because the adjustment is symmetric, the mean and median are unchanged. The range remains the same because the minimum and maximum are unchanged. The standard deviation increases because the spread of the distribution has increased.

66. Test scores continued

a) Mean will increase because we are removing the students with the lowest grades. The median might increase a bit since the position of the median moves up, but this increase in median is usually very small (median is resistant to

outliers). Standard deviation and range will decrease since what we are removing are the extreme values in the data set and so the spread will decrease. IQR will not change much (resistant to outliers).

b) Mean will decrease. Median will not change much (can decrease slightly). Spread will decrease since we are making the large values smaller, (i.e., shortening the upper tail) and so standard deviation and range will decrease. IQR will not change much (resistant to outliers).

c) Changing 35 to 3.5 and 45 to 4.5 will decrease the mean since we are making the values smaller. The range will not change since the minimum and the maximum values were not changed. Moving values to the extremes (i.e., tails of the distribution) usually increases sums of squares of deviations from the mean ($\sum(X_i - \bar{X})^2$) and so increasing the standard deviation. Changes to the median and IQR should be very slight due to their resistance to changes to very small portions of a data set.

d) (30–35) is almost the centre of the distribution, and removing values close to the centre will not much change the mean, median, or the range. The standard deviation will increase because the deletion of the values close to the centre does not much change the sums of squares of deviations from the mean ($\sum(X_i - \bar{X})^2$), but the number of values in the data set (i.e., n) decreases and so the standard deviation $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$ increases (or more intuitively, removing some very small deviations makes the average squared deviation bigger). IQR increases because Q1 decreases and Q3 increases.

e) (30–35) is almost the centre of the distribution and adding values close to the centre will not much change the mean, median, or the range. The standard deviation will decrease because adding values close to the centre does not much change the sums of squares of deviations from the mean ($\sum(X_i - \bar{X})^2$), but the number of values in the data set (i.e., n) increases and so the standard deviation $\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$ decreases (more intuitively, adding some very small deviations makes the average squared deviation smaller). IQR decreases because Q1 increases and Q3 decreases.

f) Adding an equal number of observations on the two sides of the centre (equidistant from the centre) does not change the mean or the median (only very small changes). The range does not change as the new additions do not change the minimum and maximum values. Their positions far from the centre will increase the average squared deviation from the mean , and so increase the standard deviation. Since 15 and 35 are both outside the current quartiles, adding 15s and 35s will make Q1 smaller and Q3 larger and hence make IQR larger.

g) Adding 5 to all values increases the mean and median by 5 points. Adding 5 to all values will not change the spread and so will not change range, standard deviation, or IQR. (They are measures of spread).
h) Doubling all the values in the data set will double all these statistics (i.e., mean, median, range, standard deviation, IQR).

67. Grouped data

a) The class midpoints and the mean and the standard deviation of the midpoints are shown below:

Height (class midpoint)

61	61	61	61	61	61	61	61	61	61	61	61	61	61	61
61	61	64	64	64	64	64	64	64	64	64	64	64	64	64
64	64	64	64	64	64	64	64	64	64	64	64	64	64	64
64	64	64	64	67	67	67	67	67	67	67	67	67	67	67
67	67	67	67	67	67	67	67	67	67	67	67	67	67	67
67	67	67	67	67	67	67	67	67	67	67	70	70	70	70
70	70	70	70	70	70	70	70	70	70	70	70	70	70	70
70	70	70	70	70	73	73	73	73	73	73	73	73	73	73
73	73	73	73	73	76	76	76	76	76					

Descriptive Statistics: Height (class midpoint)

Variable	N	Mean	StDev
Height (class mid)	130	67.069	3.996

The actual heights and the mean and the standard deviation of the actual values are given below. The mean and the standard deviation of the actual data are very close to those calculated using the midpoints. For grouped data, we assume that all the values in a class are equal to the midpoint. This assumption is usually reasonable unless the class width is big.

Height (actual)

60	60	61	61	61	61	61	61	62	62	62	62	62	62	62
62	62	63	63	63	63	63	63	63	64	64	64	64	64	65
65	65	65	65	65	65	65	65	65	65	65	65	65	65	65
65	65	65	65	66	66	66	66	66	66	66	66	66	66	66
66	66	66	66	66	66	66	67	67	67	67	67	67	67	68
68	68	68	68	68	68	68	68	68	68	68	69	69	69	69
69	70	70	70	70	70	70	70	70	70	70	70	71	71	71
71	71	71	71	71	72	72	72	72	72	72	72	72	72	73
73	73	73	74	74	75	75	75	75	76					

Descriptive Statistics: Height (actual)

Variable	N	Mean	StDev
Height (actual)	130	67.115	3.792

b) Mean = $\frac{\sum_i f_i m_i}{\sum_i f_i}$, $Variance = \frac{\sum_i f_i (m_i - \bar{X})^2}{\sum_i f_i - 1}$

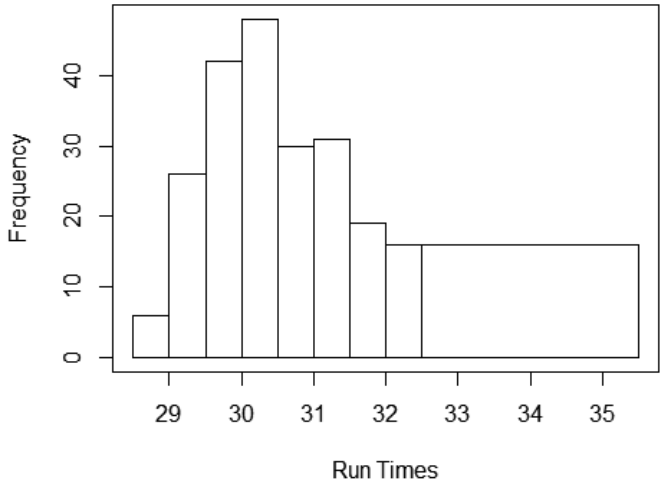
68. Grouped data II

a) 142.85

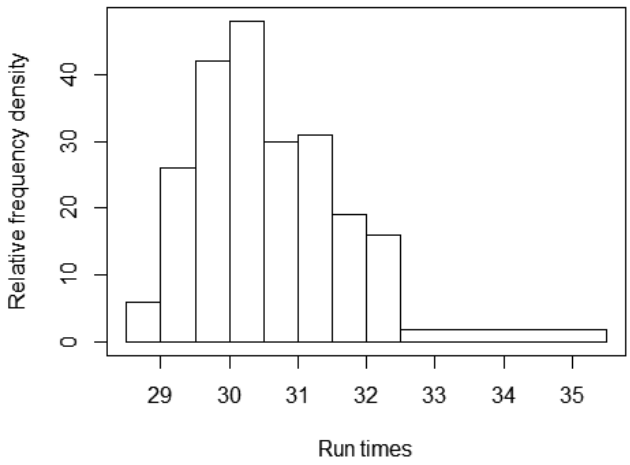
b) Computing the binned approximation to the mean, we get 145.0. This is (145.0 – 142.8)/142.8 = 2% off of the true value.

69. Unequal bin widths

a) The last wide bin now has a rather big rectangle above it, with height equal to about 13. The number of run times exceeding 32.5 appears to have grown!

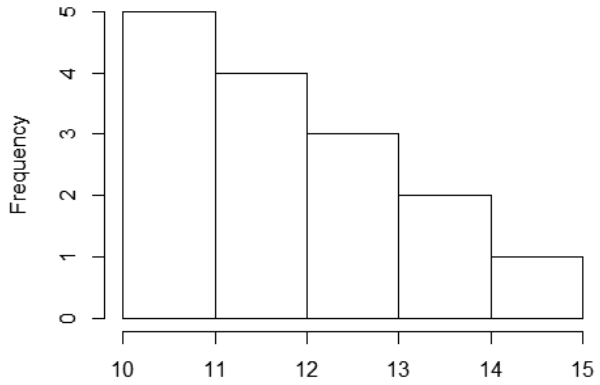


b) Now the rectangle height over the last bin equals the average of the 6 rectangle heights in the original histogram, i.e. the area of this rectangle equals the total area of the 6 rectangles it replaced. We have been true to the Area Principle, and kept area proportional to frequency. (Since area is height multiplied by width, the rectangle height must be proportional to frequency divided by bin width)

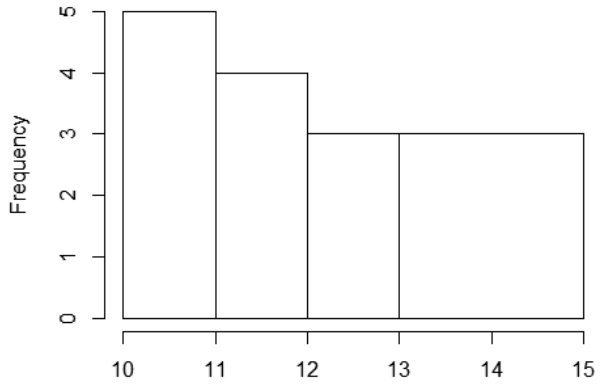


70. Unequal bin widths II

a)

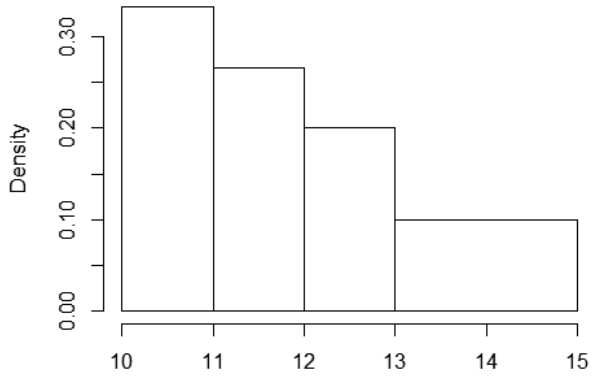


b)



With the bin widths of 1, the area of the bars corresponds to the frequency in that bin. With the last bin of length 2, the area is twice the frequency in that category. It appears there are three observations on average in the two bins 13–14 and 14–15, when in fact there are three observations in total. The Area Principle was violated.

c)



If the vertical axis is frequency divided by bin width, then the area of the last bar is 3, which corresponds to the frequency in the bin 13–15. Or, there are 1.5 observations on average in the two bins 13–14 and 14–15.

71. Rock concert accidents

a) The histogram and boxplot of the distribution of "crowd crush" victims' ages both show that a typical crowd crush victim was approximately 18–20 years of age, that the range of ages is 36 years, and that there are two outliers, one victim at age 36–38 and another victim at age 46–48. In addition, both show some right-skewness.

b) This histogram shows that there may have been two modes in the distribution of ages of "crowd crush" victims, one at 18–20 years of age and another at 22–24 years of age. Boxplots, in general, can show symmetry and skewness, but not features of shape like bimodality or uniformity.

c) Median is the better measure of centre, since the distribution of ages is right skewed and has outliers. Median is more resistant to outliers than the mean.

d) IQR is a better measure of spread, since the distribution of ages has outliers. IQR is more resistant to outliers than the standard deviation.

72. Slalom times 2010

a) The histogram and boxplot of the distribution of Men's Giant Slalom times both show that a typical time is around 165 seconds, and that the range of slalom times is about 55 seconds. Both displays also show that the distribution of slalom times is right-skewed toward the larger times.

b) Since the distribution of slalom times is skewed and contains possible outliers, the median is the better summary of the centre.

c) In the presence of skewness and possible outliers, we would prefer the IQR over the standard deviation as a measure of spread.