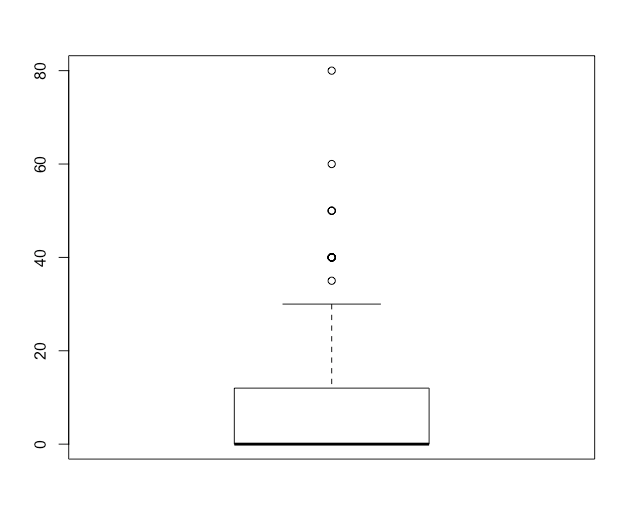
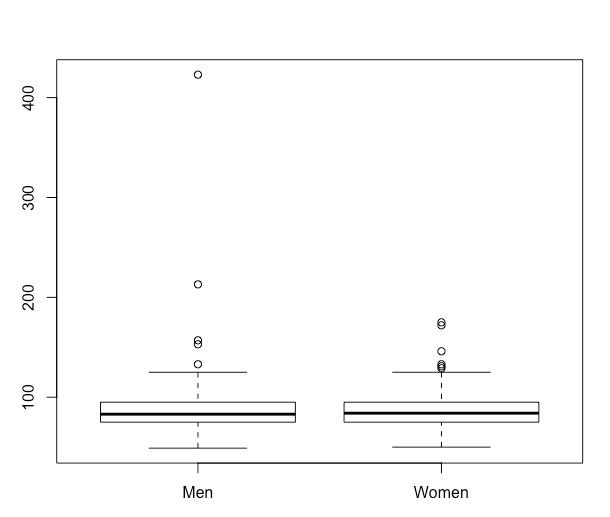
**CHAPTER 3 SOLUTIONS**

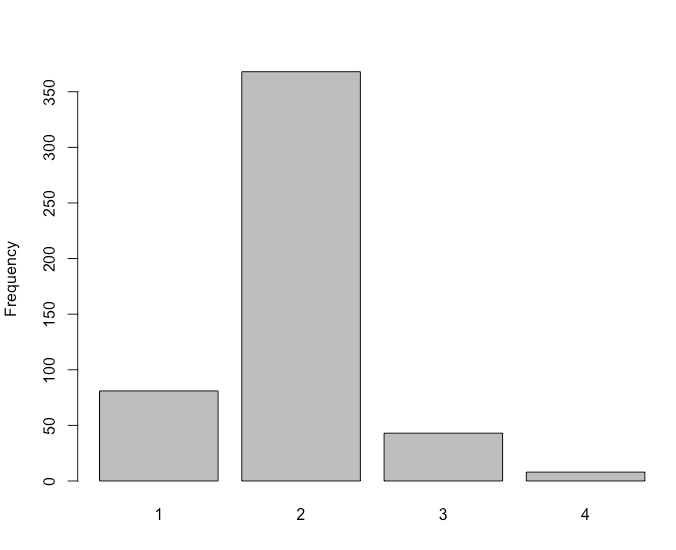
1. It is appropriate to calculate the mode, median, and mean of gender. After recoding gender by subtracting 1, the coding used is Male = 0 and Female = 1. The mode, equal to 1, indicates that most students are female. The median, equal to 1, indicates that category 1 contains the 250th score (recall *N* = 500 in our NELS dataset), which suggests, given that there are only two categories for this dichotomous variable, that at least 50% of the individuals in this dataset are female. The mean, equal to .55, indicates, for the coding used, that 55% of the individuals in this dataset are female.
2. If urban were considered to be an ordinal variable, it would be appropriate to calculate the mode and median of urban. If treated as a nominal variable, only the mode would be appropriate. The median, equal to 2, indicates that a typical student lives in a setting that may be described as having an intermediate level of urbanicity: a suburban setting. The mode, equal to 2, indicates that most of the individuals in this dataset live in suburban settings, as opposed to rural or urban settings.
3. It is appropriate to calculate only the mode of schtyp8, a nominal variable. The mode, equal to 1, indicates that most individuals in this dataset attend public school.
4. If tcherint, measured on a Likert scale, is considered to be interval-leveled, it would be appropriate to calculate the mode, median, and mean of tcherint. If considered only to be ordinal, then only the mode and median would be appropriate. The mean, equal to 1.96, and the median and mode equal to 2, all indicate that, on average, students agreed with the statement, “My teachers are interested in me.” Here is an example where all three measures of central tendency are equal or reasonably so, yet the distribution itself is not symmetric. The mode provides meaningful information in this case because tcherint, while at least ordinal, has only four categories.
5. It is appropriate to calculate the mode, median, and mean of numinst, a ratio-leveled variable. The mode, equal to 1, indicates that most students attended one post-secondary institution. The median, also equal to 1, indicates that the typical student attended only one post-secondary institution. The mean, equal to 1.21, indicates that, on average, students attended 1.21 post-secondary institutions. With more than 80% of the respondents attending one post-secondary institution, and with only one respondent attending four post-secondary institutions, the maximum number recorded, the distribution is skewed in the positive direction. Unlike the mode and median, the mean is influenced more strongly by the few people who have attended three or four institutions.
6. While it is possible to calculate the mode, median, and mean of achrdg08, an interval-leveled variable, given the continuous nature of this variable and the fact that it is reasonably symmetric, the mean would be the most appropriate measure of central tendency in this case. The mean, equal to 56.05, indicates that, on average, students scored 56.05 on reading achievement in eighth grade.
7. While it is possible to calculate the mode, median, and mean of schattrt, a ratio-leveled variable, given that schattrt is many-valued, the median and mean would be more appropriate than the mode for assessing the location of this continuous variable. The median, equal to 95, indicates that the typical student attended schools with a daily attendance rate of 95. The mean, equal to 93.65, indicates that, on average, students attended schools with a daily attendance rate of 93.65. While the mean and median are not too dissimilar in value, relative to the median, the mean has been pulled in the direction of the tail of this negatively skewed variable.
8. It is appropriate to calculate the mode and median of absent12, an ordinal variable with only six categories. Given the coding used (numbers 1-6 for the six categories), the mode, equal to 2, indicates that most students were absent from twelfth grade 1 or 2 times; and, the median, equal to 2.5, indicates that the typical student was absent from twelfth grade somewhere between 1 and 6 times.
9. 129. The R command for generating the frequency distribution table is **table(NELS$late12)**.
10. 7 to 9 times. Select the first observation using brackets: **NELS$late12[1]**, or inspect the dataset using **View(NELS)**.
11. One or two times, which is the value of both the mode and median, the appropriate measures to use with this ordinal variable. From **table(NELS$late12)**, or using **the.mode(NELS$late12)**, we see from the frequency distribution table that the mode is 2: “1-2 Times.” From **median(as.numeric(NELS$late12)**, we see the median is also 2, which represents one or two times.
12. (1)
13. (1) or (2)
14. (1)
15. (3)
16. (3)
17. (4)
18. (3)
19. (2)
20. The proportion that ever smoked is .14. The R command for generating the percentage distribution table is **percent.table(NELS$cigarett)**.
21. The mean is .14, which equals the proportion found in part (a). The R commands for recoding and calculating the mean are:   
    **NELS$re\_cig = as.numeric(NELS$cigarett) – 1  
    mean(NELS$re\_cig)**
22. A greater proportion of females (.16) report that they have ever smoked than males (.12). The R commands for generating separate percentage distribution tables for males and females are:  
    **percent.table(NELS$cigarett[NELS$gender=="Male"])  
    ../../../Desktop/Screen%20Shot%202019-06-11%20at%206.35.37%20P  
    percent.table(NELS$cigarett[NELS$gender=="Female"])  
    ../../../Desktop/Screen%20Shot%202019-06-11%20at%206.35.46%20P**
23. According to the boxplot, the distribution is severely positively skewed with multiple outliers and one particular case smoking an unusually high 80 cigarettes per day. The person who smokes the next-largest number of cigarettes smokes only 60 cigarettes per day. Further, the minimum, 25th percentile, and 50th percentile all lie at 0, indicating that at least half of the people in the Framingham dataset do not smoke at all. According to these results, it is reasonable to say that the typical person in the dataset smokes zero cigarettes per day. The R command for generating the boxplot is **boxplot(Framingham$CIGPDAY3)**.



1. The value of the mean is 7.23 whereas the value for the median is 0. The R command used to obtain these statistics is **summary(Framingham$CIGPDAY3)**. Note that if you choose to use the **mean** and **median** functions here instead, since CIGPDAY3 has missing values, these commands must include an extra argument:  
   **mean(Framingham$CIGPDAY3, na.rm = T)  
   median(Framingham$CIGPDAY3, na.rm = T)**
2. The median gives a better approximation of a typical value for the distribution because the median is robust to the presence of the outliers while the mean is much larger because it is pulled in the direction of the outliers.
3. While both distributions appear positively skewed with some positive outliers, the male distribution has a particularly more extreme outlier than the female distribution has, suggesting that the male distribution is more positively skewed than the female distribution. The extreme case in the male distribution has a glucose level above 400, far higher than any other person in the dataset. The R command for generating the boxplot is **boxplot(Framingham$GLUCOSE3~Framingham$SEX)**.

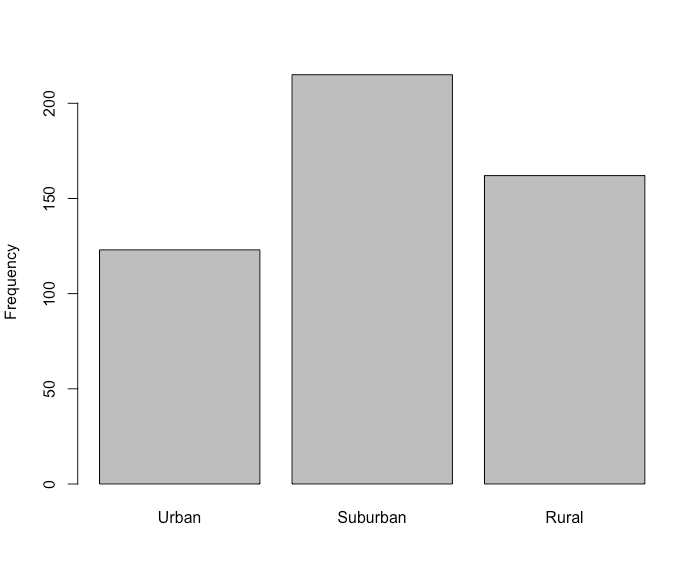


1. Because the male distribution is far more positively skewed than the female distribution, a comparison of medians for the two distributions would be more appropriate than a comparison of means in order to control for this difference in skewness. According to the medians, we conclude that women (median = 84) typically have slightly higher glucose levels than men (median = 83). However, if we chose to weight the presence of the extreme values more heavily in our analysis, we would base a comparison of glucose levels on the means. In this case, we would conclude that men (mean = 90.48), on average, have higher glucose levels than women (mean = 86.85). The R commands for generating these statistics, which require an added argument to handle missing values, are as follows:  **median(Framingham$GLUCOSE3[Framingham$SEX=="Men"],na.rm=T)  
   median(Framingham$GLUCOSE3[Framingham$SEX=="Women"],na.rm=T)  
   mean(Framingham$GLUCOSE3[Framingham$SEX=="Men"],na.rm=T)  
   mean(Framingham$GLUCOSE3[Framingham$SEX=="Women"],na.rm=T)**
2. The distribution is positively skewed. The R command for generating the bar plot is **barplot(table(as.numeric(NELS$tcherint)), ylab = "Frequency")**.

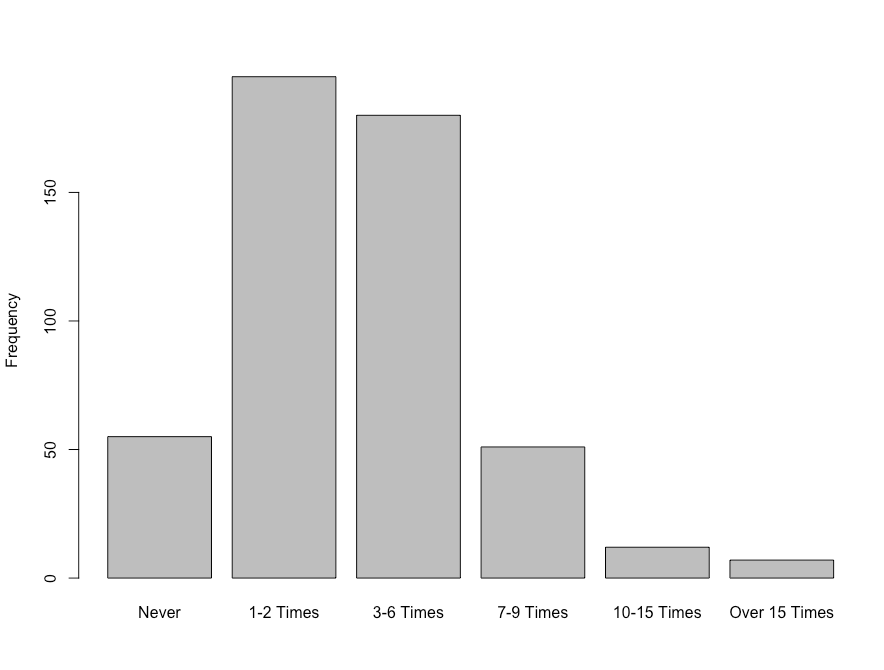


.

1. Contrary to the general rule of thumb, as it pertains particularly to the distribution of a many-valued unimodal continuous variable, in this example, the median of 2 is larger than the mean of 1.96. The R commands for generating these statistics are: **mean(as.numeric(NELS$tcherint))  
   median(as.numeric(NELS$tcherint))**
   1. In each case, the relevant summary statistics were obtained from R using the **percent.table**, **IQR,** and/or **sd** commands for measures of variability. In the cases where the skewness ratio was an appropriate calculation, we used the R commands **skew**, **se.skew**, and **skew.ratio**.
2. As a dichotomous variable, the variability of gender is best captured by the proportion of males and females in the distribution. In the NELS dataset, there are 45.4% males and 54.6% females. The greatest variability (heterogeneity) occurs when there are half males and half females. The distribution would be symmetric in this case. The greater the departure from an even split, the more skewed the distribution may be said to be. Because the mean of a dichotomous variable has meaning when it is coded with a 0 and a 1 (as the proportion of that group coded 1), the standard deviation is also often used for measuring the variability of a dichotomous variable. The standard deviation of gender, after recoding to 0 and 1, is .50. For completeness, we can compute the skewness ratio for gender as a measure of the extent to which this variable is skewed. Given that the percentage of females and males are each close to 50%, we would not expect the skewness ratio, in this case, to be large in absolute value. The skewness ratio is -1.70 (-.185/.109), indicating, as expected, a lack of skewness for this variable.
3. As only a three-valued variable (rural, suburban, urban), the variability of urban is also best captured by the proportion of cases within each of the three categories. In the NELS dataset, 24.6% of the students come from an urban area, 43% come from a suburban area, and 32.4% come from a rural area. Greatest variability (or heterogeneity) occurs when there are an equal proportion of individuals in each category (in this case, 33.3% in each). Greatest homogeneity occurs when 100% of the cases are in only one category. Given the small number of possible values for this variable, the IQR would likely not be useful, even though this variable may be considered to be ordinal. The following frequency bar graph depicts well the spread and shape of this variable. According to this bar graph, the distribution is fairly symmetric, with approximately equal numbers of students in the urban and rural categories and most students in the suburban category. The extent of the skew cannot be quantified with the skewness ratio because it does not make sense to talk about distance from the mean for an ordinal-leveled variable. The R command for generating the bar graph is **barplot(table(NELS$urban), ylab = "Frequency")**.



1. Numinst, like the previous two variables in this exercise, takes on only a small number of values (four to be exact), and so much of what has been written in connection with the solutions for these other variables may be said in relation to this one. In particular, one may note that with more than 80% of the respondents attending one post-secondary institution, and with only one respondent attending four post-secondary institutions, the maximum number recorded, the distribution may be said to be skewed in the positive direction. Given the small number of values for this variable, the IQR would add little to measuring spread over and above a simple accounting of the proportion of responses to each of the four values noted. The standard deviation could be computed as this variable is ratio-leveled, but given how skewed the variable is, the standard deviation may not be as useful as an accounting of the simple proportions.
2. Because achrdg08 is interval, many-valued, and reasonably symmetric, the standard deviation (or variance) offers an appropriate measure of spread. The value of the standard deviation is 8.83, indicating that students typically scored 8.83 points away from the mean of 56.05. The IQR would also be useful in documenting the range within which the middle 50% scored.
3. Because schattrt is ratio, many-valued, and severely negatively skewed with skewness ratio equal to -51.97(-6.21/0.12 = -51.97), the IQR, as opposed to the standard deviation, may be the most appropriate measure of spread. The value of the IQR is 3, indicating that the range of school attendance rate values for the middle 50% of the data is 3. Note that since there are missing values of schattrt, the command for the IQR must include an extra argument: **IQR(NELS$schattrt, na.rm = T)**.
4. Because absent12 is ordinal with a small number of discrete values, a bar graph may be used to show that the distribution is positively skewed, with most students being absent in 12th grade fewer than 7 times, and only a few students absent more than 20 times. The extent of the skew cannot be quantified with say, a skewness statistic, because it does not make sense to talk about distance from the mean for an ordinal leveled variable. The IQR may be used to measure the spread of this variable. The value of the IQR is 1, indicating that the range of the middle 50% of the data is only 1, that absences are highly clustered around the middle. The R command for generating the bar graph is **barplot(table(NELS$absent12), ylab = "Frequency")**.



* 1. The R command **summary(NELS$slfcnc08)** provides the following output and answers to parts (a), (b), (d), (e), (f), and (h) by calculating the maximum minus the minimum. Alternatively, the R commands **min**, **max**, **quantile**, **mean**, **median**, and **range** can be used here as well.  
       
     ../../../Desktop/Screen%20Shot%202019-06-14%20at%202.45.15%20P  
       
     The following R commands produce the remaining statistics:  
     **quantile(NELS$slfcnc08,c(.40))** produces the 40th percentile;  
     **the.mode(NELS$slfcnc08)** produces the mode;  
     **IQR(NELS$slfcnc08)** produces the IQR;  
     **var(NELS$slfcnc08)** produces the variance;  
     **sd(NELS$slfcnc08)** produces the standard deviation;  
     **skew(NELS$slfcnc08)** produces the skewness; and  
     **se.skew(NELS$slfcnc08)** produces the standard error of skewness.  
       
     Solutions to each part are listed below.

1. 0
2. 32
3. 19
4. 17
5. 21.06
6. 21
7. 17
8. 32
9. 9
10. 35.65
11. 5.97
12. -.286
13. .109  
    1. The R commands.
14. The boxplots indicate that both distributions are similarly severely negatively skewed. The skewness ratios for those whose families did not own a computer (-2.21) and for those whose families did ( -2.32) corroborate this visual impression.
15. Given the skewness of the two distributions, medians are used to answer this question. According to the respective medians, tenth-grade math achievement is higher for students whose families owned a computer when they were in eighth grade (median = 59.01) than for those whose families did not (median = 55.76). Because the distributions have similar shapes, the respective means lead to the same conclusion.
16. According to the interquartile range, tenth-grade math achievement is slightly more variable for students whose families owned a computer when they were in eighth grade (IQR = 11.7) than for those whose families did not (IQR = 10.3). Because the standard deviation is even more sensitive to outliers than the mean, the results from the respective standard deviations (for those whose families owned computers, *SD* = 7.64; for those whose families did not, *SD* = 7.69) lead to a different conclusion. Like the interquartile ranges, however, the standard deviations are close in value.
    1. To obtain the relevant statistics, we use the R commands
17. The skewness ratios for those from an urban, suburban, and rural environment are, respectively, -1.66, -.67, and -.06. We see that although all of the distributions have a degree of negative skew, none of the distributions are severely skewed.
18. Among students in the NELS dataset, according to both the mean and the median, the typical socio-economic status of those from an urban environment (mean = 20.33, median = 20.00) is higher than that of those from a suburban environment (mean = 19.23, median = 19.00), which in turn is higher than that of those from a rural environment (mean = 15.94, median = 16.00).
19. Among students in the NELS dataset, according to both the standard deviation and IQR, the socio-economic status of those from an urban environment is more dispersed (*SD* = 6.92, IQR = 10) than that of those from a suburban environment (*SD* = 6.73, IQR = 9.5), which in turn is more dispersed than that of those from a rural environment (*SD* = 6.50, IQR = 9).
20. The person with the highest socio-economic score lives in a suburban environment. The maximum socio-economic status for the students from a suburban environment is 35, while for urban and suburban it is 32.
    1. To obtain the relevant statistics, we use the R commands **skew.ratio(NELS$slfcnc08)  
       skew.ratio(NELS$slfcnc12)  
       mean(NELS$slfcnc08)  
       mean(NELS$slfcnc12)  
       median(NELS$slfcnc08)  
       median(NELS$slfcnc12)  
       sd(NELS$slfcnc08)  
       sd(NELS$slfcnc12)  
       IQR(NELS$slfcnc08)  
       IQR(NELS$slfcnc12)**
       1. The distribution of self-concept scores is severely negatively skewed in both eighth grade (skewness ratio = -2.61) and twelfth grade (skewness ratio = -3.50). The distribution is more severely skewed in twelfth grade.
       2. According to both the mean and the median, students in the NELS dataset typically have higher self-concept scores in twelfth grade (mean = 31.48, median = 31.00) than in eighth grade (mean = 21.06, median = 21.00).
       3. According to both the standard deviation and the interquartile range, the self-concept scores of students in the NELS dataset are more heterogeneous in twelfth grade (*SD* = 7.23, IQR = 10.00) than in eighth grade (*SD* = 5.97, IQR = 9.00).
21. Both the modes and medians for both males and females are identical at 2 (“Agree”). Therefore, males and females perceive their teachers to show similar amounts of interest in them. The R commands used to obtain the statistics are:

**the.mode(as.numeric(NELS$tcherint[NELS$gender=="Male"]))**

**the.mode(as.numeric(NELS$tcherint[NELS$gender=="Female"]))**

**median(as.numeric(NELS$tcherint[NELS$gender=="Male"]))**

**median(as.numeric(NELS$tcherint[NELS$gender=="Female"]))**

1. According to changes in the medians, males and females display similar changes in achievement in reading across eighth, tenth, and twelfth grades. There is an increase from eighth to tenth and then a slight decrease from tenth to twelfth. The means analysis shows a similar result that of the medians. The R commands used to obtain the statistics are:

**summary(NELS$achrdg08[NELS$gender=="Male"])**

**summary(NELS$achrdg10[NELS$gender=="Male"])**

**summary(NELS$achrdg12[NELS$gender=="Male"])**

**summary(NELS$achrdg08[NELS$gender=="Female"])**

**summary(NELS$achrdg10[NELS$gender=="Female"])**

**summary(NELS$achrdg12[NELS$gender=="Female"])**

1. According to the descriptive statistics, the distributions for males and females are quite similar, in general, across all measures of school attendance behavior. Differences may be noted in the following areas: The typical female reports being absent 3 to 6 times over the year (coded 3), while the typical male reports being absent only 1 to 2 times over the year (coded 2); also, females tend to be more variable than males with respect to being late, as noted by the difference in their IQR’s on late12. The R commands used to obtain the statistics are:

**median(as.numeric(NELS$absent12[NELS$gender=="Male"]))**

**median(as.numeric(NELS$absent12[NELS$gender=="Female"]))**

**median(as.numeric(NELS$late12[NELS$gender=="Male"]))**

**median(as.numeric(NELS$late12[NELS$gender=="Female"]))**

**median(as.numeric(NELS$cuts12[NELS$gender=="Male"]))**

**median(as.numeric(NELS$cuts12[NELS$gender=="Female"]))**

**IQR(as.numeric(NELS$absent12[NELS$gender=="Male"]))**

**IQR(as.numeric(NELS$absent12[NELS$gender=="Female"]))**

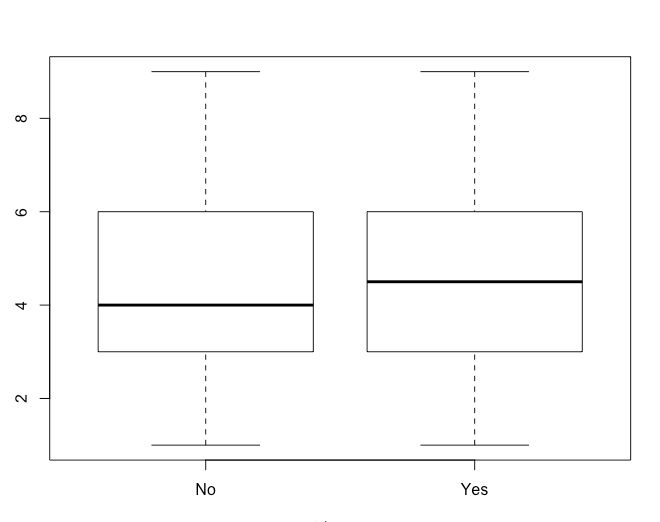
**IQR(as.numeric(NELS$late12[NELS$gender=="Male"]))**

**IQR(as.numeric(NELS$late12[NELS$gender=="Female"]))**

**IQR(as.numeric(NELS$cuts12[NELS$gender=="Male"]))**

**IQR(as.numeric(NELS$cuts12[NELS$gender=="Female"]))**

1. According to these similarly shaped boxplots, the median time spent on homework outside of school for those who did take advanced math in eighth grade is higher than for those who did not. Follow-up analyses indicate that the medians are 4.5 and 4.0 for these two groups, respectively.



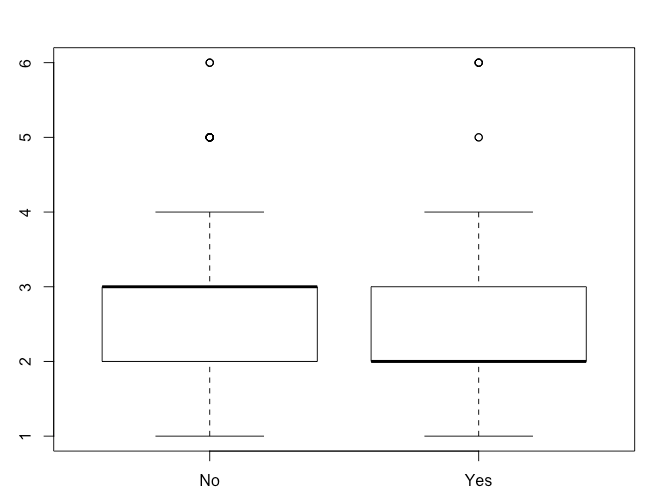
The R commands used to obtain the graph and statistics are

**boxplot(as.numeric(NELS$hwkout12)~NELS$advmath8)**

**median(as.numeric(NELS$hwkout12[NELS$advmath8=="Yes"]),na.rm=T)**

**median(as.numeric(NELS$hwkout12[NELS$advmath8=="No"]),na.rm=T)**

1. According to these similarly shaped boxplots, the median number of times absent in twelfth grade for those who did take advanced math in eighth grade is lower than for those who did not. Follow-up analyses indicate that the medians are 2.00 (1 to 2 times absent) and 3.00 (3 to 6 times absent) for these two groups, respectively.



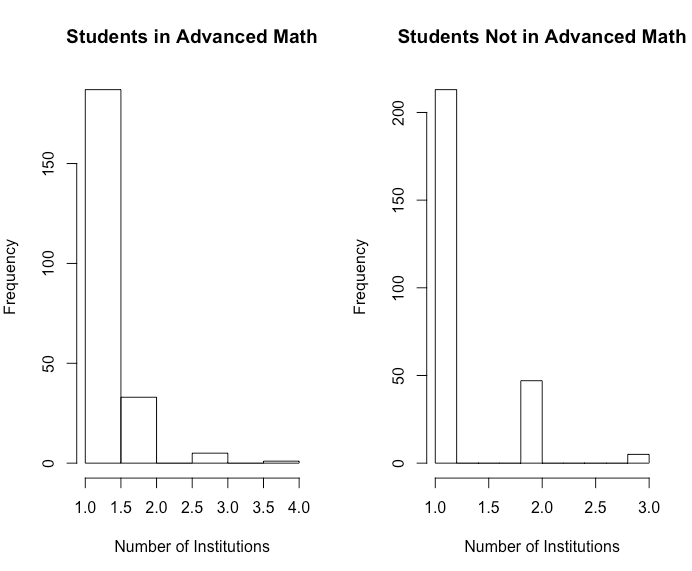
The R commands used to obtain the graph and statistics are

**boxplot(as.numeric(NELS$absent12)~NELS$advmath8)**

**median(as.numeric(NELS$absent12[NELS$advmath8=="Yes"]),na.rm=T)**

**median(as.numeric(NELS$absent12[NELS$advmath8=="No"]),na.rm=T)**

1. There is little difference in the shape of these two distributions, one for those who did and the other for those who did not take advanced math in eighth grade, although both distributions are severely positively skewed. The median and interquartile range for those who did take advanced math in eighth grade are, respectively, 1 and 0 and for those who did not take advanced math in eighth grade, the median and interquartile range are, respectively, also 1 and 0.



The R commands used to obtain the graph and statistics are

**par(mfrow=c(1,2))**

**hist(NELS$numinst[NELS$advmath8=="Yes"], main = "Students in Advanced Math", xlab = "Number of Institutions")**

**hist(NELS$numinst[NELS$advmath8=="No"], main = "Students Not in Advanced Math", xlab = "Number of Institutions")**

**par(mfrow=c(1,1))**

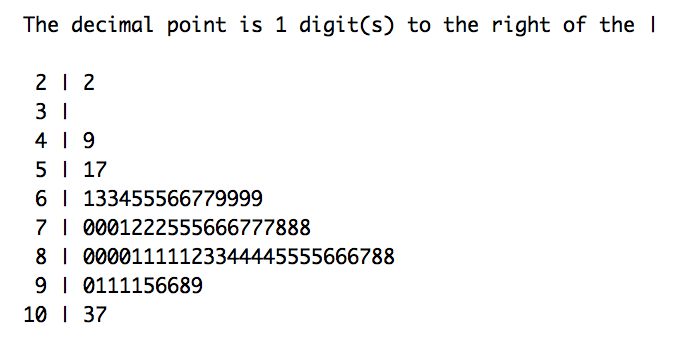
**median(NELS$numinst[NELS$advmath8=="Yes"],na.rm=T)**

**IQR(NELS$numinst[NELS$advmath8=="Yes"],na.rm=T)**

**median(NELS$numinst[NELS$advmath8=="No"],na.rm=T)**

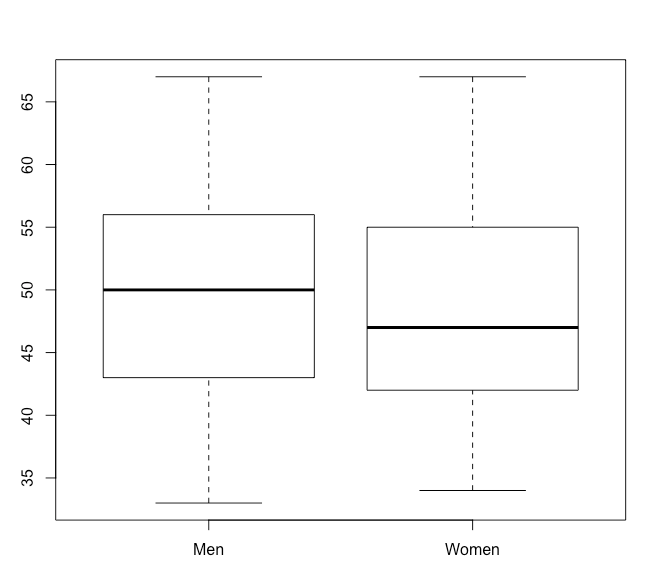
**IQR(NELS$numinst[NELS$advmath8=="No"],na.rm=T)**

Note that the command **par(mfrow=c(1,2))** splits the plotting window to display two plots side by side. Run **par(mfrow=c(1,1))** to return to the original plot settings.



The R command used to obtain the graph is **stem(Learndis$readcomp)**

1. Negatively skewed; there is one low outlier.
2. 81
3. 107
4. Probably the median. The mean is often pulled in the direction of the skew, which is negative.
5. IQR. A resistant statistic is selected because the distribution is skewed.
6. C
   1. According to the boxplot, we see that the age distribution is similar for men and women. Both distributions are fairly symmetric. The interquartile ranges are approximately equal, so the distributions are similarly spread out. The median age for men is slightly higher than it is for women. Descriptive statistics quantify these impressions. According to the skewness ratio, the age distribution for women is severely positively skewed. The skewness ratio for men is .85 and for women it is 2.28. The median age for men in the study at initial examination (median = 50) was slightly higher than for women (median = 47). Both distributions were similarly dispersed (IQR= 13). Men’s ages ranged from 33 to 67 years, while women’s ages ranged from 34 to 67 years.



The R commands used to obtain the graph and statistics are

**boxplot(Framingham$AGE1~Framingham$SEX)**

**IQR(Framingham$AGE1[Framingham$SEX=="Men"])**

**IQR(Framingham$AGE1[Framingham$SEX=="Women"])**

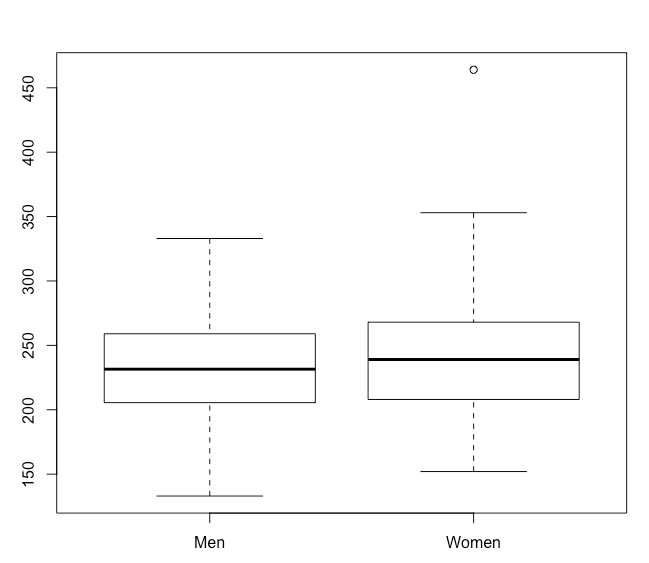
**skew.ratio(Framingham$AGE1[Framingham$SEX=="Men"])**

**skew.ratio(Framingham$AGE1[Framingham$SEX=="Women"])**

**summary(Framingham$AGE1[Framingham$SEX=="Men"])**

**summary(Framingham$AGE1[Framingham$SEX=="Women"])**

* 1. According to the boxplot, the cholesterol distribution is fairly symmetric for men, but positively skewed for women. The interquartile range for females is larger, so their cholesterol values are more heterogeneous. The median cholesterol for women is slightly higher than it is for men. Descriptive statistics quantify these impressions. According to the skewness ratio, the distribution is fairly symmetric for men (1.16) and severely positively skewed for women (4.30). The median cholesterol for men in the study at initial examination (median = 231.5) was slightly lower than for women (median = 239.0). The distribution of cholesterol for men was slightly more homogeneous (IQR = 52.8) than it was for women (IQR = 60.0). For men, the cholesterol levels ranged from 133.0 to 333.0, while for women they ranged from 152.0 to 464.0.



The R commands used to obtain the graph and statistics are

**boxplot(Framingham$TOTCHOL1~Framingham$SEX)**

**IQR(Framingham$TOTCHOL1[Framingham$SEX=="Men"])**

**IQR(Framingham$TOTCHOL1[Framingham$SEX=="Women"], na.rm = T)**

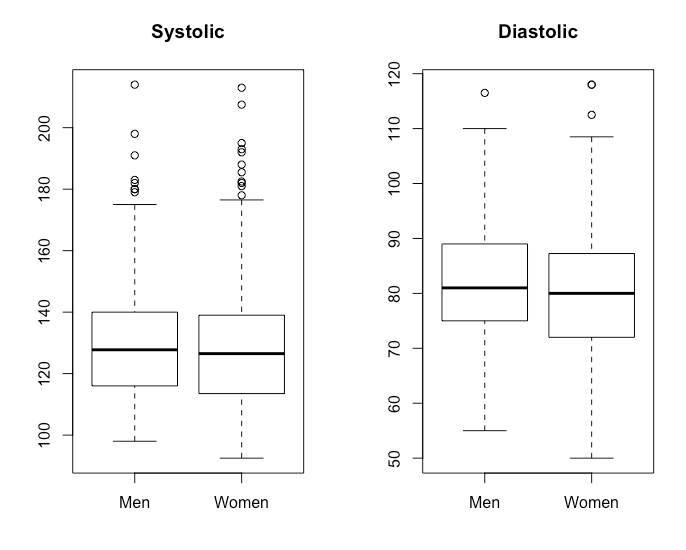
**skew.ratio(Framingham$TOTCHOL1[Framingham$SEX=="Men"])**

**skew.ratio(Framingham$TOTCHOL1[Framingham$SEX=="Women"])**

**summary(Framingham$TOTCHOL1[Framingham$SEX=="Men"])**

**summary(Framingham$TOTCHOL1[Framingham$SEX=="Women"])**

* 1. According to the boxplots, the blood pressure distributions are similar for men and women. All four distributions are positively skewed with outliers. The interquartile ranges for systolic and diastolic blood pressure are approximately equal for the sexes, so there are no gender differences in the spread of these distributions. The median systolic and diastolic blood pressures are approximately equal for men and women. Descriptive statistics quantify these impressions. According to the skewness ratio, all four distributions are severely positively skewed. The skewness ratio for the systolic blood pressure for men is 6.68, and for women it is 6.79. The skewness ratio for the diastolic blood pressure for men is 2.08, and for women it is 3.11. The most severely skewed distribution is the systolic blood pressure for women. The median blood pressure is slightly higher for men (systolic = 127.8, diastolic = 81.0) than it is for women (systolic = 126.5, diastolic = 80.0). According to the values for the interquartile range, the blood pressure scores for men (systolic IQR = 24.0, diastolic IQR = 14.0) were slightly less spread out than they were for women (systolic IQR = 25.5, diastolic IQR= 15.1). For men, the systolic blood pressures ranged from 98.0 to 214.0, while the diastolic blood pressures ranged from 55.0 to 116.5. For women, the systolic blood pressures ranged from 92.5 to 213.0, while the diastolic blood pressures ranged from 50.0 to 118.0.



The R commands used to obtain the graph and statistics are

**par(mfrow=c(1,2))**

**boxplot(Framingham$SYSBP1~Framingham$SEX, main = "Systolic", xlab="", ylab="")**

**boxplot(Framingham$DIABP1~Framingham$SEX, main = "Diastolic", xlab="", ylab="")**

**par(mfrow=c(1,1))**

**IQR(Framingham$SYSBP1[Framingham$SEX=="Men"])**

**IQR(Framingham$SYSBP1[Framingham$SEX=="Women"], na.rm = T)**

**skew.ratio(Framingham$SYSBP1[Framingham$SEX=="Men"])**

**skew.ratio(Framingham$SYSBP1[Framingham$SEX=="Women"])**

**summary(Framingham$SYSBP1[Framingham$SEX=="Men"])**

**summary(Framingham$SYSBP1[Framingham$SEX=="Women"])**

**IQR(Framingham$DIABP1[Framingham$SEX=="Men"])**

**IQR(Framingham$DIABP1[Framingham$SEX=="Women"], na.rm = T)**

**skew.ratio(Framingham$DIABP1[Framingham$SEX=="Men"])**

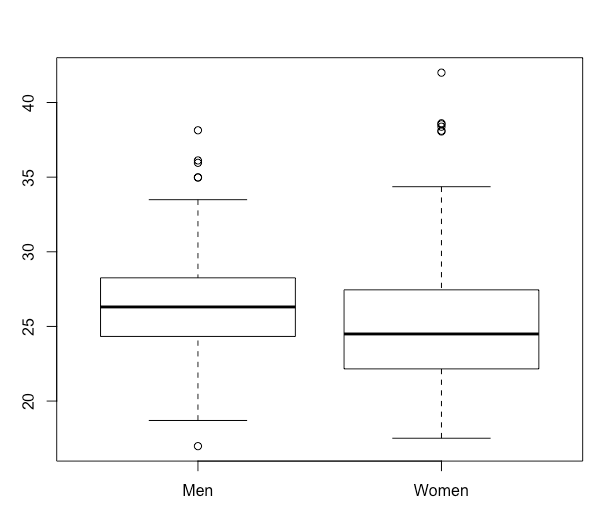
**skew.ratio(Framingham$DIABP1[Framingham$SEX=="Women"])**

**summary(Framingham$DIABP1[Framingham$SEX=="Men"])**

**summary(Framingham$DIABP1[Framingham$SEX=="Women"])**

Note that the command **par(mfrow=c(1,2))** splits the plotting window to display two plots side by side. Run **par(mfrow=c(1,1))** to return to the original plot settings.

* 1. According to the boxplot, the BMI distributions for both men and women are positively skewed. Because the interquartile range is smaller, the BMI scores are more consistent for men than for women. The median BMI for men is higher than it is for women. Descriptive statistics quantify these impressions. According to the skewness ratio, the distribution for men is fairly symmetric (1.88), while for women it is severely positively skewed (6.50). The median BMI is higher for men in the study at initial examination (median = 26.3) than for women (median = 24.5). The distribution of BMI’s was slightly more consistent for men (IQR = 3.9) than it was for women (IQR = 5.3). For men, BMI ranged from 17.0 to 38.1, while for women it ranged from 17.5 to 42.0.



The R commands used to obtain the graph and statistics are

**boxplot(Framingham$BMI1~Framingham$SEX)**

**IQR(Framingham$BMI1[Framingham$SEX=="Men"])**

**IQR(Framingham$BMI1[Framingham$SEX=="Women"], na.rm = T)**

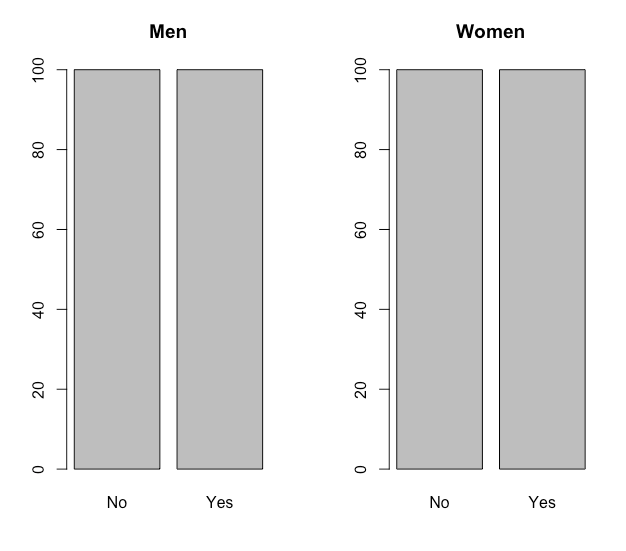
**skew.ratio(Framingham$BMI1[Framingham$SEX=="Men"])**

**skew.ratio(Framingham$BMI1[Framingham$SEX=="Women"])**

**summary(Framingham$BMI1[Framingham$SEX=="Men"])**

**summary(Framingham$BMI1[Framingham$SEX=="Women"])**

* 1. Because of the way the data were collected, of the 400 cases in the dataset, 100 were non-smoking men, 100 were non-smoking women, 100 were men who smoked, and 100 were women who smoked.



The R command used to obtain the graph is

**par(mfrow=c(1,2))**

**barplot(table(Framingham$CURSMOKE1[Framingham$SEX=="Men"]),**

**main = "Men")**

**barplot(table(Framingham$CURSMOKE1[Framingham$SEX=="Women"]),**

**main = "Women")**

**par(mfrow=c(1,1))**

Note that the command **par(mfrow=c(1,2))** splits the plotting window to display two plots side by side. Run **par(mfrow=c(1,1))** to return to the original plot settings.

|  |  |  |
| --- | --- | --- |
| *X* | *X -* | *(X - )2* |
| 7 | 2.42 | 5.856 |
| 16 | 11.42 | 130.42 |
| 0 | -4.58 | 20.976 |
| 5 | .42 | .176 |
| 0 | -4.58 | 20.976 |
| 0 | -4.58 | 20.976 |
| 7 | 2.42 | 5.856 |
| 1 | -3.58 | 12.816 |
| 6 | 1.42 | 2.016 |
| 1 | -3.58 | 12.816 |
| 0 | -4.58 | 20.976 |
| 12 | 7.42 | 55.056 |
| Total: 55 |  | Total: 308.916 |

Mean using Equation 3.1 = 

Variance using Equation 3.6 = 

Standard deviation using Equation 3.7= 

In order to find the median, mode, range, and interquartile range, it is helpful to sort the data as follows:

0 0 0 0 1 1 5 6 7 7 12 16

The median is the average of the 6th and 7th scores or (1+5) / 2 = 3.

The mode is 0.

The range is 16 – 0 = 16.

To calculate the interquartile range, we need the 75th and 25th percentiles. The 25th percentile is 0 and the 75th is 7.

The interquartile range is 7 – 0 = 7.

1. The R commands to generate the statistics are

**summary(Statisticians$AmStat)**

**the.mode(Statisticians$AmStat)**

**IQR(Statisticians$AmStat)**

**var(Statisticians$AmStat)**

**sd(Statisticians$AmStat)**

1. (0 – 4.58) + (0 – 4.58) + (0 – 4.58) + (0 – 4.58) + (1 – 4.58) + (1 – 4.58) + (5 – 4.58) + (6 – 4.58) + (7 – 4.58) + (7 – 4.58) + (12 – 4.58) + (16 – 4.58)

= (– 4.58 – 4.58 – 4.58 – 4.58 – 3.58 – 3.58 + .42 + 1.42 + 2.42 + 2.42 + 7.42 + 11.42

= -25.48 + 25.52

= .04 ≈ 0, within rounding error.

|  |  |  |
| --- | --- | --- |
| *X* | *X -* | *(X - )2* |
| 107 | -7 | 49 |
| 110 | -4 | 16 |
| 123 | 9 | 81 |
| 129 | 15 | 225 |
| 112 | -2 | 4 |
| 111 | -3 | 9 |
| 107 | -7 | 49 |
| 112 | -2 | 4 |
| 135 | 21 | 441 |
| 102 | -12 | 144 |
| 123 | 9 | 81 |
| 109 | -5 | 25 |
| 112 | -2 | 4 |
| 102 | -12 | 144 |
| 98 | -16 | 256 |
| 114 | 0 | 0 |
| 119 | 5 | 25 |
| 112 | -2 | 4 |
| 110 | -4 | 16 |
| 117 | 3 | 9 |
| 130 | 16 | 256 |
| Total: 2394 |  | Total: 1842 |

Mean using Equation 3.1 = 

Variance using Equation 3.6 = 

Standard deviation using Equation 3.7 = 

In order to find the median, mode, range, and interquartile range, it is helpful to sort the data as follows:

98 102 102 107 107 109 110 110 111 112 112 112 112 114 117 119 123 123 129 130 135

The median is 112.

The mode is 112.

The range is 135 – 98 = 37.

The interquartile range is 121 – 108 = 13.

1. The R commands to generate the statistics are

**summary(Blood$systolc1)**

**the.mode(Blood$systolc1)**

**IQR(Blood$systolc1)**

**var(Blood$systolc1)**

**sd(Blood$systolc1)**

* 1. The R command to generate the frequency distribution table is **table(Learndis$grade)**.

**../../../Desktop/Screen%20Shot%202019-06-14%20at%207.32.32%20P**

1. The 25th percentile, or *Q25* is approximately 2

The 50th percentile, or *Q50* is approximately 2

The 75th percentile, or *Q75* is approximately 3

1. The interquartile range for this distribution is approximately 3 – 2= 1.
2. The mean will decrease slightly, because the sum of all of the grades will decrease by 2. The median will not change at all. The mode will not change at all.
   1. Using Equation 3.1 for the mean, we see that because the mean is 4 and the sample size is 12, the sum of the raw scores must be 48, not 36.
   2. There are many possible answers. One is given for each case.
3. {0,0,0,0,0}
4. {0,0,0,0,5}
5. {1,2,3,5,5}
6. {0,0,0,0,0}
7. {5,5,5,5,5}
8. {-5,0,0,0,0}
   1. If the distribution of wages is $5,000, $5,000, $7,000, $10,000, and $23,000, then the mean wage is $10,000, the median wage is $7,000, and the mode wage is $5,000.
   2. In order for *X* to have a larger mean, its scores must be larger. In order for *Y* to have a larger standard deviation, its scores must be farther apart. One possible pair of distributions is *X* = {99, 100} and *Y* = {1, 10}.
   3. One possible pair of distributions is *X* = {10, 30, 50} and *Y* = {10, 35, 40}.
   4. One possible answer is *X* = {5, 5, 5, 5} and *Y* = {1, 10}.
   5. The majority means more than half. One possible distribution is {1, 2, 3, 4, 4}.
   6. No. An algebraic explanation is given. We begin by noting that a distribution with only two extreme values, one on either side of the mean, and no intermediate values between those extremes, will have a standard deviation at least as large as a distribution with both extreme and intermediate values. We show that this distribution of two extreme values has a standard deviation equal to half the range, suggesting that any other distribution of values will have a standard deviation equal to no more than half the range. We assume, without loss of generality, that the distribution has the two values, *a* and *b*. The range of this distribution is, therefore, *b - a* and half the range is, therefore, (*b - a*)/2. The mean of these two values equals (*b + a*)/2 and the standard deviation may be expressed as follows:



which proves that it is not possible for the standard deviation of a distribution to be more than half the range.

**3.28.** In the distribution depicted in the following histogram, the mean is 4 and the modes are 3 and 5.



**3.29.** One possible answer is depicted below.



**3.30.** One possible answer is *X* = {0, 10} and *Y* = {7, 9}.

**3.31.** One possible answer is *X* = {10, 50, 90} and *Y* = {50, 89, 90}.

**3.32.** One possible answer is *X* = {10, 49, 50} and *Y* = {60, 61, 62}.

**3.33.**

1. The mean. It is probably higher than the median because the distribution is so severely positively skewed.
2. 3 years. Because there is a great deal of variability in terms of the age at which people get married, we expect a large standard deviation. Note that the units (years) for the mean and standard deviation do not need to match.

**3.34.** The standard deviation is closer to 5 months. Because almost all students enter kindergarten at age 5, they are within a year of each other’s ages. The mean, however, is closer to 5 years.

**3.35.**

1. It is probably true that the bulk of the students did fairly well, while there were a few students who did very poorly.
2. Because the mean is probably lower than the median, it is preferable from a student perspective that the grade of B- be set to the mean, because a B- would then be easier to achieve.

**3.36.** A = X5, B = X1, C = X3, D = X4, E = X2.

**3.37.**

1. Positively skewed.
2. Since “most” tax breaks were less than $500, the median must be $500 or less.
3. The standard deviation will be larger than $500. Because most tax cuts are less than $500, the distance to the mean of $2,042 for most values in the distribution will be exceed $1,500 ($2042 - $500 = $1542), thus creating a standard deviation greater than $500. The presence of a small number of “huge tax cuts” given to a small number of “very wealthy businessmen” produces the very large mean relative to the bulk of the tax break values, and, in turn, a large standard deviation, greater than $500.

**3.38.** We would expect the mean to be greater than the median. The distribution is likely to be severely positively skewed because one cannot have fewer than 0 children and most women probably have 0 – 4 children, but there are some positive outliers where women have an unusually large number of children. In the case of such positive skew, the mean is likely to be larger than the median.

**3.39.** The solution to this problem can be seen as a weighted mean where each grade is weighted by the percent for which the respective exam counts. By Equation 3.4, = 83. Alternatively, this may be computed as .6\*85 + .4\*80 = 83.