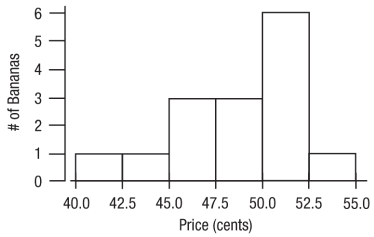
## Review of Part I – Exploring and Understanding Data

**R1.1. Bananas.**

**a)** A histogram of the prices of bananas from 15 markets, as reported by the USDA, is shown below.




**b)** The distribution of banana prices is skewed to the left, so median and IQR are appropriate measures of center and spread. The median is 49 cents per pound and the IQR is 6 cents per pound.

**c)** The distribution of the prices of bananas from 15 markets, as reported by the USDA, is unimodal and skewed to the left. The center of the distribution is approximately 50 cents, with the lowest price 42 cents per pound and the highest price 53 cents per pound.

**R1.2. Prenatal care.**

**a)** $\frac{5.4+3.9+6.1}{3} = 5.1\overline{3},$ so the overall rate of 5.1 deaths per thousand live births is equal to the average of the rates for Intensive, Adequate, and Inadequate prenatal care, when rounded to the nearest tenth. There is no reason this should be the case unless the number of women receiving each type of prenatal care is approximately the same.

**b)** Yes, the results indicate (but do not prove) that adequate prenatal care is important for pregnant women. The mortality rate is quite a bit lower for women with adequate care than for other women.

**c)** No, the results do not suggest that a woman pregnant with twins should be wary of seeking too much medical care. Intensive care is given for emergency conditions. The data do not suggest that the level of care is the cause of the higher mortality.

**R1.3. Singers by parts.**

**a)** The two statistics could be the same if enough sopranos have a height of 65 inches.

**b)** The distribution of heights of each voice part is roughly symmetric. The basses and tenors are generally taller than the altos and sopranos, with the basses being slightly taller than the tenors. The sopranos and altos have about the same median height. Heights of basses and sopranos are more consistent than altos and tenors.

**R1.4. Dialysis.**

There are only three patients currently on dialysis. With so few patients, no display is needed. We know that one patient has had his or her toes amputated and that two patients have developed blindness. What we don't know is whether or not the patient that has had his or her toes amputated has also developed blindness. Even if we wanted to, we do not have enough information to make an appropriate display.

**R1.5. Beanstalks.**

**a)** The greater standard deviation for the distribution of women's heights means that their heights are more variable than the heights of men.

**b)** The *z*-score for women to qualify is 2.40 compared with 1.75 for men, so it is harder for women to qualify.

Copyright © 2020 Pearson Education, Inc.

**R1.6. Bread.**

**a)** The distribution of the number of loaves sold each day in the last 100 days at the Clarksburg Bakery is unimodal and skewed to the right. The mode is near 100, with the majority of days recording fewer than 120 loaves sold. The number of loaves sold ranges from 95 to 140.

**b)** The mean number of loaves sold will be higher than the median number of loaves sold, since the distribution of sales is skewed to the right. The mean is sensitive to this skewness, while the median is resistant.

**c)** Create a boxplot with quartiles at 97 and 105.5, median at 100. The IQR is 8.5 so the upper fence is at 105.5 + 1.5(8.5) = 118.25. There are several high outliers. There are no low outliers because the min at 95 lies well within the lower fence at 97 – 1.5(8.5) = 84.25. Boxplots will vary.

**d)** The distribution of daily bread sales is not symmetric, but rather skewed to the right. The Normal model is not appropriate for this distribution. No conclusions can be drawn.

**R1.7. State University.**

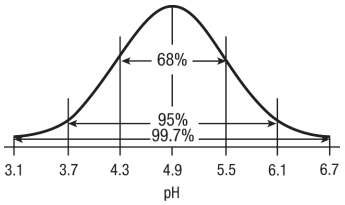
**a)** *Who* – Local residents near State University; *What* – Age, whether or not the respondent attended college, and whether or not the respondent had a favorable opinion of State University. *When* – Not specified. *Where* – Region around State University; *Why* – The information will be included in a report to the University's directors. *How* – 850 local residents were surveyed by phone.

**b)** There is one quantitative variable, age, probably measured in years. There are two categorical variables, college attendance (yes or no), and opinion of State University (favorable or unfavorable).

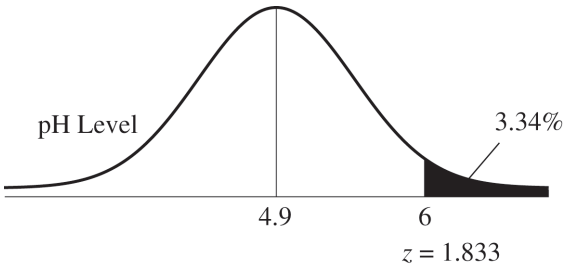
**c)** There are several problems with the design of the survey. No mention is made of a random selection of residents. Furthermore, there may be a non-response bias present. People with an unfavorable opinion of the university may hang up as soon as the staff member identifies himself or herself. Also, response bias may be introduced by the interviewer. The responses of the residents may be influenced by the fact that employees of the university are asking the questions. There may be greater percentage of favorable responses to the survey than truly exist.

**R1.8. Shenandoah rain.**

**a)** The Normal model for pH level of rainfall in the Shenandoah Mountains is shown below.




**b)** $z = \dfrac{y - \mu}{\sigma} = \dfrac{6 - 4.9}{0.6} = 1.833$

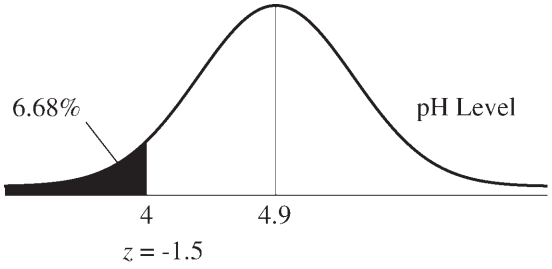



According to the Normal model, 3.3% of the rainstorms are expected to produce rainfall with pH levels above 6.

Copyright © 2020 Pearson Education, Inc.

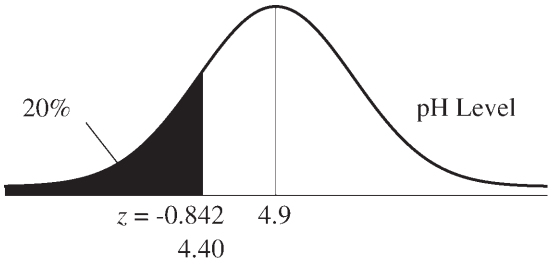
**R1.8.** (continued)

**c)** $z = \dfrac{y-\mu}{\sigma} = \dfrac{4-4.9}{0.6} = -1.5$



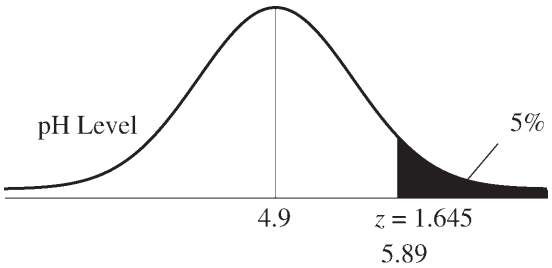

According to the Normal model, 6.7% of rainstorms are expected to produce rainfall with pH levels below 4.

**d)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow -0.842 = \dfrac{y-4.9}{0.6} \Rightarrow y = 4.40$




According to the Normal model, the most acidic 20% of storms have pH below 4.40.

**e)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow 1.645 = \dfrac{y-4.9}{0.6} \Rightarrow y = 5.89$

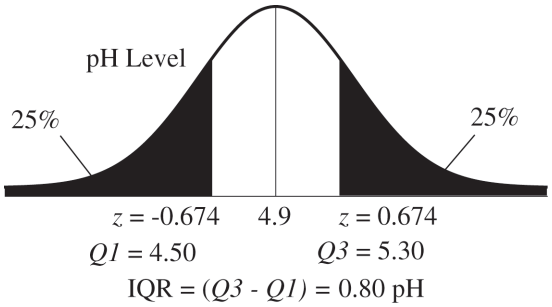



According to the Normal model, the least acidic 5% of storms have pH above 5.89.

Copyright © 2020 Pearson Education, Inc.

**R1.8.** (continued)

**f)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow -0.674 = \dfrac{\text{Q1}-4.9}{0.6} \Rightarrow \text{Q1} = 4.50$ and $z = \dfrac{y-\mu}{\sigma} \Rightarrow 0.674 = \dfrac{\text{Q3}-4.9}{0.6} \Rightarrow \text{Q3} = 5.30$




According to the Normal model, the IQR of the pH levels of the rainstorms is 5.30 – 4.50 = 0.80.

**R1.9. Fraud detection.**

**a)** Even though they are numbers, the NAICS code is a categorical variable, so mean and standard deviation are meaningless. Also, a histogram is a quantitative display, so it is not appropriate.

**b)** The Normal model will not work at all. The Normal model is for modeling distributions of unimodal and symmetric quantitative variables. NAICS code is a categorical variable.

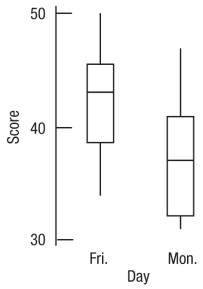
**R1.10. Streams.**

**a)** Stream Name – categorical; Substrate – categorical; pH – quantitative; Temperature – quantitative (°C); BCI – quantitative.

**b)** Substrate is a categorical variable, so a pie chart or a bar chart would be a useful display.

**R1.11. Cramming.**

**a)** Comparative boxplots of the distributions of Friday and Monday scores are shown below.

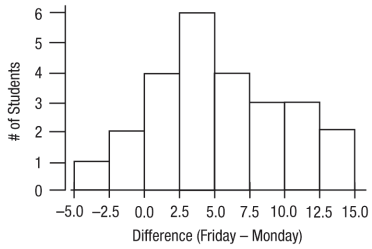



**b)** The scores on Friday were higher by about 5 points on average. This is a drop of more than 10% off the average score and shows that students fared worse on Monday after preparing for the test on Friday. The spreads are about the same, but the scores on Monday are a bit skewed to the right.

Copyright © 2020 Pearson Education, Inc.

**R1.11.** (continued)

**c)** A histogram of the distribution of change in test score is shown below.




**d)** The changes (Friday-Monday) are unimodal and centered near 4 points, with a spread of about 5 (SD). They are fairly symmetric, but slightly skewed to the right. Only 3 students did better on Monday (had a negative difference).

**R1.12. e-Books.**

The conclusion is not sound, these percentages may not be added because the groups are not disjoint. Many residents who have read a book will have also read an e-book.

**R1.13. Let's play cards.**

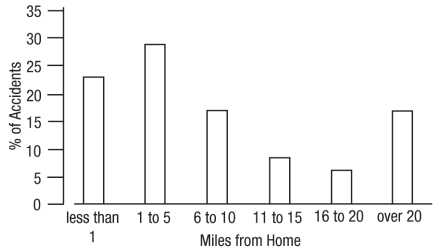
**a)** *Suit* is a categorical variable.

**b)** In the game of Go Fish, the denomination is not ordered. Numbers are merely matched with one another. You may have seen children's Go Fish decks that have symbols or pictures on the cards instead of numbers. These work just fine.

**c)** In the game of Gin Rummy, the order of the cards is important. During the game, ordered "runs" of cards are assembled (with Jack = 11, Queen = 12, King = 13), and at the end of the hand, points are totaled from the denomination of the card (face cards = 10 points). However, even in Gin Rummy, the denomination of the card sometimes behaves like a categorical variable. When you are collecting 3s, for example, order doesn't matter.

**R1.14. Accidents.**

**a)** The distances from home are organized in categories, so a bar chart is shown below. A pie chart would also be useful, since the percentages represent parts of a whole.




**b)** We are given no information about how many miles are driven in each of these categories, so we have no idea how many accidents to *expect*. We also have no information about how many accidents were involved in compiling the data.

**R1.15. Hard water.**

**a)** The variables in this study are both quantitative. Annual mortality rate for males is measured in deaths per 100,000. Calcium concentration is measured in parts per million.

**b)** The distribution of calcium concentration is skewed right, possibly bimodal. There looks to be a mode down near 12 ppm that is the center of a fairly tight symmetric distribution and another mode near 62.5 ppm that is the center of a much more spread out, symmetric (almost uniform) distribution. Mortality, however, appears unimodal and symmetric with the mode near 1500 deaths per 100,000.

Copyright © 2020 Pearson Education, Inc.

**R1.16. Hard water II.**

**a)** The overall mean mortality rate is $\frac{34(1631.59)+27(1388.85)}{34+27}=1524.15$ deaths per 100,000.

**b)** The distribution of mortality rates for the towns north of Derby is generally higher than the distribution of mortality rates for the towns south of Derby. Fully half of the towns south of Derby have mortality rates lower than any of the towns north of Derby. A quarter of the northern towns have rates higher than any of the Southern towns.

**R1.17. Seasons.**

**a)** The two histograms have different horizontal and vertical scales. This makes a quick comparison impossible.

**b)** The center of the distribution of average temperatures in January in is the low 30s, compared to a center of the distribution of July temperatures in the low 70s. The January distribution is also much more spread out than the July distribution. The range is over 50 degrees in January, compared to a range of over 20 in July. The distribution of average temperature in January is skewed slightly to the right, while the distribution of average temperature in July is roughly symmetric.

**c)** The distribution of difference in average temperature (July – January) for 60 large U.S. cities is roughly symmetric but slightly skewed to the left, with a median at approximately 40 degrees. There are several low outliers, cities with very little difference between their average July and January temperatures. The single high outlier is a city with a large difference in average temperature between July and January. The middle 50% of differences are between approximately 38 and 46 degrees.

**R1.18. Old Faithful.**

The distribution of time gaps between eruptions of Old Faithful is bimodal. A large cluster of time gaps has a mode at approximately 80 minutes and a slightly smaller cluster of time gaps has a mode at approximately 50 minutes. The distribution around each mode is fairly symmetric.

**R1.19. Old Faithful?**

**a)** The distribution of duration of the 222 eruptions is bimodal, with modes at approximately 2 minutes and 4.5 minutes. The distribution is fairly symmetric around each mode.

**b)** The bimodal shape of the distribution of duration of the 222 eruptions suggests that there may be two distinct groups of eruption durations. Summary statistics would try to summarize these two groups as a single group, which wouldn't make sense.

**c)** The intervals between eruptions are longer for long eruptions. There is very little overlap. More than 75% of the short eruptions had intervals less than about an hour (62.5 minutes), while more than 75% of the long eruptions had intervals longer than about 75 minutes. Perhaps the interval could even be used to predict whether the next eruption will be long or short.

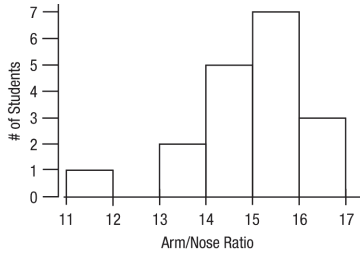
**R1.20. Teen drivers 2013.**

The chance of being a male involved in a fatal accident is different from the chance of being male. Thus, those variables are not independent. Also, the probability of being drunk in an accident is not the same for males and females, so those variables are not independent.

Copyright © 2020 Pearson Education, Inc.

**R1.21. Liberty's nose.**

**a)** The distribution of the ratio of arm length to nose length of 18 girls in a statistics class left skewed with a center around 15. There is one low outlier, a ratio of 11.8. A histogram is shown below. A boxplot or stemplot is also an appropriate display.




**b)** Even though the distribution is somewhat skewed, the mean and median are close. The mean is 15.0 and the standard deviation is 1.25.

**c)** The ratio of 9.3 for the Statue of Liberty is very low, well below the lowest ratio in the statistics class, 11.8, which is already a low outlier. Compared to the girls in the statistics class, the Statue of Liberty's nose is very long in relation to her arm.

**R1.22. Women's Short Track 2018.**

**a)** According to the Normal model, we would expect less than 68% of times to be within 4 seconds of the mean of 45.075 seconds, since 4 seconds is smaller than the standard deviation of 4.50 seconds.

**b)** In the actual data set, all but three of the times are within 4 seconds of the mean time, so about $26 / 29 = 0.90$, or 90% of the times fall within this interval.

**c)** The Normal model is not appropriate. There are three large outliers.

**R1.23. Sample.**

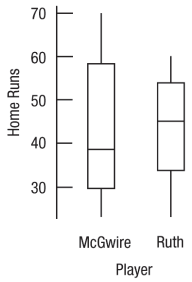
Overall, the follow-up group was insured only 11.1% of the time as compared to 16.6% for the not traced group. At first, it appears that group is associated with presence of health insurance. But for blacks, the follow-up group was quite close (actually slightly higher) in terms of being insured: 8.9% to 8.7%. The same is true for whites. The follow-up group was insured 83.3% of the time, compared to 82.5% of the not traced group. When broken down by race, we see that group is not associated with presence of health insurance for either race. This demonstrates Simpson's paradox, because the overall percentages lead us to believe that there is an association between health insurance and group, but we see the truth when we examine the situation more carefully.

**R1.24. Sluggers.**

**a)** The 5-number summary for McGwire's career is 3, 25.5, 36, 50.5, 70. The IQR is 25.

**b)** By the outlier test, 1.5(IQR) = 37.5. There are no homerun totals less than Q1 – 37.5 or greater than Q3 + 37.5. Technically, there are no outliers. However, the seasons in which McGwire hit fewer than 20 homeruns stand out as a separate group.

**c)** Parallel boxplots comparing the homerun careers of Mark McGwire and Babe Ruth are shown below.




Copyright © 2020 Pearson Education, Inc.

**R1.24.** (continued)

**d)** Without the injured seasons, McGwire and Ruth's home run production distributions look similar. (Note: Ruth's seasons as a pitcher were not included.) Ruth's median is a little higher, and he was a little more consistent (less spread), but McGwire had the two highest season totals.

**e)** A side-by-side stem-and-leaf display of the homerun careers of McGwire and Ruth is shown below.

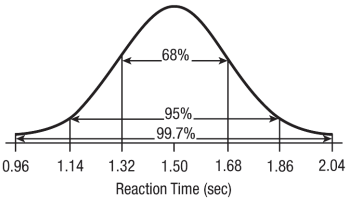
| | Stem | Leaf | |
|---|---|---|---|
| | 0 | 7 | |
| | 5 | 6 | 0 |
| | 82 | 5 | 449 |
| | 92 | 4 | 1166679 |
| McGwire | 99322 | 3 | 45 Ruth |
| | 92 | 2 | 25 |

7|0 = 70 Home Runs

f) From the stem-and-leaf display, we can see that Ruth was much more consistent. During most of his seasons, Ruth had homerun totals in the 40s and 50s. McGwire's seasons are much more spread out (not even including the three at the bottom).

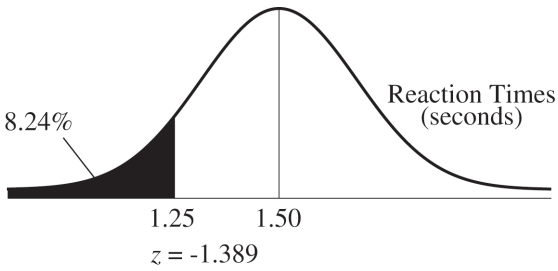
**R1.25. Be quick!**

**a)** The Normal model for the distribution of reaction times is shown below.




**b)** The distribution of reaction times is unimodal and symmetric, with mean 1.5 seconds, and standard deviation 0.18 seconds. According to the Normal model, 95% of drivers are expected to have reaction times between 1.14 seconds and 1.86 seconds.

**c)** $z = \dfrac{y - \mu}{\sigma} = \dfrac{1.25 - 1.5}{0.18} = -1.389$

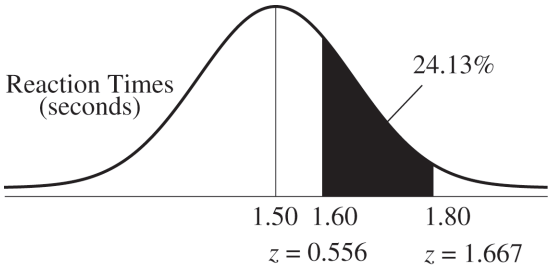



According to the Normal model, 8.2% of drivers are expected to have reaction times below 1.25 seconds.

Copyright © 2020 Pearson Education, Inc.

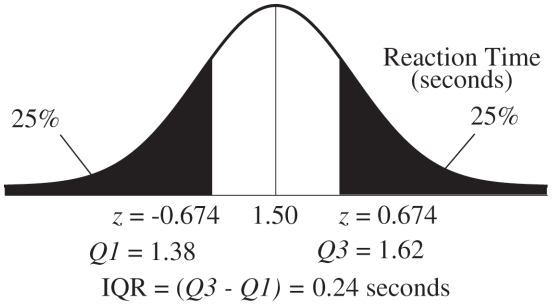
**R1.25.** (continued)

**d)** $z = \frac{y-\mu}{\sigma} = \frac{1.6-1.5}{0.18} = 0.556$ and $z = \frac{y-\mu}{\sigma} = \frac{1.8-1.5}{0.18} = 1.667$



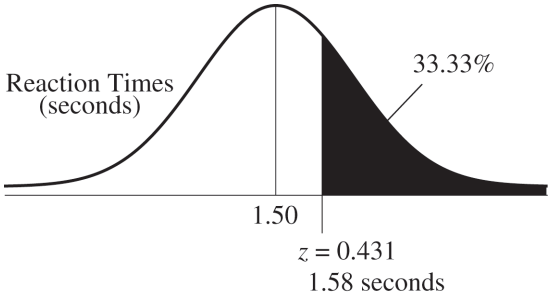

According to the Normal model, 24.1% of drivers are expected to have reaction times between 1.6 seconds and 1.8 seconds.

**e)** $z = \frac{y-\mu}{\sigma} \Rightarrow -0.674 = \frac{\text{Q1}-1.5}{0.18} \Rightarrow \text{Q1} = 1.38$ and $z = \frac{y-\mu}{\sigma} \Rightarrow 0.674 = \frac{\text{Q3}-1.5}{0.18} \Rightarrow \text{Q1} = 1.62$




According to the Normal model, the IQR of the distribution of reaction times is expected to be 1.62 – 1.38 =0.24 seconds.

**f)** $z = \frac{y-\mu}{\sigma} \Rightarrow 0.431 = \frac{y-1.5}{0.18} \Rightarrow y = 1.58$




According to the Normal model, the slowest 1/3 of all drivers are expected to have reaction times of 1.58 seconds or more. (Remember that a high reaction time is a SLOW reaction time!)

**R1.26. Music and memory.**

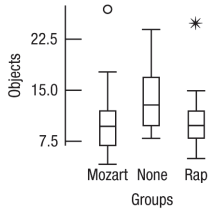
**a)** *Who* – 62 people; *What* – Type of music and number of objects remembered correctly; *When* – Not specified. *Where* – Not specified. *Why* – Researchers hoped to determine whether or not music affects memorization ability. *How* – Data were gathered in a completely randomized experiment.

**b)** Type of music (Rap, Mozart, or None) is a categorical variable. Number of items remembered is a quantitative variable.

Copyright © 2020 Pearson Education, Inc.

**R1.26.** (continued)

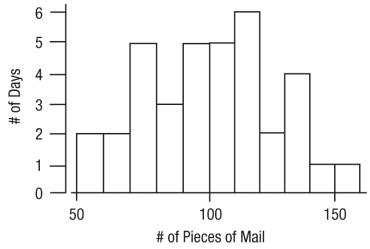
**c)** Because we do not have all the data, we can't know exactly how the boxplots look, but we do know that the minimums are all within the fences and that two groups have at least one outlier on the high side.




**d)** Mozart and Rap had very similar distributions of the number if objects remembered. The scores for None are, if anything, slightly higher than the other two groups. It is clear that groups listening to music (Rap or Mozart) did **not** score higher than those who listened to None.

**R1.27. Mail.**

**a)** A histogram of the number of pieces of mail received at a school office is shown below.




**b)** Since the distribution of number of pieces of mail is unimodal and symmetric, the mean and standard deviation are appropriate measures of center and spread. The mean number of pieces of mail is 100.25, and the standard deviation is 25.54 pieces.

**c)** The distribution of the number of pieces of mail received at the school office is somewhat symmetric and unimodal, but the center is rather flat, almost uniform. The lowest number of pieces received in a day was 52 and the highest was 151.

**d)** 23 of the 36 days (64%) had a number of pieces of mail received within one standard deviation of the mean, or within the interval 74.71 – 125.79. This is fairly close to the 68% predicted by the Normal model. The Normal model may be useful for modeling the number of pieces received by this school office.

**R1.28. Birth Order.**

**a)** There were 223 students. Of these, 113, or 50.7%, were oldest or only children.

**b)** There were 43 Humanities majors. Of these, 15, or 34.9%, were oldest or only children.

**c)** There were 113 oldest children. Of these, 15, or 13.3%, were Humanities majors.

**d)** There were 223 students. Of these, 15, or 6.7%, were oldest children majoring in Humanities.

**R1.29. Herbal medicine.**

**a)** *Who* – 100 health food store customers; *What* – Researchers asked whether or not the customer had taken the cold remedy and had customers rate the effectiveness of the remedy on a scale from 1 to 10. *When* – Not specified. *Where* – Store where natural health products are sold. *Why* – The researchers were from the Herbal Medicine Council, which sounds suspiciously like a group that might be promoting the use of herbal remedies. *How* – Researchers conducted personal interviews with 100 customers. No mention was made of any type of random selection.

**b)** "Have you taken the cold remedy?" is a categorical variable. Effectiveness on a scale of 1 to 10 is a categorical variable, as well, with respondents rating the remedy by placing it into one of 10 categories.

Copyright © 2020 Pearson Education, Inc.

**R1.29.** (continued)

**c)** Very little confidence can be placed in the Council's conclusions. Respondents were people who already shopped in a store that sold natural remedies. They may be pre-disposed to thinking that the remedy was effective. Furthermore, no attempt was made to randomly select respondents in a representative manner. Finally, the Herbal Medicine Council has an interest in the success of the remedy.

**R1.30. Birth order revisited.**

**a)** Overall, 25.6% of the students were Math/Science majors, 41.7% were Agriculture majors, 19.3% were Humanities majors, and 13.5% had other majors.

**b)** Of the oldest children, 30.1% of the students were Math/Science majors, 46.0% were Agriculture majors, 13.3% were Humanities majors, and 10.6% had other majors.

**c)** Of the second born children, 20.3% of the students were Math/Science majors, 39.1% were Agriculture majors, 24.6% were Humanities majors, and 15.9% had other majors.

**d)** No, college major does not appear to be independent of birth order. Oldest children are more likely than second born children to major in Math/Science (30% to 20%), while second born children are more likely than oldest children to major in Humanities (25% to 13%).

**R1.31. Engines.**

**a)** The count of cars is 38.

**b)** The mean displacement is higher than the median displacement, indicating a distribution of displacements that is skewed to the right. There are likely to be several very large engines in a group that consists of mainly smaller engines.

**c)** Since the distribution is skewed, the median and IQR are useful measures of center and spread. The median displacement is 148.5 cubic inches and the IQR is 126 cubic inches.

**d)** Your neighbor's car has an engine that is bigger than the median engine, but 227 cubic inches is smaller than the third quartile of 231, meaning that at least 25% of cars have a bigger engine than your neighbor's car. Don't be impressed!

**e)** Using the Outlier Rule (more than 1.5 IQRs beyond the quartiles) to find the fences:

Upper Fence: Q3 + 1.5(IQR) = 231 + 1.5(126) = 420 cubic inches.

Lower Fence: Q1 – 1.5(IQR) = 105 – 1.5(126) = - 84 cubic inches.

Since there are certainly no engines with negative displacements, there are no low outliers. Q1 + Range = 105 + 275 = 380 cubic inches. This means that the maximum must be less than 380 cubic inches. Therefore, there are no high outliers (engines over 420 cubic inches).

**f)** It is not reasonable to expect 68% of the car engines to measure within one standard deviation of the mean. The distribution of engine displacements is skewed to the right, so the Normal model is not appropriate.

**g)** Multiplying each of the engine displacements by 16.4 to convert cubic inches to cubic centimeters would affect measures of position and spread. All of the summary statistics (except the count!) could be converted to cubic centimeters by multiplying each by 16.4.

**R1.32. Engines, again.**

**a)** The distribution of horsepower is roughly uniform, with a bit of skew to the right, as the number of cars begins to taper off after about 125 horses. The center of the distribution is about 100 horsepower. The lowest horsepower is around 60 and the highest is around 160.

**b)** The interquartile range is Q3 – Q1 = 125 – 78 = 47 horses.

Copyright © 2020 Pearson Education, Inc.

**R1.32.** (continued)

**c)** Using the Outlier Rule (more than 1.5 IQRs beyond the quartiles) to find the fences:

Upper Fence: Q3 + 1.5(IQR) = 125 + 1.5(47) = 195.5 horses

Lower Fence: Q1 – 1.5(IQR) = 78 – 1.5(47) = 7.5 horses

From the histogram, we can see that there are no cars with horsepower ratings anywhere near these fences, so there are no outliers.

**d)** The distribution of horsepower is roughly uniform, not unimodal, and not very symmetric, so the Normal model is probably not a very good model of the distribution of horsepower.

**e)** Within one standard deviation of the mean is roughly the interval 75 – 125 horses. By dividing the bars of the histogram up into boxes representing one car, and taking half of the boxes in the bars representing 70–79 and 120–129, I counted 22 (of the 38) cars in the interval. Approximately 58% of the cars are within one standard deviation of the mean.

**f)** Adding 10 horses to each car would increase the measures of position by 10 horses and leave the measures of spread unchanged. Mean, median, 25th percentile and 75th percentile would each increase by 10. The standard deviation, interquartile range, and range would remain the same.

**R1.33. Age and party 2011.**

**a)** 3705 of 8414, or approximately 44.0%, of all voters surveyed were Republicans or leaned Republican.

**b)** This was a representative telephone survey conducted by Pew, a reputable polling firm. It is likely to be a reasonable estimate of the percentage of all voters who are Republicans.

**c)** 815 + 2416 = 3231 of 8414, or approximately 38.4%, of all voters surveyed were under 30 or over 65 years old.

**d)** 73 of 8414, or approximately 0.87%, of all voters surveyed were classified as "Neither" and under the age of 30.

**e)** 73 of the 733 people classified as "Neither", or 9.96%, were under the age of 30.

**f)** 73 of the 815 respondents under 30, or 8.96%, were classified as "Neither".

**R1.34. Pay.**

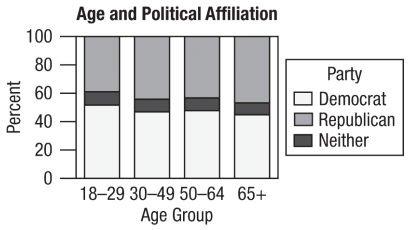
The distribution of hourly wages for Chief Executives has a mean larger than the median, indicating a distribution that is skewed to the right. Likewise, the distribution of hourly wages for General and Operations Managers has a mean higher than the median, indicating a distribution that is skewed to the right.

**R1.35. Age and party 2011 II.**

**a)** The marginal distribution of party affiliation is:

| | | | | | |
|---|---|---|---|---|---|
| As percents: | Republican: | 44.0% | Democrat: 47.3% | Neither: | 8.7% |
| As counts: | Republican: | 3705 | Democrat: 3976 | Neither: | 733 |

**b)**




Copyright © 2020 Pearson Education, Inc.

**R1.35.** (continued)

**c)** It appears that older voters are more likely to lean Republican, and younger voters are more likely to lean Democrat.

**d)** No. There is an evidence of an association between party affiliation and age. Younger voters tend to be more Democratic and less Republican.

**R1.36. Bike safety 2015.**

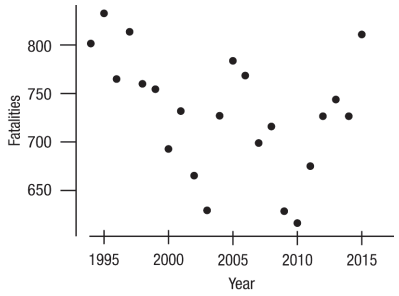
**a)** *Who* – Years from 1994 to 2015; *What* – Number of bicycle fatalities reported; *When* – 1994 to 2015; *Where* – United States; *Why* – The information was collected for a report by the Bicycle Helmet Safety Institute. *How* – Although not specifically stated, the information was probably collected from a government agency or hospital records.

**b)**

| Stem | Leaf |
|---|---|
| 6 | 233 |
| 6 | 789 |
| 7 | 0233334 |
| 7 | 56778 |
| 8 | 0113 |

6|2 = 620 - 629 Fatalities

**c)**




**d)** The stem-and-leaf display of the number of yearly bicycle fatalities reported in the United States shows that distribution is fairly symmetric. It also provides some idea about the center and spread of the annual fatalities. This is not visible on the timeplot.

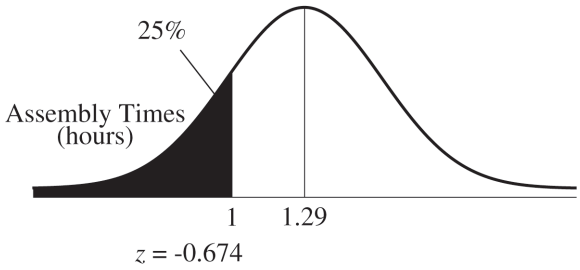
**e)** The timeplot of the number of yearly bicycle fatalities reported in the United Stats shows that the number of fatalities per year has declined until 2003, rose until 2005, fell again until 2010, but has risen sharply in recent years. The trend in the number of bicycle fatalities is unclear.

**f)** In the 10-year period from 1994 to 2003, reported bicycle fatalities decreased fairly steadily from about 800 per year to around 620 a year, then it increased sharply back to nearly 800, decreased to nearly 600 by 2010, and since then has increased back to nearly the highest levels. Overall, it's not clear what the trend is other than just random fluctuation.

Copyright © 2020 Pearson Education, Inc.

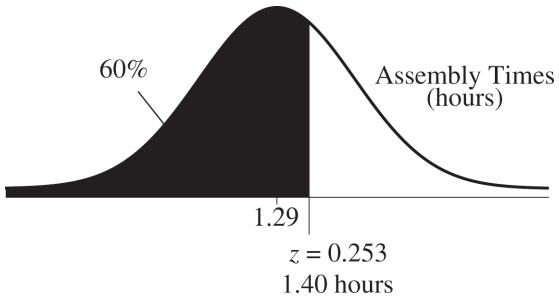
**R1.37. Some assembly required.**

**a)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow -0.674 = \dfrac{1-1.29}{\sigma} \Rightarrow \sigma = 0.43$



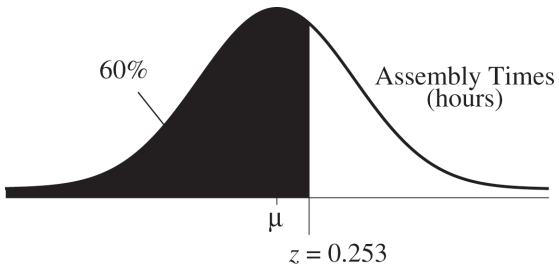

According to the Normal model, the standard deviation is 0.43 hours.

**b)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow 0.253 = \dfrac{7-1.29}{0.43} \Rightarrow y = 1.40$




According to the Normal model, the company would need to claim that the desk takes "less than 1.4 hours to assemble", not the catchiest of slogans!

**c)** $z = \dfrac{y-\mu}{\sigma} \Rightarrow 0.253 = \dfrac{1-\mu}{0.43} \Rightarrow \mu = 0.89$




According to the Normal model, the company would have to lower the mean assembly time to 0.89 hour (53.4 minutes).

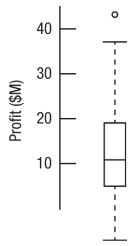
**d)** The new instructions and part-labeling may have helped lower the mean, but it also may have changed the standard deviation, making the assembly times more consistent as well as lower.

Copyright © 2020 Pearson Education, Inc.

**R1.38. Global500 2014.**

**a)** The 5-number summary of the profits (in \$M) of 30 of the *Forbes* 500 largest US corporations is: –7, 5, 11, 19, 43.

**b)** The boxplot of the distribution of the profits of 30 of the *Forbes* 500 largest US corporations is at the right.

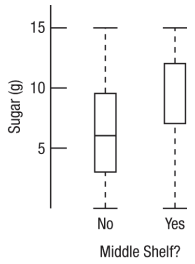



**c)** The mean profit is 13.9 \$M, and the standard deviation of the distribution of profits is 11.9 \$M.

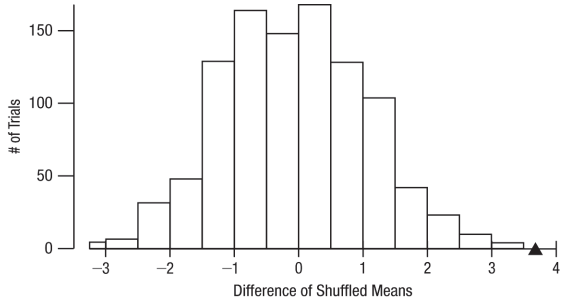
**d)** The distribution of profits is unimodal and skewed to the right with an outlier at 43 \$M. The median profit is 11 \$M. The middle 50% of companies report profit between 5 and 19 \$M. One company lost 7 \$M.

**R1.39. Shelves shuffled.**

**a)** The 21 middle shelf cereals have a mean sugar content of 9.62 g/serving, compared with 5.93 g/serving for the others, for a difference of 3.69 g/serving. The medians are 12 and 6 g/serving, respectively.



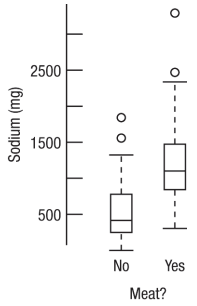

**b)** Answers will vary slightly. Because none of the 1000 shuffled differences were as large as the observed difference, we can say that a difference that large is unlikely to be produced by chance. There is evidence to suggest that the cereals on the middle shelf have a higher mean sugar content.



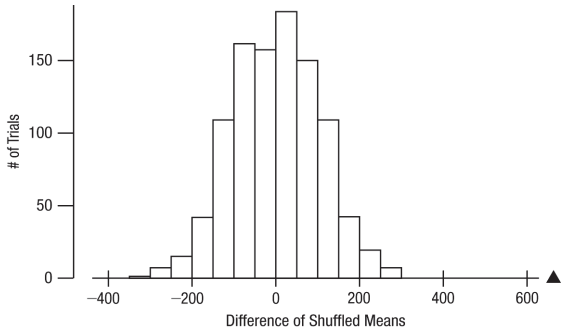

Copyright © 2020 Pearson Education, Inc.

**R1.40. Salty fries?**

**a)** The 73 meat items have a mean sodium content of 1197.4 grams compared with 532.2 grams for the non-meat items. The medians are 1110 and 400 grams, respectively. There are two mildly high outliers in each group.




**b)** Answers will vary slightly. Because none of the 1000 shuffled differences were as large as the observed difference, we can say that a difference that large is unlikely to be produced by chance. There is strong evidence to suggest that the meat items have higher mean sodium content. (The same conclusion holds with the four outliers removed.)




**R1.41. Hopkins Forest investigation.**

Answers will vary.

**R1.42. *Titanic* investigation.**

Answers will vary.

**R1.43. Student survey investigation.**

Answers will vary.

**R1.44. Movies investigation.**

Answers will vary.

Copyright © 2020 Pearson Education, Inc.