CHAPTER 1

The exercises in Chapter 1 are designed to provoke thought about issues that arise when numbers are used to communicate ideas. These questions do not have a single correct answer; therefore, solutions are not provided.

Similarly, solutions are not provided for the first few exercises in each chapter, which are intended to review concepts and definitions.

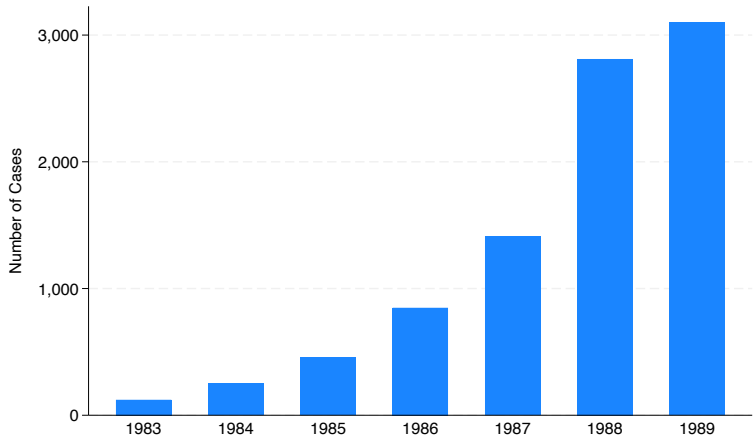
CHAPTER 2

Exercise 11
a. The number of suicides is discrete.
b. The response to treatment is ordinal.
c. The concentration of lead is continuous.
d. Political party affiliation is nominal.
e. The presence or absence of hepatitis C is discrete.
f. The length of time is continuous.
g. The number of previous miscarriages is discrete.
h. Satisfaction with care is ordinal.
i. The age of a child is continuous.

Exercise 12
The statement is not accurate. The intervals in the table are of unequal length; therefore, it does not make sense to compare the absolute frequencies within them. The interval 16–30 has length 15 minutes, for example, while the interval 11–15 is only 5 minutes long.

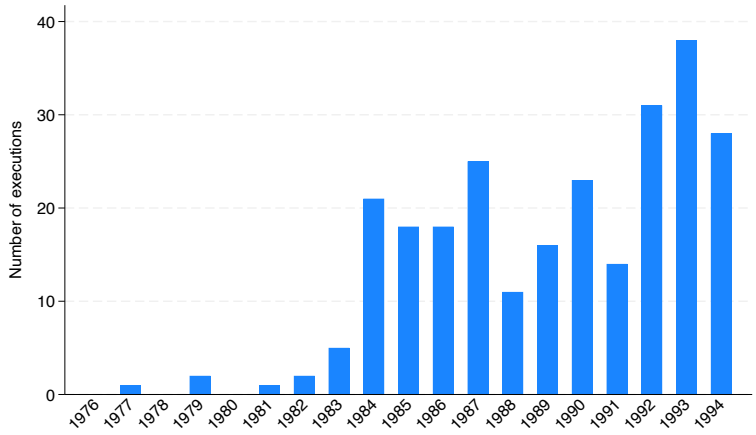
Exercise 13
A bar chart showing the numbers of cases of pediatric AIDS by year is displayed below. The graph indicates that the number of cases of AIDS increased each year between 1983 and 1989, with the biggest jump occurring between 1987 and 1988.

```
graph bar (asis) cases, over(Year)
```



Exercise 14
A bar chart of the number of executions by year is shown on the following page. There were only a few executions in the eight years immediately following the 1976 Supreme Court decision. After that, the number of executions increased, and has continued to increase over time (although not steadily; there are periodic decreases).

```
graph bar (asis) executions, over(year, label(angle(forty_five)))
```



Exercise 15

a. Serum cotinine level is continuous.

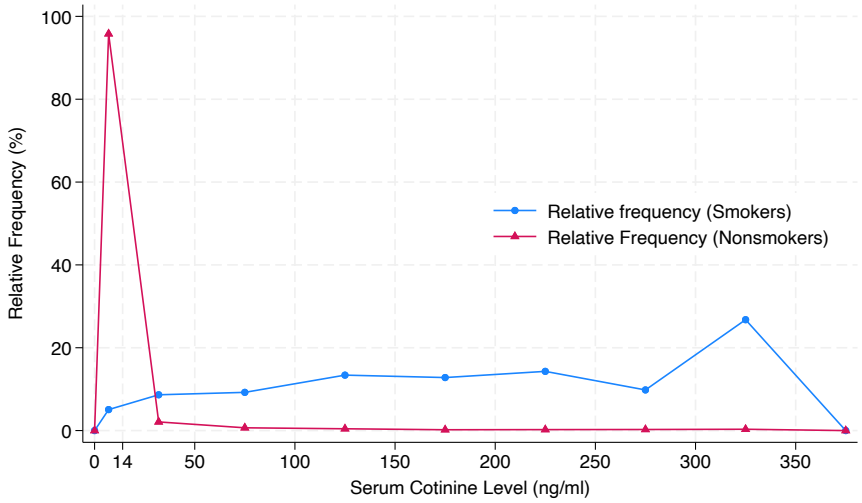
b. Because the total number of smokers is not equal to the total number of nonsmokers, it is not fair to compare the distributions of absolute frequencies for these two groups.

c. The table of relative frequencies of serum cotinine levels appears below.

Cotinine Level (ng/ml)	Smokers (%)	Nonsmokers (%)
0–13	5.1	95.8
14–49	8.6	2.1
50–99	9.2	0.7
100–149	13.4	0.4
150–199	12.8	0.2
200–249	14.3	0.2
250–299	9.8	0.3
300+	26.8	0.3

d. The frequency polygons are shown below. Note that the intervals are of unequal length. For the purposes of constructing the polygons, the last interval is assumed to be 300–349 ng/ml.

```
twoway (connected relfreq_smoke cotinine, msymbol(circle))
(connected relfreq_nonsmoke cotinine, msymbol(triangle)),
ytitle(‘"Relative Frequency (\%)"’) ylabel(0 20 40 60 80 100)
xtitle(‘"Serum Cotinine Level (ng/ml)"’) xlabel(0 14 50 100 150 200 250 300 350)
legend(position(3) ring(0))
```



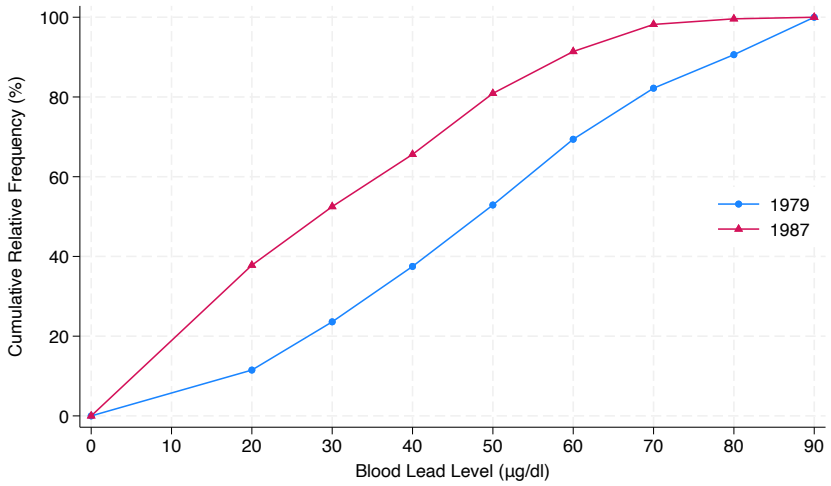
e. The distribution of smokers is fairly uniform across cotinine levels. The relative frequency is smallest in the first interval (0–13 ng/ml). It then increases, and remains consistent (hovering around 10%) across subsequent intervals up to the last (300+ ng/ml), where the relative frequency increases. For nonsmokers, nearly everyone has a cotinine level below 13 ng/ml; the relative frequency in each of the other intervals is very small.
f. Yes, it is possible that some of the subjects are misclassified. In particular, there are a number of self-reported "nonsmokers" with extremely high cotinine levels.

Exercise 16
a. The workers have lower blood lead levels in 1987.
b. The cumulative relative frequencies for each group of workers are displayed below.

Blood Lead (μg/dl)	1979 (%)	1987 (%)
< 20	11.5	37.8
20–29	23.6	52.5
30–39	37.5	65.6
40–49	52.9	80.9
50–59	69.4	91.4
60–69	82.2	98.2
70–79	90.6	99.6
≥ 80	100.0	100.0

The cumulative frequency polygons are displayed on the next page. Note that, for the purposes of constructing the polygons, the last interval is assumed to be 80–89 μg/dl.

```
twoway (connected year1 blood_lead, msymbol(circle)) (connected year2
blood_lead, msymbol(triangle)), ytitle(‘"Cumulative Relative Frequency (%)"’)
ylabel(0 20 40 60 80 100) xtitle(‘"Blood Lead Level ({&mu}g/dl)"’)
xlabel(0 10 20 30 40 50 60 70 80 90) legend(position(3) ring(0))
```



c. The distribution of blood lead levels is stochastically larger for the group of workers in 1979.

Exercise 17

a. The diagnosis of lung cancer is discrete.

b. Smoking status is discrete.

c. The graph tells us that the majority of men diagnosed with lung cancer were heavy, excessive, or chain smokers. For those who have not been diagnosed with lung cancer, the percentage of men in each category is more evenly distributed, except for heavy smokers which includes the highest percentage of men. There are almost an equal number of heavy smokers diagnosed with lung cancer as there are heavy smokers that were not diagnosed with lung cancer.

d. The table is presented below. Note that the cumulative frequencies do not add to 100% because of rounding. The values used are from the original publication [68].

	Relative Frequency		Cumulative Frequency	
Smoking status	Without lung cancer	Lung cancer	Without lung cancer	Lung cancer
None	14.6	1.3	14.6	1.3
Light	11.5	2.3	26.1	3.6
Moderately heavy	19.0	10.1	45.1	13.7
Heavy	35.6	35.2	80.7	48.9
Excessive	11.5	30.9	92.2	70.8
Chain	7.6	20.3	99.8	100.1

Exercise 18

a. The graph tells us that South Asia had by far the highest number of smallpox cases. The graph also shows that South Asia was one of the last regions to eliminate smallpox.

b. The graph tells us that the number of reported cases of smallpox was highest for Europe and Central Asia from 1920–1924, but decreased substantially after these years.

c. Based on this graph we could say that for most regions of the world, the number of smallpox cases decreases over time; however, cases in South Asia continue to spike over time until the eradication of the disease.

Exercise 19

a. The mean time to seizure is

$$\begin{aligned}\bar{x} &= \frac{\sum_{i=1}^{13} x_i}{13} \\ &= \frac{336.85}{13} \\ &= 25.9 \text{ months.}\end{aligned}$$

b. The median is the middle value, or 24 months.

c. Since two different measurements each occur twice, the data set has two modes — 12 months and 24 months.

d. The range is the largest value minus the smallest value, or

$$96 - 0.1 = 95.9 \text{ months.}$$

e. The interquartile range is the 75th percentile minus the 25th percentile, or

$$36 - 4 = 32 \text{ months.}$$

f. The standard deviation is

$$\begin{aligned}\sqrt{\frac{\sum_{i=1}^{13}(x_i - \bar{x})^2}{13-1}} &= \sqrt{\frac{\sum_{i=1}^{13}(x_i - 25.9)^2}{12}} \\ &= \sqrt{749.2} \\ &= 27.4 \text{ months.}\end{aligned}$$

Note that

$$\begin{aligned}\sum_{i=1}^{13}(x_i - \bar{x}) &= (0.10 - 25.91) + (0.25 - 25.91) + (0.50 - 25.91) \\ &\quad + (4 - 25.91) + (12 - 25.91) + (12 - 25.91) \\ &\quad + (24 - 25.91) + (24 - 25.91) + (31 - 25.91) \\ &\quad + (36 - 25.91) + (42 - 25.91) + (55 - 25.91) \\ &\quad + (96 - 25.91) \\ &= -25.81 - 25.66 - 25.41 - 21.91 - 13.91 - 13.91 \\ &\quad - 1.91 - 1.91 + 5.09 + 10.09 + 16.09 + 29.09 \\ &\quad + 70.09 \\ &= 0.02 \\ &\approx 0.\end{aligned}$$

Because we use the mean rounded to two decimal places in the calculations, the sum is not exactly equal to 0.

Exercise 20

a. The mean calcium level is

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{8} x_i}{8} \\ &= \frac{25.14}{8} \\ &= 3.14 \text{ mmol/l.} \end{aligned}$$

The median is the average of the 4th and 5th values, or

$$\frac{2.99 + 3.17}{2} = 3.08 \text{ mmol/l.}$$

The standard deviation is

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{8}(x_i - \bar{x})^2}{8-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{8}(x_i - 3.14)^2}{7}} \\ &= \sqrt{0.2608} \\ &= 0.51 \text{ mmol/l.} \end{aligned}$$

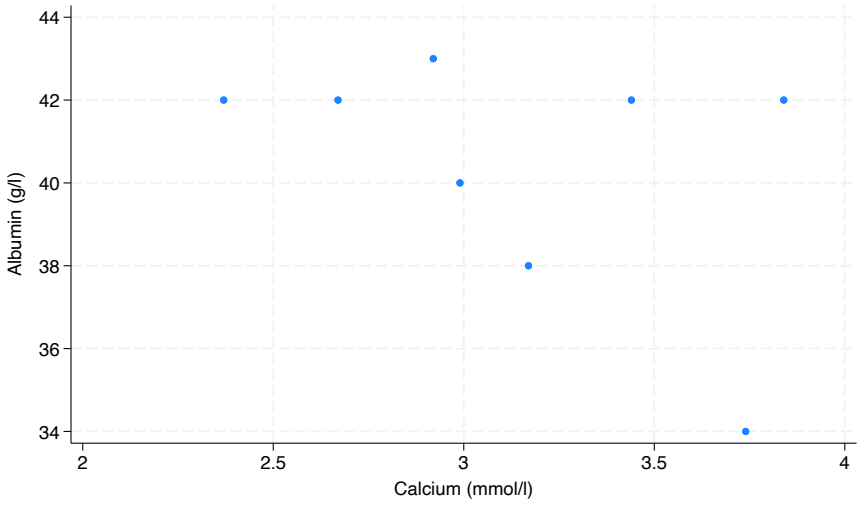
The range is the largest value minus the smallest value, or

$$3.84 - 2.37 = 1.47 \text{ mmol/l.}$$

b. The mean albumin level is 40.4 g/l. The median is 42 g/l. The standard deviation is 3.0 g/l. The range is 9 g/l.

c. A scatter plot of albumin versus calcium is displayed below.

```
twoway (scatter albumin calcium)
```



d. The patients suffering from vitamin D intoxication all have albumin levels within the normal range. However, they do not have normal blood levels of calcium. Both the mean and the median lie above the upper limit of the normal range; overall, 6 of the 8 patients have calcium levels that are above normal.

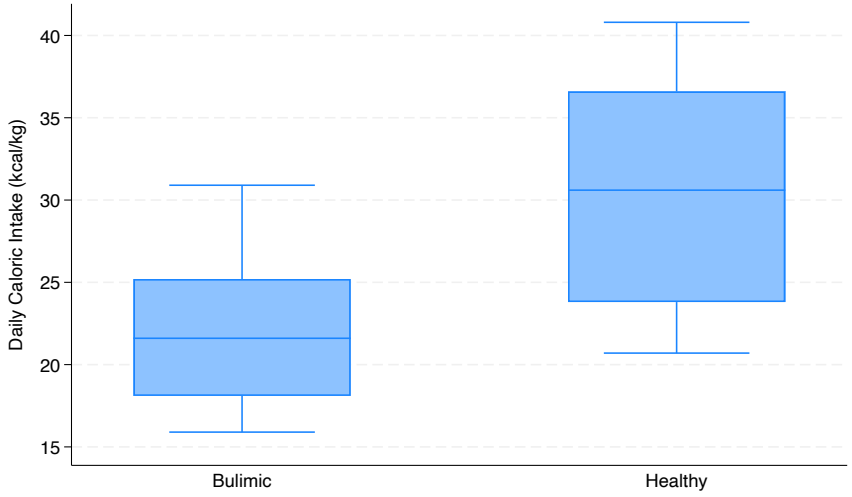
Exercise 21

a. The median daily caloric intake for the bulimic adolescents is 21.6 kcal/kg, and the median for the healthy adolescents is 30.6 kcal/kg.

b. The interquartile range for the bulimic adolescents is the 75th percentile minus the 25th percentile, or $25.2 - 18.1 = 7.1$ kcal/kg. The interquartile range for the healthy adolescents is $36.6 - 23.8 = 12.8$ kcal/kg.

c. A box plot of the daily caloric intake for bulimic and healthy adolescents is displayed below.

```
graph box intake, over(group)
```



d. Daily caloric intake tends to be higher for the healthy adolescents. This group also exhibits a greater amount of variability.

Exercise 22

a. Europe has the smallest mean; its infant mortality rates are much lower than those of either Africa or Asia.

b. The box plots indicate that Africa has the largest median; it has the greatest proportion of nations with relatively large infant mortality rates.

c. Europe has the smallest standard deviation. Its infant mortality rates are more tightly clustered about the mean than those of either Africa or Asia.

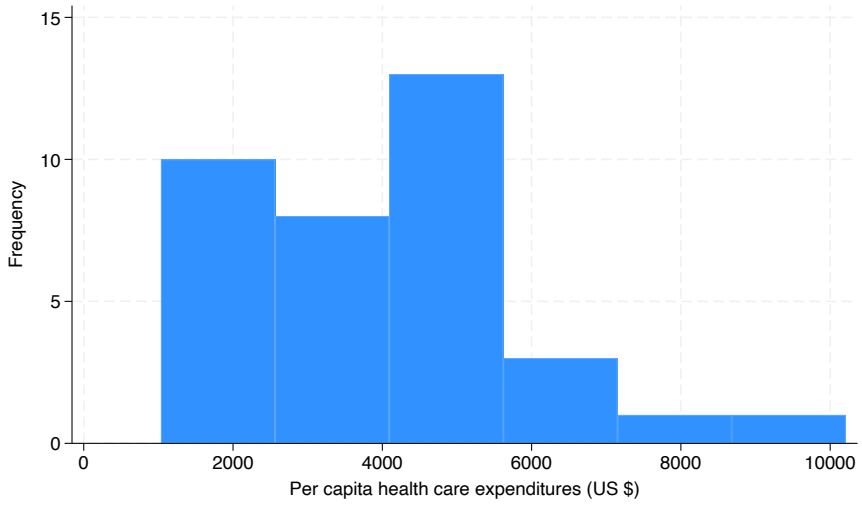
d. Since the distribution of values for Africa is unimodal and roughly symmetric, the mean and the median infant mortality rates should be fairly close.

We would not expect the mean and median to be equal for Asia, however; since the distribution is skewed to the right, the mean will be larger than the median.

Exercise 23

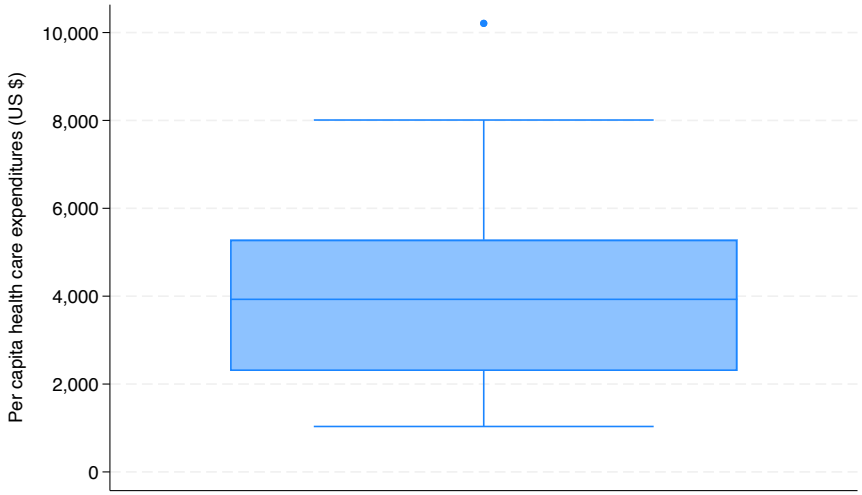
a. A histogram of per capita expenditure is displayed below.

```
histogram per_capita, frequency
```



b. A box plot of per capita expenditure is displayed below.

```
graph box per_capita
```



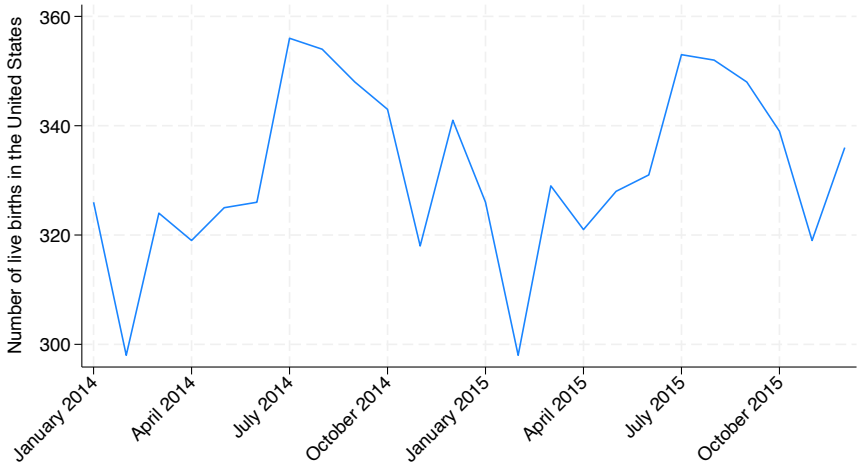
c. The figures show that the distribution of per capita health expenditure is right-skewed with a median value of approximately $4,000. There is also an outlier shown in the box plot.

d. The histogram shows how many countries have per capita values in specific ranges, and this information is not provided in the box plot. The box plot shows there is an outlier, and this information is not seen in the historgram.

Exercise 24

a. A line graph of the reported number of births over time is displayed below.

```
twoway (line births month), xtitle(, color(%0)) xlabel(1 "January 2014"
4 "April 2014" 7 "July 2014" 10 "October 2014" 13 "January 2015"
16 "April 2015" 19 "July 2015" 22 "October 2015", labels angle(forty_five))
```



b. Yes, there does seem to be a seasonal pattern of live births in the United States. Based on this two-year period, there appears to be a tendency toward more births in the spring and summer months and fewer in fall and winter.

Exercise 25

a. Serum zinc level is continuous.

b. The table of relative frequencies appears on the following page. The serum zinc levels range from 50 to 159 μg/dl; however, most of the values lie between 70 and 109 μg/dl. The intervals 70–79 and 80–89 contain the greatest numbers of observations.

c. A histogram of the data appears on the next page.

d. The histogram is unimodal and skewed to the right.

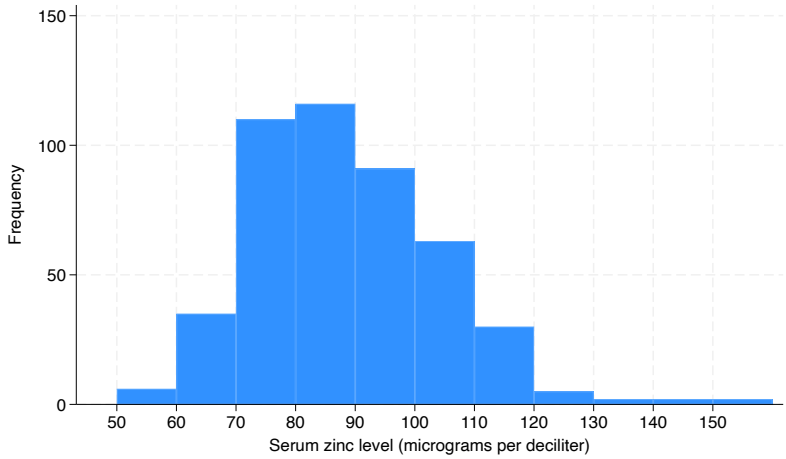
e. The output from Stata used to calculate these values is presented below. The mean is 87.9 μg/dl; the median is 6 μg/dl; the range is 153-50 = 103 μg/dl; the interquartile range is 98-76 = 22 μg/dl; the standard deviation is 16 μg/dl.

f. Since the distribution of serum zinc level is skewed, the median is a better measure of central tendency to report.

g. We would report the interquartile range and range.

Serum Zinc Level (μg/dl)	Relative Frequency
50–59	1.3%
60–69	7.6%
70–79	23.8%
80–89	25.1%
90–99	19.7%
100–109	13.6%
110–119	6.5%
120–129	1.1%
130–139	0.4%
140–149	0.4%
150–159	0.4%

```
histogram zinc, width(10) frequency xlabel(50 60 70 80 90 100 110 120 130 140 150, labels)
(bin=11, start=50, width=10)
```




```
sum zinc, detail
```

```
        Serum zinc level (micrograms per deciliter)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           56            50
 5%           64            51
10%           70            53       Obs                 462
25%           76            55       Sum of wgt.         462

50%           86                     Mean           87.93723
                      Largest        Std. dev.      16.00469
75%           98           142
90%          108           147       Variance       256.1501
95%          115           151       Skewness       .6211264
99%          135           153       Kurtosis       3.890067
```

Exercise 26
a. A table of absolute and relative frequencies of age at time of death is displayed below. The Stata code and output are also shown below.

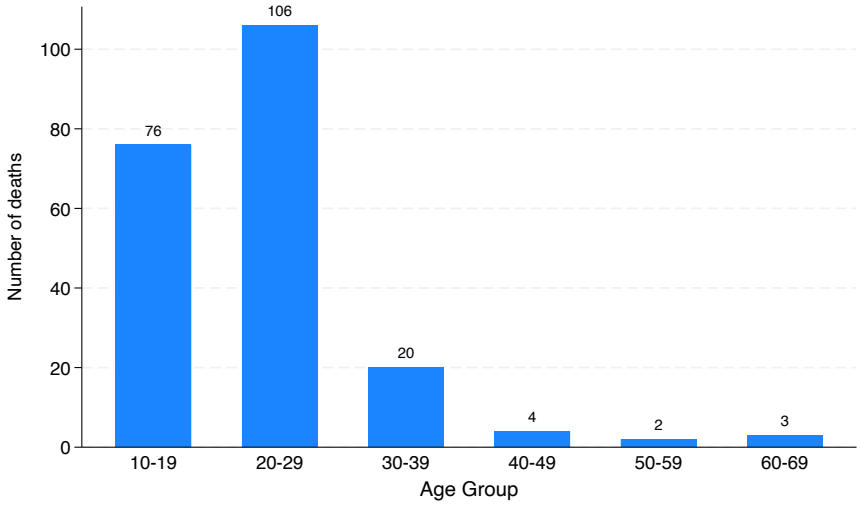
Age (Years)	Absolute Frequency	Relative Frequency
10–19	76	36.02%
20–29	106	50.24%
30–39	20	9.479%
40–49	4	1.896%
50–59	2	0.9479%
60–69	3	1.42%
Total	211	100%

```
table ( age_group ) () (), statistic(sumw) statistic(proportion)
```

```
------------------------------------------------------------
                            |  Sum of weights   Proportion
----------------------------+-------------------------------
Age group at time of death  |
  10-19                     |              76        .3602
  20-29                     |             106        .5024
  30-39                     |              20       .09479
  40-49                     |               4       .01896
  50-59                     |               2      .009479
  60-69                     |               3       .01422
  Total                     |             211            1
------------------------------------------------------------
```

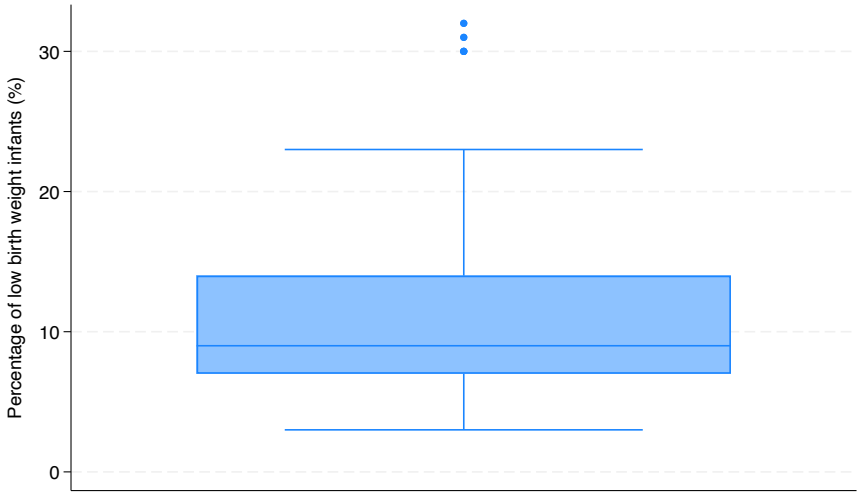
b. The table shows that most of the individuals who die while taking a selfie are between the ages of 10 and 29, with the decade 20-29 representing over 50% of deaths.
c. A bar chart of absolute number of deaths within each decade is displayed below.

```
graph bar (count), over(age_group) blabel(bar) ytitle(‘"Number of deaths"’)
b1title("Age Group")
```



Exercise 27
a. A box plot of the percentages of low birth weight infants is displayed below.

```
graph box lowbwt
```

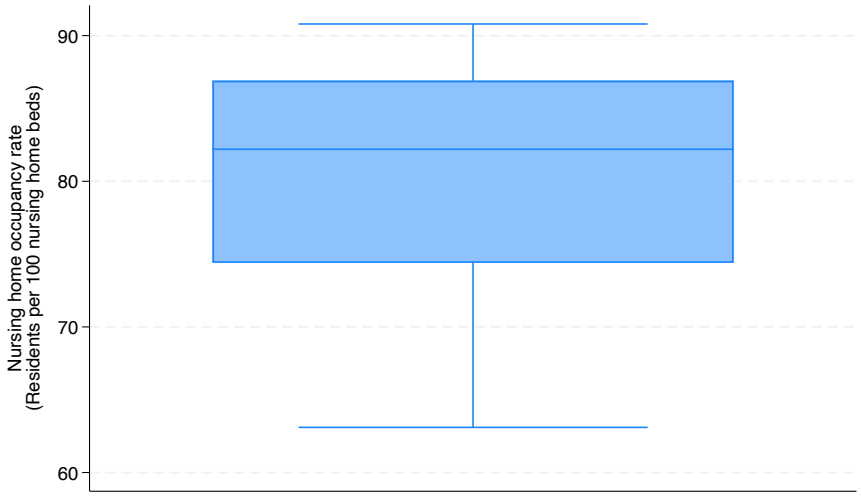


b. Albania has the lowest percentage of low birth weight infants (3%) and Yemen has the highest percentage of low birth weight infants (32%).
c. The distribution is skewed to the right; its tail extends in the direction of the higher values.
d. According to the box plot, the data contain three outlying observations. In Bangladesh and India 30% of infants are low birth weight, in Sudan 31% are low birth weight, and in Yemen 32% are low birth weight. Note that there are 3 values that are outliers, but there are 4 countries total since Bangladesh and India have the same value.

e. The mean is 10.85% and the median is 9%.
f. The median is the preferred measure of central tendency since the distribution of percentage of low birth weight infants is skewed.

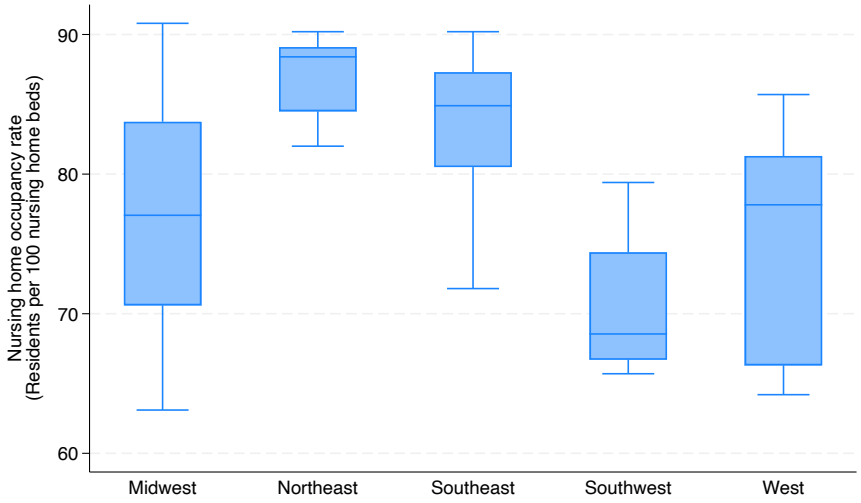
Exercise 28
a. Indiana has the smallest nursing home occupancy rate (63.1%), and South Dakota the highest number (90.8%). One factor that might influence the variability among states is the availability of support services for the elderly.
b. The box plot is shown below.

```
graph box occupancy, ytitle("Nursing home occupancy rate" "(Residents per 100 nursing home beds)")
```



c. The observations are skewed to the right. There is no state that can be categorized as an outlier.
d. The box plots are displayed below.

```
graph box occupancy, ytitle("Nursing home occupancy rate" "(Residents per 100 nursing home beds)")
over(region)
```

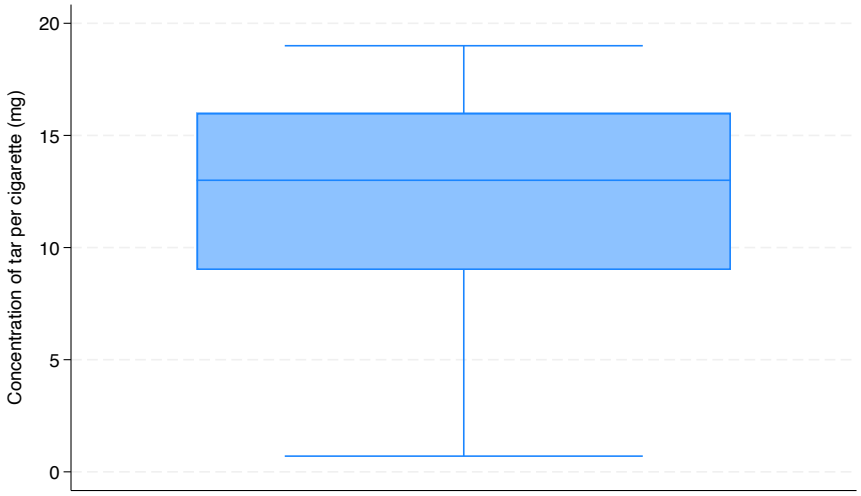


e. Yes, there are differences among the regions in terms of nursing home occupancy rate. The Northeast region has the highest median occupancy rate and the Southwest has the lowest median occupancy rate.

Exercise 29

a. A box plot of the declared concentrations of tar per cigarette is displayed below.

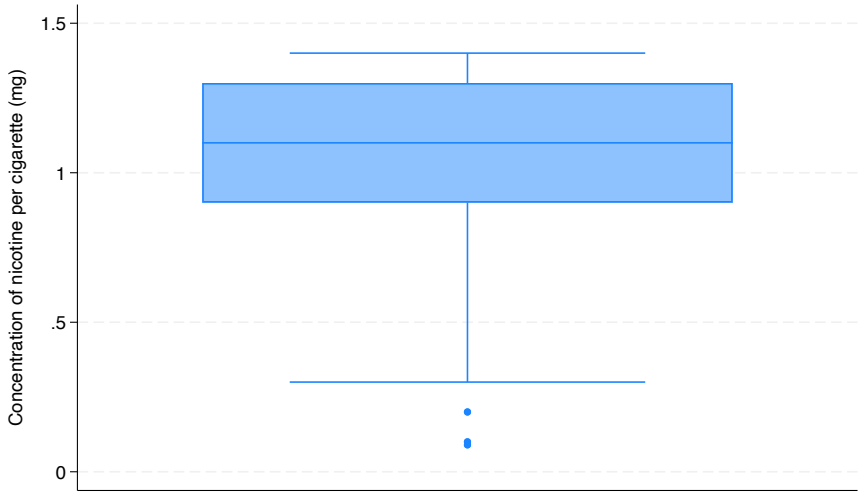
```
graph box tar
```



b. The distribution of tar concentrations is right-skewed with a median of approximately 13mg. There are no outliers.

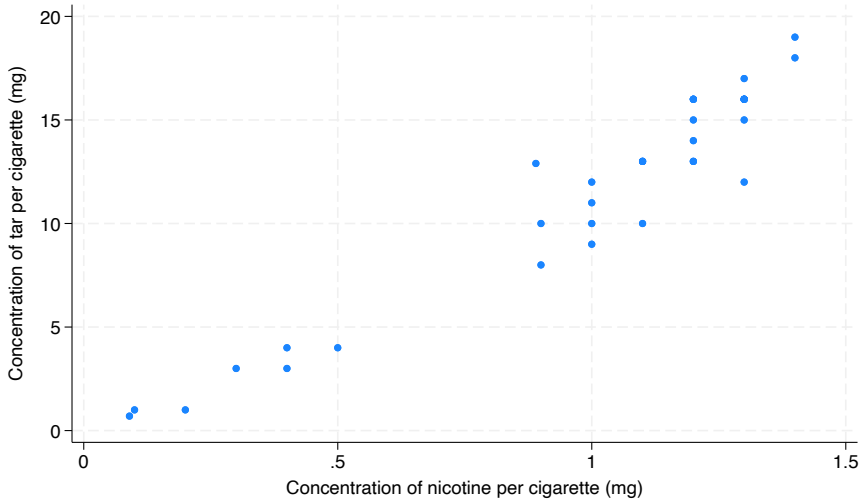
c. A box plot of the declared concentrations of nicotine per cigarette is displayed below.

```
graph box nicotine
```



d. The distribution of nicotine concentrations is right-skewed. However, there are three values that are low outliers.
e. The one-way scatter plot appears below.

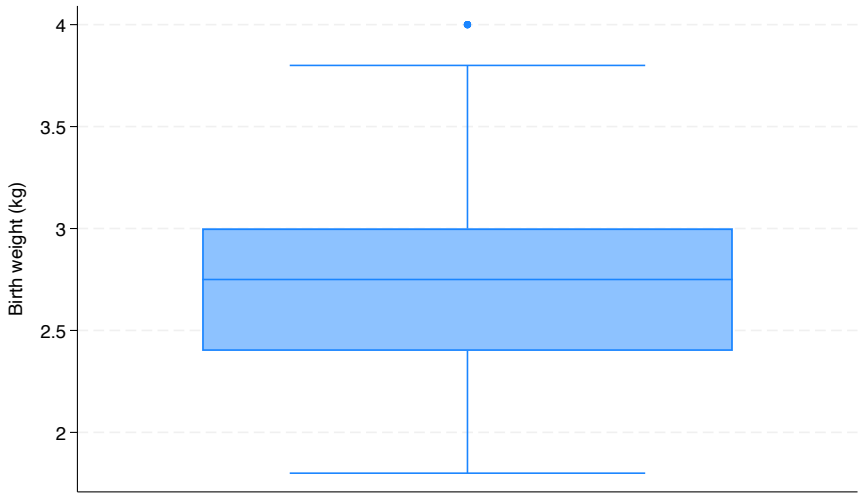
```
twoway (scatter tar nicotine)
```



f. There does appear to be a relationship between these quantities; the concentration of tar increases as the concentration of nicotine increases.

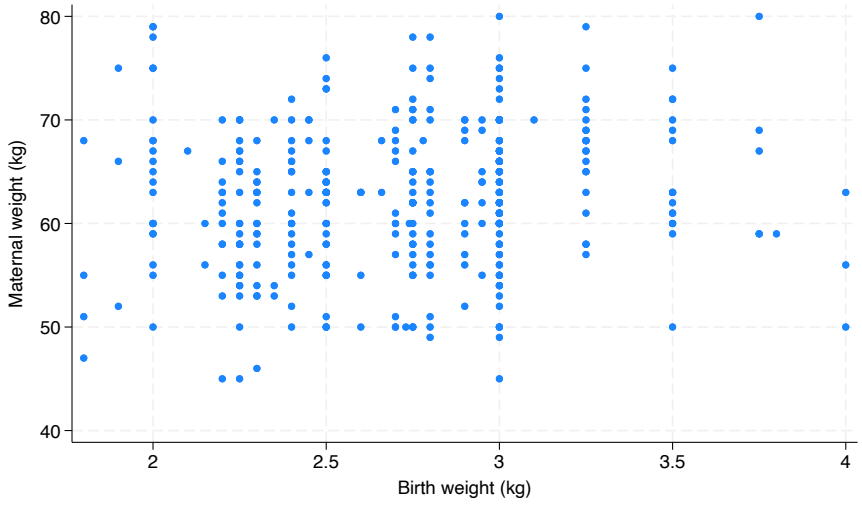
Exercise 30
a. Birth weight is continuous.
b. The box plot of birth weight is displayed below.

```
graph box bwt
```



c. Looking at the box plot from part (b), the 25th percentile is approximately 2.4 kilograms, the 50th percentile is approximately 2.75 kilograms, and the 75th percentile is approximately 3 kilograms.
d. Looking at the box plot from part (b), the minimum birth weight is approximately 1.75 kilograms and the maximum is approximately 4 kilograms. Yes, there is one outlying value at 4 kilograms.
e. The two-way scatter plot of infant birth weight versus the weight of the mother is displayed below.

```
twoway (scatter m_weight bwt)
```



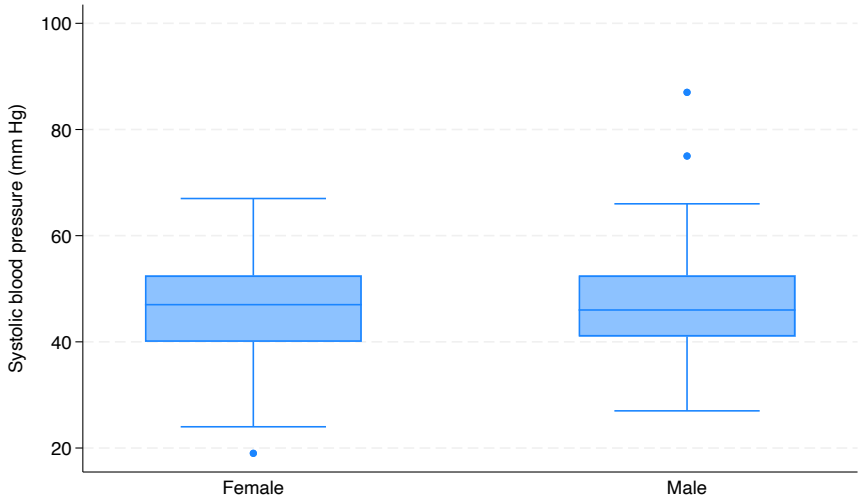
f. No, there does not seem to be a relationship between the weight of the infant and its mother's weight. There is no discernible pattern to the points.
g. Low birth weight status is discrete.
h. The table of low birth weight is presented below. The relative frequency of low birth weight is 29.25%.

```
tab low_bwt
```

Low birth weight < 2500 g	Freq.	Percent	Cum.
No	283	70.75	70.75
Yes	117	29.25	100.00
Total	400	100.00	

Exercise 31
a. Systolic blood pressure is continuous.
b. Sex is discrete.
c. The box plots of systolic blood pressure measurements are shown below. The two distributions of values are quite similar; in particular, the 25th, 50th, and 75th percentiles are nearly the same. The males have two high outlying values, while the females have one low outlier.

```
graph box sbp, over(sex)
```



d. For girls, the mean systolic blood pressure is $\bar{x}_f = 46.5$ mm Hg and the standard deviation is $s_f = 11.1$ mm Hg; for boys, the mean is $\bar{x}_m = 47.9$ mm Hg and the standard deviation is $s_m = 11.8$ mm Hg. Males have a slightly larger mean and standard deviation.

```
sort sex
by sex: summarize sbp

-> sex=   Female
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
     sbp |      56    46.46429    11.14526         19         67

-> sex=     Male
Variable |     Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
     sbp |      44    47.86364    11.80577         27         87
```

CHAPTER 3

Exercise 7
The statement is misleading. Although the number of deaths has been increasing, the population base could be increasing as well.

Exercise 8
a. The crude birth rate is

$$\frac{70,704}{6,863,560} = 10.3 \text{ per 1000 population.}$$

b. The crude death rate is

$$\frac{58,844}{6,863,560} = 8.6 \text{ per 1000 population.}$$

c. The infant mortality rate is

$$\frac{263}{70,704} = 3.7 \text{ per 1000 live births.}$$

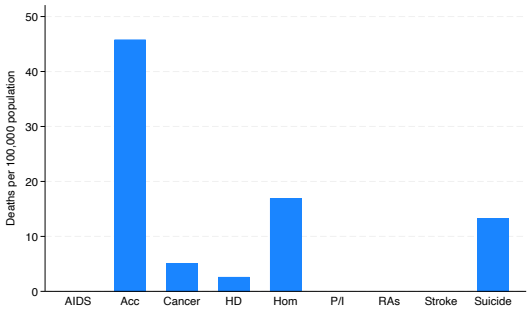
Exercise 9
a. The infant mortality rates for each category of mother's body mass index (BMI) appear in the table below.

Mother's BMI (kg/m^2)	Infant Mortality Rate (per 1000 live births)
<18.5	5.8
18.5–24.9	4.6
25.0–29.9	5.2
$\geq$ 30.0	7.1
Not stated	17.0

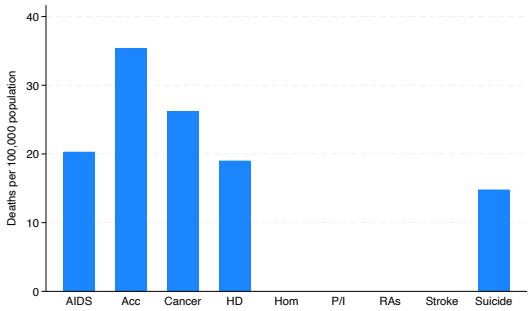
b. Infant mortality rate is the smallest for mother's in the Normal weight BMI category, and higher for those in the Underweight, Overweight, and Obese categories.
c. The infant mortality rate for children whose mother's BMI is in the Normal weight category is lower than the mortality rate for those whose mother's BMI is in a different category. Consequently, a large proportion of the infants whose mother's BMI is not stated are likely to have mothers in the Underweight, Overweight, and Obese categories, and perhaps more likely in the Obese category since this category has the highest infant mortality rate.

Exercise 10
a. The bar charts of death rates per 100,000 population by age group are displayed on the following page. Note that "Acc" represents accidents, "HD" represents heart disease, "Hom" represents homicide, "P/I" represents pneumonia/influenza, and "RAs" represents respiratory ailments.

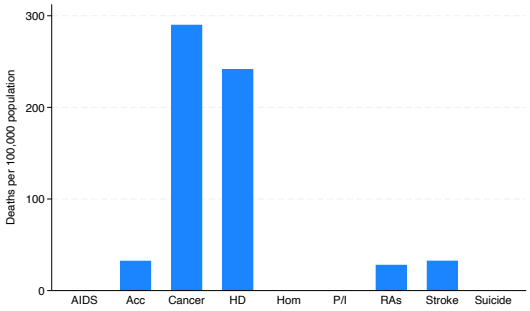
```
graph bar (asis) deaths if age == "15-24", over(cause) ytitle(‘"Deaths per 100,000 population"’)
graph bar (asis) deaths if age == "25-44", over(cause) ytitle(‘"Deaths per 100,000 population"’)
graph bar (asis) deaths if age == "45-64", over(cause) ytitle(‘"Deaths per 100,000 population"’)
graph bar (asis) deaths if age == "65+", over(cause) ytitle(‘"Deaths per 100,000 population"’)
```



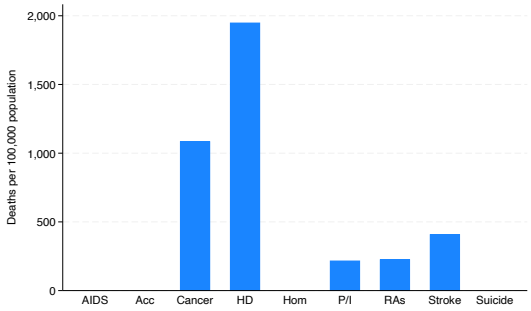
(a)



(b)



(c)



(d)

b. Cancer and heart disease are influential across all four age groups. Homicide and suicide affect primarily the younger groups; accidents also do not have a large impact on the elderly. Respiratory ailments, pneumonia/influenza and stroke affect the older groups. AIDS has a significant impact on those 25–44 years of age.

Exercise 11

a. First, sum all province populations presented in Column 2 and call this x. Then, sum the number of hospitals in each province and call this y. The aggregate country number in the last row will be $\frac{x}{y} = \frac{54{,}772{,}000}{544} = 1.00$.

b. First, sum the number of public hospitals in each province and call this x. Then, calculate the number of uninsured individuals in each province using Columns 7 and 2 and call this z, and sum these numbers and call this y:

$$\begin{aligned} z_i &= \texttt{\# uninsured in province } i = (100 - \texttt{Column 7})/100 * \texttt{Column 2} \\ y &= \sum_{i=1}^{9} z_i \end{aligned}$$

The country number (0.71 per 100,000 uninsured population) in the last row of Column 8 is $\frac{x}{\frac{y}{100{,}000}}$.

c. Similarly to answer (b) above, we can calculate the total number of private hospitals by summing the values in Column 4, and then use Columns 2 and 7 to calculate the total number of insured individuals across provinces and then divide this by 100,000. The country number (2.48 per 100,000 insured population) in the last row of Column 9 is then the total number of private hospitals divided by the total insured population divided by 100,000.
d. There is much more variation across provinces for the number of public hospitals per 100,000 uninsured in the population (Column 8) and much less across provinces for the number of private hospitals per 100,000 insured in the population (Column 9). This means individuals with insurance will have approximately the same number of hospitals available to them in any province, while the availability of hospitals can differ greatly by province for uninsured individuals.

Exercise 12
a. The rates of reported cases of polio per 100,000 children are 40.8 for the vaccine group and 80.5 for the placebo group.
The rates of true instances of polio per 100,000 children are 28.4 for the vaccine group and 70.6 for the placebo group.
The rates of incorrect diagnoses of polio per 100,000 children are 12.5 for the vaccine group and 9.9 for the placebo group.
The rates of paralytic disease per 100,000 children are 16.4 for the vaccine group and 57.1 for the placebo group.
The rates of nonparalytic disease per 100,000 children are 12.0 for the vaccine group and 13.4 for the placebo group.
b. The Salk vaccine appears to have help prevented cases of paralytic polio. It had no apparent effect on nonparalytic polio.

Exercise 13
Unlike the age-adjusted death rates, the crude death rates do not take into account the effect of age; therefore, the increasing crude rates reflect the fact that the population is growing older.

Exercise 14
a. The crude mortality rate for Maine is

$$\frac{11,082}{796,832} \quad = \quad 13.9 \text{ per } 1000 \text{ population},$$

and the crude rate for South Carolina is

$$\frac{22,401}{1,738,173} \quad = \quad 12.9 \text{ per } 1000 \text{ population}.$$

b. The proportions of the total population in each age group are displayed below.

	Proportion	
Age	Maine	South Carolina
0– 4	9.4%	11.8%
5– 9	9.3%	12.8%
10–14	8.6%	12.2%
15–19	7.6%	9.6%
20–24	13.3%	12.6%
25–34	12.7%	11.0%
35–44	11.3%	8.3%
45–54	10.0%	13.9%
55–64	9.1%	4.6%
65–74	5.8%	2.3%
75+	2.8%	1.0%
Total	100.0%	100.0%

The population in Maine is somewhat older than the population in South Carolina.
c. The age-specific mortality rates for each state are below.

	Death Rate per 1000 Population	
Age	Maine	South Carolina
0– 4	20.6	23.9
5– 9	1.9	1.9
10–14	1.4	1.8
15–19	2.2	4.3
20–24	3.7	6.5
25–34	3.9	8.7
35–44	5.5	12.4
45–54	10.8	19.9
55–64	20.4	33.1
65–74	52.2	61.5
75+	136.5	141.4

Disregarding the effect of infant mortality, mortality rate increases with age in each state.
d. It is necessary to control for the effect of age when comparing mortality rates in Maine and South Carolina. Age is a confounder; it is associated with both the population distribution and mortality rate.
e. The age-adjusted mortality rates are 12.0 per 1000 population for Maine and 17.2 per 1000 population for South Carolina.
f. The rates are now flipped; the crude mortality rate for Maine (13.9 per 1000 population) is higher than for South Carolina (12.9 per 1000 population), but the age-adjusted mortality rate for Maine (12 per 1000 population) is smaller than the age-adjusted mortality rate for South Carolina (17.2 per 1000 population). This is due to Maine having more older individuals in its population.

CHAPTER 4

Exercise 6
a. The proportion of individuals that will live to their 21st birthday is

$$\begin{aligned}\frac{l_{21}}{l_0} &= \frac{98,927}{100,000} \\ &= 0.989.\end{aligned}$$

b. The proportion of 21-year-olds that will live to their 40th birthday is

$$\begin{aligned}\frac{l_{40}}{l_{21}} &= \frac{96,537}{98,927} \\ &= 0.976.\end{aligned}$$

c. The proportion of 40-year-olds that will live to their 70th birthday is

$$\begin{aligned}\frac{l_{70}}{l_{40}} &= \frac{79,092}{96,537} \\ &= 0.819.\end{aligned}$$

c. The proportion of 70-year-olds that will live to their 90th birthday is

$$\begin{aligned}\frac{l_{90}}{l_{70}} &= \frac{28,285}{79,092} \\ &= 0.358.\end{aligned}$$

Exercise 7
a. Say you sell the policies to 346 50-year-olds who will pay you the first year on the policy. Using Halley's life table in Figure 4.3, we can see that 11 people died before reaching age 51 since there are only 335 51-year-olds. Therefore, those 11 individuals' inheritors will receive \$1,000 in the first year, and the remaining 335 will pay the second year. Similarly, 11 individuals' inheritors will receive \$1,000 in the second year and the remaining 324 will pay the third year, 11 individuals' inheritors will receive \$1,000 in the third year and the remaining 313 will pay the fourth year, and 11 individuals' inheritors will receive \$1,000 in the fourth year and the remaining 302 individuals will pay a fifth year. In the fifth year, 10 will die. So, you will pay out \$1,000 to $4 \times 11 + 10 = \$54,000$ to the inheritors, after having collected \$X (payment for a policy for a year) each year from $346 + 335 + 324 + 313 + 302 = 1,620$ payments. To break even, $1620 \times \$X = \$54,000$. Solving for X you will need to charge a 50-year-old \$33.34 for a 5-year policy in order to break even, and more than this to make a profit.
b. If you extend the calculations from part a. for another 5 years, you will pay inheritors \$104,000 while receiving $346 + 335 + 324 + 313 + 302 + 292 + 282 + 272 + 262 + 252 = 2,980$ payments. To break even, you will need to charge a 50-year-old \$34.90 for a 10-year policy.
c. Using Table 4.2 (the 2016 US life table), you would need to charge a 50-year-old \$4.50 for a 5-year policy and \$5.65 for a 10-year policy.

Exercise 8
The total person-years lived outside institutions from age x on (Column 7) is obtained by summing column 6 from the bottom up. The average remaining years outside institutions (Column 8) is calculated by dividing column 7 by column 2.

(7) Total Person-Years Outside Institutions From Age x On	(8) Average Remaining Years Outside of Institutions
7,005,116	70.1
5,532,705	56.5
4,562,720	47.3
2,671,609	28.9
982,595	13.6

Exercise 9
Life expectancy at birth has been increasing steadily since 1970, and has always been longer for females than for males for both races. Whites individuals have longer life expectancies than black individuals of the same gender.

Exercise 10
Individuals in the highest income quintile have longer life expectancies compared to individuals in lower income quintiles for both 1930 and 1960. However, the gap in life expectancy between higher-income individuals and those lower on the socioeconomic distribution expanded considerably between 1930 and 1960.

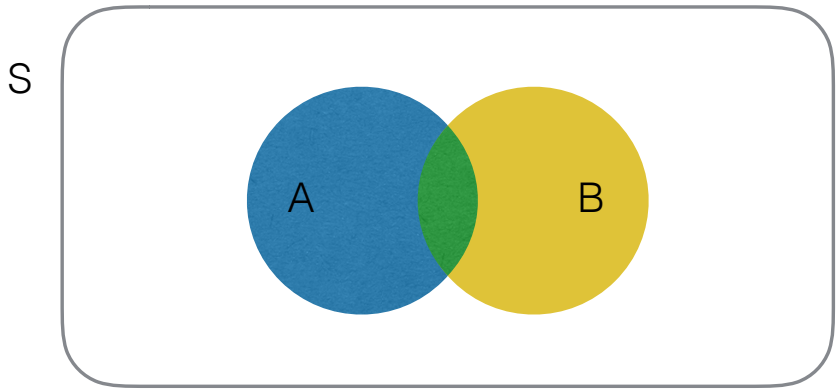
Exercise 11
a. At the ages when males have a higher rate of accidental and violent death than females, we would see higher age-specific death rates among the males.
b. Since e_x reflects the mortality rates for age x and beyond, we would expect the average life expectancies of males to be lower than those of females for these early age groups.

Exercise 12
Life expectancies at birth and at age 60 increased consistently over the centuries for both males and females. Life expectancies at age 80 also increased for females; for males, the life expectancy first dropped and then increased. Females have a longer life expectancy than males at each age and in every time period. The difference between genders is largest in the most recent time period, 1971–1975.

CHAPTER 5

Exercise 5
a. $A \cap B$ is the event that the individual is exposed to high levels of both carbon monoxide and nitrogen dioxide.
b. $A \cup B$ is the event that the individual is exposed to either carbon monoxide or nitrogen dioxide or both.
c. A^C is the event that the individual is not exposed to high levels of carbon monoxide.
d. The events A and B are not mutually exclusive.

Exercise 6
a. An example Venn diagram is displayed below.



b. Since

$$\begin{aligned} \mathrm{P}(A) \times \mathrm{P}(B) &= 0.131 \times 0.094 \\ &= 0.012 \\ &\neq \mathrm{P}(A \cap B) \\ &= 0.059, \end{aligned}$$

these two events are not independent.
c. The probability that A or B or both occur is

$$\begin{aligned} \mathrm{P}(A \cup B) &= \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) \\ &= 0.131 + 0.094 - 0.059 \\ &= 0.166. \end{aligned}$$

d. The probability that A occurs given that B occurs is

$$\begin{aligned} \mathrm{P}(A \mid B) &= \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \\ &= \frac{0.059}{0.094} \\ &= 0.628. \end{aligned}$$

Exercise 7

a.

$$\begin{aligned}\text{P(a newborn's gestational age is} \geq 37 \text{ weeks)} &= 1 - \text{P(a newborn's gestational age is} < 37 \text{ weeks)} \\ &= 1 - 0.131 \\ &= 0.869.\end{aligned}$$

b. Using the multiplicative rule of probability,

$$\begin{aligned}\text{P}(A_1 \cap B) &= \text{P}(B|A_1) \times \text{P}(A_1) \\ &= 0.448 \times 0.131 \\ &= 0.0587\end{aligned}$$

and

$$\begin{aligned}\text{P}(A_2 \cap B) &= \text{P}(B|A_2) \times \text{P}(A_2) \\ &= 0.041 \times 0.869 \\ &= 0.0356\end{aligned}$$

c. Using the total probability rule,

$$\begin{aligned}\text{P}(B) &= \text{P}(A_1 \cap B) + \text{P}(A_1 \cap B) \\ &= 0.059 + 0.036 \\ &= 0.094.\end{aligned}$$

Yes, this probability matches the probability given in Exercise 6 after rounding.

Exercise 8

a. The probability that a woman who gave birth in 2018 was 24 years of age or younger is

$$\begin{aligned}\text{P}(\leq 24) &= \text{P}(< 15 \text{ or } 15\text{–}19 \text{ or } 20\text{–}24) \\ &= \text{P}(< 15) + \text{P}(15\text{–}19) + \text{P}(20\text{–}24) \\ &= 0.0005 + 0.0474 + 0.1915 \\ &= 0.2394.\end{aligned}$$

b. The probability that the woman was 40 years of age or older is

$$\begin{aligned}\text{P}(\geq 40) &= \text{P}(40\text{–}44 \text{ or } 45\text{–}49) \\ &= \text{P}(40\text{–}44) + \text{P}(45\text{–}49) \\ &= 0.0310 + 0.0025 \\ &= 0.0335.\end{aligned}$$

c. Given that the woman was under 30 years of age, the probability that she was not yet 20 is

$$\text{P}(< 20 \,|< 30) = \frac{\text{P}(< 20 \text{ and } < 30)}{\text{P}(< 30)}$$

$$
\begin{aligned}
&= \frac{\mathrm{P}(< 20)}{\mathrm{P}(< 30)} \\
&= \frac{0.0005 + 0.0474}{0.0005 + 0.0474 + 0.1915 + 0.2899} \\
&= 0.0905.
\end{aligned}
$$

d. Given that the woman was 35 years of age or older, the probability that she was under 40 is

$$
\begin{aligned}
\mathrm{P}(< 40 \mid \geq 35) &= \frac{\mathrm{P}(< 40 \text{ and } \geq 35)}{\mathrm{P}(\geq 35)} \\
&= \frac{\mathrm{P}(35 - 39)}{\mathrm{P}(\geq 35)} \\
&= \frac{0.1495}{0.1495 + 0.0310 + 0.0025} \\
&= 0.8169.
\end{aligned}
$$

Exercise 9

a. The probability that the principal source of payment is private insurance is 0.300.

b. Since the sources of payment are all mutually exclusive, the probability that the principal source of payment is Medicare or Medicaid is

$$
\begin{aligned}
\mathrm{P}(\text{Medicare or Medicaid or other}) &= \mathrm{P}(\text{Medicare}) + \mathrm{P}(\text{Medicaid}) + \mathrm{P}(\text{other}) \\
&= 0.395 + 0.230 \\
&= 0.625.
\end{aligned}
$$

c. Given that the principal source of payment is a government program, the probability that it is Medicare is

$$
\begin{aligned}
\mathrm{P}(\text{Medicare} \mid \text{government program}) &= \frac{\mathrm{P}(\text{Medicare and government program})}{\mathrm{P}(\text{government program})} \\
&= \frac{\mathrm{P}(\text{Medicare})}{\mathrm{P}(\text{government program})} \\
&= \frac{0.395}{0.625} \\
&= 0.632.
\end{aligned}
$$

Exercise 10

a. Since these events are independent, the probability that both adults are uninsured is

$$
\begin{aligned}
\mathrm{P}(\text{both uninsured}) &= \mathrm{P}(\text{woman uninsured}) \times \mathrm{P}(\text{man uninsured}) \\
&= 0.085 \times 0.085 \\
&= 0.007.
\end{aligned}
$$

b. The probability that both adults are insured is

$$\begin{aligned} \text{P(both insured)} &= (1 - 0.085) \times (1 - 0.085) \\ &= 0.915 \times 0.915 \\ &= 0.837. \end{aligned}$$

c. The probability that all five adults are uninsured is $0.085 \times 0.085 \times 0.085 \times 0.085 \times 0.085 = 0.0000044$.

Exercise 11

a. The probability that a newborn infant will live to see his or her fifth birthday is

$$\begin{aligned} \text{P(newborn lives to age 5)} &= \frac{99,329}{100,000} \\ &= 0.993. \end{aligned}$$

b. The probability that a newborn infant will live to be 20 years of age is

$$\begin{aligned} \text{P(newborn lives to age 20)} &= \frac{98,997}{100,000} \\ &= 0.990. \end{aligned}$$

c. The probability that an individual who is 60 years old will survive for the next ten years is

$$\begin{aligned} \text{P(60-year-old lives to age 70)} &= \text{P(live to age 70} \mid \text{live to age 60)} \\ &= \frac{\text{P(live to age 70)}}{\text{P(live to age 60)}} \\ &= \frac{(79,092/100,000)}{(88,992/100,000)} \\ &= 0.889. \end{aligned}$$

d. Let A_1 represent the event that the first 60-year-old lives to age 70 and A_2 the event that the second 60-year-old lives to age 70. The probability that both will be alive on their 70th birthdays is

$$\begin{aligned} \text{P(both } A_1 \text{ and } A_2) &= \text{P}(A_1)\,\text{P}(A_2) \\ &= (0.889)(0.889) \\ &= 0.790. \end{aligned}$$

e. The probability that exactly one of the two 60-year-olds, but not both, will be alive at age 70 is

$$\begin{aligned} \text{P(either } A_1 \text{ or } A_2) &= \text{P}(A_1 \text{ and not } A_2 \text{ or } A_2 \text{ and not } A_1) \\ &= \text{P}(A_1 \text{ and not } A_2) + \text{P}(A_2 \text{ and not } A_1) \\ &= \text{P}(A_1)\,\text{P(not } A_2) + \text{P}(A_2)\,\text{P(not } A_1) \\ &= (0.889)(1 - 0.889) + (0.889)(1 - 0.889) \\ &= 2(0.889)(0.111) \\ &= 0.197. \end{aligned}$$

Exercise 12
Yes. This study only followed women who had given birth and not women in general. Therefore, the estimates of the probabilities of developing cancer only apply to women who have given birth.

Exercise 13
a. The relative risk of pregnancy during the first year of use for females who use male condoms as their method of contraception versus those who use withdrawal is calculated as

$$\begin{aligned} \text{RR} &= \frac{\text{P(pregnancy} \mid \text{male condoms)}}{\text{P(pregnancy} \mid \text{withdrawal)}} \\ &= \frac{0.126}{0.199} \\ &= 0.633. \end{aligned}$$

b. The table below shows the relative risks for each of the other methods of contraception versus withdrawal.

Method of Contraception	Relative Risk
Pill	0.362
All hormonals and IUD	0.302
Injectable	0.201

c. Each of the relative risks in the table takes a value less than 1, indicating that each of the methods carries a lower risk of pregnancy than withrawal. Injectable contraception has the smallest relative risk (least risk, relative to withdrawal), and male condoms the largest.

Exercise 14
a. The probabilities of suffering from persistent respiratory symptoms by socioeconomic status are shown below.

Socioeconomic Status	Probability
Low	0.392
Middle	0.238
High	0.141

b. Let S represent the presence of symptoms. The odds of experiencing persistent respiratory symptoms for the middle group relative to the high group are

$$\begin{aligned} \text{OR} &= \frac{\text{P}(S \mid \text{middle})/[1 - \text{P}(S \mid \text{middle})]}{\text{P}(S \mid \text{high})/[1 - \text{P}(S \mid \text{high})]} \\ &= \frac{(0.238)/(1 - 0.238)}{(0.141)/(1 - 0.141)} \\ &= 1.90, \end{aligned}$$

and for the low group relative to the high group are

$$\begin{aligned} \text{OR} &= \frac{\text{P}(S \mid \text{low})/[1 - \text{P}(S \mid \text{low})]}{\text{P}(S \mid \text{high})/[1 - \text{P}(S \mid \text{high})]} \\ &= \frac{(0.392)/(1 - 0.392)}{(0.141)/(1 - 0.141)} \\ &= 3.93. \end{aligned}$$

c. There does appear to be an association between socioeconomic status and respiratory symptoms; the odds of experiencing symptoms increase as socioeconomic status decreases.

Exercise 15
For mathematical simplicity, suppose the world's population is 100,000 and the number of confirmed cases is 100. If the US contained 4.3% of the world's population, but had 33% of the confirmed cases of COVID-19, the probability of developing COVID-19 given someone was living in the US is 33/4,300 = 0.0077. Similarly, the probability of developing COVID-19 given someone was living outside of the US is (100-33)/(100,000-4,300) = 67/95,700 = 0.0007. Therefore, the relative risk of developing COVID-19 for individuals in the United Stated versus the rest of the world is calculated as

$$\begin{aligned} \text{RR} &= \frac{\text{P}(\text{COVID-19} \mid \text{living in the US})}{\text{P}(\text{COVID-19} \mid \text{living outside of the US})} \\ &= \frac{0.0077}{0.0007} \\ &= 11. \end{aligned}$$

CHAPTER 6

Exercise 4

a. The probability of a false negative result is

$$\begin{aligned} \text{P}(-\text{ test} \mid \text{ca}) &= 1 - \text{sensitivity} \\ &= 1 - 0.869 \\ &= 0.131. \end{aligned}$$

b. The probability of a false positive result is

$$\begin{aligned} \text{P}(+\text{ test} \mid \text{no ca}) &= 1 - \text{specificity} \\ &= 1 - 0.889 \\ &= 0.111. \end{aligned}$$

c. Since $\text{P(ca)} = 0.0025$ and $\text{P(no ca)} = 0.9975$, the probability that a woman has breast cancer given that her mammogram is positive is

$$\begin{aligned} \text{P}(\text{ca} \mid +\text{ test}) &= \frac{\text{P(ca)P}(+\text{ test} \mid \text{ca})}{\text{P(ca)P}(+\text{ test} \mid \text{ca}) + \text{P(no ca)P}(+\text{ test} \mid \text{no ca})} \\ &= \frac{(0.0025)(0.869)}{(0.0025)(0.869) + (0.9975)(0.111)} \\ &= 0.0192. \end{aligned}$$

d. If we update P(ca) to be 0.025, then $\text{P(no ca)} = 0.975$, the probability that a woman has breast cancer given that her mammogram is positive is

$$\begin{aligned} \text{P}(\text{ca} \mid +\text{ test}) &= \frac{\text{P(ca)P}(+\text{ test} \mid \text{ca})}{\text{P(ca)P}(+\text{ test} \mid \text{ca}) + \text{P(no ca)P}(+\text{ test} \mid \text{no ca})} \\ &= \frac{(0.025)(0.869)}{(0.025)(0.869) + (0.975)(0.111)} \\ &= 0.167. \end{aligned}$$

When the prevalence increases, so does the probability that a female has breast cancer given that her mammogram is positive.

e. Since $\text{P(ca)} = 0.0025$ and $\text{P(no ca)} = 0.9975$, the probability that a woman does not have breast cancer given that her mammogram is negative is

$$\begin{aligned} \text{P}(\text{no ca} \mid -\text{ test}) &= \frac{\text{P(no ca)P}(-\text{ test} \mid \text{no ca})}{\text{P(no ca)P}(-\text{ test} \mid \text{no ca}) + \text{P(ca)P}(-\text{ test} \mid \text{ca})} \\ &= \frac{(0.9975)(0.889)}{(0.9975)(0.889) + (0.0025)(0.131)} \\ &= 0.9996. \end{aligned}$$

f. The positive likelihood ratio of the mammogram is

$$\frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{0.869}{1 - 0.889} = 7.83.$$

g. The negative likelihood ratio of the mammogram is

$$\frac{\text{specificity}}{1-\text{sensitivity}} = \frac{0.889}{1-0.869} = 6.79.$$

Exercise 5
a. Note that

$$\begin{aligned} \text{P}(+ \text{ test} \mid \text{cts}) &= \text{sensitivity} \\ &= 0.67 \end{aligned}$$

and

$$\begin{aligned} \text{P}(+ \text{ test} \mid \text{no cts}) &= 1 - \text{specificity} \\ &= 1 - 0.58 \\ &= 0.42. \end{aligned}$$

If the prevalence of carpal tunnel syndrome is 15%, then P(cts) = 0.15 and P(no cts) = 0.85. The predictive value of a positive test result is

$$\begin{aligned} \text{P}(\text{cts} \mid + \text{ test}) &= \frac{\text{P}(\text{cts})\,\text{P}(+ \text{ test} \mid \text{cts})}{\text{P}(\text{cts})\,\text{P}(+ \text{ test} \mid \text{cts}) + \text{P}(\text{no cts})\,\text{P}(+ \text{ test} \mid \text{no cts})} \\ &= \frac{(0.15)(0.67)}{(0.15)(0.67) + (0.85)(0.42)} \\ &= 0.22. \end{aligned}$$

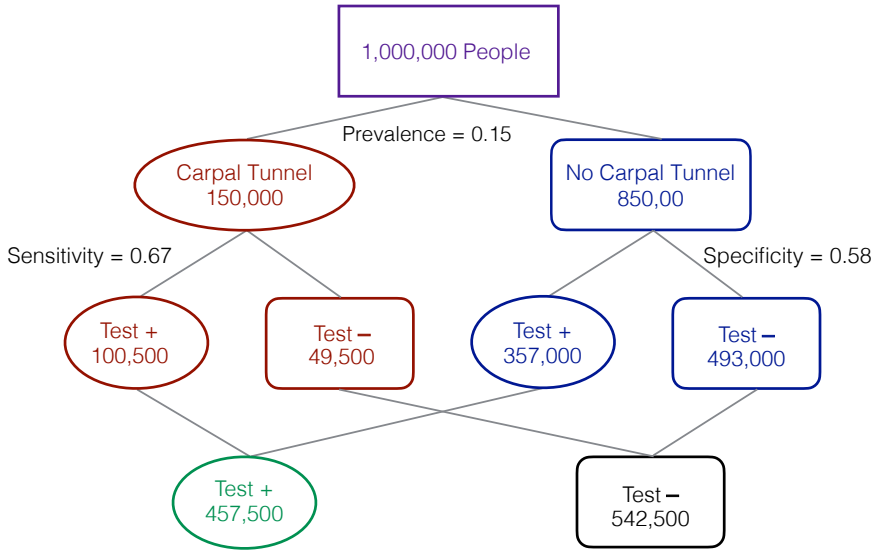
b. If the prevalence is 10%, then P(cts) = 0.10 and P(no cts) = 0.90. The predictive value of a positive test result is

$$\begin{aligned} \text{P}(\text{cts} \mid + \text{ test}) &= \frac{(0.10)(0.67)}{(0.10)(0.67) + (0.90)(0.42)} \\ &= 0.15. \end{aligned}$$

If the prevalence is 5%, then P(cts) = 0.05 and P(no cts) = 0.95, and the predictive value of a positive test result is

$$\begin{aligned} \text{P}(\text{cts} \mid + \text{ test}) &= \frac{(0.05)(0.67)}{(0.05)(0.67) + (0.95)(0.42)} \\ &= 0.08. \end{aligned}$$

As the prevalence of carpal tunnel syndrome decreases, the predictive value of a positive test decreases as well.
c. Note that in the textbook, Figure 6.3 is referenced in this question. However, this is a typo and should be Figure 6.2. A diagram illustrating the results of the diagnostic testing process is shown on the next page.



d. The positive likelihood ratio is

$$\frac{\text{sensitivity}}{1-\text{specificity}} = \frac{0.67}{1-0.58} = 1.60.$$

The negative likelihood ratio is

$$\frac{\text{specificity}}{1-\text{sensitivity}} = \frac{0.58}{1-0.67} = 1.76.$$

Exercise 6

a. The sensitivity of radionuclide ventriculography is

$$\begin{aligned} \text{P}(+\text{rv} \mid \text{cad}) &= \frac{302}{481} \\ &= 0.628, \end{aligned}$$

and its specificity is

$$\begin{aligned} \text{P}(-\text{rv} \mid \text{no cad}) &= \frac{372}{452} \\ &= 0.823. \end{aligned}$$

b. Since $\text{P}(\text{cad}) = 0.10$ and $\text{P}(\text{no cad}) = 0.90$, the probability that an individual has coronary artery disease given that he or she tests positive is

$$\begin{aligned} \text{P}(\text{cad} \mid +\text{rv}) &= \frac{\text{P}(+\text{rv} \mid \text{cad})\text{P}(\text{cad})}{\text{P}(+\text{rv} \mid \text{cad})\text{P}(\text{cad}) + \text{P}(+\text{rv} \mid \text{no cad})\text{P}(\text{no cad})} \\ &= \frac{(0.628)(0.10)}{(0.628)(0.10) + (0.177)(0.90)} \\ &= 0.283. \end{aligned}$$

c. The predictive value of a negative test is

$$\begin{aligned} \text{P(no cad} \mid -\text{rv}) &= \frac{\text{P}(-\text{rv} \mid \text{no cad})\text{P(no cad)}}{\text{P}(-\text{rv} \mid \text{no cad})\text{P(no cad)} + \text{P}(-\text{rv} \mid \text{cad})\text{P(cad)}} \\ &= \frac{(0.823)(0.90)}{(0.823)(0.90) + (0.372)(0.10)} \\ &= 0.952. \end{aligned}$$

Exercise 7

a. The sensitivity of of the PSA test is

$$\text{P}(+\text{PSA} \mid \text{cancer}) = 1 - \text{P}(-\text{PSA} \mid \text{cancer}) = 1 - 0.82 = 0.18,$$

and its specificity is

$$\text{P}(-\text{PSA} \mid \text{no cancer}) = 1 - \text{P}(+\text{PSA} \mid \text{no cancer}) = 1 - 0.02 = 0.98.$$

b. Since P(cancer) = 0.0001 and P(no cancer) = 0.9999, the predictive value of a positive test is

$$\begin{aligned} \text{P(cancer} \mid +\text{PSA}) &= \frac{\text{P}(+\text{PSA} \mid \text{cancer})\text{P(cancer)}}{\text{P}(+\text{PSA} \mid \text{cancer})\text{P(cancer)} + \text{P}(+\text{PSA} \mid \text{no cancer})\text{P(no cancer)}} \\ &= \frac{(0.18)(0.0001)}{(0.18)(0.0001) + (0.02)(0.9999)} \\ &= 0.0009. \end{aligned}$$

The predictive value of a negative test is

$$\begin{aligned} \text{P(no cancer} \mid -\text{PSA}) &= \frac{\text{P}(-\text{PSA} \mid \text{no cancer})\text{P(no cancer)}}{\text{P}(-\text{PSA} \mid \text{ no cancer})\text{P}(\text{ no cancer}) + \text{P}(-\text{PSA} \mid \text{cancer})\text{P(cancer)}} \\ &= \frac{(0.98)(0.9999)}{(0.98)(0.9999) + (0.82)(0.0001)} \\ &= 0.9999. \end{aligned}$$

c. The positive likelihood ratio is

$$\frac{\text{sensitivity}}{1 - \text{specificity}} = \frac{0.18}{1 - 0.98} = 9.$$

The negative likelihood ratio is

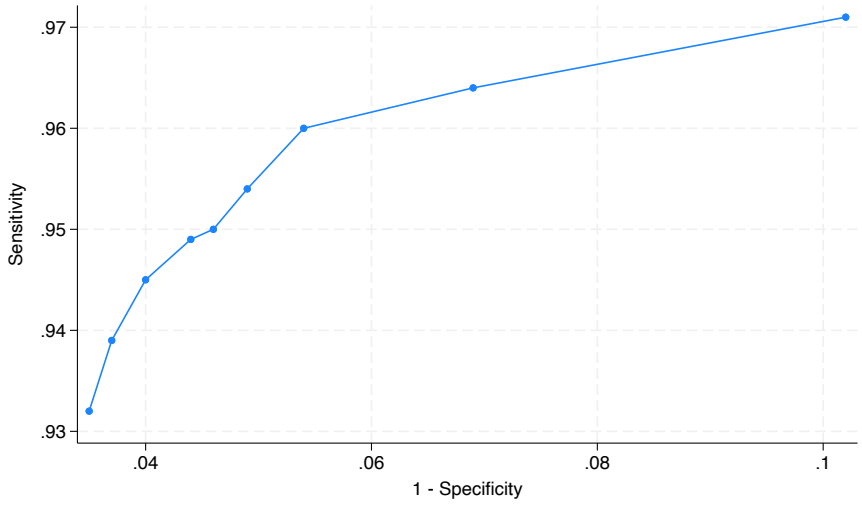
$$\frac{\text{specificity}}{1 - \text{sensitivity}} = \frac{0.98}{1 - 0.18} = 1.20.$$

Since the positive likelihood ratio is greater than the negative likelihood ratio , a positive PSA test provides more information about a patient.

Exercise 8

a. As the cutoff point is raised, the sensitivity decreases and the specificity increases.
b. As the cutoff point is raised, the probability of a false positive result decreases and the probability of a false negative result increases.
c. The ROC curve is shown on the next page.

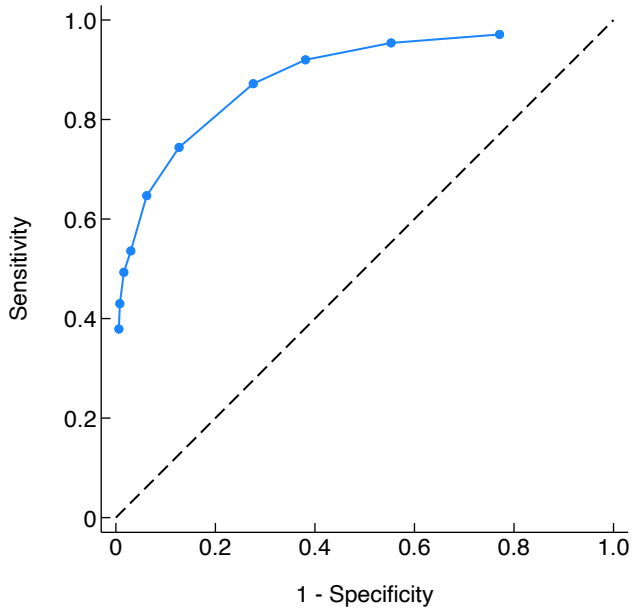
```
generate x = 1-specificity
twoway (function y = x, range(0 0.1)) (connected sensitivity x)
```



d. In this instance, the sensitivity and specificity will both be high no matter which cutoff value we select. A level of 9 ng/ml is probably best for maximizing sensitivity and specificity simultaneously; this point lies closest to the upper left-hand corner of the graph.
e. A false positive probability of 4% is equivalent to having 96% specificity. Looking at the table, 96% specificity corresponds to 94.5% sensitivity.

Exercise 9
a. As the cutoff point is raised, the sensitivity decreases and the specificity increases.
b. The ROC curve is shown on the next page.

```
generate x = 1-specificity
twoway (connected sensitivity x, msymbol(circle) lwidth(medium)) (function y = x,
range(0 1) lcolor(black) lwidth(medium) lpattern(dash)), ytitle(Sensitivity)
ytitle(, margin(medsmall)) ylabel(0 "0" .2 "0.2" .4 "0.4" .6 "0.6" .8 "0.8" 1 "1.0",
angle(horizontal) tposition(inside) nogrid) xtitle(1 - Specificity) xtitle(,
margin(medsmall)) xlabel(0 "0" .2 "0.2" .4 "0.4" .6 "0.6" .8 "0.8" 1 "1.0",
tposition(inside)) legend(off) aspect(1) xlabel(, nogrid)
```



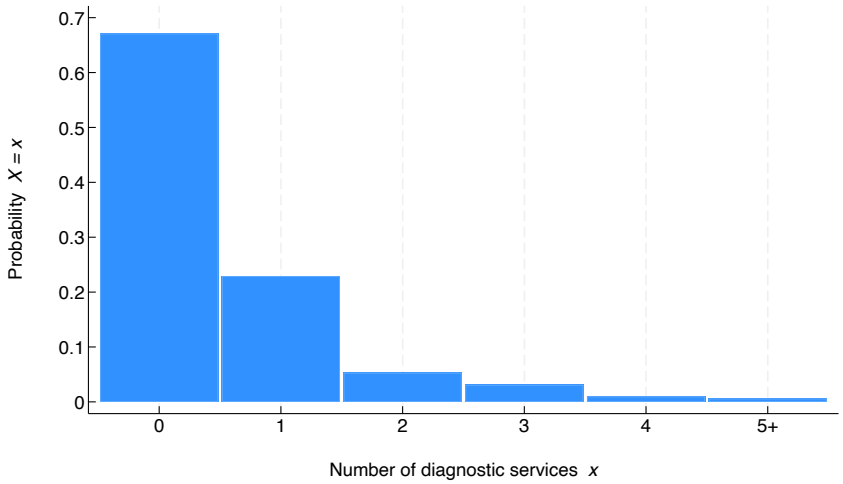
c. The point closest to the upper left-hand corner of the graph represents an FCG level of 5.6 mmol/liter; therefore, this does seem to be the best choice of a cutoff point.

CHAPTER 7

Exercise 6

a. The probability distribution of X is shown below.

```
histogram x [fweight = weights], width(1) start(-0.5) percent gap(3) ytitle(Probability {it: X = x})
ytitle(, margin(medium))  ylabel(0 "0" 10 "0.1" 20 "0.2" 30 "0.3" 40 "0.4" 50 "0.5" 60 "0.6" 70 "0.7",
angle(horizontal) tposition(inside) nogrid)  xtitle(Number of diagnostic services {it: x})
xtitle(, margin(medium)) xlabel(0 "0" 1 "1" 2 "2" 3 "3" 4 "4" 5 "5+", tposition(inside))
graphregion(margin(medium)) plotregion(margin(medsmall))
```



b. The probability that a child receives exactly 3 diagnostic services is 0.031.

c. The probability that a child receives at least 1 service is

$$\begin{aligned} \mathrm{P}(X \geq 1) &= 1 - \mathrm{P}(X = 0) \\ &= 1 - 0.671 \\ &= 0.329. \end{aligned}$$

The probability that a child receives 4 or more services is

$$\begin{aligned} \mathrm{P}(X \geq 4) &= \mathrm{P}(X = 4) + \mathrm{P}(X = 5^{+}) \\ &= 0.010 + 0.006 \\ &= 0.016. \end{aligned}$$

d. The probability that a child receives exactly 3 services given that he or she receives at least 1 service is

$$\begin{aligned} \mathrm{P}(X = 3 \mid X \geq 1) &= \frac{\mathrm{P}(X = 3)}{\mathrm{P}(X \geq 1)} \\ &= \frac{0.031}{0.329} \\ &= 0.094. \end{aligned}$$

Exercise 7

a. The probability that a child is its mother's fourth child is

$$\mathrm{P}(X = 4) = \mathrm{P}(4) \approx 0.075.$$

b. The probability that a child is its mother's first or second child is

$$\mathrm{P}(X = 1 \text{ or } X = 2) = \mathrm{P}(1) + \mathrm{P}(2) \approx 0.38 + 0.32 = 0.70.$$

c. The probability that a child is its mother's third child or higher is

$$\begin{aligned} \mathrm{P}(X \geq 3) &= \mathrm{P}(3) + \mathrm{P}(4) + \mathrm{P}(5) + \mathrm{P}(6) + \mathrm{P}(7) + \mathrm{P}(8+) \\ &= 1 - \mathrm{P}(1) - \mathrm{P}(2) \\ &\approx 1 - 0.38 - 0.32 = 0.30. \end{aligned}$$

Exercise 8

It is unlikely that X has a binomial distribution. While there are a fixed number of trials ($n = 7$) that each result in one of two mutually exclusive outcomes, the outcomes of the trials are not independent. If the concentration of carbon monoxide is exceptionally high one day, the pollution will not all disappear over night; the concentration is more likely to be high the next day as well.

Exercise 9

a. The seven individuals can be ordered in $7! = 5040$ ways.

b. Four individuals can be selected from the group in

$$\begin{aligned} \binom{7}{4} &= \frac{7!}{4!\,(7-4)!} \\ &= 35 \end{aligned}$$

different ways.

c. The probability that exactly two of the seven individuals suffer from diabetes is

$$\begin{aligned} \mathrm{P}(\text{two diabetics}) &= \binom{7}{2} (0.252)^2 (0.748)^{7-2} \\ &= 0.312. \end{aligned}$$

d. The probability that at most two of the seven have been diagnosed with diabetes is

$$\begin{aligned} \mathrm{P}(\text{at most two}) &= \mathrm{P}(\text{none}) + \mathrm{P}(\text{one}) + \mathrm{P}(\text{two}) \\ &= \binom{7}{0} (0.252)^0 (0.748)^7 + \binom{7}{1} (0.252)^1 (0.748)^6 \\ &\quad + \binom{7}{2} (0.252)^2 (0.748)^5 \\ &= 0.752. \end{aligned}$$

e. The probability that exactly four of the seven persons suffer from diabetes is

$$\begin{aligned} \mathrm{P}(\text{four diabetics}) &= \binom{7}{4} (0.252)^4 (0.748)^{7-4} \\ &= 0.059. \end{aligned}$$

Exercise 10

a. The ten persons can be ordered in $10! = 3,628,800$ different ways.

b. Four individuals can be selected in

$$\begin{aligned}\binom{10}{4} &= \frac{10!}{4!\,6!} \\ &= 210\end{aligned}$$

different ways.

c. The probability that exactly three of the ten individuals are left-handed is

$$\begin{aligned}\text{P(three left-handers)} &= \binom{10}{3}(0.098)^3(0.902)^{10-3} \\ &= 0.055.\end{aligned}$$

d. The probability that at least six of the ten persons are left-handed is

$$\begin{aligned}\text{P(at least six)} &= \text{P(six)} + \text{P(seven)} + \text{P(eight)} + \text{P(nine)} + \text{P(ten)} \\ &= \binom{10}{6}(0.098)^6(0.902)^4 + \binom{10}{7}(0.098)^7(0.902)^3 \\ &\quad + \binom{10}{8}(0.098)^8(0.902)^2 + \binom{10}{9}(0.098)^9(0.902)^1 \\ &\quad + \binom{10}{10}(0.098)^{10}(0.902)^0 \\ &= 0.0001.\end{aligned}$$

e. The probability that at most two individuals are left-handed is

$$\begin{aligned}\text{P(at most two)} &= \text{P(none)} + \text{P(one)} + \text{P(two)} \\ &= \binom{10}{0}(0.098)^0(0.902)^{10} + \binom{10}{1}(0.098)^1(0.902)^9 \\ &\quad + \binom{10}{2}(0.098)^2(0.902)^8 \\ &= 0.933.\end{aligned}$$

Exercise 11

a. Since the number of individuals in the sample who watch television for three or more hours follows a binomial distribution, the mean number per sample is $np = 20(0.207) = 4.14$ and the standard deviation is $\sqrt{np(1-p)} = \sqrt{20(0.207)(0.793)} = 1.81$.

b. The probability of obtaining results as extreme as or more extreme than those observed is

$$\begin{aligned}\text{P(18 or more)} &= \text{P(18)} + \text{P(19)} + \text{P(20)} \\ &= \binom{20}{18}(0.207)^{18}(0.793)^2 + \binom{20}{19}(0.207)^{19}(0.793)^1 \\ &\quad + \binom{20}{20}(0.207)^{20}(0.793)^0 \\ &\approx 0.00.\end{aligned}$$

c. The probability of obtaining results as extreme as or more extreme than those observed is

$$\begin{aligned}
\text{P(18 or more)} &= \text{P(18)} + \text{P(19)} + \text{P(20)} \\
&= \binom{20}{8}(0.207)^8(0.793)^1 2 + \binom{20}{9}(0.207)^9(0.793)^{11}\ldots \\
&\quad + \binom{20}{20}(0.207)^{20}(0.793)^0 \\
&= 0.039.
\end{aligned}$$

Exercise 12

a. The probability that exactly one case of tetanus will be reported is

$$\begin{aligned}
\text{P(exactly one case)} &= \frac{e^{-4.5}(4.5)^1}{1!} \\
&= 0.050.
\end{aligned}$$

b. The probability that at most two cases will be reported is

$$\begin{aligned}
\text{P(at most two cases)} &= \text{P(none)} + \text{P(one)} + \text{P(two)} \\
&= \frac{e^{-4.5}(4.5)^0}{0!} + \frac{e^{-4.5}(4.5)^1}{1!} + \frac{e^{-4.5}(4.5)^2}{2!} \\
&= 0.174.
\end{aligned}$$

c. The probability that four or more cases will be reported is

$$\begin{aligned}
\text{P(four or more cases)} &= 1 - \text{P(three or fewer cases)} \\
&= 1 - [\text{P(three)} + \text{P(two)} + \text{P(one)} + \text{P(none)}] \\
&= 1 - \left[\frac{e^{-4.5}(4.5)^3}{3!} + 0.174\right] \\
&= 1 - 0.343 \\
&= 0.657.
\end{aligned}$$

d. The mean number of cases of tetanus is $\lambda = 4.5$. The standard deviation is $\sqrt{\lambda} = \sqrt{4.5} = 2.1$.

Exercise 13

a. The probability that no suicides will be reported is

$$\begin{aligned}
\text{P(no suicides)} &= \frac{e^{-2.75}(2.75)^0}{0!} \\
&= 0.064.
\end{aligned}$$

b. The probability that at most four suicides will be reported is

$$\begin{aligned}
\text{P(at most four)} &= \text{P(none)} + \text{P(one)} + \text{P(two)} + \text{P(three)} + \text{P(four)} \\
&= \frac{e^{-2.75}(2.75)^0}{0!} + \frac{e^{-2.75}(2.75)^1}{1!} + \frac{e^{-2.75}(2.75)^2}{2!} \\
&\quad + \frac{e^{-2.75}(2.75)^3}{3!} + \frac{e^{-2.75}(2.75)^4}{4!} \\
&= 0.855.
\end{aligned}$$

c. The probability that six or more suicides will be reported is

$$\begin{aligned}
\text{P(six or more)} &= 1 - \text{P(less than six)} \\
&= 1 - [\text{P(none)} + \text{P(one)} + \text{P(two)} + \text{P(three)} \\
&\quad + \text{P(four)} + \text{P(five)}] \\
&= 1 - \left[0.855 + \frac{e^{-2.75}(2.75)^5}{5!}\right] \\
&= 1 - 0.939 \\
&= 0.061.
\end{aligned}$$

Exercise 14

a. Use the Poisson approximation to the binomial distribution. The mean number of infants who would die in the first year is $\lambda = np = 2000(0.0059) = 11.8$.

b. The probability that at most 5 infants die in the first year of life is

$$\begin{aligned}
\text{P(at most 5)} &= \text{P}(0) + \text{P}(1) + \text{P}(2) + \text{P}(3) + \text{P}(4) + \text{P}(5) \\
&= \frac{e^{-11.8}(11.8)^0}{0!} + \frac{e^{-11.8}(11.8)^1}{1!} + \frac{e^{-11.8}(11.8)^2}{2!} + \frac{e^{-11.8}(11.8)^3}{3!} \\
&\quad + \frac{e^{-11.8}(11.8)^4}{4!} + \frac{e^{-11.8}(11.8)^5}{5!} \\
&= 0.02304.
\end{aligned}$$

c. The probability that between 15 and 20 infants die in the first year of life is

$$\begin{aligned}
\text{P(between 15 and 20)} &= \text{P}(15) + \text{P}(16) + \text{P}(17) + \text{P}(18) + \text{P}(19) + \text{P}(20) \\
&= \frac{e^{-11.8}(11.8)^{15}}{15!} + \frac{e^{-11.8}(11.8)^{16}}{16!} + \frac{e^{-11.8}(11.8)^{17}}{17!} \\
&\quad + \frac{e^{-11.8}(11.8)^{18}}{18!} + \frac{e^{-11.8}(11.8)^{19}}{19!} + \frac{e^{-11.8}(11.8)^{20}}{20!} \\
&= 0.13169.
\end{aligned}$$

Exercise 15

a. The probability that z is greater than 2.60 is 0.005.

b. The probability that z is less than 1.35 is $1 - 0.089 = 0.911$.

c. The probability that z is between -1.70 and 3.10 is $1 - 0.045 - 0.001 = 0.954$.

d. The value $z = 1.04$ cuts off the upper 15% (actually, 14.9%) of the standard normal distribution.

e. The value $z = -0.84$ cuts off the lower 20% of the distribution.

Exercise 16

a. The probability that a randomly selected woman has a diastolic blood pressure less than 60 mm Hg is

$$\text{P}(X < 60) = \text{P}\left(\frac{X - 77}{11.6} < \frac{60 - 77}{11.6}\right)$$

$$\begin{aligned} &= \mathrm{P}(Z < -1.47) \\ &= 0.071. \end{aligned}$$

b. The probability that she has a diastolic blood pressure greater than 90 mm Hg is

$$\begin{aligned} \mathrm{P}(X > 90) &= \mathrm{P}\left(\frac{X-77}{11.6} > \frac{90-77}{11.6}\right) \\ &= \mathrm{P}(Z > 1.12) \\ &= 0.131. \end{aligned}$$

c. The probability that she has a diastolic blood pressure between 60 and 90 mm Hg is

$$\begin{aligned} \mathrm{P}(60 \leq X \leq 90) &= 1 - \mathrm{P}(X < 60) - \mathrm{P}(X > 90) \\ &= 1 - 0.071 - 0.131 \\ &= 0.798. \end{aligned}$$

d. A woman with a z-score of 2.15 tells us that her diastolic blood pressure is 101.94 mm Hg. This can be found by solving the following equation for X:

$$\begin{aligned} 2.15 &= \frac{X-77}{11.6} \\ X &= 2.15(11.6) + 77 \\ X &= 101.94. \end{aligned}$$

Exercise 17

a. The z-score associated with a weight of 130 pounds is

$$\frac{130 - 172.2}{29.8} = -1.42$$

b. The probability that a randomly selected man weighs less than 130 pounds is

$$\begin{aligned} \mathrm{P}(X < 130) &= \mathrm{P}\left(\frac{X-172.2}{29.8} < \frac{130-172.2}{29.8}\right) \\ &= \mathrm{P}(Z < -1.42) \\ &= 0.078. \end{aligned}$$

c. The probability that he weighs more than 210 pounds is

$$\begin{aligned} \mathrm{P}(X > 210) &= \mathrm{P}\left(\frac{X-172.2}{29.8} > \frac{210-172.2}{29.8}\right) \\ &= \mathrm{P}(Z > 1.27) \\ &= 0.102. \end{aligned}$$

d. First, the probability that a male weighs less than 130 pounds or more than 210 pounds is $\mathrm{P}(X < 130) + \mathrm{P}(X > 210) = 0.078 + 0.102 = 0.18$. We now have the value of p and can use the binomial distribution to calculate the probability as

$$\mathrm{P}(X = 2) = \binom{5}{2} (0.18)^2 (0.82)^3 = 0.178.$$

e. Among five males selected at random, the probability that at least one will have a weight outside the range 130 to 210 pounds is

$$\begin{aligned}
\text{P(at least one} < 130 \text{ or} > 210) &= 1 - \text{P(none} < 130 \text{ or} > 210) \\
&= 1 - \text{P(all between 130 and 210)} \\
&= 1 - [\text{P}(130 \leq X \leq 210)]^5 \\
&= 1 - [1 - 0.078 - 0.102]^5 \\
&= 1 - (0.820)^5 \\
&= 0.629.
\end{aligned}$$

Exercise 18

a. Since the mean of this distribution is 100, the probability that an adult selected from the general population has an IQ score above 100 is 0.50.

b. The probability that an adult has an IQ score above 130 is

$$\begin{aligned}
\text{P}(X > 130) &= \text{P}\left(\frac{X - 100}{15} > \frac{130 - 100}{15}\right) \\
&= \text{P}(Z > 2) \\
&= 0.02275.
\end{aligned}$$

c. The probability that an adult has an IQ score below 80 is

$$\begin{aligned}
\text{P}(X < 80) &= \text{P}\left(\frac{X - 100}{15} < \frac{80 - 100}{15}\right) \\
&= \text{P}(Z < -1.33) \\
&= 0.09176.
\end{aligned}$$

d. The probability that an adult has an IQ score between 85 and 115 is

$$\begin{aligned}
\text{P}(85 \leq X \leq 115) &= 1 - \text{P}(X < 85) - \text{P}(X > 115) \\
&= 1 - \text{P}\left(\frac{X - 100}{15} < \frac{85 - 100}{15}\right) - \text{P}\left(\frac{X - 100}{15} > \frac{115 - 100}{15}\right) \\
&= 1 - \text{P}(Z < -1) - \text{P}(Z > 1) \\
&= 1 - 0.15866 - 0.15866 \\
&= 0.68268.
\end{aligned}$$

Exercise 19

a. The probability of correctly predicting coronary heart disease for a man who will develop it is

$$\begin{aligned}
\text{P}(X_d \geq 260) &= \text{P}\left(\frac{X_d - 244}{51} \geq \frac{260 - 244}{51}\right) \\
&= \text{P}(Z \geq 0.31) \\
&= 0.378.
\end{aligned}$$

b. The probability of predicting heart disease for a man who will not develop it, or the probability of a false positive, is

$$\begin{aligned} \mathrm{P}(X_{nd} \geq 260) &= \mathrm{P}\left(\frac{X_{nd} - 219}{41} \geq \frac{260 - 219}{41}\right) \\ &= \mathrm{P}(Z \geq 1.00) \\ &= 0.159. \end{aligned}$$

c. The probability of failing to predict heart disease for a man who will develop it, or the probability of a false negative, is

$$\begin{aligned} \mathrm{P}(X_d < 260) &= 1 - \mathrm{P}(X_d \geq 260) \\ &= 1 - 0.378 \\ &= 0.622. \end{aligned}$$

d. If the cutoff point is lowered to 250 mg/100 ml, the probability of a false positive result would increase while the probability of a false negative would decrease.
e. Initial serum cholesterol level is not very useful for predicting coronary heart disease in this population. Because the normal curves for men who develop disease and those who do not have a great deal of overlap, the probabilities of false positive and false negative outcomes are both very high.

CHAPTER 8

Exercise 8

a. The mean of the sample means would be $\mu = 29.5$ mg/100 ml.

b. The standard deviation would be $\sigma/\sqrt{n} = 9.25/\sqrt{20} = 2.07$ mg/100 ml. This standard deviation is also called the standard error of the mean.

c. The standard deviation of the sample means is smaller than the standard deviation of the albumin levels themselves, by a factor of $1/\sqrt{20}$.

d. The sample means would be approximately normally distributed.

e. The proportion of the means that are larger than 33 mg/100 ml is

$$\begin{aligned} \mathrm{P}(\overline{X} > 33) &= \mathrm{P}\left(\frac{\overline{X} - 29.5}{2.07} > \frac{33 - 29.5}{2.07}\right) \\ &= \mathrm{P}(Z > 1.69) \\ &= 0.046 \\ &= 4.6\%. \end{aligned}$$

f. The proportion of the means that are less than 28 mg/100 ml is

$$\begin{aligned} \mathrm{P}(\overline{X} < 28) &= \mathrm{P}\left(\frac{\overline{X} - 29.5}{2.07} < \frac{28 - 29.5}{2.07}\right) \\ &= \mathrm{P}(Z < -0.72) \\ &= 0.236 \\ &= 23.6\%. \end{aligned}$$

g. The proportion of the means that are between 29 and 31 mg/100 ml is

$$\begin{aligned} \mathrm{P}(29 < \overline{X} < 31) &= \mathrm{P}\left(\frac{29 - 29.5}{2.07} < \frac{\overline{X} - 29.5}{2.07} < \frac{31 - 29.5}{2.07}\right) \\ &= \mathrm{P}(-0.24 < Z < 0.72) \\ &= 1 - 0.236 - 0.405 \\ &= 0.359 \\ &= 35.9\%. \end{aligned}$$

Exercise 9

a. The distribution of means of samples of size 10 has mean $\mu = 0$, standard error $\sigma/\sqrt{n} = 1/\sqrt{10} = 0.32$, and is normally distributed. (Since the underlying population is itself normal, this is true for any sample size n.)

b. The proportion of means that are greater than 0.60 is

$$\begin{aligned} \mathrm{P}(\overline{X} > 0.60) &= \mathrm{P}\left(\frac{\overline{X} - 0}{0.32} > \frac{0.60 - 0}{0.32}\right) \\ &= \mathrm{P}(Z > 1.87) \\ &= 0.031 \\ &= 3.1\%. \end{aligned}$$

c. The proportion of means that are less than -0.75 is

$$\begin{aligned} \mathrm{P}(\overline{X} < 0.75) &= \mathrm{P}\left(\frac{\overline{X} - 0}{0.32} < \frac{0.75 - 0}{0.32}\right) \\ &= \mathrm{P}(Z < -2.34) \\ &= 0.010 \\ &= 1.0\%. \end{aligned}$$

d. The value $Z = 0.84$ cuts off the upper 20% of the standard normal distribution. Therefore, $\overline{X} = 0.84(0.32) + 0 = 0.27$ cuts off the upper 20% of the distribution of sample means.
e. The value $Z = -1.28$ cuts off the lower 10% of the standard normal distribution, and $\overline{X} = -1.28(0.32) + 0 = -0.41$ cuts off the lower 10% of the distribution of sample means.

Exercise 10
The distribution of means of samples of size 40 has mean $\mu = 1.81$ μg/m^3, standard error $\sigma/\sqrt{n} = 2.25/\sqrt{40} = 0.36$ μg/m^3, and, assuming that $n = 40$ is large enough, is approximately normally distributed. The central limit theorem applies even though the underlying population of measurements is skewed to the right.

Exercise 11
a. The probability that the newborn's birth weight is less than 2500 grams is

$$\begin{aligned} \mathrm{P}(X < 2500) &= \mathrm{P}\left(\frac{X - 3500}{430} < \frac{2500 - 3500}{430}\right) \\ &= \mathrm{P}(Z < -2.34) \\ &= 0.010. \end{aligned}$$

b. The value $Z = -1.645$ cuts off the lower 5% of the standard normal curve. Therefore, $X = (-1.645)(430) + 3500 = 2793$ cuts off the lower 5% of the distribution of birth weights.
c. The distribution of means of samples of size 5 had mean $\mu = 3500$ grams, standard error $\sigma/\sqrt{n} = 430/\sqrt{5} = 192$ grams, and is approximately normally distributed.
d. The value $\overline{X} = (-1.645)(192) + 3500 = 3184$ cuts off the lower 5% of the distribution of samples of size 5.
e. The probability that the sample mean is less than 2500 grams is

$$\begin{aligned} \mathrm{P}(\overline{X} < 2500) &= \mathrm{P}\left(\frac{\overline{X} - 3500}{192} < \frac{2500 - 3500}{192}\right) \\ &= \mathrm{P}(Z < -5.21) \\ &= 0.000. \end{aligned}$$

f. The number of newborns with a birth weight less than 2500 grams follows a binomial distribution with $n = 5$ and $p = 0.01$. Therefore, the probability that only one of the 5 newborns has a birth weight less than 2500 grams is

$$\begin{aligned} \mathrm{P}(X = 1) &= \binom{5}{1}(0.01)^1(0.99)^4 \\ &= 0.048. \end{aligned}$$

Exercise 12

a. Note that the distribution of means of samples of size 15 is approximately normal with mean $\mu = 13.3$ g/100 ml and standard error $\sigma/\sqrt{n} = 1.12/\sqrt{15} = 0.29$ g/100 ml. Therefore,

$$\begin{aligned} \mathrm{P}(13.0 \leq \overline{X} \leq 13.6) &= \mathrm{P}\left(\frac{13.0 - 13.3}{0.29} \leq \frac{\overline{X} - 13.3}{0.29} \leq \frac{13.6 - 13.3}{0.29}\right) \\ &= \mathrm{P}(-1.03 \leq Z \leq 1.03) \\ &= 1 - 0.152 - 0.152 \\ &= 0.696. \end{aligned}$$

Approximately 69.6% of the samples will have a mean hemoglobin level between 13.0 and 13.6 g/100 ml.

b. The means of samples of size 30 are approximately normally distributed with mean $\mu = 13.3$ g/100 ml and standard error $\sigma/\sqrt{n} = 1.12/\sqrt{30} = 0.20$ g/100 ml. Since

$$\begin{aligned} \mathrm{P}(13.0 \leq \overline{X} \leq 13.6) &= \mathrm{P}\left(\frac{13.0 - 13.3}{0.20} \leq \frac{\overline{X} - 13.3}{0.20} \leq \frac{13.6 - 13.3}{0.20}\right) \\ &= \mathrm{P}(-1.50 \leq Z \leq 1.50) \\ &= 1 - 0.067 - 0.067 \\ &= 0.866, \end{aligned}$$

about 86.6% of the samples will have a mean hemoglobin level between 13.0 and 13.6 g/100 ml.

c. We are interested in the sample size n for which

$$\mathrm{P}(\mu - 0.2 \leq \overline{X} \leq \mu + 0.2) = 0.95,$$

or, equivalently,

$$\mathrm{P}(-0.2 \leq \overline{X} - \mu \leq 0.2) = 0.95.$$

Dividing all three terms of the inequality by the standard error $1.12/\sqrt{n}$ and substituting $Z = (\overline{X} - \mu)/(1.12/\sqrt{n})$, we have

$$\mathrm{P}\left(\frac{-0.2}{1.12/\sqrt{n}} \leq Z \leq \frac{0.2}{1.12/\sqrt{n}}\right) = 0.95.$$

Since 95% of the area under the standard normal curve lies between -1.96 and 1.96, we must solve the equation

$$\begin{aligned} 1.96 &= \frac{0.2}{1.12/\sqrt{n}} \\ &= \frac{0.2\sqrt{n}}{1.12}. \end{aligned}$$

Multiplying both sides of the equality by $1.12/0.2 = 5.6$, we have

$$\begin{aligned} \sqrt{n} &= 1.96(5.6) \\ &= 10.976 \end{aligned}$$

and

$$n = 120.5.$$

Therefore, samples of size 121 would be required.
d. Here we are interested in the sample size n for which

$$\mathrm{P}(\mu - 0.1 \leq \overline{X} \leq \mu + 0.1) \quad = \quad 0.95,$$

or

$$\mathrm{P}(-0.1 \leq \overline{X} - \mu \leq 0.1) \quad = \quad 0.95.$$

Dividing all three terms by $1.12/\sqrt{n}$ and substituting $Z = (\overline{X} - \mu)/(1.12/\sqrt{n})$, we have

$$\mathrm{P}\left(\frac{-0.1}{1.12/\sqrt{n}} \leq Z \leq \frac{0.1}{1.12/\sqrt{n}}\right) \quad = \quad 0.95.$$

Solving the equation

$$\begin{aligned} 1.96 &= \frac{0.1}{1.12/\sqrt{n}} \\ &= \frac{0.1\sqrt{n}}{1.12} \end{aligned}$$

and multiplying both sides of the equality by $1.12/0.1 = 11.2$, we have

$$\begin{aligned} \sqrt{n} &= 1.96(11.2) \\ &= 21.952 \end{aligned}$$

and

$$n \quad = \quad 481.9.$$

Samples of size 482 would be required.

Exercise 13
a. Note that

$$\begin{aligned} \mathrm{P}(300 \leq X \leq 400) &= \mathrm{P}\left(\frac{300 - 341}{79} \leq \frac{X - 341}{79} \leq \frac{400 - 341}{79}\right) \\ &= \mathrm{P}(-0.52 \leq Z \leq 0.75) \\ &= 1 - 0.302 - 0.227 \\ &= 0.471. \end{aligned}$$

Approximately 47.1% of the males have a serum uric acid level between 300 and 400 μmol/l.
b. The distribution of means of samples of size 5 is normal with mean $\mu = 341$ μmol/l and standard error $\sigma/\sqrt{n} = 79/\sqrt{5} = 35.3$ μmol/l. Therefore,

$$\begin{aligned} \mathrm{P}(300 \leq \overline{X} \leq 400) &= \mathrm{P}\left(\frac{300 - 341}{35.3} \leq \frac{\overline{X} - 341}{35.3} \leq \frac{400 - 341}{35.3}\right) \\ &= \mathrm{P}(-1.16 \leq Z \leq 1.67) \\ &= 1 - 0.123 - 0.047 \\ &= 0.830. \end{aligned}$$

Approximately 83.0% of the samples have a mean serum uric acid level between 300 and 400 μmol/l.
c. The distribution of means of samples of size 10 is normal with mean $\mu = 341$ μmol/l and standard error $\sigma/\sqrt{n} = 79/\sqrt{10} = 25.0$ μmol/l. Therefore,

$$\begin{aligned} \mathrm{P}(300 \leq \overline{X} \leq 400) &= \mathrm{P}\left(\frac{300-341}{25.0} \leq \frac{\overline{X}-341}{25.0} \leq frac400-34125.0\right) \\ &= \mathrm{P}(-1.64 \leq Z \leq 2.36) \\ &= 1 - 0.051 - 0.009 \\ &= 0.940. \end{aligned}$$

Approximately 94.0% of the samples have a mean serum uric acid level between 300 and 400 μmol/l.
d. For the standard normal distribution, the interval $(-1.96, 1.96)$ contains 95% of the observations. The corresponding values of $\overline{X}$ are $\overline{X} = -1.96(25.0) + 341 = 292$ and $\overline{X} = 1.96(25.0) + 341 = 390$. Therefore, the interval $(292, 390)$ encloses 95% of the means of samples of size 10. This symmetric interval is shorter than an asymmetric one.

Exercise 14
a. The means of samples of size 25 are approximately normally distributed with mean $\mu = 172.2$ lb and standard error $\sigma/\sqrt{n} = 29.8/\sqrt{25} = 6.0$ lb.
b. The value $Z = 1.28$ is an upper bound for 90% of the observations in the standard normal distribution. Therefore, $\overline{X} = (1.28)(6.0) + 172.2 = 179.9$ lb is an upper bound for the distribution of sample means.
c. The value $Z = -0.84$ is a lower bound for 80% of the observations in the standard normal distribution. Therefore, $\overline{X} = (-0.84)(6.0) + 172.2 = 167.2$ lb is a lower bound for the distribution of sample means.
d. Note that

$$\begin{aligned} \mathrm{P}(\overline{X} \geq 190) &= \mathrm{P}\left(\frac{\overline{X}-172.2}{6.0} \geq \frac{190-172.2}{6.0}\right) \\ &= \mathrm{P}(Z \geq 2.97) \\ &= 0.001. \end{aligned}$$

A sample mean this large or larger occurs only 0.1% of the time. This is a very unusual result.

Exercise 15
The probability that a sample mean lies in the interval $(195.9, 226.1)$ is

$$\begin{aligned} \mathrm{P}(195.9 \leq \overline{X} \leq 226.1) &= \mathrm{P}\left(\frac{195.9-211}{9.2} \leq \frac{\overline{X}-211}{9.2} \leq \frac{226.1-211}{9.2}\right) \\ &= \mathrm{P}(-1.64 \leq Z \leq 1.64) \\ &= 1 - 0.051 - 0.051 \\ &= 0.898. \end{aligned}$$

CHAPTER 9

Exercise 5

a. A two-sided 95% confidence interval for μ_s is

$$\left(130 - 1.96\,\frac{11.8}{\sqrt{10}},\ 130 + 1.96\,\frac{11.8}{\sqrt{10}}\right)$$

or

$$(122.7,\ 137.3).$$

b. The interval may be described in one of the following ways: we are 95% confident that this interval covers the true mean systolic blood pressure μ_s, or there is a 95% chance that this interval covers μ_s before a sample is selected, or approximately 95 out of 100 intervals constructed in this way will cover μ_s.

c. A two-sided 90% confidence interval for μ_d is

$$\left(84 - 1.645\,\frac{9.1}{\sqrt{10}},\ 84 + 1.645\,\frac{9.1}{\sqrt{10}}\right)$$

or

$$(79.3,\ 88.7).$$

d. A two-sided 99% confidence interval for μ_d is

$$\left(84 - 2.58\,\frac{9.1}{\sqrt{10}},\ 84 + 2.58\,\frac{9.1}{\sqrt{10}}\right)$$

or

$$(76.6,\ 91.4).$$

e. The 99% confidence interval is wider than the 90% interval. The smaller the range of values that is considered, the less confident we are that the interval covers μ_d.

Exercise 6

a. For the t distribution with 5 degrees of freedom, 5% of the area lies to the right of $t = 2.015$.
b. Only 1% of the area lies to the left of $t = -3.365$.
c. Since 0.5% of the area lies to the left of $t = -4.032$ and another 0.5% lies to the right of $t = 4.032$, 99% of the area lies between the two values.
d. The value $t = 2.571$ cuts off the upper 2.5% of the distribution.

Exercise 7

a. For the t distribution with 21 degrees of freedom, 1% of the area lies to the left of $t = -2.518$.
b. 10% of the area lies to the right of $t = 1.323$.
c. Since 5% of the area lies to the left of $t = -1.721$ and another 0.5% lies to the right of $t = 2.831$, 94.5% of the area lies between the two values.
d. The value $t = -2.080$ cuts off the lower 2.5% of the distribution.

Exercise 8

a. Since the population standard deviation is unknown, we use the t distribution with 11 df rather than the normal distribution. A two-sided 95% confidence interval for μ_1 is

$$\left(4.49 - 2.201\,\frac{0.83}{\sqrt{12}},\ 4.49 + 2.201\,\frac{0.83}{\sqrt{12}}\right)$$

or

$$(3.96,\ 5.02).$$

b. A two-sided 90% confidence interval for μ_1 is

$$\left(4.49 - 1.796\,\frac{0.83}{\sqrt{12}},\ 4.49 + 1.796\,\frac{0.83}{\sqrt{12}}\right)$$

or

$$(4.06,\ 4.92).$$

The 90% confidence interval is shorter than the 95% interval.
c. A two-sided 95% confidence interval for μ_2 is

$$\left(3.71 - 2.201\,\frac{0.62}{\sqrt{12}},\ 3.71 + 2.201\,\frac{0.62}{\sqrt{12}}\right)$$

or

$$(3.32,\ 4.10).$$

d. It is assumed that the underlying distributions of FVC and FEV_1 are approximately normally distributed.

Exercise 9
a. Since the population standard deviation σ is unknown, we use the t distribution with 13 df rather than the normal distribution. A two-sided 95% confidence interval for μ is

$$\left(29.6 - 2.160\,\frac{3.6}{\sqrt{14}},\ 29.6 + 2.160\,\frac{3.6}{\sqrt{14}}\right)$$

or

$$(27.5,\ 31.7).$$

b. The length of this interval is $31.7 - 27.5 = 4.2$ weeks.
c. Since the interval is centered around the sample mean $\bar{x} = 29.6$ weeks, we are interested in the sample size necessary to produce the interval

$$(29.6 - 1.5,\ 29.6 + 1.5)$$

or

$$(28.1,\ 31.1).$$

We know that the 95% confidence interval is of the form

$$\left(29.6 - 1.96\,\frac{3.6}{\sqrt{n}},\ 29.6 + 1.96\,\frac{3.6}{\sqrt{n}}\right).$$

To find n, therefore, we must solve the equation

$$1.5 \quad = \quad \frac{1.96(3.6)}{\sqrt{n}}$$

or

$$\begin{aligned} n &= \left[\frac{1.96(3.6)}{1.5}\right]^2 \\ &= 22.1. \end{aligned}$$

A sample of size 23 is required.
d. Here we are interested in the sample size necessary to produce the interval

$$(29.6 - 1,\ 29.6 + 1)$$

or

$$(28.6,\ 30.6).$$

The 95% confidence interval takes the form

$$\left(29.6 - 1.96\,\frac{3.6}{\sqrt{n}},\ 29.6 + 1.96\,\frac{3.6}{\sqrt{n}}\right).$$

To find n, therefore, we solve the equation

$$1 = \frac{1.96(3.6)}{\sqrt{n}}$$

or

$$\begin{aligned} n &= \left[\frac{1.96(3.6)}{1}\right]^2 \\ &= 49.8. \end{aligned}$$

A sample of size 50 is required.

Exercise 10
a. First note that

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^{18} x_i}{18} \\ &= 112.8\% \end{aligned}$$

and

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{18}(x_i - 112.8)^2}{18 - 1}} \\ &= 14.4\%. \end{aligned}$$

Because the population standard deviation σ is unknown, we use the t distribution with 17 df rather than the normal distribution. A 95% confidence interval for the mean percentage of ideal body weight is

$$\left(112.8 - 2.110\,\frac{14.4}{\sqrt{18}}\,,\ 112.8 + 2.110\,\frac{14.4}{\sqrt{18}}\right)$$

or

$$(105.6\,,\ 120.0).$$

b. The confidence interval does not contain the value 100. As a result, we conclude that the mean percentage of ideal body weight for the population of insulin-dependent diabetics is different from 100%; the true percentage is higher.

Exercise 11
Note: the table used for this exercise is missing a row. There should be an eighth row with a value of 2.37 for Calcium and 42 for Albumin. a. Because the population standard deviation is unknown, we use the t distribution with 7 df rather than the normal distribution. The sample mean calcium level is $\bar{x}_c = 3.14$ mmol/l and the standard deviation is $s_c = 0.51$ mmol/l. A one-sided lower 95% confidence bound for the true mean calcium level μ_c is $3.14 - 1.895(0.51/\sqrt{8}) = 2.80$ mmol/l.
b. The sample mean albumin level is $\bar{x}_a = 40.4$ g/l and the standard deviation is $s_a = 3.0$ g/l. A one-sided lower 95% confidence bound for the true mean albumin level μ_a is $40.4 - 1.895(3.0/\sqrt{8}) = 38.4$ g/l.
c. The lower 95% confidence bound for the mean calcium level does not lie within the normal range of values; this suggests that calcium levels are elevated for this group. There is no evidence that albumin levels differ from the normal range.

Exercise 12
a. The figure shows a slow increase of the estimate for the annual number of nonfatal firearm injuries, with a large increase in 2015.
b. Th 95% confidence interval gets wider over time which suggests that the precision of the estimate of the annual number of nonfatal firearm injuries is decreasing over time.

Exercise 13
a. A two-sided 95% confidence interval for μ is $(86.5, 89.4)$.

```
. ci zinc, level(95)

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
    zinc |     462     87.93723    .7446055        86.47399    89.40047
```

b. We are 95% confident that this interval covers the true mean serum zinc level for the population. Equivalently, if 100 random samples of size 462 are selected from the population and 100 confidence intervals calculated in this manner, approximately 95 of the intervals will contain the true mean μ and 5 will not.
c. A 90% confidence interval for μ is $(86.7, 89.2)$.

```
. ci zinc, level(90)

Variable |     Obs         Mean    Std. Err.       [90% Conf. Interval]
---------+--------------------------------------------------------------
    zinc |     462     87.93723    .7446055           86.71    89.16446
```

d. The 90% confidence interval is shorter than the corresponding 95% interval.

Exercise 14
a. A 95% confidence interval for the true mean systolic blood pressure of male low birth weight infants is $(44.3, 51.5)$.

```
. ci sbp if sex==1
```

```
Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
     sbp |      44     47.86364    1.779788        44.27435    51.45292
```

b. A 95% confidence interval for the true mean systolic blood pressure of female low birth weight infants is $(43.5, 49.4)$.

```
. ci sbp if sex==0

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
     sbp |      56     46.46429    1.489348        43.47956    49.44901
```

c. It is possible that males and females have the same mean systolic blood pressure. There is a great deal of overlap between the two confidence intervals.

Exercise 15
a. The point estimate and 95% confidence interval for the true mean PDI score is 94.78 and $(92.2, 97.4)$, respectively.

```
. ci pdi

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
     pdi |     143     94.78322    1.325531        92.16289    97.40354
```

b. The point estimate and 95% confidence interval for the true mean MDI score is 104.74 and $(102.2, 107.3)$, respectively.

```
. ci mdi

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------
     mdi |     144     104.7361    1.300364        102.1657    107.3065
```

c. Neither of these intervals contains the value 100. This tells us that the children born with congenital heart disease who undergo reparative heart surgery during the first three months of life have have mean PDI and MDI scores that are different than normal healthy infants at one year of age.

CHAPTER 10

Exercise 9
a. The null hypothesis of the test is

$$H_0 : \mu = 74.4 \text{ mm Hg.}$$

b. The alternative hypothesis is

$$H_A : \mu \neq 74.4 \text{ mm Hg.}$$

c. The test statistic is

$$\begin{aligned} \text{z} &= \frac{\bar{x}_d - \mu_0}{\sigma_d/\sqrt{n}} \\ &= \frac{84 - 74.4}{9.1/\sqrt{10}} \\ &= 3.34. \end{aligned}$$

d. The probability distribution of the test statistic is a standard Normal distribution, Z.
e. The area to the right of $z = 3.34$ is less than 0.001, and the area to the left of $z = -3.34$ is less than 0.001 as well; therefore, $p < 0.002$.
f. Since $p < 0.05$, we reject H_0.
g. We conclude that the mean diastolic blood pressure for the population of female diabetics between the ages of 30 and 34 is not equal to 74.4 mm Hg. In fact, it is higher.
h. Since $p < 0.01$, the conclusion would have been the same.

Exercise 10
a. The null hypothesis is

$$H_0 : \mu \geq 7250/\text{mm}^3,$$

and the alternative hypothesis is

$$H_A : \mu < 7250/\text{mm}^3.$$

b. Since the population standard deviation is unknown, we use the t-test rather than the z-test. The test statistic is

$$\begin{aligned} \text{t} &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{4767 - 7250}{3204/\sqrt{15}} \\ &= -3.00. \end{aligned}$$

c. The probability distribution of the test statistic is a t distribution with $15 - 1 = 14$ degrees of freedom.
d. For a t_{14} distribution, $0.0005 < p < 0.005$.
e. $p < 0.05$, therefore, we reject H_0.
f. We conclude that the mean white blood cell count of humans infected with *E. canis* is lower than $7250/\text{mm}^3$, the mean of the general population.

Exercise 11

a. Since the population standard deviation is unknown, we use the t distribution with $58 - 1 = 57$ df rather than the normal. A t distribution with 57 df can be approximated by a t distribution with 60 df; in this case, 95% of the observations lie between -2.000 and 2.000. (More accurately, if df $= 57$ then 95% of the observations lie between -2.002 and 2.002.) A two-sided 95% confidence interval for μ is

$$\left(25.0 - 2.000\frac{2.7}{\sqrt{58}},\ 25.0 + 2.000\frac{2.7}{\sqrt{58}}\right)$$

or

$$(24.3,\ 25.7).$$

b. The null hypothesis for this test is

$$H_0 \colon \mu = 24.0 \text{ kg/m}^2$$

and the alternative hypothesis is

$$H_A \colon \mu \neq 24.0 \text{ kg/m}^2.$$

The test statistic is

$$\begin{aligned} \text{t} &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{25.0 - 24.0}{2.7/\sqrt{58}} \\ &= 2.82. \end{aligned}$$

The probability distribution of the test statistics is a t distribution with 57 degrees of freedom. For this probability distribution, $2(0.0005) < p < 2(0.005)$ or $0.001 < p < 0.01$. Therefore, we reject H_0.

c. We conclude that the mean baseline body mass index for the population of men who later develop diabetes mellitus is not equal to 24.0 kg/m^2, the mean for the population of men who do not. In fact, it is higher.

d. Since the value 24.0 does not lie inside the 95% confidence interval for μ, we should have expected that the null hypothesis would be rejected.

Exercise 12

a. The null hypothesis for this test is

$$H_0 \colon \mu = 136 \text{ mm Hg}$$

and the alternative hypothesis is

$$H_A \colon \mu \neq 136 \text{ mm Hg}.$$

The test statistic is

$$\begin{aligned} \text{t} &= \frac{\bar{x}_s - \mu_{0s}}{s_s/\sqrt{n}} \\ &= \frac{143 - 136}{24.4/\sqrt{86}} \\ &= 2.66. \end{aligned}$$

For a t distribution with 85 degrees of freedom, $2(0.0005) < p < 2(0.005)$ or $0.001 < p < 0.01$. Therefore, we reject H_0 at the 0.10 level of significance. The mean systolic blood pressure for workers who have experienced a major coronary event is not equal to the mean for workers who have not.

b. The null hypothesis is

$$H_0 : \mu = 84 \text{ mm Hg}$$

and the alternative hypothesis is

$$H_A : \mu \neq 84 \text{ mm Hg}.$$

The test statistic is

$$\begin{aligned} \text{t} &= \frac{\bar{x}_d - \mu_{0d}}{s_d/\sqrt{n}} \\ &= \frac{87 - 84}{16.0/\sqrt{86}} \\ &= 1.74. \end{aligned}$$

For a t distribution with 85 degrees of freedom, $2(0.025) < p < 2(0.05)$ or $0.05 < p < 0.10$. Therefore, we again reject H_0 at the 0.10 level.

c. The workers who have experienced a major coronary event have a higher mean systolic blood pressure and also a higher mean diastolic blood pressure than the workers who have not.

Exercise 13

a. The null hypothesis for this test is

$$H_0 : \mu = 100\%$$

and the alternative hypothesis is

$$H_A : \mu \neq 100\%.$$

For the values listed, the mean ($\bar{x}$) is 112.78 and the standard deviation (s) is 14.42. The test statistic is

$$\begin{aligned} \text{t} &= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \\ &= \frac{112.78 - 100}{14.42/\sqrt{18}} \\ &= 3.76. \end{aligned}$$

For a t distribution with $18 - 1 = 17$ degrees of freedom, $2(0.0005) < p < 2(0.005)$ or $0.001 < p < 0.01$. Therefore, we reject H_0 at the 0.05 level of significance. We conclude that the percentage of ideal body weight for the population of insulin-dependent diabetics is not equal to 100. In fact, it is higher.

Exercise 14

It would be impossible for the FDA to completely eliminate the occurrence of type II errors. The probability of committing a type II error is the probability of failing to reject the null hypothesis when it is false; the only way to make this probability equal to 0 is to **always** reject every null hypothesis.

Exercise 15
a. Since $\alpha = 0.05$, H_0 would be rejected for $z \leq -1.96$. Writing

$$\begin{aligned} z &= -1.96 \\ &= \frac{\bar{x} - 3500}{430/\sqrt{n}} \end{aligned}$$

and solving for $\bar{x}$,

$$\bar{x} = 3500 - 1.96\left(\frac{430}{\sqrt{n}}\right).$$

The null hypothesis would be rejected for this value. The value of z that corresponds to $\beta = 0.10$ for a two-sided test is 1.645; for the distribution centered at $\mu_1 = 3200$ grams,

$$1.645 = \frac{\bar{x} - 3200}{430/\sqrt{n}}$$

and

$$\bar{x} = 3200 + 1.645\left(\frac{430}{\sqrt{n}}\right).$$

Equating the two expressions for $\bar{x}$,

$$\begin{aligned} n &= \left[\frac{(1.96 + 1.645)(430)}{(3500 - 3200)}\right]^2 \\ &= 26.7. \end{aligned}$$

A sample of size 27 would be required.
b. Following the same logic from part a., the value of z that corresponds to $\beta = 0.05$ for a two-sided test is 1.96; for the distribution centered at $\mu_1 = 3200$ grams,

$$1.96 = \frac{\bar{x} - 3200}{430/\sqrt{n}}$$

and

$$\bar{x} = 3200 + 1.96\left(\frac{430}{\sqrt{n}}\right).$$

Equating the two expressions for $\bar{x}$,

$$\begin{aligned} n &= \left[\frac{(1.96 + 1.96)(430)}{(3500 - 3200)}\right]^2 \\ &= 31.6. \end{aligned}$$

A sample of size 32 would be required.
c. Using the same formulas as in part a., for the distribution centered at $\mu_1 = 3300$ grams,

$$\begin{aligned} n &= \left[\frac{(1.96 + 1.645)(430)}{(3500 - 3300)}\right]^2 \\ &= 60.1. \end{aligned}$$

A sample of size 61 would be required.
d. To calculate the power we will first calculate β and then calculate the power as $1 - \beta$. Since we are conducting a two-sided test at the 0.05 level of significance, the null hypothesis

$$H_0 \colon \mu = 3500\text{g}$$

would be rejected for $z \leq -1.96$. Solving for $\bar{x}$,

$$\begin{aligned} -1.96 &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\bar{x} - 3500}{430/\sqrt{30}} \end{aligned}$$

and

$$\bar{x} = 3346.1.$$

Consequently, the null hypothesis would be rejected if the sample has a mean that is less than or equal to 3346.1. We must now find the proportion of the area under the curve centered at the true mean $\mu = 3300$ that lies to the right of 3346.1. This is the area where H_0 fails to be rejected. Solving for z in this case,

$$\begin{aligned} z &= \frac{3346.1 - 3300}{430/\sqrt{30}} \\ &= 0.59. \end{aligned}$$

The area to the right of $z = 0.59$ is 0.278. Therefore, the power of the test is

$$\begin{aligned} \text{power} &= \text{P(reject } H_0 \mid H_0 \text{ is false)} \\ &= 1 - \beta \\ &= 0.722. \end{aligned}$$

Exercise 16
a. Since we are conducting a two-sided test at the 0.05 level of significance, the null hypothesis

$$H_0 \colon \mu = 0\%$$

would be rejected for $z \leq -1.96$. Solving for $\bar{x}$,

$$\begin{aligned} -1.96 &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\bar{x} - 0}{0.02/\sqrt{25}} \end{aligned}$$

and

$$\bar{x} = 0.00784.$$

Consequently, the null hypothesis would be rejected if the sample has a mean that is less than or equal to 0.00784. We must now find the proportion of the area under the curve centered at

the true mean $\mu = 0.01$ that lies to the right of 0.00784. This is the area where H_0 fails to be rejected. Solving for z in this case,

$$\begin{aligned} z &= \frac{0.00784 - 0.01}{0.02/\sqrt{25}} \\ &= -0.54. \end{aligned}$$

The area to the left of $z = -0.54$ is 0.295. Therefore, the power of the test is

$$\begin{aligned} \text{power} &= \text{P(reject } H_0 \mid H_0 \text{ is false)} \\ &= 1 - \beta \\ &= 0.705. \end{aligned}$$

b. Following the same logic as in part a above and changing 25 to 40, the null hypothesis

$$H_0 : \mu = 0\%$$

would be rejected for $z \leq -1.96$. Solving for $\bar{x}$,

$$\begin{aligned} -1.96 &= \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\bar{x} - 0}{0.02/\sqrt{40}} \end{aligned}$$

and

$$\bar{x} = 0.0062.$$

Consequently, the null hypothesis would be rejected if the sample has a mean that is less than or equal to 0.0062. We must now find the proportion of the area under the curve centered at the true mean $\mu = 0.01$ that lies to the right of 0.0062. This is the area where H_0 fails to be rejected. Solving for z in this case,

$$\begin{aligned} z &= \frac{0.0062 - 0.01}{0.02/\sqrt{40}} \\ &= -1.2. \end{aligned}$$

The area to the left of $z = -0.54$ is 0.115. Therefore, the power of the test is

$$\begin{aligned} &= 1 - \beta \\ &= 0.885. \end{aligned}$$

c. The value of z that corresponds to $\beta = 0.10$ (90% power) is 1.28. In this case, therefore,

$$\begin{aligned} n &= \left[\frac{(1.96 + 1.28)(0.02)}{(0.01 - 0)}\right]^2 \\ &= 41.99. \end{aligned}$$

A sample of size 42 would be required.

Exercise 17

a. The p-value of the two-sided test is 0.0001. Therefore, we reject the null hypothesis that the mean PDI score is equal to 100 in favor of the alternative hypothesis that it is not. In fact, the mean for these children is less than 100.

```
. ttest pdi = 100

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     pdi |     143    94.78322    1.325531    15.85104    92.16289    97.40354
------------------------------------------------------------------------------
Degrees of freedom: 142

                            Ho: mean(pdi) = 100

    Ha: mean < 100               Ha: mean ~= 100               Ha: mean > 100
       t =  -3.9356                  t =  -3.9356                  t =  -3.9356
   P < t =   0.0001          P > |t| =   0.0001          P > t =   0.9999
```

b. The p-value of the two-sided test is 0.0004. Therefore, we reject the null hypothesis that the mean MDI score is equal to 100 in favor of the alternative hypothesis that it is not. The mean for these children is greater than 100.

```
. ttest mdi = 100

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     mdi |     144    104.7361    1.300364    15.60437    102.1657    107.3065
------------------------------------------------------------------------------
Degrees of freedom: 143

                            Ho: mean(mdi) = 100

    Ha: mean < 100               Ha: mean ~= 100               Ha: mean > 100
       t =   3.6421                  t =   3.6421                  t =   3.6421
   P < t =   0.9998          P > |t| =   0.0004          P > t =   0.0002
```

c. The 95% confidence interval for the true mean PDI score is $(92.2, 97.4)$, and the 95% confidence interval for the true mean MDI score is $(102.2, 107.3)$. Neither of these intervals contains the value 100. Because both null hypotheses were rejected, this is what we expected.

```
. ci pdi

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
```

```
---------+--------------------------------------------------------------------
     pdi |     143     94.78322    1.325531        92.16289    97.40354

. ci mdi

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     mdi |     144     104.7361    1.300364        102.1657    107.3065
```

CHAPTER 12

Exercise 5
a. For $F_{8,16}$, 10% of the area under the curve lies to the right of $F = 2.09$.
b. The value $F = 3.89$ cuts off the upper 1% of the distribution.
c. Since 0.5% of the area under the curve lies to the right of $F = 4.52$, 99.5% lies to the left of this value.

Exercise 6
a. For $F_{3,30}$, 0.5% of the area under the curve lies to the right of $F = 5.24$.
b. Since 5% of the area under the curve lies to the right of $F = 2.92$, 95% lies to the left of this value.
c. The value $F = 3.59$ cuts off the upper 2.5% of the distribution.
d. The value $F = 7.05$ cuts off the upper 0.1%.

Exercise 7
a. The estimate of the within-groups variance is

$$\begin{aligned} s_W^2 &= \frac{(n_1-1){s_1}^2 + (n_2-1){s_2}^2 + (n_3-1){s_3}^2 + (n_4-1){s_4}^2}{n_1+n_2+n_3+n_4-4} \\ &= \frac{(268)(13.4)^2 + (52)(10.1)^2 + (27)(11.6)^2 + (8)(12.2)^2}{269+53+28+9-4} \\ &= 164.1. \end{aligned}$$

b. Since the grand mean of the data is

$$\begin{aligned} \bar{x} &= \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + n_3\bar{x}_3 + n_4\bar{x}_4}{n_1+n_2+n_3+n_4} \\ &= \frac{269(115) + 53(114) + 28(118) + 9(126)}{269+53+28+9} \\ &= 115.36, \end{aligned}$$

the estimate of the between-groups variance is

$$\begin{aligned} s_B^2 &= \frac{n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + n_3(\bar{x}_3-\bar{x})^2 + n_4(\bar{x}_4-\bar{x})^2}{4-1} \\ &= \frac{269(-0.36)^2 + 53(-1.36)^2 + 28(2.64)^2 + 9(10.64)^2}{4-1} \\ &= 449.0. \end{aligned}$$

c. To test the null hypothesis that the mean systolic blood pressures of the four groups are identical versus the alternative hypothesis that they are not, we first calculate the test statistic

$$\begin{aligned} \mathrm{F} &= \frac{{s_B}^2}{{s_W}^2} \\ &= \frac{449.0}{164.1} \\ &= 2.74. \end{aligned}$$

d. The distribution of the test statistic is an F distribution with $4-1=3$ and $359-4=355$ degrees of freedom.
e. Looking at the table, $0.025 < p < 0.050$. Therefore, we reject H_0 at the 0.05 level of significance.
f. We conclude that there is a difference among the mean systolic blood pressures of the four groups.
g. To conduct $\binom{4}{2} = 6$ tests and keep the overall probability of committing a type I error at 0.05, we set the significance level for each individual comparison at

$$\begin{aligned} \alpha^* &= \frac{0.05}{6} \\ &= 0.0083. \end{aligned}$$

To compare the mean systolic blood pressures of nonsmokers and current smokers, we calculate the test statistic

$$\begin{aligned} \mathrm{t}_{12} &= \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_w^2\left[(1/n_1)+(1/n_2)\right]}} \\ &= \frac{115-114}{\sqrt{164.1\left[(1/269)+(1/53)\right]}} \\ &= 0.52. \end{aligned}$$

A t distribution with $359-4=355$ df can be approximated by the standard normal distribution. In this case, $p > 0.0083$; therefore, we are unable to reject the null hypothesis that the two means are identical. Similarly, we can calculate the appropriate test statistics for each of the other five comparisons.

$$\begin{aligned} \mathrm{t}_{13} &= \frac{115-118}{\sqrt{164.1\left[(1/269)+(1/28)\right]}} \\ &= -1.18 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{14} &= \frac{115-126}{\sqrt{164.1\left[(1/269)+(1/9)\right]}} \\ &= -2.53 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{23} &= \frac{114-118}{\sqrt{164.1\left[(1/53)+(1/28)\right]}} \\ &= -1.34 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{24} &= \frac{114-126}{\sqrt{164.1\left[(1/53)+(1/9)\right]}} \\ &= -2.60 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{34} &= \frac{118-126}{\sqrt{164.1\left[(1/28)+(1/9)\right]}} \\ &= -1.63 \end{aligned}$$

Two of the comparisons result in a p-value that is less than 0.0083 — the comparison of group 1 with group 4, and the comparison of group 2 with group 4. Therefore, we conclude that the mean systolic blood pressures of both nonsmokers and current smokers are lower than the mean systolic blood pressure of tobacco chewers.

Exercise 8

a. To test the null hypothesis that the mean LDL cholesterol levels are identical for each of the four populations, we must first calculate estimates of the within-groups and between-groups variances. Note that

$$\begin{aligned} s_W^2 &= \frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2+(n_3-1){s_3}^2+(n_4-1){s_4}^2}{n_1+n_2+n_3+n_4-4} \\ &= \frac{(72)(1.62)^2+(104)(1.43)^2+(239)(1.24)^2+(1079)(1.31)^2}{73+105+240+1080-4} \\ &= 1.75. \end{aligned}$$

Since

$$\begin{aligned} \bar{x} &= \frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3+n_4\bar{x}_4}{n_1+n_2+n_3+n_4} \\ &= \frac{73(6.22)+105(5.81)+240(5.77)+1080(5.47)}{73+105+240+1080} \\ &= 5.58, \end{aligned}$$

we have that

$$\begin{aligned} s_B^2 &= \frac{n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2+n_3(\bar{x}_3-\bar{x})^2+n_4(\bar{x}_4-\bar{x})^2}{4-1} \\ &= \frac{73(0.64)^2+105(0.23)^2+240(0.19)^2+1080(-0.11)^2}{4-1} \\ &= 19.06. \end{aligned}$$

Therefore, the test statistic is

$$\begin{aligned} \mathrm{F} &= \frac{s_B^2}{s_W^2} \\ &= \frac{19.06}{1.75} \\ &= 10.89. \end{aligned}$$

For an F distribution with $4-1=3$ and $1498-4=1494$ df, $p<0.001$. Therefore, we reject the null hypothesis.

b. We conclude that mean LDL cholesterol level is not the same for each of the four groups.

c. In order to use the one-way analysis of variance technique, the four populations must be at least approximately normally distributed, and their variances must all be the same.

d. It is necessary to take an additional step in this analysis. We have concluded that the mean LDL levels are not the same for all four groups, but we cannot yet say which group means differ from which others. In order to do this, we need to use the Bonferroni method of multiple comparisons. (It would tell us that patients with intermittent claudication and those with minor asymptomatic disease have higher mean LDL cholesterol levels than patients with no disease.)

Exercise 9

a. The average minutes of individual therapy per session is longest for the private not-for-profit centers and shortest for the for-profit centers; the average minutes of group therapy is longest for the for-profit centers and shortest for the public centers.

b. To test the null hypothesis that the mean minutes of individual therapy per session are identical for each type of center, we first calculate estimates of the within-groups and between-groups variances. Note that

$$\begin{aligned} {s_W}^2 &= \frac{(n_1-1){s_1}^2+(n_2-1){s_2}^2+(n_3-1){s_3}^2}{n_1+n_2+n_3-3} \\ &= \frac{(36)(15.47)^2+(311)(11.41)^2+(168)(11.08)^2}{37+312+169-3} \\ &= 135.4. \end{aligned}$$

Since

$$\begin{aligned} \bar{x} &= \frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3}{n_1+n_2+n_3} \\ &= \frac{37(49.46)+312(54.76)+169(53.25)}{37+312+169} \\ &= 53.88, \end{aligned}$$

we have that

$$\begin{aligned} {s_B}^2 &= \frac{n_1(\bar{x}_1-\bar{x})^2+n_2(\bar{x}_2-\bar{x})^2+n_3(\bar{x}_3-\bar{x})^2}{3-1} \\ &= \frac{37(-4.42)^2+312(0.88)^2+169(-0.63)^2}{3-1} \\ &= 515.8. \end{aligned}$$

Therefore, the test statistic is

$$\begin{aligned} \mathrm{F} &= \frac{{s_B}^2}{{s_W}^2} \\ &= \frac{515.8}{135.4} \\ &= 3.81. \end{aligned}$$

For an F distribution with $3-1=2$ and $518-3=515$ df, $0.01 < p < 0.025$. We reject the null hypothesis at the 0.05 level of significance and conclude that mean minutes of individual therapy per session are not the same for all three types of centers.

c. To apply the Bonferroni method of multiple comparisons, we conduct three pairwise tests at the $0.05/3 = 0.0167$ level of significance. The test statistics are:

$$\begin{aligned} \mathrm{t}_{12} &= \frac{49.46 - 54.76}{\sqrt{135.4\,[(1/37) + (1/312)]}} \\ &= -2.62 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{13} &= \frac{49.46 - 53.25}{\sqrt{135.4\,[(1/37) + (1/169)]}} \\ &= -1.79 \end{aligned}$$

$$\begin{aligned} \mathrm{t}_{23} &= \frac{54.76 - 53.25}{\sqrt{135.4\,[(1/312) + (1/169)]}} \\ &= 1.36 \end{aligned}$$

All test statistics have a t distribution with 515 df. The comparison of private for-profit and not-for-profit centers results in $p = 0.008$; the p-values for the other two comparisons are both greater than 0.0167. Therefore, we conclude that the mean minutes of individual therapy per session is shorter for for-profit centers than for not-for-profit centers.

d. To test the null hypothesis that the mean minutes of group therapy per session are identical for each type of center, we again calculate estimates of the within-groups and between-groups variances. Note that

$$\begin{aligned} s_W^2 &= \frac{(29)(42.91)^2 + (295)(31.27)^2 + (164)(27.12)^2}{30 + 296 + 165 - 3} \\ &= 947.7. \end{aligned}$$

Since

$$\begin{aligned} \bar{x} &= \frac{30(105.83) + 296(98.68) + 165(94.17)}{30 + 296 + 165} \\ &= 97.60, \end{aligned}$$

we have that

$$\begin{aligned} s_B^2 &= \frac{30(8.23)^2 + 296(1.08)^2 + 165(-3.43)^2}{3 - 1} \\ &= 2159.2. \end{aligned}$$

The test statistic is

$$\begin{aligned} \mathrm{F} &= \frac{s_B^2}{s_W^2} \\ &= \frac{2159.2}{947.7} \\ &= 2.28. \end{aligned}$$

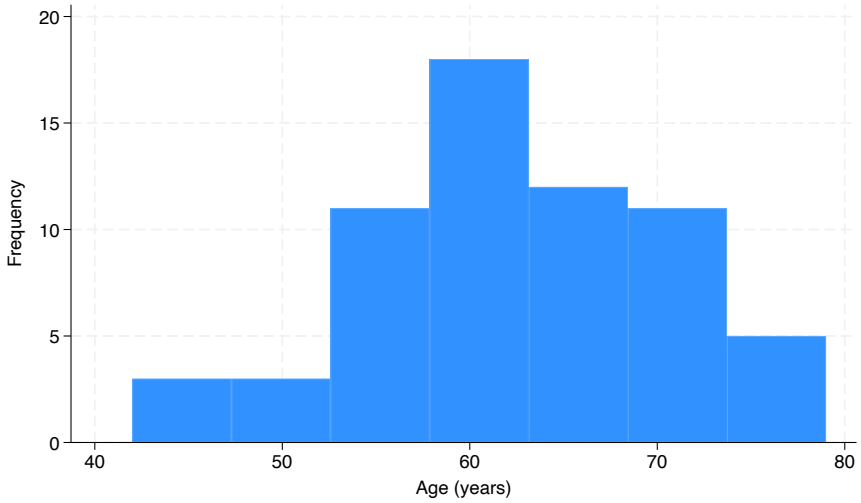
For an F distribution with $3 - 1 = 2$ and $491 - 3 = 488$ df, $p > 0.10$. Therefore, we are unable to reject the null hypothesis that mean minutes of group therapy per session are identical.

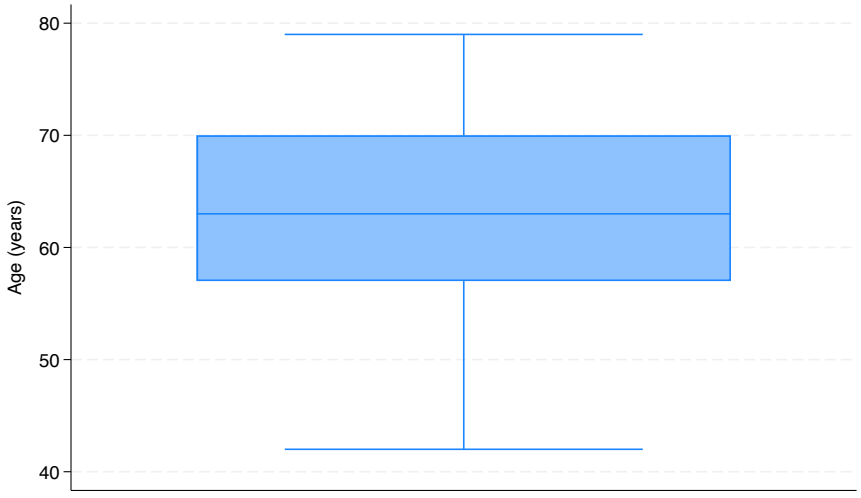
e. Since we did not reject the null hypothesis in part d., we do not have to carry out a multiple comparisons procedure.
f. While private for-profit centers have shorter individual therapy sessions than not-for-profit centers, on average, there are no significant differences in the length of group therapy sessions.

Exercise 10
a. Yes, age seems to be approximately normally distributed. The histogram and box plot of age are shown below.

```
histogram age, frequency

graph box age
```





b. The sample mean ages for the three centers are 62.5, 63.3, and 60.8 years. The sample standard deviations are 8.7, 7.8, and 8.0 years.

```
. sort center
. by center: summarize age

-> center=        1
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
     age |      22    62.54545   8.672492         41         76

-> center=        2
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
     age |      18    63.27778    7.78993         47         75

-> center=        3
Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+--------------------------------------------------------
     age |      23    60.82609   8.003705         42         73
```

c. The code and output to test the null hypothesis that the mean ages for the centers are identical is shown below. The test statistic is $F = 0.50$ and has an $F_{2,60}$ distribution.
d. From the output we see that $p = 0.6108$. We are unable to reject the null hypothesis.
e. We conclude that there is no evidence of an age difference among the three medical centers.

```
. oneway age center

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      66.6141226      2   33.3070613      0.50     0.6108
 Within groups         4020.37     60   67.0061667
------------------------------------------------------------------------
    Total           4086.98413     62   65.9190988

Bartlett's test for equal variances:  chi2(2) =   0.2430  Prob>chi2 = 0.886
```

Exercise 11

a. Using the two-sample t test that assumes equal variances, the test statistic is t=−0.6072 and $p = 0.5451$. We are unable to reject the null hypothesis that mean systolic blood pressure is identical for girls and boys.

```
. ttest sbp, by(sex)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Female |      56    46.46429    1.489348    11.14526    43.47956    49.44901
    Male |      44    47.86364    1.779788    11.80577    44.27435    51.45292
---------+--------------------------------------------------------------------
combined |     100       47.08    1.140324    11.40324    44.81735    49.34265
---------+--------------------------------------------------------------------
    diff |           -1.399351    2.304609               -5.972771     3.17407
------------------------------------------------------------------------------
Degrees of freedom: 98

                 Ho: mean(Female) - mean(Male) = diff = 0

     Ha: diff < 0                 Ha: diff ~= 0                 Ha: diff > 0
       t =  -0.6072                 t =  -0.6072                   t =  -0.6072
   P < t =   0.2726           P > |t| =   0.5451           P > t =   0.7274
```

b. Using the one-way analysis of variance, the test statistic is F=0.37 and $p = 0.5451$. Again we are unable to reject the null hypothesis that mean systolic blood pressure is identical for girls and boys.

```
. oneway sbp sex

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      48.2496104      1   48.2496104      0.37     0.5451
 Within groups      12825.1104     98   130.868473
------------------------------------------------------------------------
    Total             12873.36     99   130.033939

Bartlett's test for equal variances:  chi2(1) =   0.1590  Prob>chi2 = 0.690
```

c. The F-test and the t-test do appear to be mathematically equivalent. Note that the p-values are exactly the same.