**CHAPTER 2 SOLUTIONS**

* 1. The R code used to generate the frequency table below is **table(NELS$absent12)**

Never 1-2 Times 3-6 Times 7-9 Times 10-15 Times Over 15 Times

55 195 180 51 12 7

The R code used to generate the percent table below is **percent.table(NELS$absent12)**

Never 1-2 Times 3-6 Times 7-9 Times 10-15 Times Over 15 Times

11.0 39.0 36.0 10.2 2.4 1.4

1. 180.
2. It is impossible to know the exact number. However, we can estimate the value by assuming that equal numbers of students were absent 3, 4, 5, and 6 times. Accordingly, the total number absent exactly 3 times in twelfth grade is estimated to be 180/4 = 45.
3. 36 percent.
4. 86 percent.
5. Positively skewed. The bulk of students have six or fewer absences, and there are only a few students with an unusually high number of absences.
   1. The R code used to generate the frequency table below is **table(NELS$parmarl8)**

Divorced Widowed Separated Never-Married Marriage-Like Married

26 6 8 5 5 427

The R code used to generate the percent table below is **percent.table(NELS$parmarl8)**

Divorced Widowed Separated Never-Married Marriage-Like Married

5.2 1.2 1.6 1.0 1.0 85.4

1. 8.
2. Of all students that reported their parents’ marital status in eighth grade, 1.6 percent reported that their parents were separated.
3. Married.
   1. The R code used to generate the frequency table below is **table(NELS$famsize)**

2 3 4 5 6 7 8 9

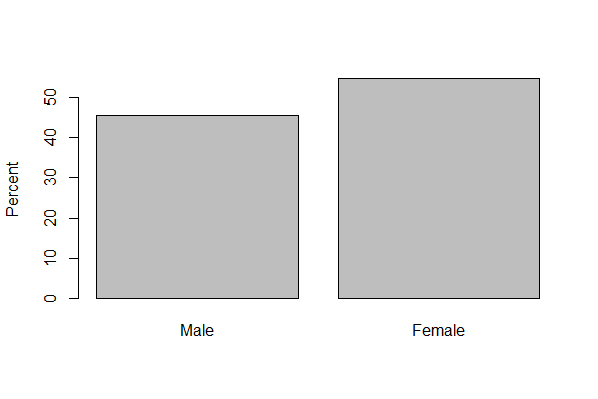
9 52 199 142 55 21 9 13

The R code used to generate the percent table below is **percent.table(NELS$famsize)**

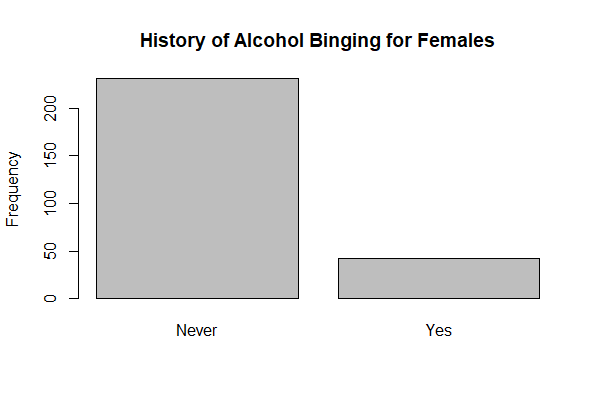
2 3 4 5 6 7 8 9

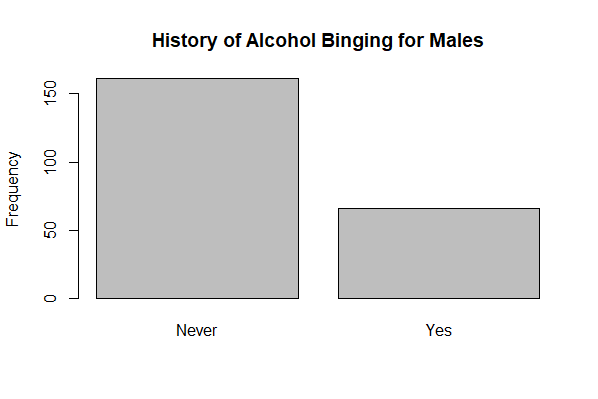
1.8 10.4 39.8 28.4 11.0 4.2 1.8 2.6

1. 9.
2. 52.
3. 9 + 52 = 61.
4. 10.4 percent.
5. 12.2 percent.
6. *Q2* is approximately 4.
   1. The R code used to generate the graph below is **barplot(percent.table(NELS$gender), ylab = "Percent")**



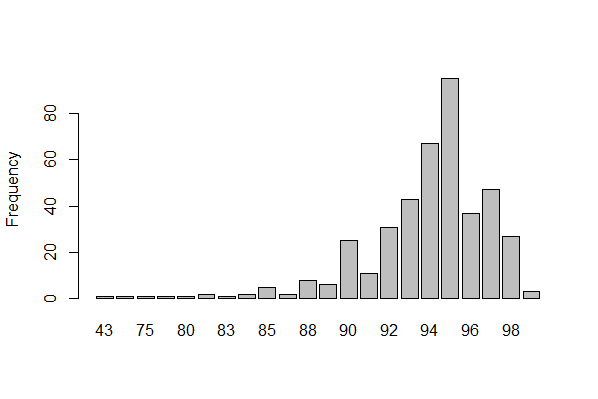
1. Gender is a nominal variable.

1. There are more females.
2. Just over 50 percent.
3. Yes. Every person in the data set is classified as either male or female, so the categories are mutually exclusive and exhaustive.
   1. The first graph below is that for females, the second is for males. 

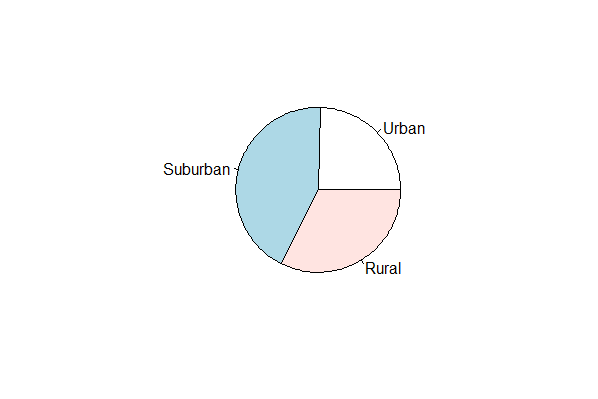


1. For males the answer is approximately: (70/230) x 100 or 30.4 percent

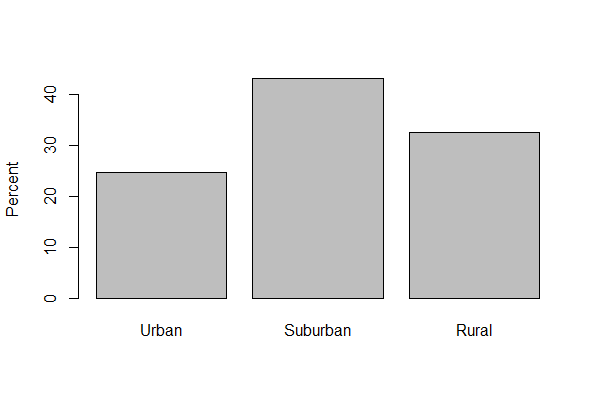
1. For females the answer is approximately: (40/270) x 100 or 14.8 percent. Males have a greater tendency to have ever binged on alcohol.
   1. The R command used to generate the graph is **barplot(table(NELS$schattrt), ylab = "Frequency")**



1. The values are not equally-spaced, making it difficult to judge the shape of the distribution, in general, and the distance of the values in the lower tail from the bulk of the distribution, in particular.
2. A stem-and-leaf, histogram, or interactive line graph.
3. The shape of the distribution is negatively skewed, even more negatively skewed than it appears to be given the unequal spacing of values along the x-axis. Most schools have attendance rates above 90 percent; however, there are two schools with attendance rates below 50 percent.
   1. The R command is **pie(table(NELS$urban))**



1. Approximately a quarter, or 25%,, come from an urban area.
2. The R command to generate the graph is **barplot(percent.table(NELS$urban), ylab = "Percent")**



1. There is no “correct” answer to this question. It is a matter of preference.
2. Most students are from suburban areas, followed by rural, and the fewest are from urban areas.
   1. The R command to generate this graph is **stem(Learndis$mathcomp, scale=2)**

6 | 12

6 | 6678999

7 | 002223334

7 | 5566777777888888889

8 | 000122444444

8 | 56667789

9 | 011222333333

9 | 5566777788

10 | 001333

10 | 8

11 | 223

11 | 5678

12 | 1

1. The variable is interval, so a line graph, histogram, or boxplot, could also be used.
2. 2 + 7 + 9 + 19 + 12 + 8 + 12 + 10 + 6 + 1 + 3 + 4 + 1 = 94.
3. 9 + 19 = 28.
4. No.
5. It is reasonably symmetric because the two tails are about equal in size and there are no outliers. It is also slightly negatively skewed because the tail associated with the larger values is longer.
6. Because there are an even number of scores (94), the 50th percentile is the average of the two middle scores, which are in the 47th and 48th positions. Both of these values are 84, so the 50th percentile is 84.
7. 78; it occurs 8 times.
8. 121.
9. Yes.
   1. The R command to generate this graph is **stem** **(Statisticians$Birth, scale = 2)**

182 | 2

183 |

184 |

185 | 7

186 |

187 | 6

188 |

189 | 03345

190 | 069

191 | 5

1. 1.
2. Yes. One statistician was born in 1822.
3. The distribution is negatively skewed because of the low outlier.
4. Two statisticians were born in 1893. No other pair of statisticians was born in the same year.
5. 1915.
6. Yes.
   1. The R command to generate this graph is **stem** **(Framingham$TOTCHOL1, scale = 2)**

13 | 3

14 |

15 | 2579

16 | 2233444557789

17 | 12334566779

18 | 01135555556666667799

19 | 00000224455555556666677778889999

20 | 000112223344445566666778888

21 | 00001122233333444444444556667788999

22 | 000001111223333334555555677888899999

23 | 00011112222233445555677778899999

24 | 000000111122233333444555556678888999

25 | 00000022223333333445556667777888888999

26 | 0000001123445556666677888999

27 | 000123455557888999

28 | 00012233556678

29 | 00001112222255579

30 | 00345588

31 | 02335578

32 | 23458

33 | 0233

34 | 056

35 | 3

36 |

37 |

38 |

39 |

40 |

41 |

42 |

43 |

44 |

45 |

46 | 4

1. 133.
2. Yes. One person’s initial cholesterol was 464.
3. Positively skewed due to the two high outliers.
4. 235.
5. 464.
6. Yes.
   1. The R command to generate this graph is **stem(NELS$achmat08, scale = 2)**

36 | 6

37 | 122

38 | 478

39 | 0357

40 | 0225555799

41 | 34689

42 | 0335577

43 | 000133445689

44 | 00012244467788

45 | 334467778899

46 | 3445556667788889

47 | 0022344456677888

48 | 00000133667889

49 | 02223444466889

50 | 001112356677888899

51 | 0000111233344455779

52 | 111112223334455668889

53 | 000022456677889

54 | 00011223334445555778889999

55 | 00000003345678899

56 | 112233467778888999

57 | 0000112233344444778

58 | 01113334677888899999

59 | 034446679

60 | 001223334556777888

61 | 000112333345689

62 | 000111123334555889

63 | 2223677777779

64 | 01122566788999

65 | 223455677889

66 | 0001122333444667799

67 | 122333456677889

68 | 0022235666689

69 | 00122446779

70 | 0114899

71 | 499

72 | 146

73 | 3345788

74 | 02233335

75 | 00014

76 |

77 | 222222

* 1. Approximately 36.6. The scores in the data set are given to two places after the decimal, but the stem and leaf plot reports only the first place after the decimal.
  2. One.
  3. No.
  4. Approximately symmetric.
  5. 55 and 63.7 occur 7 times.
  6. For general questions about the shape, level, and spread of a distribution, a stem-and-leaf plot provides a better visual summary of the distribution. However, when exact frequencies and variable values are desired, the frequency distribution table often provides a greater level of detail.
  7. The R command used to generate the graph below, looking at men, is **stem(Framingham$SYSBP1[Framingham$SEX=="Men"], scale = 2)**

9 | 89

10 | 012

10 | 5555778888899

11 | 00000111122222333344444

11 | 5555555666677888999999

12 | 00001111122223333444444

12 | 5556667777788888899

13 | 000000000111112223333333344

13 | 5567777788889999

14 | 00011112223444

14 | 566899

15 | 0023

15 | 56789999

16 | 01233

16 | 78

17 | 133

17 | 559

18 | 0023

18 |

19 | 1

19 | 8

20 |

20 |

21 | 4

The R command used to generate the graph below, looking at women, is **stem(Framingham$SYSBP1[Framingham$SEX=="Women"], scale = 2**

9 | 3

9 | 56788

10 | 0000002233334

10 | 55556778888888899

11 | 000001122233344444

11 | 555566666688888899

12 | 000000012223333344444

12 | 55666667777888888888999999

13 | 0000111222222333333444

13 | 556677899999

14 | 00000111224

14 | 7889

15 | 02334

15 | 5589

16 | 001

16 | 5558

17 | 033

17 | 678

18 | 123

18 | 68

19 | 23

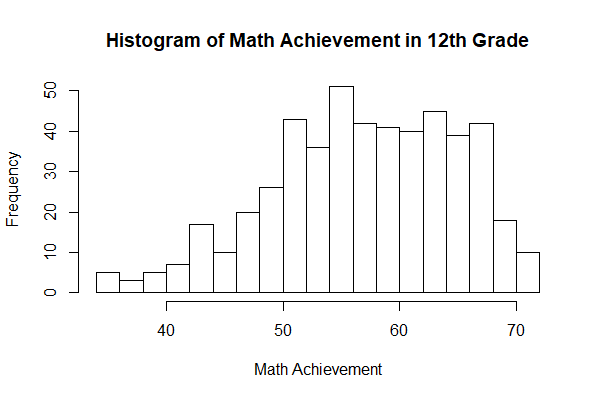
19 | 5

20 |

20 | 8

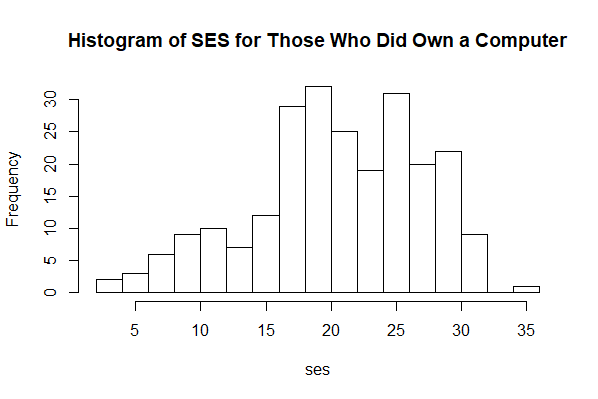
21 | 3

1. Female. For males, the lowest systolic blood pressure value is 98, while for females it is 93.
2. Male. In the female distribution, the most frequently occurring systolic blood pressure is 128, while in the male distribution, it is 130.
3. The distributions appear to be quite similar in spread.
4. The shapes of the distributions appear to be quite similar; they are both similarly positively skewed.
   1. The R command to generate this graph is **hist(NELS$achmat12, main = "Histogram of Math Achievement in 12th Grade", xlab = "Math Achievement", breaks = 20, xlim=c(30,75))**

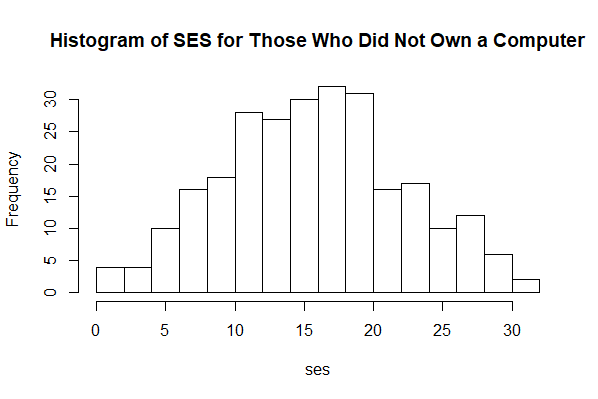


1. Estimating from the histogram, the lowest score is approximately 34 and it appears about 5 times.
2. Between 54 and 55.
3. It is slightly negatively skewed. There is a longer tail on the left than on the right.

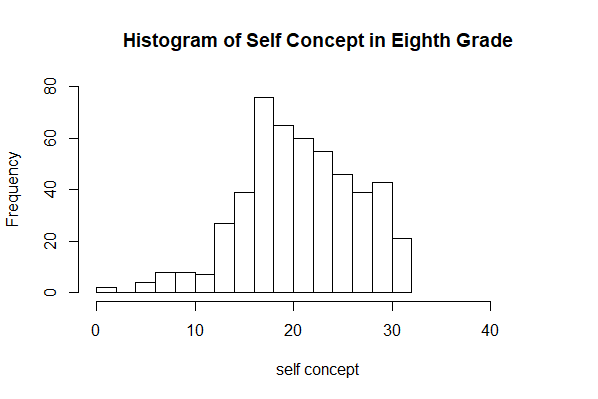
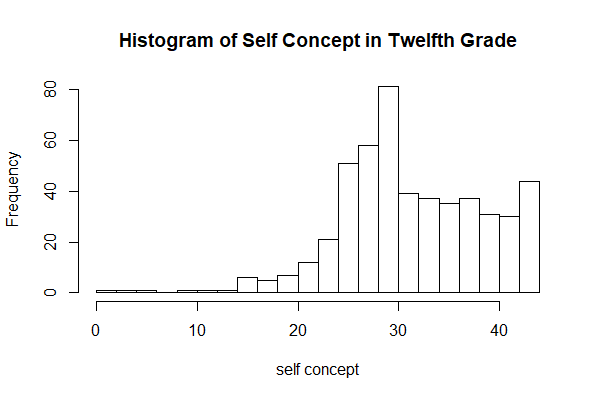
* 1. The R command used to generate the graph for those who did have a computer in eighth grade, below, is **hist(NELS$ses[NELS$computer=="Yes"], main = "Histogram of SES for Those Who Did Own a Computer", xlab = "ses", breaks = 20)**



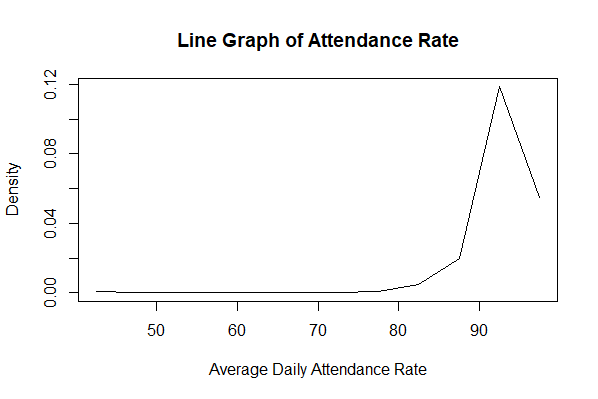
The R command used to generate the graph for those who did not have a computer in eighth grade, below, is **hist(NELS$ses[NELS$computer=="No"], main = "Histogram of SES for Those Who Did Own a Computer", xlab = "ses", breaks = 20**



1. The group that owned a computer in eighth grade.
2. Yes, it appears to be higher for the group that owns a computer as the distribution for this group begins and ends at a higher point along the ses scale and the bulk of the scores in that distribution is located higher up on the ses scale than for the other group.
3. No. The spread of the two distributions appears to be quite similar.
   1. The R code used to create the histogram for eighth graders is **hist(NELS$slfcnc08, main = "Histogram of Self Concept in Eighth Grade", xlab = "self concept", breaks = 20, xlim=c(0, 45), ylim = c(0,80))** and for twelfth graders is **hist(NELS$slfcnc12, main = "Histogram of Self Concept in Twelfth Grade", xlab = "self concept", breaks = 20)**. The **xlim** and **ylim** arguments set the limits of the x-axis and y-axis, respectively, and the **c** function is used to combine the lower and upper limit for each.

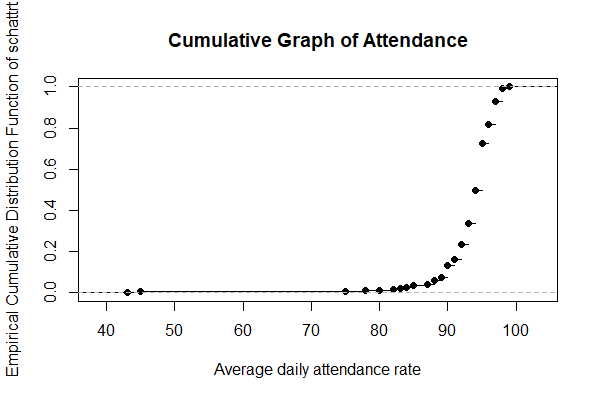


1. The level of self-concept is higher for the twelfth grade – the bulk of the scores in the histogram for twelfth grade is further to the right on the *x*-axis than for eighth grade.
2. In eighth grade the scores are more closely clustered.
3. The scores are more negatively skewed in twelfth grade – the tail is longer in the negative direction or to the left.
   1. The graph below is created using the R command **line.graph(NELS$schattrt, main = "Line Graph of Attendance Rate", xlab = "Average Daily Attendance Rate", ylab = "Density")**

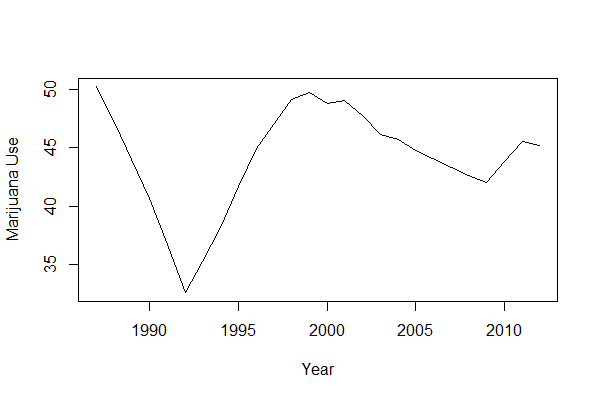


The distribution is negatively skewed.

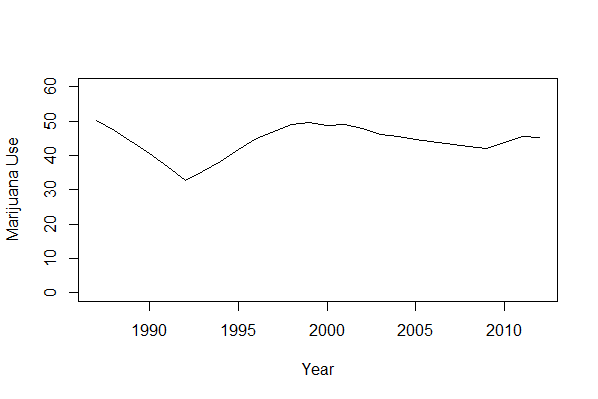
* 1. The R command to generate this graph is **plot(ecdf(NELS$schattrt), main = "Cumulative Graph of Attendance", xlab = "Average daily attendance rate", ylab = "Empirical Cumulative Distribution Function of schattrt")**



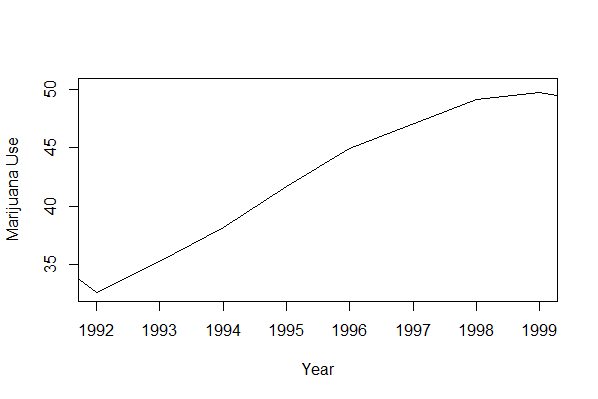
1. Approximately .15\*500 = 75.
2. Negatively skewed.
3. The R command to generate the graph is **plot(Marijuana$MarijuanaUse~Marijuana$Year, type = "l", xlab = "Year", ylab = "Marijuana Use")**



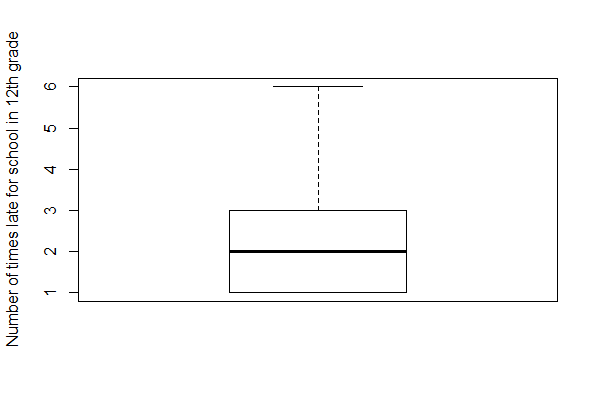
1. The compressed scale reduces the appearance of a change in the percentages of students who use marijuana over time. Said differently, student behavior with respect to smoking marijuana appears to be more consistent over time with the compressed scale.



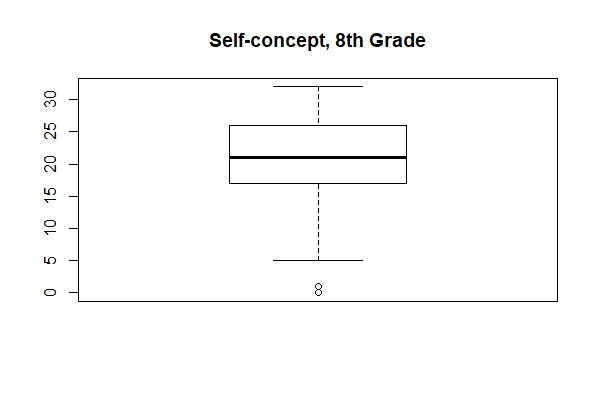
1. The percentage of seniors who have tried marijuana increases rather dramatically from 1992 through 1999.

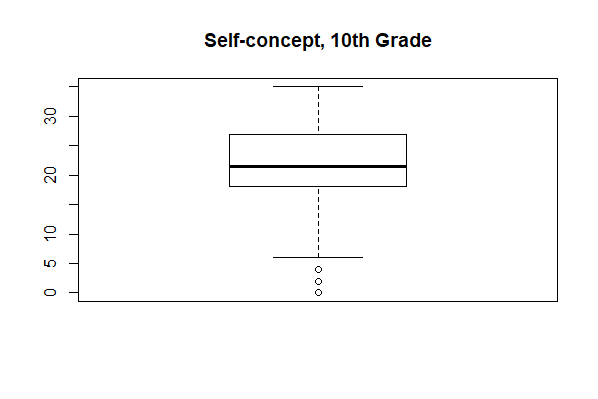


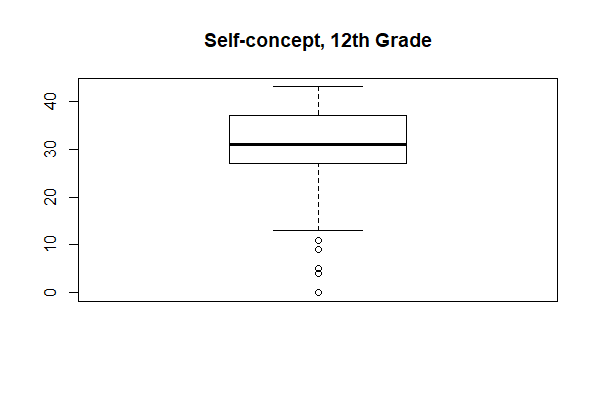
1. $30,000. This means that 15 percent of the students are expecting to be earning less than $30,000 at age 30.
2. $40,000.
3. 73.86.
   1. Because parmarl8 is only nominal. Variables must be at least ordinal-leveled for percentiles to be meaningfully calculated.
4. 102.
5. 25.
6. According to the frequency distribution table, the percentile rank of his eighth grade self-concept score is 74.8. While higher than almost three-fourths of the individuals in the data set, one might have expected this person’s self-concept to be even higher.
7. He was least self-confident in twelfth grade. In eighth grade, his self-concept score of 25, gave him a percentile rank of 74.8 percent. In tenth grade, his self-concept score of 32, gave him a percentile rank of 88 percent. In twelfth grade, his self-concept score of 22, gave him a percentile rank of 7.2 percent.
   1. The R command for generating the graph is **boxplot(as.numeric(NELS$late12), ylab = "Number of times late for school in 12th grade")**



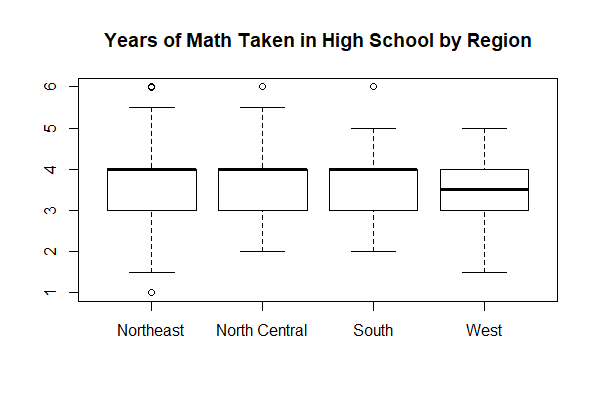
* 1. The value of the 50th percentile is 1. Since the value 1 corresponds to being late one or two times, we know that half the students in the NELS study were late to school no more than only one or two times.
  2. The value of the *IQR* is the box height, or 2 – 0 = 2.
  3. The graph is positively skewed because there is a long whisker above the box and no whisker at all below the box.
  4. The minimum score is the same as the 25th percentile score. That is, at least 25% of students in the NELS data set were never late in twelfth grade.
  5. The R command to generate these graphs are **boxplot(NELS$slfcnc08, main = "Self-concept, 8th Grade")**, **boxplot(NELS$slfcnc10, main = "Self-concept, 10th Grade")**, and **boxplot(NELS$slfcnc12, main = "Self-concept, 12th Grade")**.



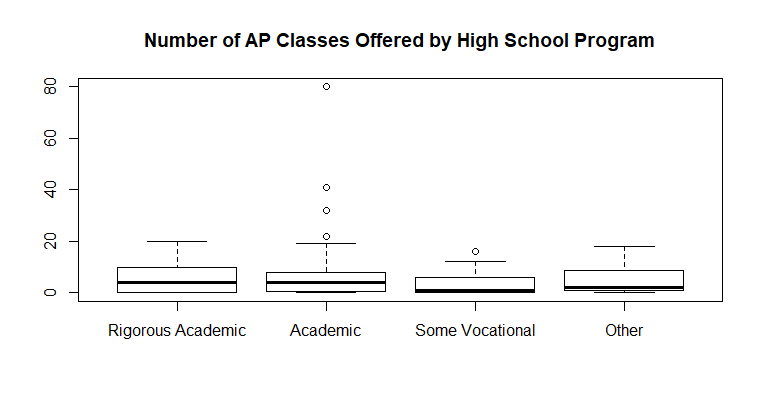




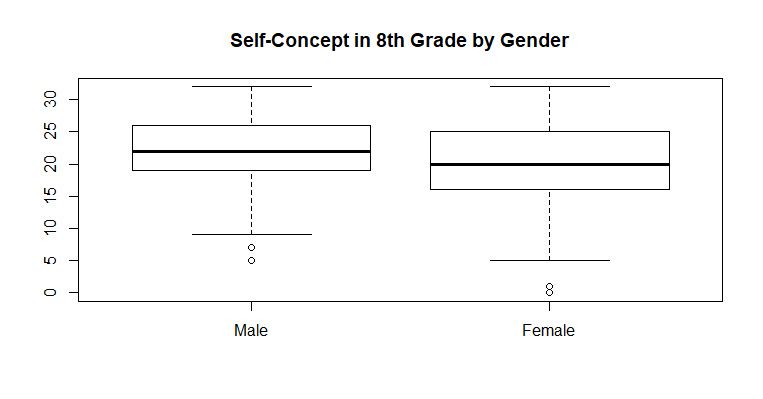
1. Negatively skewed. All of the outliers are relatively low scores in the distribution.
2. Eighth grade.
3. Twelfth grade as measured by the 50th percentile.
4. The twelfth grade distribution contains the highest self-concept score, which is approximately 43.
5. Twelfth grade.
   1. The R command to generate this graph is **boxplot(NELS$unitmath~NELS$region, main = "Years of Math Taken in High School by Region")**



1. West.
2. West.
3. There is a student in the Northeast that took only one year of math in high school.
4. The distributions of all regions appear to be similar in spread according to the interquartile range.
5. Because 25 percent of the scores in any distribution fall between *Q1* and *Q2,* the number ofstudents in the Northeast whose scores fall between these two quartiles will be .25x106 or approximately 27 for those in the Northeast regardless of the variable in question.
   1. The R command to generate this graph is **boxplot(NELS$apoffer~NELS$hsprog, main = "Number of AP Classes Offered by High School Program")**



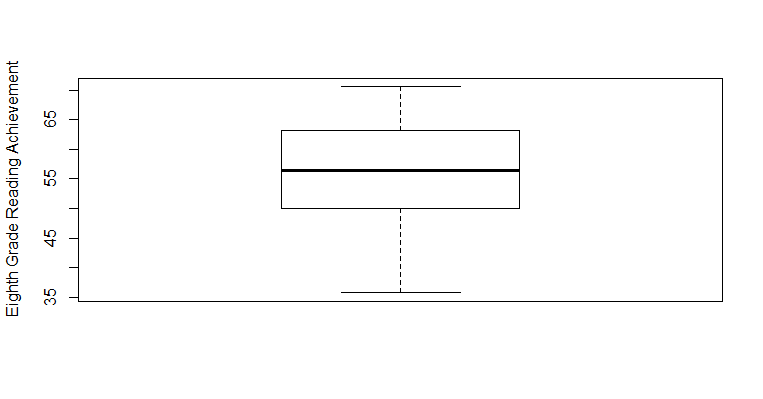
1. Schools that offer only rigorous high school academic programs are likely to be smaller in scale (e.g., they may be non-public schools) than those that offer academic programs, which could explain the relatively smaller number of AP course offerings overall.
2. By using the percentage of all course offerings per school.
   1. The R command used to generate these graphs are **boxplot(NELS$slfcnc08~NELS$gender, main = "Self-Concept in 8th Grade by Gender")**, **boxplot(NELS$slfcnc10~NELS$gender, main = "Self-Concept in 10th Grade by Gender")** and **boxplot(NELS$slfcnc12~NELS$gender, main = "Self-Concept in 12th Grade by Gender")**.



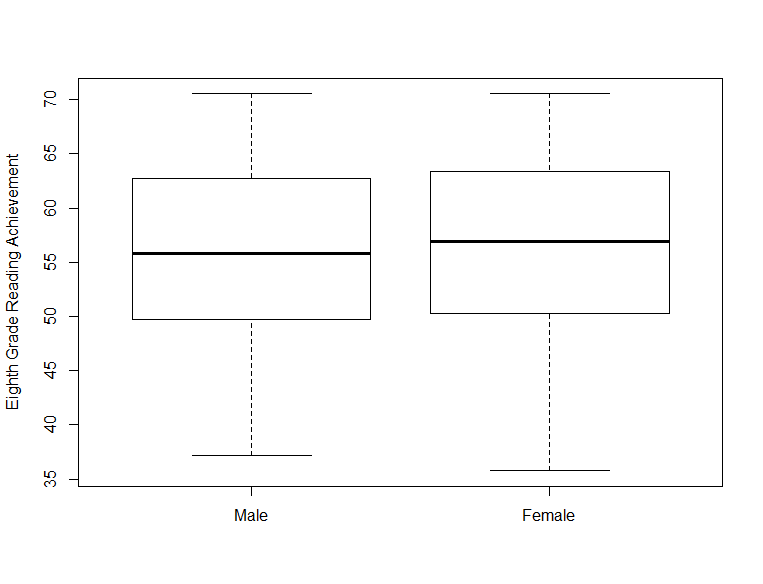




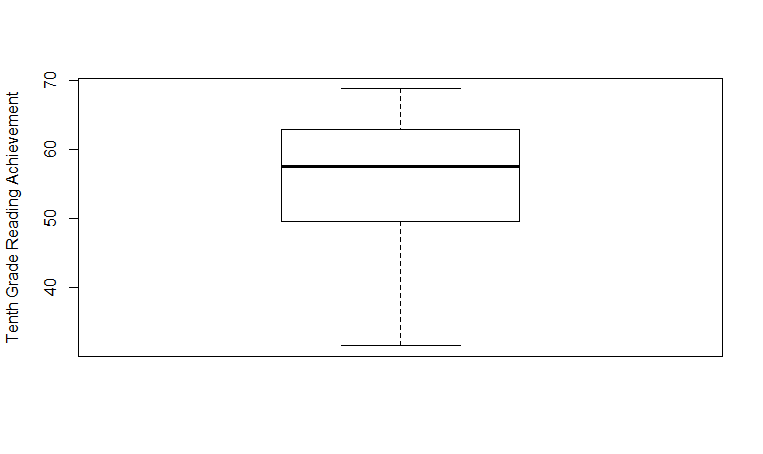
1. Yes.
2. The large presence of extremely low self-concept scores for females as compared to males.
3. Yes. Box and whiskers are themselves comparable for males and females.
4. **boxplot(NELS$achrdg08, ylab = "Eighth Grade Reading Achievement")**



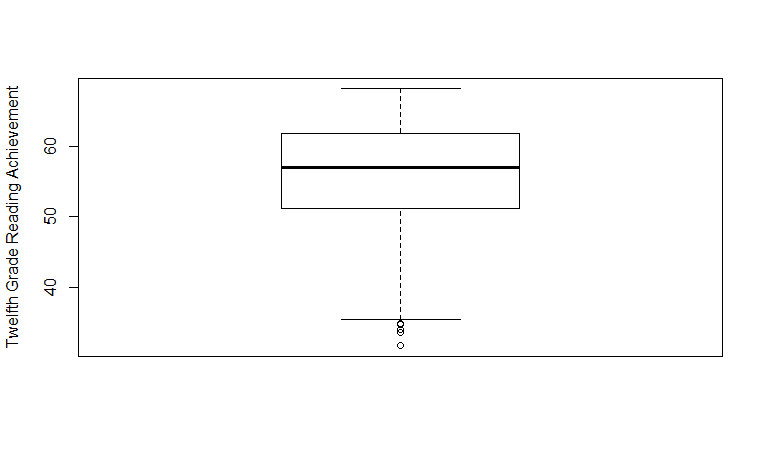
1. **boxplot(NELS$achrdg08~NELS$gender, ylab = "Eighth Grade Reading Achievement")**



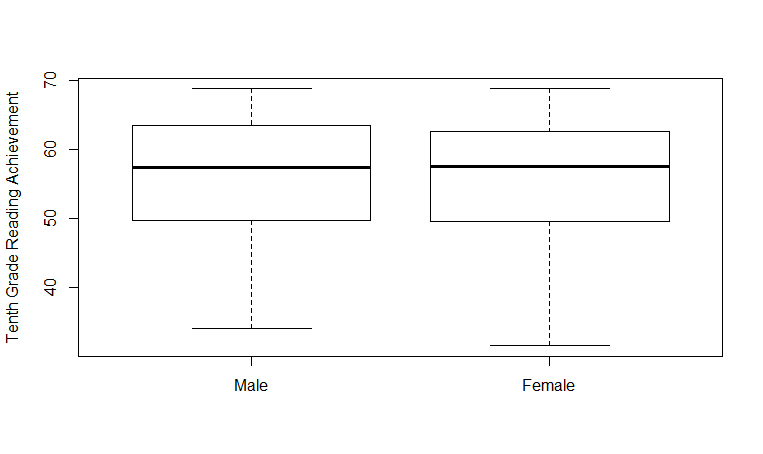
1. **boxplot(NELS$achrdg10, ylab = "Tenth Grade Reading Achievement")**



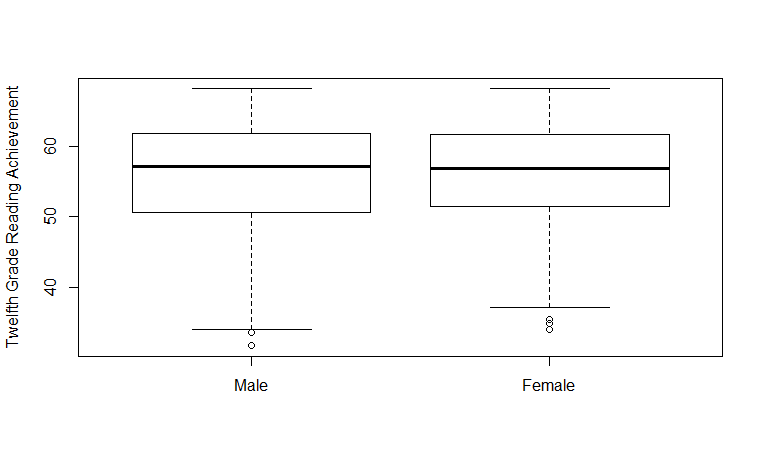
**boxplot(NELS$achrdg12, ylab = "Twelfth Grade Reading Achievement")**



1. **boxplot(NELS$achrdg10~NELS$gender, ylab = "Tenth Grade Reading Achievement")**



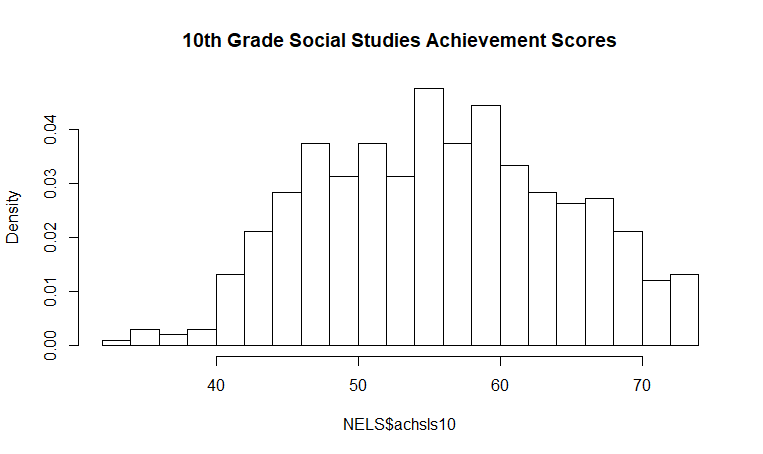
**boxplot(NELS$achrdg12~NELS$gender, ylab = "Twelfth Grade Reading Achievement")**



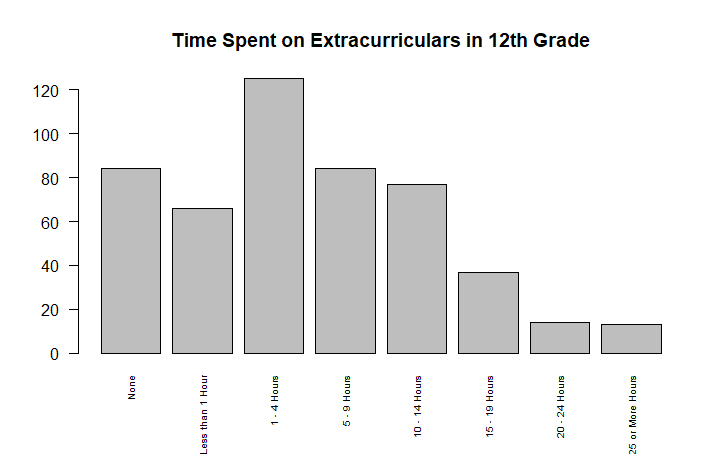


1. Positively skewed.
2. Negatively skewed.
3. Slightly positively skewed.
4. Positively skewed.
5. Positively skewed.
6. Positively skewed.

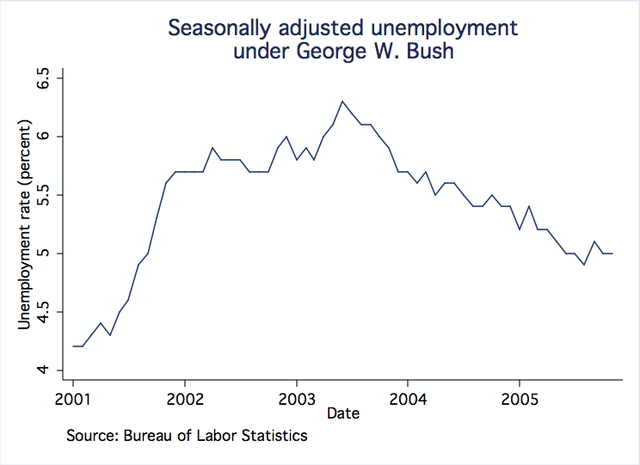
1. South
2. Northeast
3. West
4. Midwest
5. (1) The variable is nominal.
6. (1) The variable is ordinal with only three categories.
7. (1) The variable is nominal.
8. (1) or (2) The variable is interval, so (2) is appropriate, but it takes on only four values, so (1) is also appropriate.
9. (1) or (2) The variable is ratio, so (2) is appropriate, but there are only four categories, so (1) is also appropriate.
10. (2) The variable is interval and takes on many values.
11. (2) The variable is ratio and takes on many values.
12. (1) or (2) The variable is ordinal, so (2) is appropriate, but it takes on only six values, so (1) is also appropriate.
13. 72.89.
14. 13 students scored 72.89.
15. 72.89.
16. 55.7.
17. 59.9.
18. 94.9 percent.
19. According to the histogram, the shape of the distribution is reasonably symmetric. The R command to generate the histogram is **hist(NELS$achsls10, freq = F, main = "10th Grade Social Studies Achievement Scores", breaks = 20)**



1. 125.
2. The numeric value is 3, which corresponds to the third category of 1 – 4 hours.
3. Slightly positively skewed. The bulk of the students spend less than 20 hours per week on extracurricular activities, but a few students spend quite a bit more than that. The R code for the plot below is **barplot(table(NELS$excurr12), main = "Time Spent on Extracurriculars in 12th Grade", cex.names = .65, las=2)**.



1. May 2003. It is highlighted because it is when President Bush signed the Jobs and Growth Act.
2. 130,000,000 and 5.7 percent.
3. August 2005.
4. It is not fair to conclude that. Among the most serious flaws with the graph is the restriction of the time range. It depicts fewer than two of the eight years President Bush was in office. In a graph created by Brendan Nyhan, we see that a longer time frame shows that President Bush’s term began with an unemployment rate that was lower than it was even after the Jobs and Growth Act took effect.



Even though there is only a relatively small difference between the velocities, the graphic accentuates the difference because the vertical axis does not start at 0, one bar is about twice the height of the other, making it appear that Dickey’s knuckleball was twice as fast in 2012 than it was in 2013. When the graph is edited so that the vertical axis starts at 0, the comparison is more accurate.



1. 31% of the students in one of the top six most in-demand college majors majored in business.
2. The pie graph implies that the percentages are out of a meaningful whole. Instead, a bar graph of counts could give us a sense of the relative sizes of these majors without implying that there are not any others.
3. 10.5%.
4. Assuming that recent graduates and experienced graduates can be combined to represent all graduates in a certain time period it makes sense for this bar graph to be stacked.
   1. *C4 D1  Q1  D3  C50  Q3*
   2. The statement is wrong. For example, if everyone scores 10, then a raw score of 75 is not surpassed by any of the scores in the distribution. A raw score does not, in general, give information about percentages.
   3. The statement is wrong. For example, if 100 people take the test, then the highest percentile rank possible is 99, that is, the highest score will have 99 percent of the scores falling below it, unless there is a tie.
   4. The statement is wrong. This depends on the percentage of people below Alice’s and Ellen’s scores. For example, if there were 10 people taking the exam and eight people scored 40, Alice scored 50, and Ellen scored 100, then the percentile rank of Alice’s score is 80, while the percentile rank of Ellen’s score is 90, not twice that of Alice’s score. Of course, a percentile rank can never be 100 or above.
   5. The statement is wrong. Percentiles are not necessarily interval.
   6. The statement is wrong. If the students at this student’s school are exceptionally good in math, having a percentile rank of 85 in math relative to this school could be better than having a percentile rank of 95 in science relative to the entire city.
   7. The statement is wrong. The validity of the conclusion depends on the number of people questioned. Whereas 5 out of 5 is 100 percent, 5 out of 100 is only 5 percent. A description in terms of relative frequency or percentage would probably be more appropriate in this situation than a description in terms of frequency.
   8. They are all possible values for percentiles. Only 50 is a possible value for a percentile rank.
   9. A distribution with outliers that are balanced in both the positive and negative directions can be symmetric.
   10. A boxplot would have only one whisker, for example, if the 75th percentile was the same as the maximum score, that is, if the top 25 percent of the scores in the distribution were all the same value.



