Elementary Statistics

Brian E. Blank

March 5, 2016

FIRST PRESIDENT UNIVERSITY PRESS

Solutions to the Exercises

Chapter 1

1. The population under study is composed of all licensed drivers in the United States in the year 2009. The two categorical variables that give rise to the table are *Age Group* and *Gender*.

2. The table is indeed a contingency table. That is because all the categories of one of the variables are represented by the rows, all the categories of the other variable are represented by the columns, and every cell is filled with a joint frequency.

3. If, for each row, we add all the entries of that row, then we obtain the right margin, which is the (marginal in this context) distribution of *Age Group*:

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
9,932,441	35,924,231	36,637,530	40,555,334	38,484,123	26,237,607	14,152,955	7,694,165

   If, for each column, we add all the entries of that column, then we obtain the bottom margin, which is the (marginal in this context) distribution of *Gender*:

Male	Female
104,261,813	105,356,573

   The number of licensed drivers in the USA in 2009 can be obtained by adding the entries of either marginal distribution. The sum comes to 209,618,386 whichever way it is calculated.

4. The three percentages associated with the number 18,024,284 are the row percentage $(18,024,284/35,924,231) \times 100\%$, or 50.17 %, the column percentage $(18,024,284/104,261,813) \times 100\%$, or 17.29 %, and the table percentage $(18,024,284/209618386) \times 100\%$, or 8.60 %. (The divisors are respectively, the row sum of the row that contains 18,024,284, the column sum of the column that contains 18,024,284), and the sum of the entries of each of the two marginal distributions. The values of all three divisors were calculated in the preceding exercise.) The row percentage is the percentage of males among all licensed drivers aged 20–29. The column percentage is the percentage of male licensed drivers aged 20-29 among all licensed male drives. The table percentage is the percentage of licensed male drivers aged 20-29 among all licensed drivers.

5. The conditional distributions that involve the number 7,271,261 are the row and the column contaiing that number. That is

Male	Female
6,881,694	7,271,261

and

< 20	20–29	30–39	40–49	50–59	60–69	70–79	> 79
4,853,746	17,899,947	18,291,443	20,248,986	19,382,381	13,255,192	7,271,261	4,153,617

These conditional distributions are, respectively, the gender distribution of licensed drivers in the 70-79 age group, and the age group distribution of licensed female drivers.

6. Let us express in terms of percentages the conditional distributions of gender for each age group category. In other words, we will present the table of row percentages.

Age Group	Male	Female
< 20	51.13%	48.87%
20–29	50.17%	49.83%
30–39	50.07%	49.93%
40–49	50.07%	49.93%
50–59	49.64%	50.36%
60–69	49.48%	50.52%
70–79	48.62%	51.38%
> 79	46.02%	53.98%

If the population under consideration consisted only of licensed drivers between 20 and 69, then I'd declare Gender and Age Group to be independent. But the three pairs of numbers, 51.13 and 48.87, 48.62 and 51.38, and especially 46.02 and 53.98 show two gaps that are large enough that it is likely that Gender and Age Group *are* dependent in the youngest and oldest sectors of the population. That is enough to deduce dependence for the population under study.

7. Let a be the number of surveyees that agreed. Then the number that disagreed is $a + 1077$. The total number of surveyees, namely 2625, is therefore $a + (a + 1077) + 78$. We solve the equation $a + (a + 1077) + 78 = 2625$ to find $a = 735$ and $a + 1077 = 1812$. We can fill in the right margin of the table:

	Male	Female	Total
Agree			735
Disagree			1812
Do not know			78
Total			2625

8. In total, 78 surveyees did not know. Therefore, $78 - 34$, or 44 females did not know. Similarly, $735 - 363$, or 372 males agreed, and $1812 - 1005$, or 807 males disagreed. The frequency table is

	Male	Female	Total
Agree	372	363	735
Disagree	807	1005	1812
Do not know	34	44	78
Total	1213	1412	2625

9. Proportion of females who agree: 363/1412, or 0.2571.
   Proportion of surveyees who agree that are female: 363/735, or 0.4939.
   Proportion of surveyees who are female and agree: 373/2625, or 0.1421.

10. Row percentage: $100 \cdot 807/1812$, or 44.54—the percentage of disagreeing surveyees who are male.
    Column percentage: $100 \cdot 807/1213$, or 66.53—the percentage of male surveyees who disagree.
    Table percentage: $100 \cdot 807/2625$, or 30.74—the percentage of surveyees who are male and disagree.

11. Simpson's Paradox! The explanation for the reversal of the combined group means is that the proportion of group A in NE is greater than the proportion of group A in NJ. That is,

$$\frac{\text{NJ(A)}}{\text{NJ(A)} + \text{NJ(B)} + \text{NJ(C)}} < \frac{\text{NE(A)}}{\text{NE(A)} + \text{NE(B)} + \text{NE(C)}}.$$

Chapter 2

1: There are four values in the first bin and their average is 5. Therefore, the sum of the values in the first bin is 4×5, or 20. There are six values in the second bin and their average is 14. Therefore, the sum of the values in the second bin is 6×14, or 84. Continuing, we see that the sums of the data values in the remaining two bins are 10×24, or 240, and 4×34, or 136. Therefore, the sum of all the data values is $20 + 84 + 240 + 136$, or 480. Adding the frequencies in the second row of the table, we see that the size of the data set is 24. The mean of the data set is therefore 480/24, or 20.

2. First we need to determine the size of the data set: $N = (225 - 126) + 1$, or $N = 100$. (Why has the number 1 been added to the difference $(225 - 126)$? What if the data set were 3,4 or 7,8,9? Are the sizes of these data sets $4 - 3$ and $9 - 7$, or are they $(4 - 3) + 1$ and $(9 - 7) + 1$?) Because N is even, we split the ordered data set into two subsets of an equal size, namely 50:

$$126, 127, 128, \ldots 173, 174, 175, \qquad 176, 177, 178, \ldots 223, 224, 225.$$

   (Note: the median, 175.5, is not included with either group, N being even.) The lower group has an even number of terms. So its median, namely Q_1, is obtained by averaging the two middle values, 150 and 151: $Q_1 = 150.5$. An analogous calculation with the upper group results in $Q_3 = 200.5$. The requested IQR is $IQR = 200.5 - 150.5$, or $IQR = 50$.

3. a) Several data values occur twice, but only one, 34, occurs more than two times. It is the mode, according to the original definition of *mode*. For the binned data, the mode is the bin [80,90), even though no individual datum in that bin occurs more than once.

   b) The distribution is unimodal and negatively skewed: the longer tail extends in the direction of *decreasing* data values.
   c) The bins are [0,10), [10,20), ..., [100,110). The class width is 10.

   d) There are 20 observations from 2 to 69 inclusive. Likewise, there are 20 observations from 80 to 106, inclusive. These groups completely balance each other on opposite sides of the median. Therefore, the median of the entire data set will be the median of the [79,80) bin. The number of terms in that bin is 8, an even number, so the median is the average of the two middle numbers, 76 and 76. Thus, $Q_2 = 76$. The lower half of the full data set is

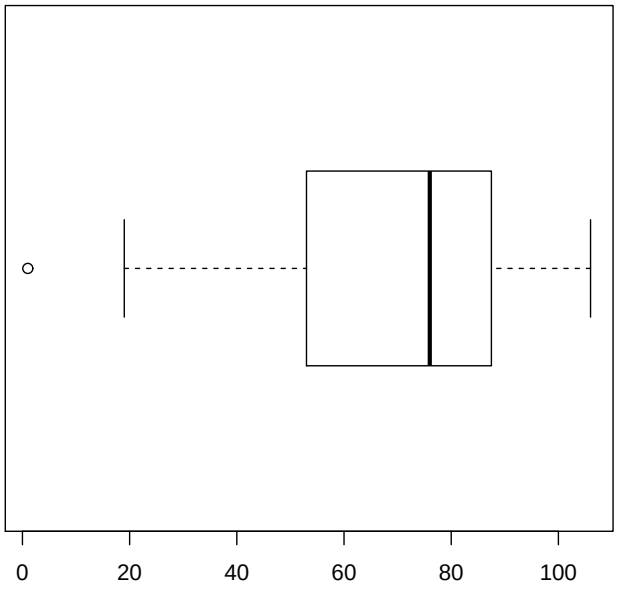
$$1, 19, 22, 25, 34, 34, 34, 40, 49, 47, 50, 56, \qquad 58, 58, 60, 61, 62, 63, 66, 39, 74, 75, 75, 76.$$

The median of this set is the average, $(56 + 58)/2$, or 57, of the two middle numbers. Thus, $Q_1 = 57$. The upper half of the full data set is

$$76, 77, 77, 79, 80, 81, 82, 83, 84, 85, 86, 87, \qquad 88, 90, 92, 92, 94, 94, 97, 99, 100, 103, 105, 106.$$

The median of this set is the average, $(87+88)/2$, or 87.5, of the two middle numbers. Thus, $Q_3 = 87.5$. Consequently, IQR $= 87.5 - 57$, or 30.5.

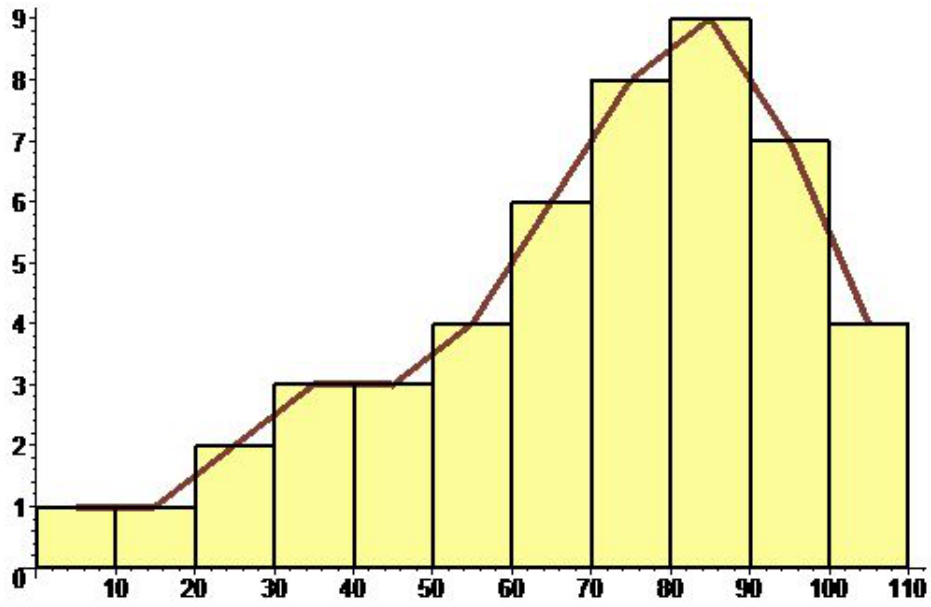
e) We will draw the boxplot horizontally. The left side of the box is at Q_1, namely 57. The right side of the box is at Q_3, namely 87.5. A line cutting the box into two is dran at the median, 76. The left fence is at $Q_1 - 1.5 \times IQR$, or $57 - 1.5 \times 30.5$, or 11.25. The right fence is at $Q_3 + 1.5 \times IQR$, or $87.5 + 1.5 \times 30.5$, or 133.25. The smallest datum that is not to the left of the left fence is 19. Thus, 19 is the terminal point on the of the left whisker. The largest datum that is not to the right of the right fence is 106. Thus, 106 is the terminal point on the of the right whisker. We complete the boxplot by plotting the smallest datum 2, which is an outlier. The boxplot that follows was produced with two lines in R. In the first line the data set was assigned to `stem_and_leaf_data`. The second line was the call `boxplot(stem_and_leaf_data, range = 1.5, horizontal=TRUE)`. Notice that R, on its own (without any user prompting), determined the datum 2 to be an outlier.



f) The frequency distribution is obtained by counting in each row the digits that appear in the right column:

[0,10)	[10,20)	[20,30)	[30,40)	[40,50)	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)	[100,110)
1	1	2	3	3	4	6	8	9	7	4

Next we plot the points (5,1), (15,1), (25,2), (35,3), (45,3), (55,4), (65,6), (75,8), (85,9), (95,7), and (105,4). The first coordinates are the class marks (midpoints) of the bins and the second coordinates are the frequencies for the bins. Finally, consecutive points are connected with line segments. (A histogram had been included with the frequency polygon in the figure below, but it is not part of the frequency polygon.)



4. Let $x_1, x_2, x_3, \ldots x_N$ be the actual number of defects per item. The recorded values were $y_1, y_2, y_3, \ldots y_N$ where $y_1 = 10\,x_1, y_2 = 10\,x_2, \ldots, y_N = 10\,x_N$. Then

$$\begin{aligned} 100 &= \bar{y} \\ &= \frac{1}{N}\,(y_1 + y_2 + \cdots + y_N) \\ &= \frac{1}{N}\,(10\,x_1 + 10\,x_2 + \cdots + 10\,x_N) \\ &= \frac{10}{N}\,(x_1 + x_2 + \cdots + x_N) \\ &= 10\,\bar{x}. \end{aligned}$$

It follows that $\bar{x} = 10$.

We will use the equation $\bar{y} = 10\,\bar{x}$, just established, in the standard deviation correction. Let s_y be the standard deviation of the recorded y-values and let s_x be the standard deviation of the x-values. Then

$$
\begin{aligned}
10 &= s_y \\
&= \sqrt{\frac{1}{N-1}\big((y_1-\bar{y})^2+(y_2-\bar{y})^2+\cdots+(y_N-\bar{y})^2\big)} \\
&= \sqrt{\frac{1}{N-1}\big((10\,x_1-10\,\bar{x})^2+(10\,x_2-10\,\bar{x})^2+\cdots+(10\,x_N-10\,\bar{x})^2\big)} \\
&= \sqrt{\frac{1}{N-1}\big(100(x_1-\bar{x})^2+100(x_2-\bar{x})^2+\cdots+100(x_N-\bar{x})^2\big)} \\
&= \sqrt{\frac{100}{N-1}\big((x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2\big)} \\
&= 10\sqrt{\frac{1}{N-1}\big((x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_N-\bar{x})^2\big)} \\
&= 10\,s_x.
\end{aligned}
$$

Thus, $s_x = 1$.

5. From the data, we see that $Q_1 = 1, Q_2 = 2, Q_3 = 3$, and IQR = 2. Therefore, the lower fence is at $Q_1 - 1.5 \times 2$, or -2. The upper fence is at $Q_3 + 1.5 \times 2$, or 6. No datum is outside a fence. Therefore the lower whisker extends downward from 1 to the smallest datum, 1. The upper whisker extends upward from 3 to the largest datum, 3. In other words, each whisker is a single point on an edge of the box. (The question on the exam was asked in multiple choice form; the correct answer choice was, *The boxplot has no whiskers.*)

6. a) The IQR is $60 - 24$, or 36. Therefore, the class width IQR/3 (given) is 12. b) Because $x_{[1]}$ is the lowest class limit (given) and $x_{[1]} = 10$ (given as the first number of the five-number summary), we see that the first bin is [10,22). The first class mark is therefore $(10 + 22)/2$, or 16. c) Each class limit is 12 units to the right of the preceding class limit. The last greatest class limit cannot be less than the greatest datum, 80 (given as the fifth number of the five-number summary), we see that we must have 10, 22, 34, 46, 58, 70, 82 as class limits. The are no class limits to the right of 82, because then we would have a empty bin and there are none (given). The number of bins is always one less than the number of class limits, of which we count 7. Therefore, there are 6 bins. d) As we have already reasoned, the largest class limit is 82. e) The median is 46 ((given as the third number of the five-number summary). Thus, the median is the class limit separating the bin [34,46) from [46,58). Our convention, as indicated by the mathematical notation we use to write the bins, is that a value that is a class limit belongs to the larger bin. Thus, 46 belongs to [46,58). The bins that precede this one are [10,22), [22,34), [34,46). Therefore, 46 belongs to the fourth bin, counting from the left. f) The lower fence of a Tukey boxplot of X is $Q_1 - 1.5 \times \text{IQR}$, or $24 - 1.5 \times 36$, or -30. The lower whisker of a vertical Tukey boxplot extends downward from Q_1 to the smallest datum not less than the lower fence. Thus, the lower whisker extends downward from 24 to 10. It has length 14. The upper fence of a Tukey boxplot of X is $Q_3 + 1.5 \times \text{IQR}$, or $60 + 1.5 \times 36$, or 114. The upper whisker of a vertical Tukey boxplot extends upward from Q_3 to the greatest datum not greater than the upper fence. Thus, the lower whisker extends upward from 60 to 80. It has length 20.

7. Conclusion (i) is incorrect: an outlier of 13 hours has been recorded. (It is not clear why it has been recorded as an outlier, because the boxplot for the group shows that 12 hours is a data value, but never mind.) Therefore, assessments (A) and (F) are false. Conclusion (ii) is correct: the upper middle quartile (part of the box above the median line) stretches over more values than the lower middle quartile (part of the box below the median line) and the upper whisker is longer than the lower whisker. Conclusion (iii) is incorrect by similar reasoning. Conclusion (iv), *The minimum sleep duration at home was 6*

hours, is correct, but understanding why is a bit more complicated. The shortest duration that appears in the boxplot is, indeed, 6 hours. But a boxplot does not always reveal the smallest data value. How can we be sure there is no data value less than 6? Well, 6 was the smallest value inside the lower fence (not visible). Could there be a value outside the fence that has not been plotted. Such a value could not be an outlier (because we display outliers in boxplots). Thus, if there were a value lower than the invisible fence, then there would have to be other data values not too far away. Now, look at the length of the upper whisker: 3 units. (Alternatively, the vertical side length of the box, namely 2, tells us that the IQR is 2, and therefore the fences are 1.5×2 units from the box. With a lower fence at 4, we see that a value smaller than 6 would have to be plotted as an outlier.

8. The author does not ask questions of this nature. Nor does he approve of them. However, to avoid having to answer ths question multiple times orally, he will set down his thoughts in writing. The correct answer is supposed to be (D). The author doubts that it is. We can rule out (A). Eye color is a categorical variable. No doubt one color is found more frequently than the others, but it is not called a "mode", that term being reserved for numerical variables. Let us reject (B) too: it is difficult to imagine circumstances in which the number of TV sets at home is bimodal. As for (C), (D), and (E), take your pick. No doubt, the number of cigarettes smoked daily is bimodal for most populations that come to mind. There is one mode for smokers, and 0 is another mode, because there are plenty of non-smokers in the sort of population that comes readily to mind. This is the answer the author would have chosen had he been subjected to the exam from which this question has been taken. But there is an assumption about the population that has to be made: if there is no cigarette consumption, for example at a cigar club, then the only value of the variable might be 0. As for (E), it again depends on the population and some knowledge. For example, if the population comprises adults in the U.S., then it is reasonable to expect the distribution to be bimodal: a mode for males and another for females. As it turns out, the average circumferences are close enough (55 cm for women and 57 cm for men) and the standard deviations are large enough (1.61 for women and 1.85 for men), that the distribution is unimodal. Who knew? As for (C), the hours of homework last week, the author certainly expects that to be bimodal if the population is the usual collection of students, which typically divides into two types: those who spend some time on homework, and those who spend no time on homework. If the author had paid to take the exam, he would have asked for his money back.

9. The answer is quickly seen to be (B). As discussed in the notes, a stem-and-leaf plot displays all the data values. A dotplot, by contrast, displays nothing but piled up dots. The author is not a proponent of dotplots and consequently omitted them from discussion in the notes. Nothing more will be said here.

10. Chebyshev's lower bound refers to data values in an open interval centered about the mean. So $\bar{X} = 28$. Chebyshev's lower bound then refers to the number of observations in the open interval $(28 - \lambda \times s_X, 28 + \lambda \times s_X)$. Therefore, $28 + \lambda \times s_X = 38$, or $\lambda \times s_X = 10$. Thus, $s_X = 10/\lambda$. We will put this equation in the background until we find the value of λ. For this unknown λ, Chebyshev's inequality asserts that, for any positive λ, the number $N_c(\lambda)$ of observations in the open interval $(28 - \lambda \times s_X, 28 + \lambda \times s_X)$ satisfies $N_c(\lambda)/25 > 1 - 1/\lambda^2$. From the given information, $N_c(\lambda) > 16/25$, we infer that $1 - 1/\lambda^2 = 16/25$, or $9/25 = 1/\lambda^2$, or $\lambda^2 = 25/9$, or $\lambda = 5/3$. Because $s_X = 10/\lambda$, we have $s_X = 10/(5/3) = 6$.

Chapter 3

1. a) The fraction of the observations of X that are smaller than 0.684 is $\Phi(0.684)$, which is calculated as follows:

$$\begin{aligned}
\Phi(0.684) &= \Phi(0.68 + 0.004) \\
&= \Phi\left(0.68 + \frac{4}{10}\,0.01\right) \\
&= \Phi\left(0.68 + \frac{4}{10}\,(0.69 - 0.68)\right) \\
&\approx \Phi\,(0.68) + \frac{4}{10}\,\left(\Phi\,(0.69) - \Phi\,(0.68)\right) \\
&= 0.7517 + \frac{4}{10}\,(0.7549 - 0.7517) \\
&= 0.75298.
\end{aligned}$$

In `R`, $\Phi(0.684)$ is obtained by the command `pnorm(0.684)`. `R`'s return is 0.7530124. Two numbers u and v are said to agree to d decimal places if $|u - v| < 5 \times 10^{-(d+1)}$. Because $|0.7530124 - 0.75298| = 3.24171\text{e} - 05$, our approximation using the table is correct to four decimal places.

b) The fraction of the observations of X that are smaller than -0.684 is $\Phi(-0.684)$, which is calculated using the result of part (a) as follows:

$$\Phi(-0.684) = 1 - \Phi(0.684) = 1 - 0.75298 = 0.24702.$$

In `R`, $\Phi(-0.684)$ is obtained by the command `pnorm(-0.684)`. `R`'s return is 0.2469876. Our approximation using the table differs from this by 3.24e-05.

c) The fraction of the observations of X that are greater than 0.553 is $1 - \Phi(0.553)$. Because 0.553 is between 0.55 and 0.56, we find $\Phi(0.55) = 0.7089$ and $\Phi(0.56) = 0.7123$ from the table and we calculate $\Phi(0.553)$ as follows:

$$\begin{aligned}
\Phi(0.553) &= \Phi(0.55 + 0.003) \\
&= \Phi\left(0.55 + \frac{3}{10}\,0.01\right) \\
&= \Phi\left(0.55 + \frac{3}{10}\,(0.56 - 0.55)\right) \\
&\approx \Phi\,(0.55) + \frac{3}{10}\,\left(\Phi\,(0.56) - \Phi\,(0.55)\right) \\
&= 0.7089 + \frac{3}{10}\,(0.7123 - 0.7089) \\
&= 0.70992.
\end{aligned}$$

Thus, the answer, $1 - \Phi(0.553)$, is 1 - 0.70992, or 0.29008.

In `R`, the input `1 - pnorm(0.553)` mimics our use of the table. That is because the given table tabulates areas of left tails. The default of `pnorm` is also to return areas of left tails. That is, the command `pnorm(z)` is equivalent to the explicit call `pnorm(z, lower.tail = TRUE)`. For the area of the tail to the right of z, the direct command `pnorm(z, lower.tail = FALSE)` does the job. Both `1 - pnorm(0.553)` and `pnorm(0.553, lower.tail = FALSE)` result in the return 0.2901317. The difference between this answer and the approximation we obtained from the table is 5.17e-05, which means we have almost

four decimal places of accuracy. Another measure of our accuracy is to calculate the percentage error, $(0.2901317 - 0.29008)/0.2901317 \times 100\%$, or about 0.018%.

d) The fraction of the observations of X that are greater than -1.234 is $1-\Phi(-1.234)$, or $1-\big(1-\Phi(1.234)\big)$, or $\Phi(1.234)$. Because 1.234 is between 1.23 and 1.24, we find $\Phi(1.23) = 0.8907$ and $\Phi(1.24) = 0.8925$ from the table and we calculate as follows:

$$\begin{aligned}
\Phi(1.234) &= \Phi(1.23 + 0.004) \\
&= \Phi\left(1.23 + \frac{4}{10}\,0.01\right) \\
&= \Phi\left(1.23 + \frac{4}{10}\,(1.24 - 1.23)\right) \\
&\approx \Phi\,(1.23) + \frac{4}{10}\,\left(\Phi\,(1.24) - \Phi\,(1.23)\right) \\
&= 0.8907 + \frac{4}{10}\,(0.8925 - 0.8907) \\
&= 0.89142.
\end{aligned}$$

In `R`, both the command `1 - pnorm(-1.234)` and the command `pnorm(-1.234, lower.tail = FALSE)` do the job; they return 0.8913985. The difference between this answer and the approximation we obtained from the table is 2.15e-05, which means we have four decimal places of accuracy.

e) The fraction of the observations of X that are between -0.85 and 0.85 is equal to $\Phi(0.85) - \Phi(-0.85)$, or $\Phi(0.85) - \big(1 - \Phi(0.85)\big)$, or $2\Phi(0.85) - 1$. That the fraction of the observations between -z and z in a standard normal distribution is $2\Phi(z) - 1$ is a basic fact that should be known. In this case, we have $2\Phi(0.85) - 1 = 2 \times 0.8023 - 1 = 0.6046$. In `R`, the command `pnorm(0.85) - pnorm(-0.85)` returns 0.6046749.

f) The fraction of the observations of X that are between 0.57 and 0.75 is $\Phi(0.75) - \Phi(0.57)$, or $0.7734 - 0.7157$, or 0.0577. In `R`, the command `pnorm(0.75) - pnorm(0.57)` returns 0.0577115.

g) The fraction of the observations of X that are between -0.94 and 0.57 is $\Phi(0.57) - \Phi(-0.94)$, or $\Phi(0.57) - \big(1 - \Phi(0.94)\big)$, or $\Phi(0.57) + \Phi(0.94) - 1$, or 0.7157 + 0.8264 - 1, or 0.5421. In `R`, the command `pnorm(0.57) - pnorm(-0.94)` returns 0.5420524.

h) The fraction of the observations of X that are between -1.75 and -0.94 is $\Phi(-0.94) - \Phi(-1.75)$, or $\big(1 - \Phi(0.94)\big) - \big(1 - \Phi(1.75)\big)$, or $\Phi(1.75) - \Phi(0.94)$, or 0.9599 - 0.8264, or 0.1335. In `R`, the command `pnorm(-0.94) - pnorm(-1.75)` returns 0.1335496.

i) The fraction of the observations of X that are between 0 and 1.87 is $\Phi(1.87) - \Phi(0)$, or $\Phi(1.87) - 0.5$, or $0.9693 - 0.5$, or 0.4693. In `R`, the command `pnorm(1.87) - pnorm(0)` returns 0.4692581.

j) The fraction of the observations of X that are between -1.66 and 0 is $\Phi(0) - \Phi(-1.66)$, or $\Phi(0) - \big(1 - \Phi(1.66)\big)$, or $0.5 - 1 + 0.9515$, or 0.4515. In `R`, the command `pnorm(0) - pnorm(-1.66)` returns 0.4515428.

2. Think of the Φ table as a tool for solving the equation $p = \Phi(z)$. The preceding exercise was about solving for p given z. Thius exercise is about solving for z given p. When solving for p, which involves a direct use of the Φ table, a value of z is given. We locate the row and column that together give z to an accuracy of two decimal places. The entry in that row and column is the value p for which $p = \Phi(z)$. In

a reverse-lookup, the value of p is given. We locate this value in the table, if it is there. Taken together, the row and column in which p is found combine to form a number z to two decimal places for which $p = \Phi(z)$. If p is not found in the table, we find the row that has two entries, p_1 and p_2, and use those entries to approximate z.

a) To solve $\Phi(z) = 0.8$, we identify two consecutive values in one row that "bracket" 0.8: this means that one cell value is a bit smaller than p and the next cell entry in the row is a bit larger than p. For $p = 0.8$, we find the cell entries $p_1 = 0.7995$ and $p_2 = 0.8023$. They are in the row that has the p-values for $z = 0.80$ through $z = 0.89$. (It is just by chance that these z-values are so close to the given p-value.) The first of these cell entries, namely $p_1 = 0.7995$, is in the 0.04 column and the second, $p_2 = 0.8023$, is in the 0.05 column. So the z-value we seek is between $z_1 = 0.84$ and $z_2 = 0.85$. If we did not need great accuracy and wanted to save time, we could split the difference and go with 0.845 as our answer. On multiple choice exams, answers are normally spread far enough apart so the time-saving simple-averaging approach should be your first approach, if you are using the table. For a more accurate third decimal place, we will calculate a weighted average. The gap, $p_2 - p_1$, between the consecutive cell-entry p-values we found is $0.8023 - 0.7995$, or 0.0028. The gap between p_1 and the given p-value is $0.8000 - 0.7995$, or 0.0005. We will find the value of z that is in the same relative position with respect to $z_1 = 0.84$ and $z_2 = 0.85$ as p has with respect to p_1 and p_2:

$$z = z_1 + \frac{p - p_1}{p_2 - p_1}(z_2 - z_1) = 0.84 + \frac{0.0005}{0.0028}(0.01) = 0.8417857.$$

In `R`, the appropriate call is `qnorm(0.8)`, which returns 0.8416212. On a multiple choice exam, once you located 0.84 and 0.85 directly in the table, you almost certinly would be able to answer with a unique answer choice between 0.84 and 0.85.

b) To solve $\Phi(z) = 0.75$, we bracket $p = 0.75$ with the one-row cell entries $p_1 = 0.7486$ and $p_2 = 0.7517$, which correspond to $z_1 = 0.67$ and $z_2 = 0.68$, respectively. Our approximation is

$$z = z_1 + \frac{p - p_1}{p_2 - p_1}(z_2 - z_1) = 0.67 + \frac{0.75 - 0.7486}{0.7517 - 0.7486}(0.01) = 0.6745161.$$

In `R`, the appropriate call is `qnorm(0.75)`, which returns 0.6744898. The error of our approximation is 2.63e-05, so we have four decimal places of accuracy.

c) To solve $\Phi(z) = 2/3 = 0.6667$, we bracket $p = 0.6667$ with the one-row cell entries $p_1 = 0.6664$ and $p_2 = 0.6700$, which correspond to $z_1 = 0.43$ and $z_2 = 0.44$, respectively. Our approximation is

$$z = z_1 + \frac{p - p_1}{p_2 - p_1}(z_2 - z_1) = 0.43 + \frac{0.6667 - 0.6664}{0.6700 - 0.6664}(0.01) = 0.4308333.$$

In `R`, the appropriate call is `qnorm(0.6667)`, which returns 0.430819. The error of our approximation is 1.43e-05, so we have four decimal places of accuracy.

d) Solving $\Phi(z) = 0.4$ using the given table involves a step not needed in parts (a)-(c). The difference is that, in this problem, the given p-value is less than 0.5. However, for conciseness, the given table is limited to the p-range between 0.5 and 1.0. That results in a positive z-range. A p-value that is less than 0.5 corresponds to a negative z-value. Our adaptation is to use the given table and find the positive value z for which $-z$ is the required answer. The key to this is the identity $\Phi(-z) = 1 - \Phi(z)$, found in Section 3.1 with a geometric figure explaining it. Set $\Phi(-z) = 0.4$. Then $0.4 = 1 - \Phi(z)$, or $\Phi(z) = 1 - 0.4 = 0.6$. We solve for z using the method shown in detail in parts (a)-(c). We find that in the $z = 0.2...$ row, the cell entry 0.5987, which corresponds to $z = 0.25$, and the cell entry 0.6026, which corresponds to $z = 0.26$, bracket 0.6. Thus

$$z = 0.25 + \frac{0.6 - 0.5987}{0.6026 - 0.5987}(0.01) = 0.2533333.$$

Our answer is -0.2533333. That is, $\Phi(-0.2533333) = 0.4$. In `R`, the call `qnorm(0.4)` returns -0.2533471.

e) To find the z such that $\Phi(z) = 1/3$ using the given table, we follow the procedure of the preceding problem and find the solution of the equation $\Phi(z) = 1 - 1/3 = 2/3$ instead. We found the answer to this in part (c): $z = 0.4308333$. Our answer is therefore $z = -0.4308333$. In `R`, the call `qnorm(1/3)` returns -0.4307273.

f) To find the z such that $\Phi(z) = 1/4$ using the given table, we first solve for the z for which $\Phi(z) = 1 - 1/4 = 3/4 = 0.75$. We did this in part (b) and found $z = 0.6745161$. Our answer is -0.6745161: that is, $\Phi(-0.6745161) = 1/4$. In `R`, the call `qnorm(1/4)` returns -0.6744898.

3. a) From part (a) of the preceding exercise, we know that 20% of the observations of X are more than 0.8417857 standard deviations above the mean. Therefore, of the 50% of the observations in X that are above the mean, 30% are above the mean but not more than 0.8417857 standard deviations above the mean. We fill in the blank with 0.8417857.

   b) From part (a) of the this exercise, we know that 30% of the observations of X are above the mean but not more than 0.8417857 standard deviations above the mean. By symmetry, we see that 30% of the observations of X are below the mean but not more than 0.8417857 standard deviations below the mean. It follows that 60% of the observations of X are within 0.8417857 standard deviations of the mean.

   c) From part (b) of the preceding exercise, we see that about half the observations in X are within z standard deviations of the mean for $z = 0.6745161$. (25% of the observations are more than 0.6745161 standard deviations above the mean and, by symmetry, 25% of the observations are more than 0.6745161 standard deviations below the mean. If we exclude these 50% of the observations, we are left with 50% of the observations and they must lie within 0.6745161 standard deviations of the mean.

   d) From part (e) of the preceding problem, we see that 1/3 of the observations of X are more than 0.4308333 standard deviations below the mean. Because half of the observations of X are below the mean, we deduce that 1/2 - 1/3, or 1/6 of the observations are below the mean but not more than 0.4308333 standard deviations below the mean.

4. a) $x = \overline{X} + z \cdot (4.8) = 65 + (-0.37)(4.8) = 63.224$.
   b) $x = \overline{X} + z \cdot (5/7) = 3.91 + (1.43)(5/7) = 4.931429$.

5. Zoe's IQ z-score is $(125-100)/15$, or 1.666667. This is the same as her z-score on the SAT. Destandardizing, we find her raw SAT score x from the equation $(x-500)/100 = 1.666667$, or $x = 500+1.666667\times100$, or $x = 666.6667$. This rounds to 667. (SAT scores are integers.)

6. a) Zenobia's z-score is $(755-700)/40$, or 1.375. Because $\Phi(1.375) = 0.9154343$, Zenobia's LSAT score is higher than 91.54% of her fellow law students at the school. That means that 8.46% have higher scores.
   b) Zelda's LSAT z-score is -1.34. We calculate $\Phi(-1.34)$ with either a calculator or a statistical computer program such as `R`. Or, we can calculate $1 - \Phi(1.34)$ using the table in this chapter. However we do the calculation, we get 0.09012267, which means that 9.01% of law students at the university have lower LSAT scores than Zelda's.
   c) Here we do a reverse lookup to solve for the z-score z for which $\Phi(z) = 0.95$. We find $z = 1.644854$, which means that Ziva's raw LSAT score is 1.644854 standard deviations above the school's mean. Therefore, it is $700 + 1.644854 \times 40$, or 766 (on rounding to the nearest integer).
   d) Here we do a reverse lookup to solve for the z-score z for which $\Phi(z) = 0.13$. In `R`, the command `qnorm(0.13)` returns the desired z-score: -1.126391. This corresponds to a raw LSAT score of $700 -$

1.126391 × 40, or 655 (on rounding to the nearest integer). Using the table instead, we first solve $\Phi(z) = 1 - 0.13 = 0.87$. The z-score we seek will be the negative of this solution. From the table, we see that $\Phi(1.12) = 0.8686$ and $\Phi(1.13) = 0.8708$. Our approximation of the z-score is

$$z = 1.12 + \frac{0.8700 - 0.8686}{0.8708 - 0.8686} \times 0.01 = 1.126364.$$

The z-score we seek is therefore -1.126364, which leads to the answer $700 - 1.126364 \times 40$, or 655.

7. a) The z-score of 772 is given by $z = (772 - 500)/100$, or $z = 2.72$. The corresponding raw ACT score is $21 + 2.72 \times 5$, or 34.6, which rounds up to 35.
b) The z-score of 28 is given by $z = (28 - 21)/5$, or $z = 1.4$. The corresponding raw SAT score is $500 + 1.4 \times 100$, or 640.
c) The z-score of 24 is $(24-21)/5$, or $z = 0.6$. The z-score of 30 is $(30-21)/5$, or $z = 1.8$. The fraction of ACT scores between these two z-scores is $\Phi(1.8) - \Phi(0.6)$, or $0.9640697 - 0.7257469$, or 0.2383228. Thus, 23.83228% of ACT test takers have scores between 24 and 30. Next we calculate the z-score of 540. It is $(540 - 500)/100$, or 0.4. The Φ-value of 0.4 is 0.6554217. This means that 65.54217% of SAT takers have scores lower than 540. We are interested in the SAT score b for which $(65.54217 + 23.83228)$% of SAT test takers have scores less than b. Because $65.54217 + 23.83228 = 89.37445$, we seek the value of z such that $\Phi(z) = 0.8937445$. We solve that $z = 1.246691$. This z-score corresponds to the raw SAT score $b = 500 + 1.246691 \times 100$, or $b = 625$. We will check this answer using `R` as follows: the commands `pnorm(30, mean = 21, sd = 5) - pnorm(24, mean = 21, sd = 5)` and `pnorm(625, mean = 500, sd = 100) - pnorm(540, mean = 500, sd = 100)` return 0.2389285 and 0.2383228, which differ by only a small rounding error.

8. Let m and s be the mean and standard deviation of X. From the lines for observations 4 and 5, we obtain the two equations $0.7385489 = (20 - m)/s$ and $0.9847319 = (21 - m)/s$. After clearing the denominators, we have $20 - m = 0.7385489\,s$ and $21 - m = 0.9847319\,s$. Subtract the left side of the first equation from the left side of the second, and the right side of the first equation from the right side of the second. The resulting equation is $1 = (0.9847319 - 0.7385489)\,s$, or $s = 4.062018905$. It follows that $m = 20 - (0.7385489)(4.062018905)$, or $m = 17$. We now have the last two entries of the Raw column. The z-score z_1 of observation x_1 is immediate: $z_1 = (11 - 17)/4.062018905$, or $z_1 = -1.477098$. Also, we can transform the z-score of observation 3 back to a raw score x_3: $x_3 = 17 + (0.246183)(4.062018905)$, or $x_3 = 18$. Next we tansform the T-score $t_2 = 45.07634$ of observation 2 to a z-score z_2: $z_2 = (45.07634 - 50)/10 = -0.492366$. We next transform z_2 to the raw observation x_2: $x_2 = 17 + (-0.492366)(4.062018905)$, or $x_2 = 15$. We now have the entire first column. We still have four z-scores, z_1, z_3, z_4, z_5, to transform to T-scores: $t_1 = 50 + 10 \cdot (-1.477098) = 35.22902$, $t_3 = 50+10\cdot(0.246183) = 52.46183$, $t_4 = 50+10\cdot(0.7385489) = 57.38549$, $t_5 = 50+10\cdot(0.9847319) = 59.84732$. All that remains are the means and standard deviations of the z-scores and T-scores. We know that there are supposed to be 0 & 1 and 50 & 10. Because we have all the data, verification is routine. The completed table is

Observation	Raw	z-score	T-score
1	11	-1.477098	35.22902
2	15	-0.492366	45.07634
3	18	0.246183	52.46183
4	20	0.7385489	57.38549
5	21	0.9847319	59.84732
Mean	17	0	50
Standard Devtn	4.062018905	1	10

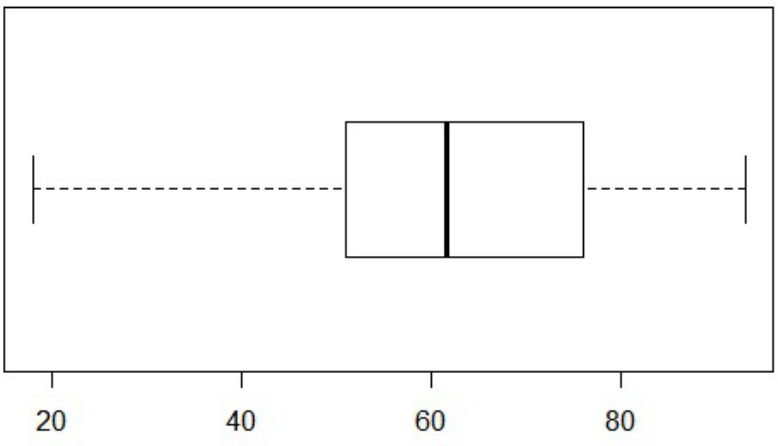
9. a) For every x, its z-score is $z = (x - 62.02083)/19.21711$. The quartiles are $Q_1 = 51, Q_2 = 61.5, Q_3 = 76$. Their z-scores are -0.5734905, -0.02710241, and 0.7274335, respectively.

   b) We would expect the z-score of Q_3 to be the solution z of the equation $\Phi(z) = 0.75$. We calculated this to be z = 0.6745161 in part (b) of Exercise 2. In a symmetric distribution, the median equals the mean. Subtracting the mean from the median results in 0. That would be the z-score of the median in a distribution that can be modelled by a normal distribution. The z-score of Q_1 would be the negative of the z-score of Q_3, i.e., -0.6745161. We see that the z-scores of the first and third quartiles in a normal distribution are not really close to the observed z-scores for these quartiles.

   c) The z-score of the largest observation is (93-62.02083)/19.21711, or 1.612062.

   d) The fraction of observations less than or equal to an observation with z-score 1.612062 is $\Phi(1.612062)$, or 0.9465258. For a distribution of 48 observations, that would come to 45 or 46 observations. In this distribution, there were 48.

   A boxplot shows at a glance that the distribution under consideration cannot be accurately modelled by a normal distribution. The median is visably noncentral in the box and the whiskers are far from having equal length.



Boxplot of Neonate Mother-Gazing Distribution

10. Let X denote the given distribution. The mean of X is 20 and the standard deviation is 7.438638. Let $z = (x - 20)/7.438638$ for every x in the distribution X. The z-scores that result from this transformation are, in order, -0.5377329, -0.2688664, 1.4787654, -0.4032996, 1.2098989, -1.3443321, -0.1344332. Routine calculations of the mean and standard deviation of the transformed distribution are 0 and 1.

11. Let X be the distribution of the copy machine's lifespan. The mean and standard deviation are, respectively, 42 and 7. The z-score of 28 is $(28 - 42)/7$, or -2. Let z_2 be the z-score of the still unknown number n. Because $190/200 = 0.95$, the equation we must solve is $\Phi(z_2) - \Phi(-2) = 0.95$, or $\Phi(z_2) = 0.95 + \Phi(-2)$, or $\Phi(z_2) = 0.95 + (1 - \Phi(2))$, or $\Phi(z_2) = 0.95 + 1 - 0.9773$, $\Phi(z_2) = 0.9727$. From the table of Φ-values, we see that z_2 is between 1.92, for which $\Phi = 0.9726$, and 1.93, for which $\Phi = 0.9732$. We calculate the following approximation of the required z-score: $z_2 = 1.92 + 0.01 \times (9727 - 9726)/(9732 - 9726) = 1.9217$.

The number of months t of a copier lifespan which $z_2 = 1.9217$ is the z-score is $(t - 42)/7 = 1.9217$, or 55.45. (Only one choice, 56, in the multiple choice answer list is close. In fact, the answer 55 for n would render the salesman's assertion untruthful, and we all know that salesmen are never untruthful.

12. From class lectures, we know that the third quartile is about 2/3 of a standard deviation above the mean. That is, the z-score of the raw value x that we seek is approximately 2/3. Destandardizing, we solve the equation $2/3 = (x - 42)/7$ for x. We obtain the approximate answer $42 + (2/3) \times 7$, or 46.67, which rounds up to 46.7, the correct answer on the exam. A more accurate approximation can be obtained by using the Phi table and interpolation. Because $\Phi(0.67) = 0.7486$ and $\Phi(0.68) = 0.7517$, we obtain $\Phi(z) = 0.75$ for $z = 0.67 + ((7500 - 7486)/(7517 - 7486)) \cdot 0.01$, or $z = 0.67452$. The raw score x is then the solution of $0.67452 = (x - 42)/7$, or $x = 46.72164$, which also rounds to 46.7. For those of you accustomed to the greater accuracy of your calculator, the z-score we seek is obtained in R via the command `qnorm(0.75)` and is 0.6744898. The raw score we seek, which is the answer to the problem, can be obtained from one R command, `qnorm(0.75, 42, 7)`, which gives 46.72143.

13. The z-score of 36 is $(36 - 42)/7$, or -0.8571. The fraction we seek is an area of a left (good!) tail for a negative (bad!) z-score, so we covert it to the area of a right (bad!) tail of a positive z-score, and then we subtract the fraction from 1: $1 - \Phi(0.8571)$, or $1 - \Phi(0.8571)$. From the table of Φ-values, we see that $\Phi(0.85) = 0.8023$ and $\Phi(0.86) = 0.8051$. Therefore, to a reasonably good approximation,

$$\Phi(0.8571) = 0.8023 + 0.71 \times (0.8051 - 0.8023) = 0.804288.$$

The fraction that will fail before 36 months is approximately $1 - 0.804288$, or 0.1957. As a percentage, about 19.6%.

14. The10 manufacturer wants to reduce the 36-month failure rate to only 10%.

Let μ denote the desired mean lifespan. With that as the average, the z-score of 36 becomes $z_0 = (36 - \mu)/7$, assuming the standard deviation does not change. It is desired that $\Phi(z_0) = 0.10$, or $1 - \Phi(-z_0) = 0.10$, or $\Phi(-z_0) = 0.90$. (Because 0.10 is less than 1/2, we see that z_0 is negative, and we must convert to the positive z-value, $-z_0$.) From the table of Φ-values, we see that $\Phi(1.28) = 0.8997$ and $\Phi(1.29) = 0.9015$. Our interpolated approximation of z_0 is

$$-z_0 = 1.28 + 0.01 \times \frac{9000 - 8997}{9015 - 8997} = 1.2817.$$

We substitute -1.2817 into the z-score formula, $z_0 = (36 - \mu)/7$, and solve: $-1.2817 = (36 - \mu)/7$, or $\mu = 36 + 7 \times 1.2817$, or 44.97 months.

15. A competing manufacturer says that not only will 90% of their copiers

There are two unknown parameters to find: the mean μ and the standard deviation σ. We expect to need two equations to solve (simultaneously) for two unknowns. That is why two facts are given. We must translate each fact into an equation involving μ and σ. To that end, the z-score of 36 is $z_1 = (36 - \mu)/\sigma$ and the equation for that score is $\Phi(z_1) = 1 - 0.9$, or $\Phi(z_1) = 0.1$. The z-score of 42 is $z_2 = (42 - \mu)/\sigma$ and the equation for that score $\Phi(z_2) = 1 - 0.65$, or $\Phi(z_2) = 0.35$. Both Φ-values are less than 1/2, which means the z-scores are negative, so we must convert the equations to equivalent ones in terms of positive z-scores: $\Phi(-z_1) = 1 - 0.1$, or $\Phi(-z_1) = 0.9$, and $\Phi(-z_2) = 1 - 0.35$, or $\Phi(-z_2) = 0.65$. From the table of Φ-values we find $-z_1 = 1.2817$ (details of this very calculation are in the preceding exercise) and $-z_2 = 0.38 + 0.01 \times (6500 - 6480)/(6517 - 6480)$, or $= -z_2 = 0.3854$. The equations we must solve are $-1.2817 = (36 - \mu)/\sigma$ and $-0.3854 = (42 - \mu)/\sigma$ Solving these two equations simultaneously, we find $\mu = 44.58$ and $\sigma = 6.69$. The answer is N(44.6, 6.7).

16. Suppose a normal model has mean $\mu = 5$ and standard deviation $\sigma = 3$.

We first calculate the z-scores z_1 and z_2 of 11 and -1 respectively. We have $z_1 = (-1-5)/3$, or $z_1 = -2$, and $z_2 = (11-5)/3$, or $z_1 = 2$. So, we are being for the percentage of data within 2 standard deviations from the mean. The reason the examiner added "to 2 decimal places" is because the 68-95-99.7 Rule (Empirical Rule) tells us to expect *about* 95% of the data within two sigs of the mean. Good question so far, but the answer list of multiple choices offers increasing numbers that conclude with ... d) 86.64So, only two answers are plausible, (e) and (f). However, the ballpark Empirical Rule approximation, 95%, is certainly not correct to two decimal places. That leaves only one likely choice, but let's continue and verify its correctness. For any nonnegative number z, the exact value for the percentage of data that is expected to be within z standard deviations of the mean is $100 \cdot (2\,\Phi(z) - 1)$. Setting $z = 2$, we see that the percentage of data expected between -1 and 11 is $100 \cdot (2\,\Phi(2) - 1)$, or $100 \cdot (2 \times 0.9772 - 1)$, or 95.44. As often happens when using tables, the last digit is suspect. The listed answer is more accurate.

17. The author does not ask questions like this. In fact, he strenuously objects to them. However, to avoid having to answer ths question multiple times orally, he will set down his thoughts in writing. The correct answer is supposed to be (C). We can rule out (A). The number of cigarettes smoked daily is bimodal: 0 is the mode for nonsmokers, and there is some positive mode for smokers. A normal model is unimodal, so (A) is out. (Furthermore, this variable certainly skews right and is not symmetric.). We can reject (B) for the same reason: the variable skews right. The author believes that the variable of choice (D) is bimodal because there are two groups of students: those who spend a positive amount of time on homework and those who spend no time. If *Hours of Homework Last Week* is bimodal, then we can rule out a normal model. In any event, it seems likely that this variable skews right and is therefore asymmetric. We reject answer (D). Of course, categorical variable *Eye Color* may be instantly dismissed. That leaves (C) as the only choice, so it must be the correct answer, right? It does seem likely that *Head circumference* follows a normal model, but it requires spacialized knowledge that the distribution is not bimodal. (Women's 'head circumferences tend to be smaller than men's, but not enough to make the distribution bimodal.)

18. We first calculate the z-value for which $\Phi(z) = 0.90$. Using the R code `qnorm(0.90)` we find $z = 1.281552$. Maple code for this is `statevalf[icdf,normald[0,1]](0.90)`, resulting in $z = 1.281551566$. Using tables we obtain a reasonably accurate reverse lookup approximation. We see that 0.90 is between the last two entries, 0.8997 and 0.9015, of the row beginning with $z = 1.2$. Thus, the z-value we seek is between 1.28 and 1.29. Our interpolation gives us

$$z = 1.28 + \frac{0.9000 - 0.8997}{0.9015 - 0.8997} \times (1.29 - 1.28) = 1.28167.$$

(Our approximation is off by only 12/100,000.) Next we destandardize: we find the number of words x in the female cohort that corresponds to this z-score. From $(x - 14297)/6441 = 1.281552$, we obtain $x = 22551.47$. (Nadia was pronouncing the name "Otto" but only finished the first syllable when the clock struck midnight.) From the given information—Aidan's daily word count is the same as Nadia's, we have answered the first question.

To continue with the male calculation, we reverse everything. (The author did warn of perceptions about Aidan.) The two steps for Nadia were reverse lookup and destandardization, in that order. For Aidan, the two steps will be standardization and direct lookup, in that order. Reverse processes, reverse order. (This problem has everything.)

The z-score of 22551.47 in the male cohort is $z = (22551.47 - 14060)/9065$, or $z = 0.9367313845$. We finish with a direct lookup of $\Phi(0.9367313845)$. In R, `pnorm(0.936731)` yields 0.8255516. In Maple, `statevalf[cdf,normald[0,1]](0.9367313845)` yields 0.8255516259. Using the looked-up values $\Phi(0.93) = 0.8238$ and $\Phi(0.94) = 0.8264$, our approximation from the Phi table gives

$$\Phi(0.9367) \approx \Phi(0.93) + \frac{0.9367 - 0.93}{0.94 - 0.93}\big(\Phi(0.94) - \Phi(0.93)\big) = 0.8238 + \frac{0.9367 - 0.93}{0.94 - 0.93} \times (0.8264 - 0.8238) = 0.825542.$$

All roads have led to Rome: no matter how we calculate it, Aidan is more voluble than about 82.55% of his cohort. In other words, although only 10% of women are more talkative than Nadia (and therefore Aidan), 17.5% of men are. How can that be? According to the same study, women are, on average, more talkative. Still, the means for men and women are not that far apart. On the other hand, the standard deviation (spread from the mean) for men is significantly greater. Although Aidan has the same nonstandardized score as Nadia, the greater spread in the male cohort means that there will be more men further from the mean, i.e., men with greater word counts than Aidan, hence his lower percentile.

19. Let m and s be the unknown mean and standard deviation of X. We begin by solving the equation $\Phi(z) = 0.05$. The solution is the negative number z such that $\Phi(-z) = 0.95$. We find that $-z = 1.644854$, or $z = -1.644854$. This leads to the equation $(10 - m)/s = -1.644854$, or $m = 10 + 1.644854\,s$. We will come back to this equation in the two unknowns m and s after we derive a second equation that involves them. For that, we solve the equation $\Phi(z) = 0.75$, finding $z = 0.6744898$. This leads to the equation $(13 - m)/s = 0.6744898$, or $m + 0.6744898\,s = 13$. If we replace m in this equation with the value of m found in the first equation, then we have $(10 + 1.644854\,s) + 0.6744898\,s = 13$, or $2.319344\,s = 3$, or $s = 1.293469$. Thus $m = 10 + 1.644854 \times 1.293469 = 12.12757$. The average age at which a baby learns to walk is a few days after the babies first birthday.

20. Dear First President University Student:
Thank you for this queestion. It indicates that the method of interpolation shown in the notes was not fully explained. The answer to your question will attempt to do better. No promises.
The calculation you have photographed is not actually an interpolation. The first line is correct. The second line is not incorrect, but it represents the first step toward going off the rails. A more meaningful way to express 0.004 would be $\left(\frac{4}{10}\right) \times (0.01)$ instead of $\left(\frac{1}{10}\right)(0.04)$. The third line is also not incorrect, but it does indicate that the calculation is doomed to fly off the tracks. Writing $0.04 = 0.09 - 0.05$ is arbitrary. For example, it is also true that $0.04 = 7772777.09 - 7772777.05$, which is also arbitrary and, truth to tell, much more bizarre than your difference. The fourth line shows that the calculation has entirely derailed. Even if all was well, the equality sign on this line should be an approximation sign, but we will not quibble much over that. The problem is that $\Phi(0.09)$ and $\Phi(0.05)$ have nothing whatsovever to do with values of $\Phi(z)$ near $z = 1.684$. The calculation should begin

$$\begin{aligned} \Phi(0.684) &= \Phi(0.68 + 0.004) \\ &= \Phi\left(0.68 + \frac{4}{10}\,(0.01)\right) \\ &= \Phi\left(0.68 + \frac{4}{10}\,(0.69 - 0.68)\right). \end{aligned}$$

Notice how the numbers used, 0.68 and 0.69, are the numbers that bracket the z-value we seek. The next step is to turn the left side of the last line into an approximation that can be evaluated from the table:

$$\Phi(0.684) \approx \Phi\,(0.68) + \frac{4}{10}\left(\Phi(0.69) - \Phi(0.68)\right) = 0.7517 + \frac{4}{10}\,(0.7549 - 0.7517) = 0.75298.$$

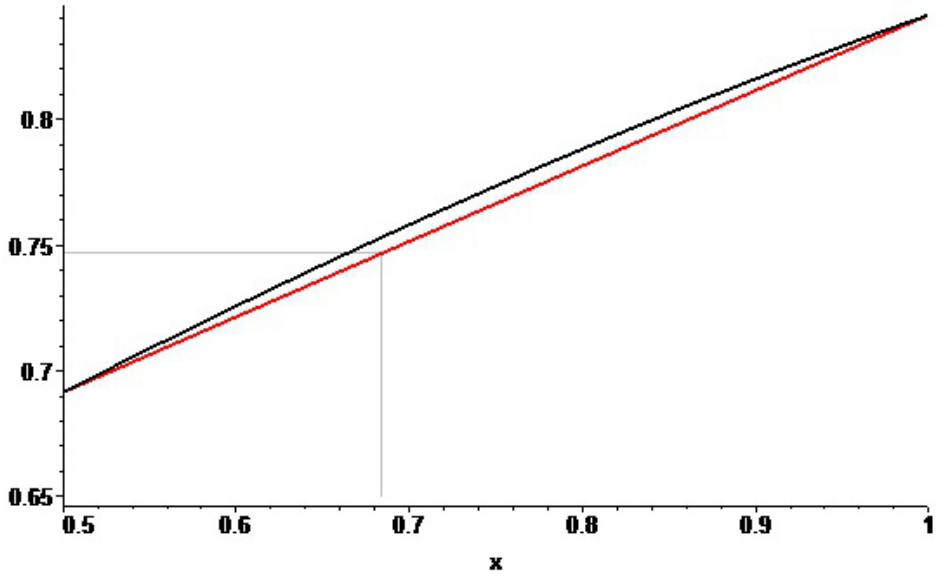
Let us look at the geometry of this approximating interpolation. In an interpolation, we use known values $\Phi(a)$ and $\Phi(b)$ to approximate $\Phi(z)$ for a value of z between a and b. The approximation is to use the line segment that joins the point $\big(a, \Phi(a)\big)$ to $\big(b, \Phi(b)\big)$. If (z, p) is a point on that connecting line segment, then we use $p \approx \Phi(z)$. The equation for p is

$$p = \Phi(a) + \frac{\big(\Phi(b) - \Phi(a)\big)}{(b - a)}\,(z - a).$$

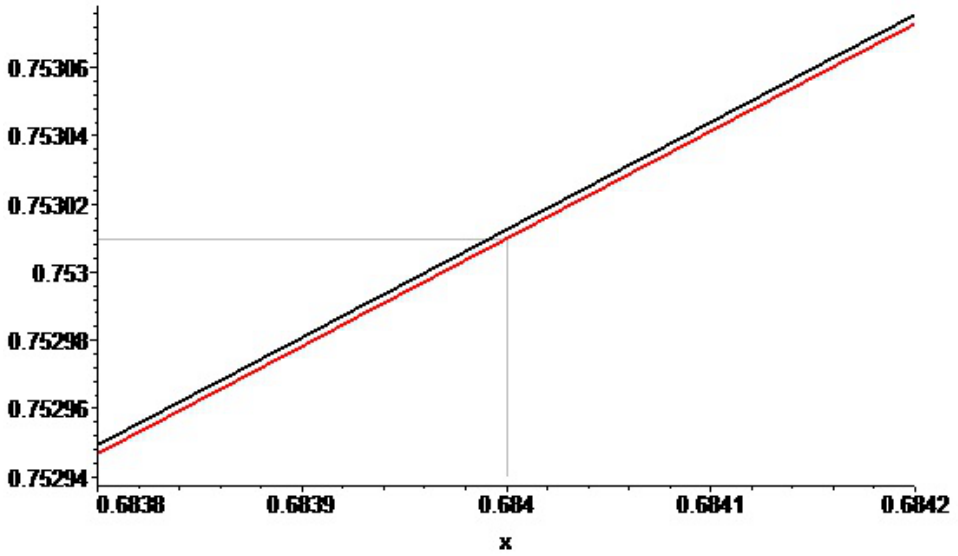
Thus,

$$\Phi(z) \approx \Phi(a) + \frac{\big(\Phi(b) - \Phi(a)\big)}{(b-a)}\,(z-a).$$

Suppose that the table of values that we had was not very refined and went up by increments of 0.5. We would bracket 0.684 with $a = 0.5$ and $b = 1.0$. The approximating line segment is shown in red in the figure below. The ordinate over 0.684 on the line segment is 0.746619, which is reasonably close to the actual value, 0.753012, but not an extremely precise approximation.



Now let us use the bracketing values $a = 0.68$ and $b = 0.69$ that are available in our table. We have not plotted the entire approximating line segment over the interval from $a = 0.68$ to $b - 0.69$—had we done so, we would not have been able to distinguish the graph of $\Phi(z)$ from the line segment approximating it. Instead, we put a very narrow interval centered at 0.684 under the microscope. The figure shows that the approximation 0.75298 is extremely close to the actual value.



Chapter 4.

1. $r = 0.8839$

2. $r = 0.8868, \quad \tau = 0.6, \quad \rho = 0.7714$

3. Let us reorder the first set X to be in increasing order. Reorder Y in the same way (so that there are no changes in the paired bivariate data, only in the positions that they occupy).

   Then,
   For Y $= y_{[1]}, y_{[2]}, y_{[3]}$, we calculate $\rho = 1.0$
   For Y $= y_{[1]}, y_{[3]}, y_{[2]}$, we calculate $\rho = 0.5$
   For Y $= y_{[2]}, y_{[1]}, y_{[3]}$, we calculate $\rho = 0.5$
   For Y $= y_{[2]}, y_{[3]}, y_{[1]}$, we calculate $\rho = -0.5$
   For Y $= y_{[3]}, y_{[1]}, y_{[2]}$, we calculate $\rho = -0.5$
   For Y $= y_{[3]}, y_{[2]}, y_{[1]}$, we calculate $\rho - 1.0$.

4. The paired rankings are (1,2), (2,1), (3,4), and (4,3) (where, as usual in statistics, the ranking numbers are lowest to highest). Comparing (1,2) to the three pairs that follow, there are two concordant pairs and one discordant pair. Comparing (2,1) to the two pairs that follow, there are two concordant pairs. Comparing (3,4) to the pair that follows, there is one discordant pair. Therefore,

   $$\tau = \frac{(2-1)+2-1}{6} = \frac{1}{3}.$$

   Pearson's r is routinely calculated to be 0.6. Spearman's ρ is *exactly* the same thing, 0.6, (because ranking rankings doesn't change the values at all).

5. We are to correlate the following data:

Bowtie neutral	1	2	3	4	5
Bowtie reviler	5	2	3	4	1

   The paired rankings are (1,5), (2,2), (3,3), (4,4), and (5,1) (where, as usual in statistics, the ranking numbers are lowest to highest). Comparing (1,5) to the four pairs that follow, there are four discordant pairs. Comparing (2,2) to the three pairs that follow, there are two concordant pairs and one discordant pair. Comparing (3,3) to the two pairs that follow, there is one concordant pair and one discordant pair. Comparing (4,4) to the pair that follows, there is one discordant pair. Therefore,

   $$\tau = \frac{-4+(2-1)+(1-1)-1}{\frac{5\cdot 4}{2}} = \frac{-4}{10} = -0.4.$$

   Pearson's r is routinely calculated to be -0.6. Spearman's ρ is *exactly* the same thing, -0.6, (because ranking rankings doesn't change the values at all).

6. We are to correlate the following data:

Bowtie neutral	1	2	3	4	5
Bowtie reviler	80	50	60	70	0

   Kendall's *tau* is unchanged from the previous problem because the first rows are identical and the second rows, though changed, are in the same order: $\tau = -0.4$. Pearson's r is routinely calculated: $r = -0.71074$. Spearman's ρ is the Pearson's r of the preceding exercise: $\rho = -0.6$.

7. How ever we think to calculate it, Kendall's τ comes to 0.2. We could correlate the sequence 7, 8, 9, 10, 11 with the sequence 9, 7, 12, 8, 10. Or we could correlate the rankings of these sequences: 1, 2, 3, 4, 5 and 3, 1, 5, 2, 4. Or we could correlate the rankings in the usual low-to-high order of statistics: 5, 4, 3, 2, 1 and 3, 5, 1, 4, 2. The Kendall τ correlations all some to 0.2.
   The Pearson's r correlation of the sequence 7, 8, 9, 10, 11 with the sequence 9, 7, 12, 8, 10 comes to 0.246598.
   In this exercise, Spearman's ρ is *not* the same as Pearson's r. That is because the entries in the table are not the rankings of the cases in the lists that are being correlated: they are rankings in different lists. To calculate Spearman's ρ, we must rank them within their lists. Doing so results in the sequences 1, 2, 3, 4, 5 and 3, 1, 5, 2, 4. Pearson's r correlation for these sequences is 0.3 and that is Spearman's ρ for the tabulated entries.

Chapter 5.

1. (i) $\bar{x} = 3, \bar{y} = 10, s_X = 2.645751311$, and $s_Y = 9.165151390$.
   (ii) $r = 0.9897433186$
   (iii) $y = 3.428571429\,x - .2857142857$
   (iv) 1.142857143, 1.428571428, 0.28571428
   (v) 168.0000000, 164.5714287, 3.428571431
   (vi) 0.9795918367

2. The regression line for z-scores passes through the origin. Using the origin and the given point (1,0.42) to calculate the slope of the regression line, we see that the slope is 0.42. But theory tells us that the slope is Pearson's r. Therefore $r = 0.42$. The amount of variation of Y that can be accounted for by the linear model is r^2, or 0.42^2, or 0.1764.

3. From the coefficient of x in the equation $y = 2x+1$, we obtain $r \cdot s_Y/s_X = 2$. Similarly, $r \cdot s_X/s_Y = 0.6$. If we multiply together corresponding sides of these two equations, then the unknown standard deviations cancel and we get $r^2 = 1.2$. On square-rooting this becomes $r = \pm 1.095$. But $r > 0$ (because the slopes of the regression lines are positive). Therefore, $r = 1.095$. Of course, this answer is baloney because $-1 \leq r \leq 1$.

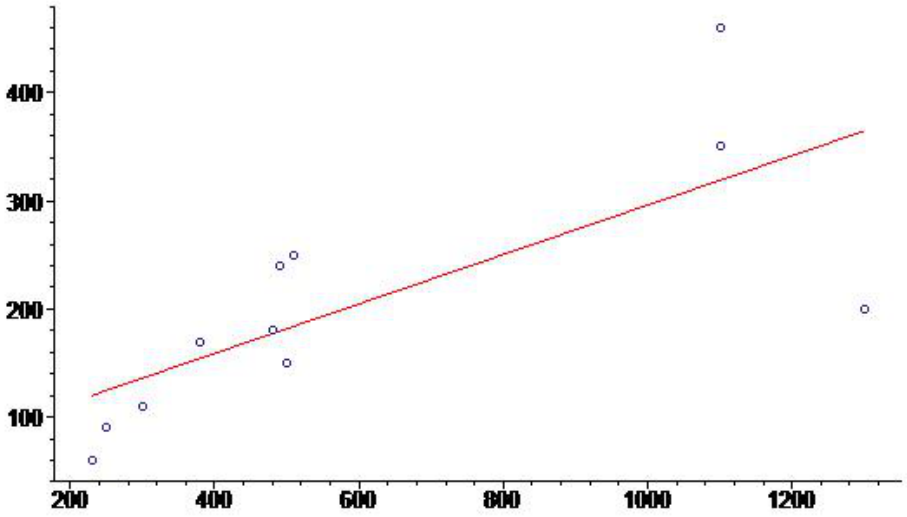
4. The graph is $\mathrm{Y} = \mathrm{X}^2$, or $(\text{square-root of Y})^2 = \mathrm{X}^2$. The relationship between X and square-root of Y is linear. (The scatter would be linear but it would only occupy the first quadrant.)

5. Not influential with low leverage and low residual.

   First, let's get the terms straight. The point P at the far right is not a *regression line outlier*. Had it been included in the plot, the regression line would pass close to the point in question. That point *is* a scatter plot outlier and each of its coordinates is an outlier for the variable for which it is an observed datum. According to the definition we are using, P is a "high leverage point" because its x-coordinate is far from $\bar{x}$. In our definition, which is a standard one, we do not say such a point has high or low leverage. Indeed, it would be rather silly if we said that a high leverage point had low leverage. The reason a point is said to be high leverage is because it has the *potential* to greatly influence the parameters of the regression line. Presumably, the instructor who made up this problem uses the term "high leverage" to mean that the point greatly affects the regression line equation and "low leverage" to mean little affect. We use the terms "influential" and "not influential." In this example, imagine the regression line with and without P. Not much difference. So P is not influential ("low leverage"). Also, because the regression line passes nearby, the residual is small.

6. (i) $y = 0.2284\,x + 67.56$
   (ii) Without Finland, the regression line is $y = 0.2138\,x + 72.56$. Finland is not influential. Without

Great Britain, the regression line is $y = 0.1621\,x + 90.21$. Great Britain is influential. Without U.S., the regression line is $y = 0.3687\,x + 9.139$. U.S. is influential.
(iii) Here is the scatter plot of the 11 countries together with the regression.

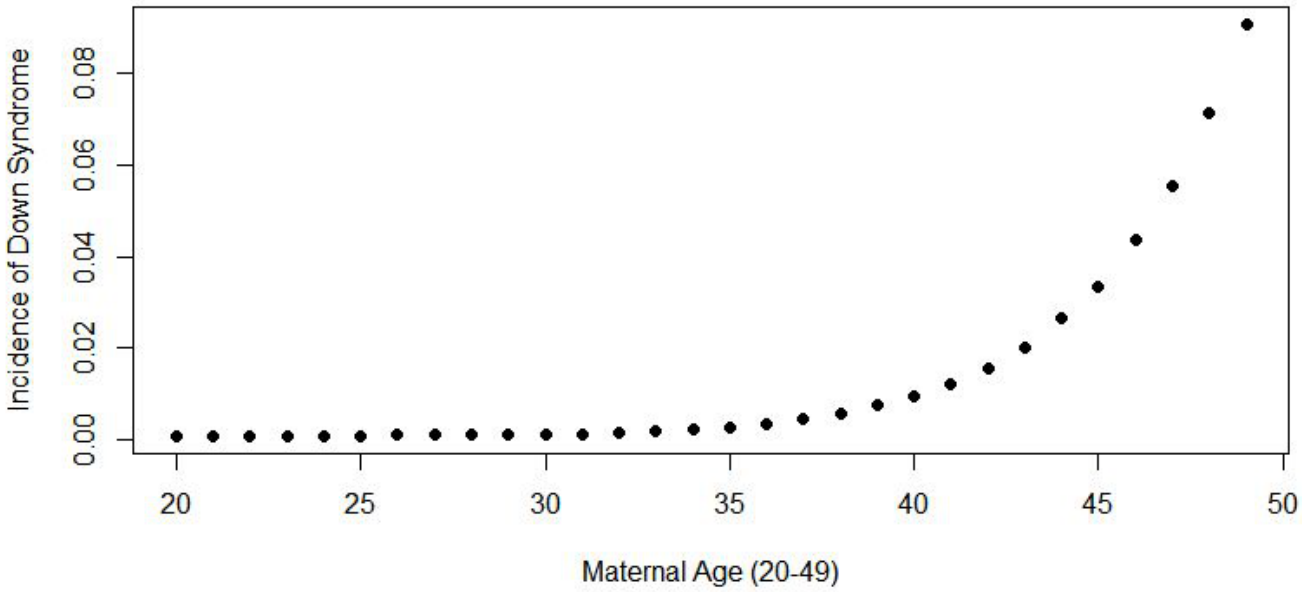


The points corresponding to Great Britain and the United States are regression line outliers. You will see in the list of all residuals that they have the largest absolute residuals.

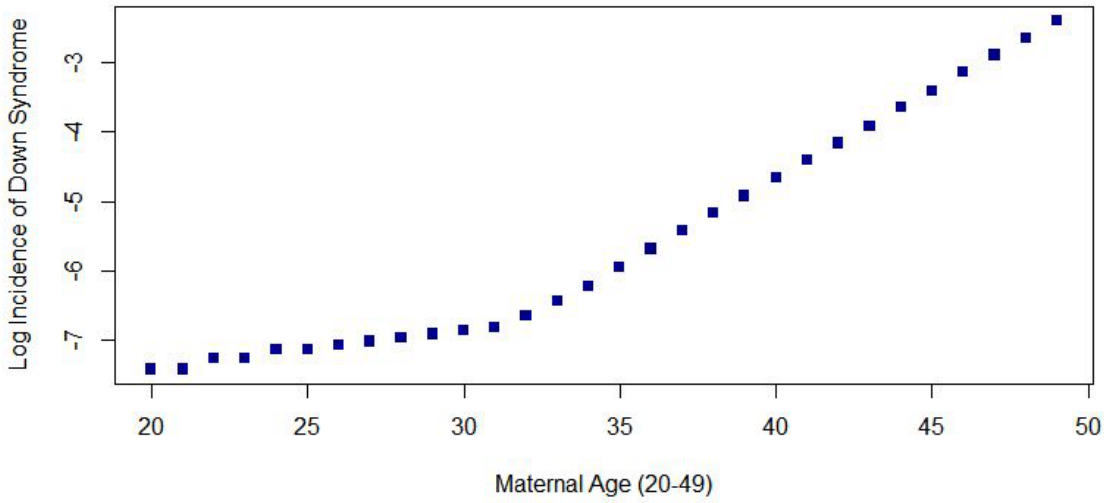
Country	Absolute Residual
Australia	2.8
Canada	31.8
Denmark	15.6
Finland	31.2
Great Britain	141.2
Iceland	60.1
Netherlands	60.5
Norway	34.7
Sweden	26.1
Switzerland	65.9
United States	164.5

Eleven Country Lung Cancer Mortality, 1950—Residuals

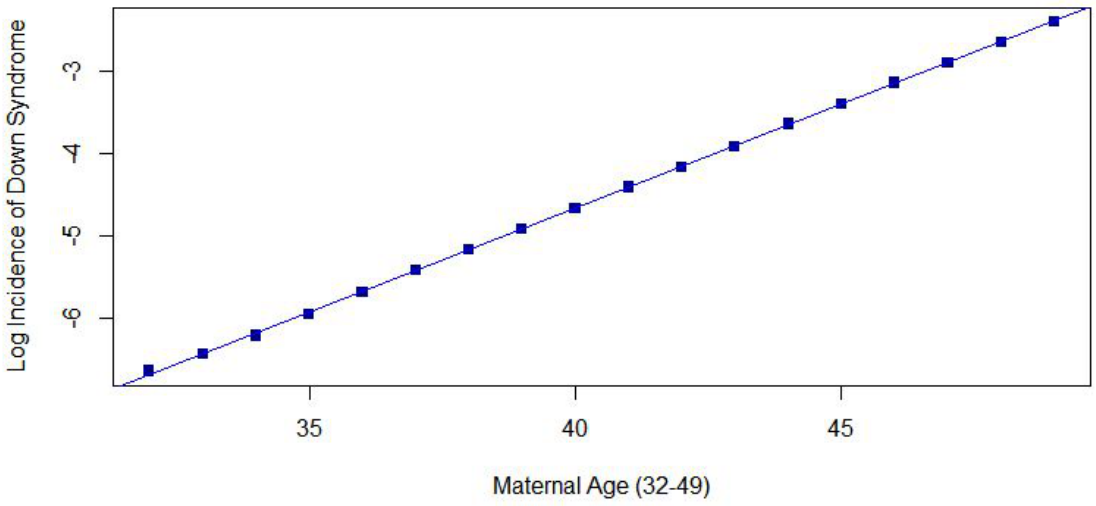
7. a) 0.7610731
b) A scatterplot should have been the first step. If a scatterplot does not suggest that the data follows a linear model, then a calculated value of r is likely to have little meaning. The following figure shows that the relationship between Maternal Age and Risk is not linear at all.



c) 0.9684005
d) We are not there yet. The transformed data is still nonlinear.



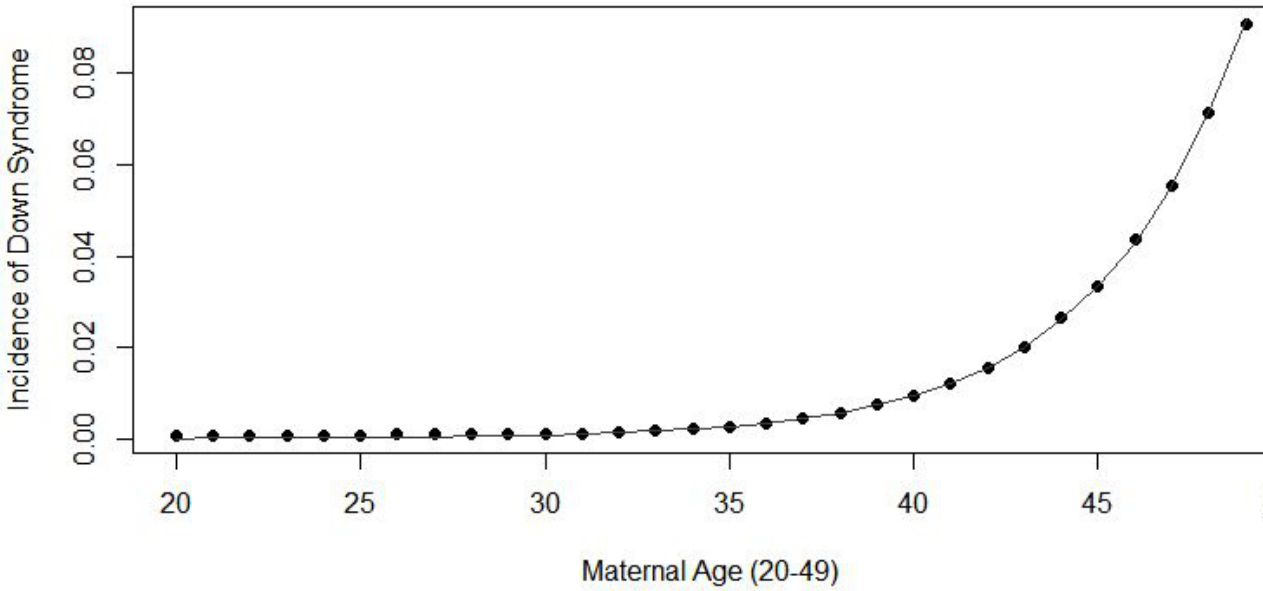
e) 0.9999339 (If you ever find bivariate data in the wild with a greater value of r, then kindly inform the author.)
f) The regression line clearly passes very close to every plotted point.



g) The regression line is $\ln(y) = 0.2531\,x - 14.7885$, or $y = \exp(0.2531\,x - 14.7885)$, or

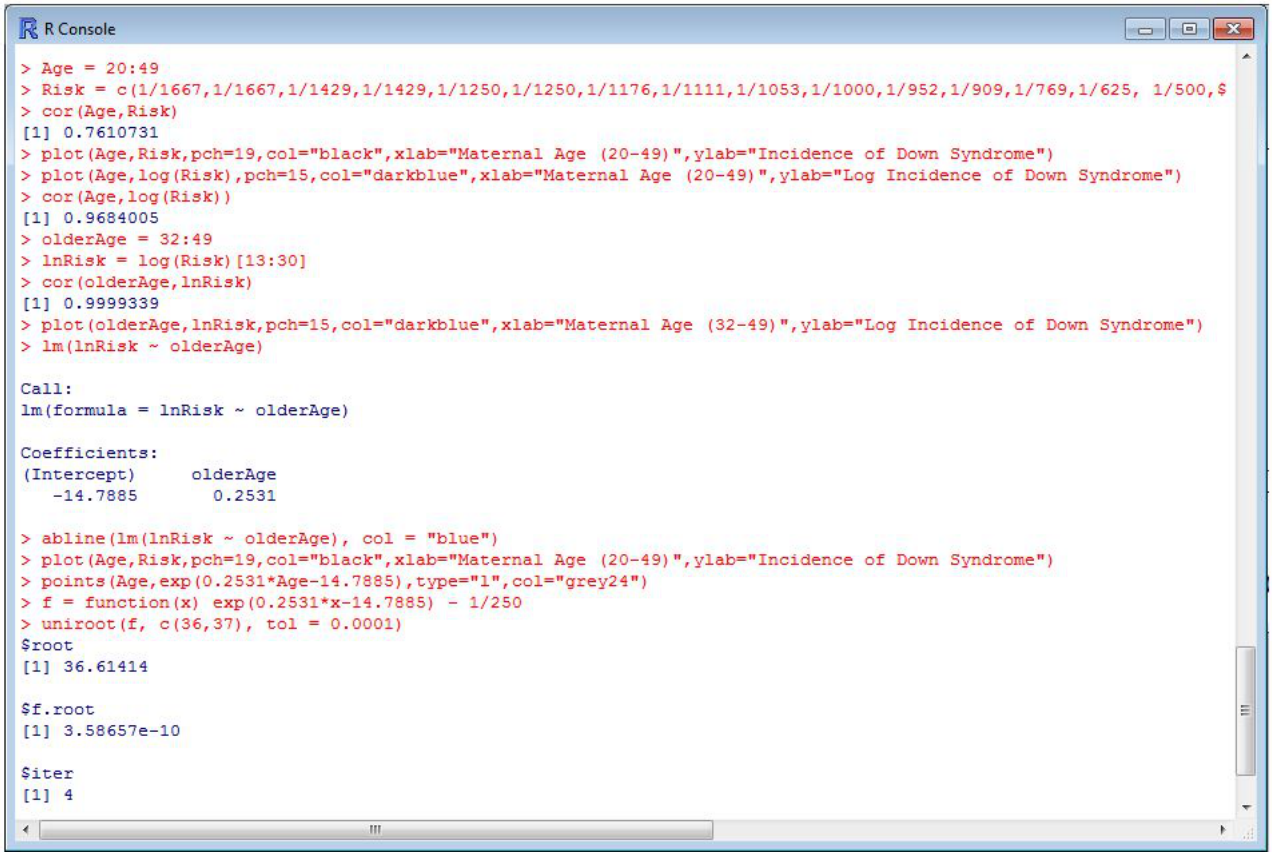
$$y = 0.37795\,10^{-6}\,e^{0.2531\,x}.$$

The requested plots appear in the next figure.



h) We solve $\exp(0.2531\,x - 14.7885) = 1/250$, or $x = (14.7885 + \ln(1/250))/0.2531$, or $x = 36.6141$.

It may be useful to show a screen capture of the R session used to answer this question.



Some points to notice. The colon operator in R has been used via the call `i:j` to create the vector of consecutive integers beginning with i and ending with j. The natural logarithm $\ln(x)$ is written as `log(x)` in R. Like other functions, it can be applied to an entire vector. Thus, `log(c(x1,x2))` is the vector `c(log(x1),log(x2))`. In the R session shown, `log(Risk)` is the vector with entries that are the logarithms of the entries of `Risk`. The ability in R to apply transformations to entire vectors is also used later in the session: the code

```
exp(0.2531*Age-14.7885)
```

results in the vector

```
c(exp(0.2531*x1-14.7885), exp(0.2531*x2-14.7885), ...)
```

where `Age` is the vector `c(x1, x2, ...)`. Another line of interest is

```
f = function(x) exp(0.2531*x-14.7885) - 1/250
```

which creates the function $f(x) = \exp(0.2531\,x - 14.7885) - 1/250$. The last line in the session,

```
uniroot(f, c(36,37), tol = 0.0001)
```

solves the equation $f(x) = 0$ for a value x in the interval $[36, 37]$. (We can see from the given risk data that the solution we seek is in this interval.)

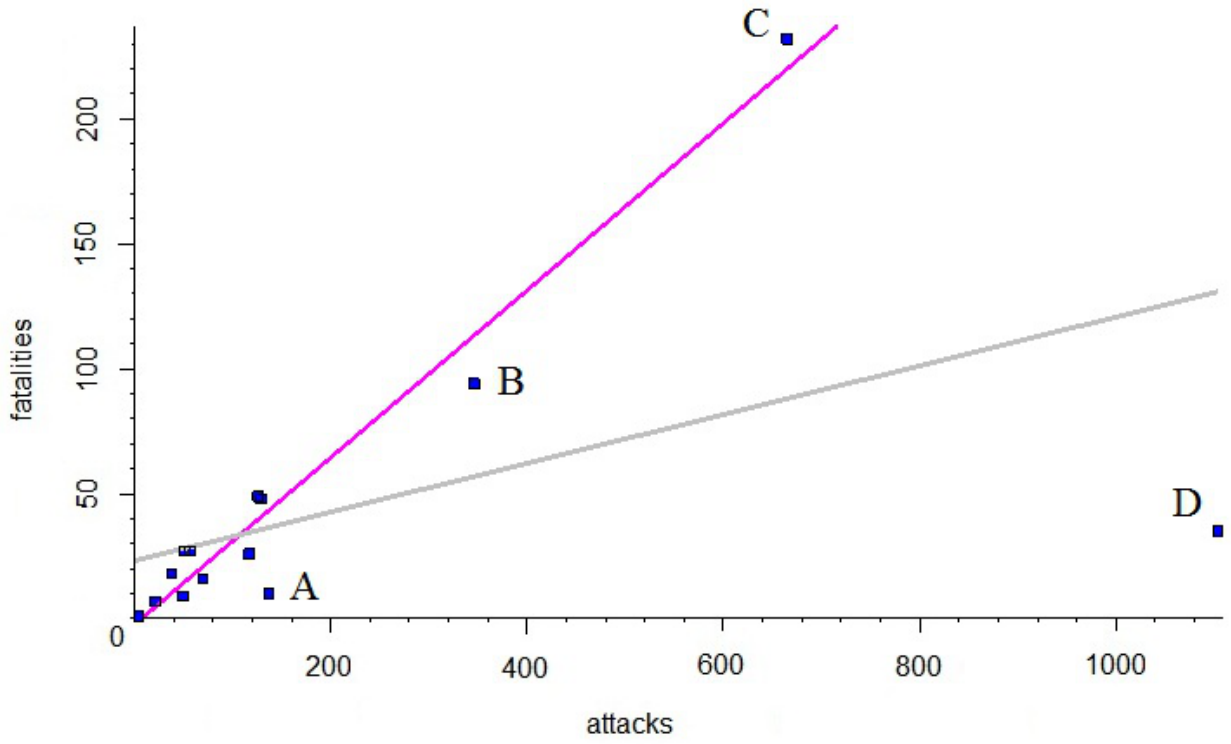
8. The scatter plot was created using the data from the following table:

Region	**Attacks**	**Fatalities**
United States	1104	35
Australia	665	232
Africa	346	94
Asia	129	48
Hawaii	137	10
Pacific Islands	126	49
South America	117	26
Antilles/Bahamas	70	16
Middle America	56	27
Europe	51	27
New Zealand	49	9
Réunion Island	38	18
Open Ocean	21	7
Bermuda	3	0
	2899	548

Shark Attacks! Worlwide 1958–2014
Source: Wikipedia `https://en.wikipedia.org/wiki/Shark_attack`
Retrieved 2016-02-21

There are some problems with the data entries in Wikipedia's table. Both column totals in the marginal row are wrong. Top to bottom, the regions were (apparently) listed by decreasing number of attacks. However, if so, Asia is out of place. But never mind.

When you first look at the figure below, disregard the magenta line not given in the statement of the exercise. Four points have been labeled: A, B, C, and D.



The four labeled points contribute to the requested identifications as follows:

Property	**Points**
Predictor Variable Outlier	B, C, D
Response Variable Outlier	B, C
Regression line outlier	A, B, C, D
High leverage point	B, C, D
Influential point	D

The x-coordinates of points B, C, and D are much greater than the x-coordinates of the other eleven points. At the other extreme of the predictor observations, the minimum x-coordinate is not particularly distant from the main cluster. Hence, the x-coordinates of B, C, and D are the predictor variable outliers.

The y-coordinates of points B and C are much greater than the y-coordinates of the other twelve points. These are the response variable outliers. Notice that the y-coordinate of point D is *not* a response variable outlier.

The vertical distances of points A, B, C, and D to the regression line (in grey) are all much greater than the other vertical distances. Hence, these points are the regression line outliers.

Points with x-coordinates that are distant from $\bar{x}$ are high leverage points. Predictor variable outliers are often the x-coordinates of high leverage points, but a point may be high leverage without having an x-coordinate that is a predictor variable outlier. In this case, the high leverage points are exactly the points with x-coordinates that are predictor variable outliers.

From the initial figure, it is clear that point D is strikingly fishy. In the jargon of statistics, it is influential. It clearly drags the regression substantially down from what it would have been without point D. Even had the magenta line not been added to the figure given above, you could imagine it as a regression line for the plotted points minus point D. The equation for the (grey) regression line based on all 14 points is $y = 0.09811\,x + 22.30839$. The (magenta) regression line based on the 13 points that remain when point D is removed is $y = 0.3344\,x - 2.4280$. The significant algebraic change to the equation of the regression line is as evident as the visual change. Points B and C are *not* influential. They lie quite close to the magenta regression line, which would have been the regression line without point D. It is clearly point D that is doing the major damage. Point A is an interesting point. Unlike points B and C, it remains a regression line outlier when point D is removed. In fact, it is even more of a regression line outlier after the removal of point D. However, if we remove point A after having removed point D, then the regression line changes from $y = 0.3344\,x - 2.4280$ to $y = 0.3342\,x - 0.4607$. The slope has barely changed and the y-intercept change is small with respect to the range of y-values. Point A is not influential.

Chapter 6.

1. The first group lacks Lance, the second lacks Tiger, the third lacks Tiger, the seventh lacks Tiger, the eighth lacks Tiger, and the tenth also lacks Tiger. There are four complete sets, so this small simulation leads to the approximation 4/10, or 0.4, for the probability of acquiring the set of three by means of 5 purchases.

2. In the thirty trials, the number of matches are

   1, 2, 3, 0, 1, 1, 2, 0, 1, 0, 0, 1, 1, 0, 0, 2, 0, 1, 0, 0, 0, 1, 0, 1, 2, 2, 1, 0, 0, 1

   for a total of 24 matches. The average number of matches per trial is 0.8

3. These are the simulated births:

1031	G	G	BG	G	G	BBB	G	BBB
1032	BBB	BBG	BG	G	G	BG	G	BBG
1033	G	G	BBB	G	G	BG	G	G
1034	G	BBG	G	G	G	G	G	BBB

   There are 52 births, of which 27 are girls and 25 are boys. The average number of children is 52/32, or 13/8, or 1.625. The average number of girls is 27/32, or 0.84375. The average number of boys is 25/32, or 0.78125.

4 The following trials result in complete sets: 95034, 28713, 96409 01927, 42648, 82425, 71709, 00095, 32863, 29485, 82226, 93074, 85848, 48767, 95593, 94007, 69971, 60779, 53791, 17287. That is a success ratio of 20/35, or 4/7, or 0.5714. (In a computer simulation of 1,000,000 trials, the success ratio was 0.507131.)

5. We will list by row and column the blocks composed of 5 different digits:
   (1,2), (1,4), (2,2), (2,4), (2,8), (3,6), (4,3), (5,6), (6,1), (7,1), (7,2), (7,8), (8,5), (8,7), (9,2), (10,4), (11,2), (11,3), (11,6), (11,7), (12,5), (12,6), (13,3), (14,5), (14,7), (14,8), (15,4), (15,5), (15,6). Because 29 of the 120 blocks of five contain five different digits, we estimate $p \approx 29/120 \approx 0.24167$.

6. The experiment is not completely randomized. If it were, either group could have as few as 0 dogs or as many as 20 dogs, depending on the chance assignment of every element. The experiment is blocked by species: 10 animals of each species will go to each group. Subject to that constraint, the assignments are randomized. Hence, the experiment is randomized, blocked by species. That answers the question, namely C. But let's make sure that D is not also an acceptable answer. The factor (and there is only one) is the flea collar. The flea collar is the treatment. Another treatment (for the control group) is the lack of collar. Teatments are not experimental units. Blocking is a division of the set of experimental units.

7. The effect of cold weather on battery performance is being tested. The car is just a device to implement the test. As long as it will start when the battery delivers the requisite current, it serves its purpose. By using two different brands, the experimenters eliminate the possibility that one particular brand is sensitive to temperature, as opposed to batteries in general. It may have been desirable to include additional brands, but it was not absolutely necessary. Similarly, additional temperature testing could have provided additional information, but the two temperatures chosen would have sufficed to establish an effect of temperature. The problem with the experiment was that the treatment (temperature) was confounded by what should have been a blocking variable (brand). One treatment (moderate temperature for the control group) was applied to one brand and the treatment to the other. If the second test terminated in less time than the first, the reduced cranking time might have been due to the colder weather or it might have been due to a less effective battery produced by the second manufacturer.

8. Maternal age. This variable is strongly correlated with the risk of Down syndrome. Moreover, it is associated with order of birth: A mother's age for a second birth is greater than her age for a first birth, a mother's age for a third birth is greater than her age for a second birth, and so on.

9. Gender can be associated with the choice of treatment. For example, doxazosin mesylate, a blood pressure medication, is prescribed more frequently to men because it can also provide some relief for BPH, benign prostatic hyperplasia.

Chapter 7

1. We have

$$\mathrm{P}(F) = \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) = 0.4 + 0.5 - 0.1 = 0.8,$$

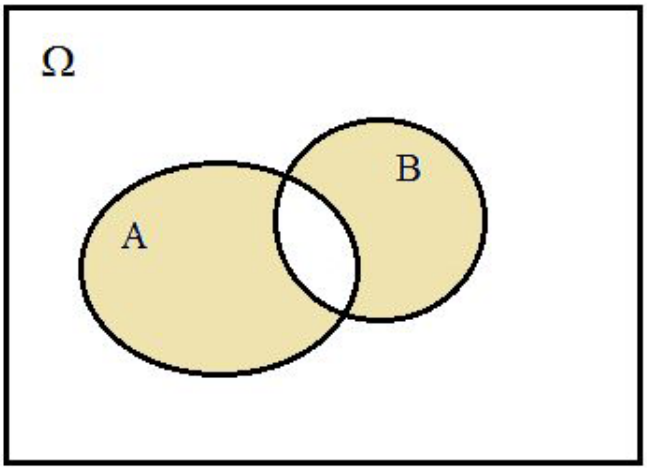
and, using this result,

$$\begin{aligned} \mathrm{P(E)} &= \mathrm{P(F)} - \mathrm{P(A \cap B)} \\ &= 0.8 - 0.1 \\ &= 0.7, \end{aligned}$$

and

$$\mathrm{P}(G) = \mathrm{P}\left((A \cup B)^{\mathsf{c}}\right) = 1 = \mathrm{P}(A \cup B) = 1 - 0.8 = 0.2.$$

2. Refer to the Venn diagram.



We see that

$$P(E) = P(A \cup B) - P(A \cap B) = \Big(P(A) + P(B)\Big) - P(A \cap B) - P(A \cap B) = P(A) + P(B) - 2\,P(A \cap B).$$

3. We have A = {(1,1), (1,3), (2,2), (3,1), (1,5), (2,4), (3,3), (4,2), (5,1), (2,6), (3,5), (4,4), (5,3), (6,2), (4,6), (5,5), (6,4), (6,6)}, B = { (1,1), (1,2), (1,3), (1,4), (1,5), (1,6), (2,1), (2,3), (2,5), (3,1), (3,2), (3,3), (3,4), (3,5), (3,6), (4,1), (4,3), (4,5), (5,1), (5,2), (5,3), (5,4), (5,5), (5,6), (6,1), (6,3), (6,5)}, and A ∩ B = {(1,1), (1,3), (3,1), (1,5), (3,3), (5,1), (3,5), (5,3), (5,5)}. It follows that P(A) = 18/36 = 1/2, P(B) = 27/36 = 3/4, and P(A ∩ B) = 9/36 = 1/4 ≠ P(A) · P(B). The events are not independent.

4. Let A be the event of a college student having an accident within one year. Let B be the event that a random selection of a college student results in the selection of a binge-drinker. Then B and B^c partition the sample space, so

$$P(A) = P(A|B)\,P(B) + P\,(A|B^c)\;P\,(B^c) = (0.184)(0.444) + (0.059)(0.556) = 0.1145.$$

5. Using Bayes's Rule, we have

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P\,(B^c)} = \frac{(0.184) \cdot (0.444)}{(0.184) \cdot (0.444) + (0.059) \cdot (0.556)} = 0.7135.$$

6. Let S be the event that a baby will live to age 60. Let E be the event that a baby will live to age 80.
   a) P(E) = 0.312 + 0.147 + 0.013 = 0.472.
   b) P(S) = 0.142 + 0.242 + P(E) = 0.142 + 0.242 + 0.472 = 0.856.
   c) Notice that E ⊂ S. Therefore, E ∩ S = E. It follows that

$$P(E|S) = \frac{P(E \cap S)}{P(S)} = \frac{P(E)}{P(S)} = \frac{0.472}{0.856} = .5514.$$

7. Let M_1 be the event that the first selected passenger is from the Middle East and let M_2 be the event that the second selected passenger is from the Middle East. The question asks for $P\,(M_1 \cap M_2)$. If desired, the probability of the intersection of two events can be expressed using conditional probability:

$$P\,(M_1 \cap M_2) = P\,(M_2|M_1)\;P\,(M_1) = \frac{5}{29}\,\frac{6}{30} = \frac{1}{29} = 0.03448.$$

8. Often, a problem that asks for the probability of E = "at least one ..." can most easily be done by finding the probability of F = "no ..." and using the set equation $E = F^c$. Then P(E) = 1 - P(F). The probability of the event F that no groupmate has had more than one semester of calculus (i.e., no groupmate has had two or more semesters of calculus) is $(0.15 + 0.72)^3$, or $(0.87)^3$, or 0.658503. Therefore, the probability that, of your other three groupmates, at least one has had more than one semester of Calculus is 1 - P(F), or 1 - 0.658503, or 0.341497. (Note that this probability is an approximation. The key to the accuracy of the approximation is that the class size is presumably large, given that the lecture hall is large. Suppose that there are 200 students in the class, of which 87% have had at most one semester of calculus. That means 174 students who have had at most one semester of calculus. When you are assigned your first groupmate, the pool is reduced to 173 out of 199. However, 173/189 is 0.8693467337, which is very close to .87. Similarly, 172/198 is 0.8686868687, which is also very close to 0.87.)

9. The requested probability is

$$\frac{305}{1000} \cdot \frac{695}{999} + \frac{695}{1000} \cdot \frac{305}{999}, \text{ or } 0.4244.$$

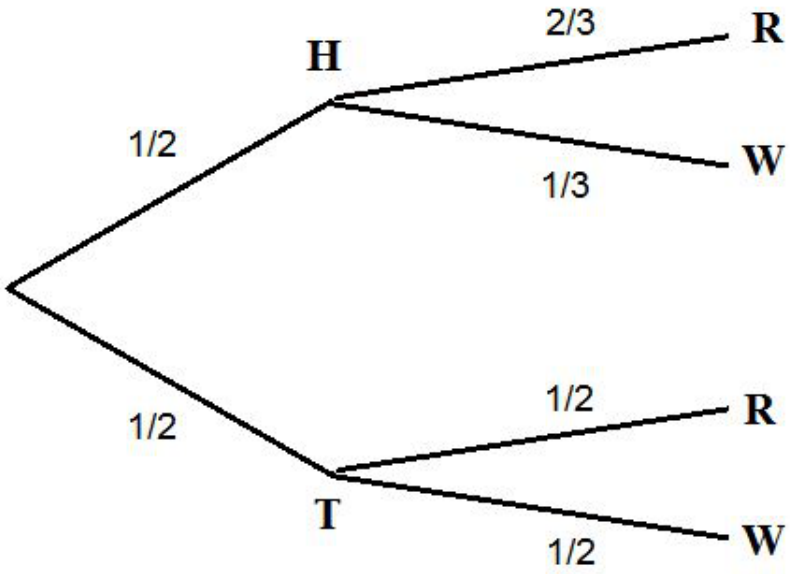
10. We have

$$\begin{aligned} \mathrm{P}(\mathrm{S} \,|\, \mathrm{POS}) &= \frac{\text{sensitivity} \cdot \text{prevalence}}{\text{sensitivity} \cdot \text{prevalence} + (1 - \text{specificity})(1 - \text{prevalence})} \\ &= \frac{0.946 \cdot 0.01}{0.946 \cdot 0.01 + (1 - 0.941)(1 - 0.01)} \\ &= 0.1394. \end{aligned}$$

11. We are asked for P(S | NEG), where NEG refers to the result of the Pap test. Using Bayes's Rule, we have

$$\begin{aligned} \mathrm{P}(\mathrm{S}|\mathrm{NEG}) &= \frac{\mathrm{P}(\mathrm{NEG} \,|\, \mathrm{S}) \cdot \mathrm{P}(\mathrm{S})}{\mathrm{P}(\mathrm{NEG} \,|\, \mathrm{S}) \cdot \mathrm{P}(\mathrm{S}) + \mathrm{P}\left(\mathrm{NEG} \,|\, \mathrm{S}^c\right) \cdot \mathrm{P}\left(\mathrm{S}^c\right)} \\ &= \frac{(1 - 0.544) \cdot (0.01)}{(1 - 0.544) \cdot (0.01) + (0.968) \cdot (0.99)} \\ &= 0.004736. \end{aligned}$$

12. Refer to the tree diagram.



We have

$$P(R) = \mathrm{P(R\,|\,H)} \cdot \mathrm{P(H)} + \mathrm{P(R\,|\,T)} \cdot \mathrm{P(T)} = \frac{2}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{7}{12}.$$

13. We are asked for $\mathrm{P(H\,|\,R)}$. We can calculate this using Bayes's Rule:

$$\mathrm{P(H|R)} = \frac{\mathrm{P(R\,|\,H)P(H)}}{\mathrm{P(R\,|\,H)P(H) + P(R\,|\,T)P(T)}} = \frac{\frac{2}{3} \cdot \frac{1}{2}}{\frac{2}{3} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}} = \frac{4}{7}.$$

14. This experiment is a sequence of Bernoulli trials. Technically, the trials are not identically distributed. Once a potential interviewee has been contacted and found to be an internet user, the name is taken out of the pool. However, the pool of teens is so teeming that removal of a few names does not change the probability of success being 0.08 and failure 0.92 (at least not to any decimal place we are keeping tack of). So, if X is the trial on which the first success occurred, then X is a geometric random variable. The question asks for $f_X(6)$ and that probability is $(0.92)^5(0.08)$, or 0.52727.

15. Although the wording of this problem was not explicit, the intent is that the sampling is done without replacement. (It would be pointless to put the dead batteries back in the box. Doing so could clearly result in trying the same dead battery multiple times.) Thus, this experiment is a sequence of Bernoulli trials that are *not* identically distributed. The selection of a working battery is a success. The probability of a success on the k^{th} trial given the the first $k-1$ trials were failures is $8/(12-(k-1))$. Therefore the probability of failure on the first four trials and success on the fifth is

$$\left(1 - \frac{8}{12}\right) \cdot \left(1 - \frac{8}{11}\right) \cdot \left(1 - \frac{8}{10}\right) \cdot \left(1 - \frac{8}{9}\right) \cdot 1 = 0.002020.$$

16. Of the 48 animals, $(20-7)+(28-20)$, or $13+8$, or 21 are female. Thus $\mathrm{P(F)} = 21/48 = 7/16$. Of the 48 animals, 8 are female cats. Therefore $\mathrm{P(C \cap F)} = 8/48 = 1/6$, and

$$\mathrm{P(C\,|\,F)} = \frac{\mathrm{P(C \cap F)}}{\mathrm{P(F)}} = \frac{\frac{1}{6}}{\frac{7}{16}} = \frac{16}{6 \cdot 7} = \frac{8}{21} = 0.38095.$$

17. One aspect of this problem should be compared with the corresponding aspect of the preceding problem. In the preceding problem, two members of the subpopulation who participated in the survey were selected. The survey size, 1000, was relatively large, but not so large that we were careless about the possible effect on frequencies that removing one member from the pool might cause. In this problem, we are selecting 15 members from a large population. By the time we get to the 15th person, we have removed 14 persons from the pool, but, because the population is large, the frequencies, 0.305 and 0.695, will apply equally well (any changes occuring beyond the third digit following the decimal). If we call the selection of a likely reader a success and let X be the number of successes in 15 trials, then X is a binomial random variable and

$$\begin{aligned}
\mathrm{P}(\mathrm{X} \geq 5) &= \sum_{k=5}^{15} \binom{15}{k} (0.305)^k (0.695)^{15-k} \\
&= 1 - \sum_{k=0}^{4} \binom{15}{k} (0.305)^k (0.695)^{15-k} \\
&= 1 - \binom{15}{0}(0.305)^0(0.695)^{15} - \binom{15}{1}(0.305)^1(0.695)^{14} \\
&\qquad - \binom{15}{2}(0.305)^2(0.695)^{13} - \binom{15}{3}(0.305)^3(0.695)^{12} \\
&\qquad - \binom{15}{4}(0.305)^4(0.695)^{11} \\
&= 1 - 0.004263557 - 0.02806586 - 0.08621671 - 0.1639565 - 0.2158564 \\
&= 0.502.
\end{aligned}$$

18. Let X be the number of frogs in the sample of 120 that have the trait. Then $f_X(k) = \binom{120}{k}(1/8)^k(7/8)^{120-k}$ for $0 \leq k \leq 120$. The requested probability is

$$\sum_{k=15}^{120} \binom{120}{k} (1/8)^k (7/8)^{120-k}, \text{ or } 0.541256.$$

Note: This *type* of question is fine. However, the author would not include a problem of this type on an exam with a parameter anywhere near as large as 120. Exercise 10 covers similar ground, but the parameters involved are more manageable.

19. (i) The probability that the first Vitamin D-deficient child is the 6th one tested is $q^5 \cdot p$, where $p = 0.2$ and $q = 0.8$. This comes to 0.065536.
(ii) The event that no more than two of the first 12 children tested have Vitamin D deficiency is the disjoint union of three events: 0 have Vitamin D deficiency, exactly one has Vitamin D deficiency, and exactly two have Vitamin D dficiency. The probability of these events are obtained from the distribution of a binomial random variable X with $N = 12$ and $p = 0.2$: $f_X(0) = \binom{12}{0}p^0q^{12}$, $f_X(1) = \binom{12}{1}p^1q^{11}$, $f_X(2) = \binom{12}{2}p^2q^{10}$. These probabilities are 0.06871947674, 0.2061584302, 0.2834678415, and their sum is 0.5583457485. That is the answer to (ii). The answer to the multiple choice question is 0.065536 + 0.5583457485, or .6238817485. (Answer (E))

20. There are four outcomes. Writing Alex's roll first, they are $\Omega = \{(2,1),(2,3),(8,1),(8,3)\}$. Alex wins with three of the four outcomes, but loses if the outcome is (2,3). Because the die rolls of Alex and Bella are independent, the probability of the outcome (2,3) is (5/6) × (4/6), or 5/9. Suppose that if Alex loses, then Alex pays Bella m. Let X be the amount that Alex wins (or loses if $\mathrm{X} < 0$). Then $\mathrm{E}(\mathrm{X}) = 10 \cdot (1-5/9) + (-m) \cdot 5/9$. For a fair game, $\mathrm{E}(\mathrm{X}) = 0$. We solve $\mathrm{E}(\mathrm{X}) = 10 \cdot (1-5/9) + (-m) \cdot 5/9 = 0$ and find $m = 8$.

21. As in the preceding exercise, there are four outcomes. Writing Alex's roll first, they are $\Omega = \{(2,1),(2,3),(8,1),(8,3)\}$. Let X be Alex's winnings (losings if $X < 0$). Then $X = 2$ if the outcome is (2,1), -3 if the outcome is (2,3), and 8 if the outcome is (8,1) or (8,3). Using independence, we calculate $P(X = 2) = (5/6)(2/6) = 10/36$, $P(X = -3) = (5/6)(4/6) = 20/36$, and $P(X = 8) = (1/6)(2/6) + (1/6)(4/6) = 6/36$. Therefore $E(X) = 2 \cdot (10/36) + (-3) \cdot (20/36) + 8 \cdot (6/36)$, or 8/36, or $2/9 (\approx 0.22)$.

22. There are four outcomes. They have the following probabilities: $P(G) = 1/2$, $P(BG) = 1/4$, $P(BBG) = 1/8$, and $P(BBB) = 1/7$. Let X be the number of children. Then

$$E(X) = 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + 3 \cdot P(X = 3) \ = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \left(\frac{1}{8} + \frac{1}{8}\right) = 1.75.$$

Let Y be the number of girls and let Z be the number of boys. The possible values of Y are 0 and 1. The possible values of Z are 0, 1, 2, and 3. We have

$$E(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{7}{8} = \frac{7}{8} = 0.875$$

and

$$E(Z) = 0 \cdot P(Z = 0) + 1 \cdot P(Z = 1) + 2 \cdot P(Z = 2) + 3 \cdot P(Z = 3) = \ = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{8} + 3 \cdot \frac{1}{8} == \frac{7}{8} = 0.875.$$

These expectations may be compared with the averages found in the small simulation of Exercise 3 in Chapter 6.

23. Let π denote the event that a teen passes the written driving test. Let S be the event that a teen studies for the written driving test. Then S and S^c partition the sample space. We are given that $P(S) = 0.7$, from wjich it follows that $P(S^c) = 0.3$. We are also given that $P(\pi \mid S) = 0.95$ and $P(\pi \mid S^c) = 0.6$. By Bayes's Law, we have

$$\begin{aligned} P(S^c \mid \pi) &= \frac{P(\pi \mid S^c)\ P(S^c)}{P(\pi \mid S^c)\ P(S^c) + P(\pi \mid S)\ P(S)} \\ &= \frac{(0.6) \cdot (0.3)}{(0.6) \cdot (0.3) + (0.95) \cdot (0.7)} \\ &= 0.213. \end{aligned}$$

24. The probability that the first digit is not a repeat of a previously selected digit in the block is 1: there are no previously selected digits. The probability that the second digit is not a repeat of a previously selected digit in the block is 9/10. Given that the first two digits are distinct, the probability that the third digit is not a repeat of a previously selected digit in the block is 8/10. Given that the first three digits are distinct, the probability that the fourth digit is not a repeat of a previously selected digit in the block is 7/10. Given that the first four digits are distinct, the probability that the fifth digit is not a repeat of a previously selected digit in the block is 6/10. The probability we seek is $1 \times (9/10) \times (8/10) \times (7/10) \times (6/10)$, or 0.3024.

25. Let G be the event of graduation. Let PUB be the event of being an entering freshman from a public high school. The given informations is $P(PUB) = 0.65$, $P(G \mid PUB) = 0.80$, and $P(G \mid PUB^c) = 0.90$. From this information we deduce that $P(PUB^c) = 1 - 0.65 = 0.35$, $P(G^c \mid PUB) = 1 - 0.80 = 0.20$, and $P(G^c \mid PUB^c) = 1 - 0.90 = 0.10$. By the Law of Total Probability, we have

$$P(G^c) = P(G^c \mid PUB)\, P(PUB) + P(G^c \mid PUB^c)\, P(PUB^c) = (0.20)(0.65) + (0.10)(0.35) = 0.165.$$

26. Let X be the amount the horse will sell for after the two races. The value we seek is E(X) - 25000. There are three possible values for X: 150,000, 60,000, and 5,000. There are four outcomes in the sample space $W_1\,W_2$, $W_1\,L_2$, $L_1\,W_2$, $L_1\,L_2$. Using the independence property, we have

$$\begin{aligned}
\mathrm{P}(W_1\,W_2) &= \frac{15}{100} \times \frac{25}{100} = \frac{375}{10,000},\\
\mathrm{P}(W_1\,L_2) &= \frac{15}{100} \times \frac{75}{100} = \frac{1125}{10,000},\\
\mathrm{P}(L_1\,W_2) &= \frac{85}{100} \times \frac{25}{100} = \frac{2125}{10,000},\\
\mathrm{P}(L_1\,L_2) &= \frac{85}{100} \times \frac{75}{100} = \frac{6375}{10,000}.
\end{aligned}$$

Thus,

$$f_{\mathrm{X}}(150,000) = \frac{375}{10,000}, \quad f_{\mathrm{X}}(60,000) = \frac{1125}{10,000} + \frac{2125}{10,000} = \frac{3250}{10,000}, \quad f_{\mathrm{X}}(5,000) = \frac{6375}{10,000},$$

and

$$E(X) - 25,000 = 150,000 \times \frac{375}{10,000} + 60,000 \times \frac{3250}{10,000} + 5,000 \times \frac{6375}{10,000} - 25,000 = 3312.50.$$

27. The fraction of brown candies is $1 - (0.2 + 0.2 + 3 \times 0.1)$, or 0.3. The probability of picking 4 brown candies in a row is $(0.3)^4$, or 0.0081.

28. Let R be the event that when one candy is selected it is red. Let J be the event that when one candy is selected it is orange. These are clearly mutually exclusive events. Therefore $\mathrm{P}(\mathrm{R} \cup \mathrm{J}) = \mathrm{P}(\mathrm{R}) + \mathrm{P}(\mathrm{J})$, or $\mathrm{P}(\mathrm{R} \cup \mathrm{J}) = 0.2 + 0.1 = 0.3$. When one candy is drawn, the probability $\mathrm{P}\left((\mathrm{R} \cup \mathrm{J})^{\mathsf{c}}\right)$ that it is not a red or an orange is 1 - 0.3, or 0.7. When four candies are drawn, the probability that they do not include a red or an orange is $(0.7)^4$, or 0.2401. The probability that there is at least one red or orange in four drawings is 1 - 0.2401, or 0.7599.

29. Let W_j be the event that a white is drawn on the j^{th} selection. Then $p = \mathrm{P}(\mathrm{W}_j) = 0.25$ and $q = \mathrm{P}\left(\mathrm{W}_j^{\mathsf{c}}\right) = 0.75$. The problem asks for $\mathrm{P}(\mathrm{W}_1^{\mathsf{c}} \cap \mathrm{W}_2^{\mathsf{c}} \cap \mathrm{W}_3)$, which, by the independence of the selections, is $\mathrm{P}(\mathrm{W}_1^{\mathsf{c}}) \times \mathrm{P}(\mathrm{W}_2^{\mathsf{c}}) \times \mathrm{P}(\mathrm{W}_3)$, or $0.75 \times 0.75 \times 0.25$, or 0.140625. (You may have recognized that this problem asks for $f_{\mathrm{X}}(3)$ where X is a geometric random variable.)

30. The deck of 52 cards is divided into 4 suits of 13 cards each. Let S be the thirteen spades, D the thirteen diamonds, and D^{c} be the 39 Spades, Hearts, and Clubs. Notice that S is contained in D^{c}. As a result, we have $\mathrm{S} \cap \mathrm{D}^{\mathsf{c}} = \mathrm{S}$. Thus,

$$\mathrm{P}(\mathrm{S} \,|\, \mathrm{D}^{\mathsf{c}}) = \frac{\mathrm{P}(\mathrm{S} \cap \mathrm{D}^{\mathsf{c}})}{\mathrm{P}(\mathrm{D}^{\mathsf{c}})} = \frac{\mathrm{P}(\mathrm{S})}{\mathrm{P}(\mathrm{D}^{\mathsf{c}})} = \frac{13/52}{39/52} = \frac{1}{3}.$$

31.

31. The expected number is $2 \times (0.08) + 3 \times (0.30) + 4 \times (0.37) + 5 \times (0.18) + 6 \times (0.07)$, or 3.86.

32. Let X_j be the number of gray73 buttons drawn from the j^{th} urn. Then the expected total number of buttons selected is $\mathrm{E}(\mathrm{X}) = \mathrm{E}(\mathrm{X}_1) + \mathrm{E}(\mathrm{X}_2) + \cdots + \mathrm{E}(\mathrm{X}_{14}) = 14\,\mathrm{E}(\mathrm{X})$. We have dropped the subscript of X in the expression at the right because the urns are identical and it does not matter which one we consider. The table below shows the 8 outcomes for each urn with the likelihood of each outcome. These are taken from a tree diagram, but it is easier to include a table than a tree diagram. In the table, G stands for gray73 and R stands for the rainbow coalition of the other 4 colors.

Outcome	Probability
GGG	(3/11)(2/10)(1/9)
GGR	(3/11)(2/10)(8/9)
GRG	(3/11)(8/10)(2/9)
GRR	(3/11)(8/10)(7/9)
RGG	(8/11)(3/10)(2/9)
RGR	(8/11)(3/10)(7/9)
RRG	(8/11)(7/10)(3/9)
RRR	(8/11)(7/10)(6/9)

These 8 probabilities allow us to calculate the p.f. f_X. There are four values X can assume: $V = \{0, 1, 2, 3\}$. There is no need for us to calculate $f_X(0)$ because this probability is multiplied by 0 in the formula for expectation. There are three outcomes that result in 1 G: GRR, RGR, RRG. The sum of the likelihoods in those three rows give $f_X(1)$:

$$f_X(1) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right)\left(\frac{7}{9}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right)\left(\frac{7}{9}\right) + \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)\left(\frac{3}{9}\right) = \frac{28}{55} = 0.5090909.$$

There are three outcomes that result in 2 G: GGR, GRG, RGG. The sum of the likelihoods in those three rows give $f_X(1)$:

$$f_X(2) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right)\left(\frac{8}{9}\right) + \left(\frac{3}{11}\right)\left(\frac{8}{10}\right)\left(\frac{2}{9}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right)\left(\frac{2}{9}\right) = \frac{8}{55} = 0.1454545.$$

Finally, there is only one outcome with 3 G:

$$f_X(3) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right)\left(\frac{1}{9}\right) = \frac{1}{165} = 0.006060606.$$

Thus,

$$\mathrm{E(X)} = 0 \times f_X(0) + 1 \times f_X(1) + 2 \times f_X(2) + 3 \times f_X(3) = 0 + \frac{28}{55} + 2 \times \frac{8}{55} + 3 \times \frac{1}{165} = \frac{9}{11}.$$

Therefore, the expected number of gray73 buttons selected is $14 \times (9/11)$, or 126/11, or 11.45454545.

33. (a) Let X_1 and X_2 be a random sample of yearly rainfall totals in Los Angeles. (Whether or not the two years selected are consecutive is not germane—totals for any two years are assumed independent.) Then $\mathrm{X}_1 + \mathrm{X}_2$ is normal with mean 24.12 and standard deviation $\sqrt{(3.1)^2 + (3.1)^2}$, or 4.3841. We calculate

$$\begin{aligned}
\mathrm{P}\left(\mathrm{X}_1 + \mathrm{X}_2 > 26\right) &= \mathrm{P}\left(\mathrm{X}_1 + \mathrm{X}_2 - 24.12 > 26 - 24.12\right) \\
&= \mathrm{P}\left(\frac{\mathrm{X}_1 + \mathrm{X}_2 - 24.12}{4.3841} > \frac{26 - 24.12}{4.3841}\right) \\
&= \mathrm{P}\left(\mathrm{Z} > 0.42882\right) \\
&= 1 - \mathrm{P}\left(\mathrm{Z} \leq 0.42882\right) \\
&= 1 - \Phi(0.42882) \\
&= 1 - 0.66598 \\
&= 0.334.
\end{aligned}$$

(b) The random variable $\mathrm{X}_1 - \mathrm{X}_2$ is normal with mean 0 and standard deviation $\sqrt{(3.1)^2 + (3.1)^2}$, or

4.3841. We have

$$\begin{aligned} \mathrm{P}\left(\mathrm{X}_1 - \mathrm{X}_2 > 4\right) &= \mathrm{P}\left(\frac{\mathrm{X}_1 - \mathrm{X}_2}{4.3841} > \frac{4}{4.3841}\right) \\ &= \mathrm{P}\left(\mathrm{Z} > 0.91239\right) \\ &= 1 - \mathrm{P}\left(\mathrm{Z} \leq 0.91239\right) \\ &= 1 - \Phi(0.91239) \\ &= 1 - 0.81922 \\ &= 0.181. \end{aligned}$$

34. Let X_1 and X_2 be a random sample of two adult women in the United States. Then $\bar{\mathrm{X}} = \left(\mathrm{X}_1 + \mathrm{X}_2\right)/2$ is a normal random variable with mean 64.5 inches and variance given by

$$\begin{aligned} \mathrm{Var}\left(\bar{\mathrm{X}}\right) &= \mathrm{Var}\left(\frac{1}{2}\,\mathrm{X}_1 + \frac{1}{2}\,\mathrm{X}_2\right) \\ &= \left(\frac{1}{2}\right)^2 \mathrm{Var}\left(\mathrm{X}_1\right) + \left(\frac{1}{2}\right)^2 \mathrm{Var}\left(\mathrm{X}_2\right) \qquad \text{by (7.4.10)} \\ &= \frac{1}{4}\left((2.4)^2 + (2.4)^2\right) \\ &= 2.88. \end{aligned}$$

The standard deviation is equal to $\sqrt{2.88}$, or 1.697. Letting Z denote a standard normal random variable, we have

$$\begin{aligned} \mathrm{P}\left(\bar{\mathrm{X}} > 67\right) &= \mathrm{P}\left(\bar{\mathrm{X}} - 64.5 > 67 - 64.5\right) \\ &= \mathrm{P}\left(\frac{\bar{\mathrm{X}} - 64.5}{1.697} > \frac{67 - 64.5}{1.697}\right) \\ &= \mathrm{P}\left(\mathrm{Z} > 1.4732\right) \\ &= 1 - \mathrm{P}\left(\mathrm{Z} \leq 1.4732\right) \\ &= 1 - \Phi(1.4732) \\ &= 1 - 0.92965 \\ &= 0.07. \end{aligned}$$

35. Let U be the amount of cereal in the new box, X the amount of cereal poured into the small bowl, and Y the amount of cereal poured into the large bowl. Then W = U - (X + Y), the weight of the remaining cereal, is a normal r.v. with mean 16.3 - (2.0 + 3.0), or 11.3, and standard deviation $\sqrt{0.15^2 + 0.25^2 + 0.3^2}$, or 0.41833. It follows that Z = (U - 11.3)/0.41833 is a standard normal random variable. Therefore,

$$\begin{aligned} \mathrm{P}(\mathrm{W} > 12) &= \mathrm{P}\left(\frac{\mathrm{W} - 11.3}{0.41833} > \frac{12 - 11.3}{0.41833}\right) \\ &= \mathrm{P}(\mathrm{Z} > 1.6733) \\ &= 1 - \mathrm{P}(\mathrm{Z} \leq 1.6733) \\ &= 1 - \Phi(1.6733) \\ &= 1 - 0.95287 \\ &= 0.04713. \end{aligned}$$

36. Let U and X be, respectively, the weight and cost of a watermelon purchased at the first store. Let W and Y be, respectively, the weight and cost of a watermelon purchased at the second store. Then X = 30 U and Y = 26 W and X - Y is a normal random variable with mean $30 \cdot 21 - 26 \cdot 19$, or 136, and standard deviation $\sqrt{30^2 \cdot 3^2 + 26^2 \cdot 2^2}$, or 103.94. It follows that Z = ((X - Y) - 136)/103.94 is a standard normal random variable. Thus,

$$\begin{aligned} \mathrm{P}(\mathrm{X} < \mathrm{Y}) &= \mathrm{P}(\mathrm{X} - \mathrm{Y} < 0) \\ &= \mathrm{P}\big((\mathrm{X} - \mathrm{Y}) - 136 < -136\big) \\ &= \mathrm{P}\left(\frac{(\mathrm{X} - \mathrm{Y}) - 136}{103.94} < -\frac{136}{103.94}\right) \\ &= \mathrm{P}(\mathrm{Z} < -1.30845) \\ &= 0.09536. \end{aligned}$$

37. Let Z be standard normal and let Y be the mean of the sample of 45 pregnant women. Then Y is normal with mean 268 and standard deviation $17/\sqrt{45}$, or 2.5342. We have

$$\begin{aligned} \mathrm{P}(Y < 265) &= \mathrm{P}\left(\frac{\mathrm{Y} - 268}{2.5342} < \frac{265 - 268}{2.5342}\right) \\ &= \mathrm{P}(\mathrm{Z} < -1.18381) \\ &= \mathrm{P}(\mathrm{Z} > 1.18381) \\ &= 1 - \mathrm{P}(\mathrm{Z} \leq 1.18381) \\ &= 1 - \Phi(1.18381) \\ &= 1 - 0.881391 \\ &= 0.1186. \end{aligned}$$

38. Let X be Alice's half-marathon time, and let Y be Sharon's. The information about E(X) and E(Y) is not needed. Per haps it has been given as a red herring. Perhaps E(X + Y) = E(X) + E(Y) is supposed to prompt you into thinking that $\sigma_{X+Y} = \sigma_X + \sigma_Y$, which is false. Perhaps E(X + Y) = E(X) + E(Y) is meant to remind you that Var(X +Y) = Var(X) + Var(Y), which is true because X and Y are independent. In fact, this equation is the key. We have

$$\mathrm{Var}(\mathrm{X} + \mathrm{Y}) = \mathrm{Var}(\mathrm{X}) + \mathrm{Var}(\mathrm{Y}) = \sigma_X^2 + \sigma_Y^2 = 4^2 + 2^2 = 20.$$

Therefore, $\sigma_{X+Y} = \sqrt{\mathrm{Var}(\mathrm{X} + \mathrm{Y})} = \sqrt{20} = 4.47$.

39. Using $\lambda = 82/52 = 1.576923$, we complete the table as follows:

Earthquakes (k)	Number n of Weeks with k Earthquakes	Frequency $n/52$ (Empirical Probability)	$\exp(-\lambda)\,\lambda^k/k!$
0	13	0.25	0.21
1	17	0.32	0.33
2	13	0.25	0.26
3	5	0.10	0.14
4+	4	0.08	0.06

40. (a) We have

$$\begin{aligned}
\mathrm{P}\left(X > t_0 + \tau | X > t_0\right) &= \frac{\mathrm{P}\left(X > t_0 + \tau \text{ and } X > t_0\right)}{\mathrm{P}\left(X > t_0\right)} \qquad \text{(Definition of conditional probability)} \\
&= \frac{\mathrm{P}\left(X > t_0 + \tau\right)}{\mathrm{P}\left(X > t_0\right)} \qquad \text{(Because } X > t_0 + \tau \text{ implies } X > t_0\text{)} \\
&= \frac{\exp\left(-\lambda(t_0 + \tau)\right)}{\exp\left(-\lambda(t_0)\right)} \qquad \text{(using the given formula for right tails)} \\
&= \frac{\exp\left(-\lambda \cdot t_0\right)\exp\left(-\lambda \cdot \tau\right)}{\exp\left(-\lambda(t_0)\right)} \qquad \text{(exponential of a sum = product of exponentials)} \\
&= \exp\left(-\lambda \cdot \tau\right).
\end{aligned}$$

(b) Part (a) tells us that $\mathrm{P}\left(X > t_0 + \tau | X > t_0\right) = \mathrm{P}\left(X > \tau\right)$. The meaning of this is that the probability of not failing in the next τ minutes is the same no matter how long the unit has been in service.

41. In a sequence of i.i.d. Bernoulli trials, the number X of the trial on which the first success occured is a geometric random variable. If p is the probability of success on a trial, then $\mathrm{E(X)} = 1/p$. Here $p = 0.03$ and $\mathrm{E(X)} = 1/0.03 - 33.3$. Round up to the nearest integer: 34.

42. The number of elephants struck by lightning in one year is Poisson with parameter 1.8. The number of elephants struck by lightning in a five year period is Poisson with parameter 5×1.8, or 9. The probability of at least 10 lightning strikes in that period is

$$1 - \sum_{k=0}^{9} \exp(-9)\frac{9^k}{k!},$$

or 0.41259.

43. Let $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{12}$ be the weights of the contents of the 12 boxes in the case. Let $\mathrm{Y} = \mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{12}$. Then Y is normal with mean 12×13, or 156, and variance $12 \times (0.5)^2$, or 3. The standard deviation of Y is $\sqrt{3}$. We calculate

$$\begin{aligned}
\mathrm{P}\left(\mathrm{Y} > 160\right) &= \mathrm{P}\left(\mathrm{Y} - 156 > 160 - 156\right) \\
&= \mathrm{P}\left(\frac{\mathrm{Y} - 156}{\sqrt{3}} > \frac{160 - 156}{\sqrt{3}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} > 2.3094\right) \\
&= 1 - \mathrm{P}\left(\mathrm{Z} \leq 2.3094\right) \\
&= 1 - \Phi\left(2.3094\right) \\
&= 0.01046.
\end{aligned}$$

44. The inspection of each apple is a Bernoulli trial with probability $p = 0.07$ of success and $q = 0.93$ of failure. Notice that we are using a "success" to describe an apple that fails the test and contributes toward the rejection of the shipment. We made this choice because it is the threshold for rejection that has been stated. Thus, a target has been given and it is natural to think of succeeding if we reach the target. If this chosen usage of "success" seems perverse, then p and q can be swapped with the proviso that the only thing that matters is matching "success" to the acceptance/rejection decision that is appropriate for the chosen meaning.

It is implicitly assumed in this problem that the 140 Bernoulli trials that we have just described are independent. That assumption should have been stated because it is not at all evident that the tests

are actually independent. The apples might have mostly come from one corner of an orchard ravaged by disease. Or, rough handling that results in one apple being badly bruised may well cause neighboring apples to be bruised. There is even an old saying in English that disclaims independence. As the proverb warns, *One rotten apple spoils the barrel.*

It is further implicitly assumed that the quantity of apples in the truck is so large that the one-by-one removal of 140 apples from the population does not alter (or, more precisely, alters negligibly) the subsequent probabilities of selecting good or bad apples. In other words, p and $q = 1-p$ remain constant throughout the trials.

Because 5% (1/20) of 140 is 7, the shipment is rejected if 8 or more apples among the 140 inspected are unacceptable. The shipment is accepted if 0, 1, 2, 3, 4, 5, 6, or 7 of the 140 apples are satisfactory. It will be noted that, just as in screening for a state (disease, presence of a substance, ...), two types of error are possible. Chance may result in the acceptance of an unsatisfactory shipment because two few bad apples are sampled (a false negative), or a satisfactory shipment might be rejected because two many bad apples are in the sample (a false positive). The problem, in effect, asks for the probability of a false negative.

After all this lead-up the solution is brief. Let X be the number of bad apples found in the sample of 140. Then $\mathrm{X} \sim \mathrm{N}(140, 0.07)$, whence

$$f_{\mathrm{X}}(k) = \binom{140}{k} p^k \, q^{140-k} \qquad (0 \le k \le 140)$$

and

$$\mathrm{P}(\text{Accept shipment}) = \sum_{k=0}^{7} f_{\mathrm{X}}(k) = \sum_{k=0}^{7} \binom{140}{k} (0.07)^k \, (0.93)^{140-k} = 0.2290887114.$$

The calculation of the sum, involving 8 summands as it does, is somewhat tedious using a basic scientific calculator. In R, the code for the answer is `pbinom(7,140,0.07)`.

Chapter 8

1. a) We have

$$\mu_{\mathrm{X}} = \frac{12880580 + 11594163 + 9909877 + 6596855 + 6063589 + 5757564 + 5457173 + 3107126}{8} = 7670866.$$

As the first step toward determining the variance, we first calculate

$$\begin{aligned} E(\mathrm{X}^2) &= \frac{12880580^2 + 11594163^2 + 9909877^2 + 6596855^2 + 6063589^2 + 5757564^2 + 5457173^2 + 3107126^2}{8} \\ &= 6.892622 \times 10^{13}. \end{aligned}$$

Then, making use of a helpful formula for the variance, we have

$$\begin{aligned} \mathrm{Var}(\mathrm{X}) &= E(\mathrm{X}^2) - (E(\mathrm{X}))^2 \\ &= 6.892622 \times 10^{13} - (7670866)^2 \\ &= 6.892622 \times 10^{13} - 5.884219 \times 10^{13} \\ &= 1.008403 \times 10^{13}, \qquad \text{and} \end{aligned}$$

$$\sigma_{\mathrm{X}} = \sqrt{1.008403 \times 10^{13}} = 3\,175\,536.$$

b) We have $\mu_{\overline{X}} = \mu_X = 7670866$ and

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{3}} = 1,833,397.$$

c) The observed sample means are 8 133 781, 6 139 336, 7 528 187, 6 360 197, 1 0357 199, 5 937 197, and 6 719 487. The standard deviation of this dataset is 1,555,502

2. a,b) The means of X and $\overline{X}_9$ are both 15.54545.
c) The standard deviation of $\overline{X}_9$ is $\sigma_X/\sqrt{9}$, or 14.49575/3, or 4.831917.
d) The fifteen observed sample means are 26.33333, 15.55556, 10.88889, 11.44444, 13.88889, 15.33333, 21.33333, 20.11111, 11.33333, 12.66667, 14.44444, 13.11111, 12.55556, 15.66667, and 18.11111. The standard deviation of this dataset is 4.31532.

3. a) The exact probability is

$$\sum_{k=14}^{19} \binom{25}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{25-k},$$

or 0.34298 to five decimal places.

b) The normal approximation without correction for continuity is

$$\begin{aligned}
P(14 \le X \le 19) &= P\left(\frac{14}{25} \le \frac{X}{25} \le \frac{19}{25}\right) \\
&= P\left(0.56 \le \overline{X} \le 0.76\right) \\
&= P\left(\frac{0.56 - 0.5}{\sqrt{(0.5)(0.5)/25}} \le \frac{\overline{X} - 0.5}{\sqrt{(0.5)(0.5)/25}} \le \frac{0.76 - 0.5}{\sqrt{(0.5)(0.5)/25}}\right) \\
&= P\left(\frac{0.06}{0.1} \le Z \le \frac{0.26}{0.1}\right) \\
&= P\left(0.6 \le Z \le 2.6\right) \\
&= \Phi(2.6) - \Phi(0.6) \\
&= 0.26959.
\end{aligned}$$

c) The normal approximation with correction for continuity is

$$\begin{aligned}
P(14 \le X \le 19) &= P(13.5 \le X \le 19.5) \\
&= P\left(\frac{13.5}{25} \le \frac{X}{25} \le \frac{19.5}{25}\right) \\
&= P\left(0.54 \le \overline{X} \le 0.78\right) \\
&= P\left(\frac{0.54 - 0.5}{\sqrt{(0.5)(0.5)/25}} \le \frac{\overline{X} - 0.5}{\sqrt{(0.5)(0.5)/25}} \le \frac{0.78 - 0.5}{\sqrt{(0.5)(0.5)/25}}\right) \\
&= P\left(\frac{0.04}{0.1} \le Z \le \frac{0.28}{0.1}\right) \\
&= P\left(0.4 \le Z \le 2.8\right) \\
&= \Phi(2.8) - \Phi(0.4) \\
&= 0.34202.
\end{aligned}$$

4. a) The exact probability is $\sum_{k=160}^{500} \binom{500}{k}(0.3)^k(0.7)^{500-k}$, or 0.177 to three decimal places. If X is the number of regular alcohol consumers in a random sample of size 500, then, using $n = 500$ and $p = 0.3$, the normal approximation is

$$\begin{aligned}
\mathrm{P}(\mathrm{X} \geq 160) &= \mathrm{P}\left(\frac{\mathrm{X}}{500} \geq \frac{160}{500}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} \geq 0.32\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}} - 0.30}{\sqrt{(0.3)(0.7)/500}} \geq \frac{0.32 - 0.30}{\sqrt{(0.3)(0.7)/500}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq \frac{0.32 - 0.30}{0.020494}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq 0.9759\right) \\
&= 1 - \Phi(0.9759) \\
&= 1 - 0.8354 \\
&= 0.1646.
\end{aligned}$$

The normal approximation with correction for continuity is

$$\begin{aligned}
\mathrm{P}(\mathrm{X} \geq 160) &= \mathrm{P}\left(\frac{\mathrm{X}}{500} \geq \frac{159.5}{500}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} \geq 0.319\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}} - 0.30}{\sqrt{(0.3)(0.7)/500}} \geq \frac{0.319 - 0.30}{\sqrt{(0.3)(0.7)/500}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq \frac{0.319 - 0.30}{0.020494}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq 0.9271\right) \\
&= 1 - \Phi(0.9271) \\
&= 1 - 0.8230 \\
&= 0.177.
\end{aligned}$$

b) The exact probability is $\sum_{k=0}^{74} \binom{300}{k}(0.3)^k(0.7)^{300-k}$, or 0.02388 to five decimal places. If X is the number of regular alcohol consumers in a random sample of size 300, then, using $n = 300$ and $p = 0.3$,

the normal approximation is

$$\begin{aligned}
\mathrm{P}(\mathrm{X} < 75) &= \mathrm{P}(\mathrm{X} \leq 74) \\
&= \mathrm{P}\left(\frac{\mathrm{X}}{300} \leq \frac{74}{300}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} \leq 0.2467\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}} - 0.30}{\sqrt{(0.3)(0.7)/300}} \leq \frac{0.2467 - 0.30}{\sqrt{(0.3)(0.7)/300}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \leq \frac{0.2467 - 0.30}{0.0264575}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \leq -2.01455\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq 2.01455\right) \\
&= 1 - \Phi(2.01455) \\
&= 1 - 0.978024 \\
&= 0.021976.
\end{aligned}$$

The error with this approximation is 0.00191.

The normal approximation with correction for continuity is

$$\begin{aligned}
\mathrm{P}(\mathrm{X} < 75) &= \mathrm{P}(\mathrm{X} \leq 74.5) \\
&= \mathrm{P}\left(\frac{\mathrm{X}}{300} \leq \frac{74.5}{300}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} \leq 0.24833\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}} - 0.30}{\sqrt{(0.3)(0.7)/300}} \leq \frac{0.24833 - 0.30}{\sqrt{(0.3)(0.7)/300}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \leq \frac{0.24833 - 0.30}{0.0264575}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \leq -1.9529\right) \\
&= \mathrm{P}\left(\mathrm{Z} \geq 1.9529\right) \\
&= 1 - \Phi(1.9529) \\
&= 1 - 0.97458 \\
&= 0.02542.
\end{aligned}$$

The error with this approximation is 0.00154. (A bit less than the error that results without the correction for continuity.)

5. Let $\mathrm{X}_j = 1$ if the j^{th} passenger insists on flying first class and $\mathrm{X}_j = 0$ otherwise. We are to find an integer m such that $\mathrm{P}\left(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{250} \leq m\right) = 0.975$. First observe that, for $p = 0.1$, we have

$q = 1 - p = 0.9$ and $\sqrt{p\,q/250} = 0.0189737$. Thus,

$$\begin{aligned}
0.975 &= \mathrm{P}\left(\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{250} \le m\right) \\
&= \mathrm{P}\left(\frac{\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{250}}{250} \le \frac{m}{250}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} \le \frac{m}{250}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}} - p \le \frac{m}{250} - p\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}} - p}{\sqrt{p\,q/250}} \le \frac{m/250 - p}{\sqrt{p\,q/250}}\right) \\
&= \mathrm{P}\left(\mathrm{Z} \le \frac{m/250 - 0.1}{0.0189737}\right) \\
&= \Phi\left(\frac{m/250 - 0.1}{0.0189737}\right)
\end{aligned}$$

Because $\Phi(1.96) = 0.975$, we have $(m/250 - 0.1)/0.0189737 = 1.96$, or $m = 34.297$. We round up the fractional seat for an answer of 35.

6. Let X be the weight of food product, if that is what you call it, in a jar of Ozziemite. Let $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{36}$ be a random sample of Ozziemite jars. Let us pause to contemplate so awesome a collection.

   There is a wrong way to proceed. Most students will find this path on their own with ho aid from the author, but he wants to help the others. The error is subtle, but, once pointed out, obvious. Nevertheless, even an individual who has been set straight by wise counsel may easily blunder into the error again. Here goes: We calculate

$$\begin{aligned}
\mathrm{P}(36\,\mathrm{X} > 585) &= \mathrm{P}(\mathrm{X} > 585/36) \\
&= \mathrm{P}(\mathrm{X} > 16.25) \\
&= \mathrm{P}\left(\frac{\mathrm{X} - 16}{0.6} > \frac{16.25 - 16}{0.6}\right) \\
&= \mathrm{P}\left(Z > 0.4166667\right) \\
&= 1 - \Phi(0.4166667) \\
&= 0.3384611.
\end{aligned}$$

   Did you spot the error of our ways. Two, actually. Near the end, we replaced (X - 16)/0.6 with Z. On what authority? The fraction (X - 16)/0.6 is the standardization of X, but where in the statement of the problem was it said that X has a normal distribution? Actually, the computation went off the rails from the get-go. Instead of using $36\,\mathrm{X}$, we should have been using $\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{36}$. There is a big difference! Even though $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{36}$ are identically distributed, they are not the same random variables, and so their sum is *not* $36\,\mathrm{X}$. (If the grades, Y_1 and Y_2, of two students on an exam are randomly selected, then the two grades are drawn from a single distribution F_Y. But we don't expect $\mathrm{Y}_1 + \mathrm{Y}_2 = 2\,\mathrm{Y}$: there is no reason for the observed values of the students' exam scores to be the same.) By using $\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_{36}$ we also avoid the other error we mentioned: this sum *is* approximately

normal according to the Central Limit Theorem. Thus

$$\begin{aligned}
\mathrm{P}\left(\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}>585\right) &= \mathrm{P}\left(\frac{\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}}{36}>\frac{585}{36}\right)) \\
&= \mathrm{P}(\overline{\mathrm{X}}>16.25) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}}-16}{0.6/\sqrt{36}}>\frac{16.25-16}{0.6/\sqrt{36}}\right) \\
&\approx \mathrm{P}\left(Z>2.5\right) \\
&= 1-\Phi(2.5) \\
&= 0.0062.
\end{aligned}$$

7. The mean of a uniform distribution on an interval is the midpoint of the interval: in this case 0.5. The variance of a uniform distribution on the interval $[a,b]$ is $(b-a)^2/12$: in this case 1/12, or 0.8333. The standard deviations of the uniform distributions in this exercise are therefore 0.2886751.

   Using the Central Limit Theorem, we calculate

$$\begin{aligned}
\mathrm{P}\left(\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}>22\right) &= \mathrm{P}\left(\frac{\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}}{36}>\frac{22}{36}\right) \\
&= \mathrm{P}\left(\overline{\mathrm{X}}>0.6111111\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}}-0.5}{0.2886751/6}>\frac{0.6111111-0.5}{0.2886751/6}\right) \\
&= \mathrm{P}\left(\frac{\overline{\mathrm{X}}-0.5}{0.2886751/6}>2.309401\right) \\
&\approx \mathrm{P}\left(\mathrm{Z}>2.309401\right) \\
&= 1-\Phi\left(2.309401\right) \\
&= 0.01046067.
\end{aligned}$$

   The author ran a simulation to reality-check this answer. In an R-session, samples of size $n=36$ were drawn from a uniform distribution on the interval [0,1] one thousand times. For those 1000 trials, $\mathrm{P}\left(\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}\right)$ exceeded 22 seven times; the maximum value of $\mathrm{P}\left(\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{36}\right)$ was 22.70965. The frequency of sums over 22, namely, 7/1000, or 0.007, is in the ballpark of the theoretical calculated probability, 0.01046067—the error is about 0.003. For those interested, the R code used for the simulation was

```
> trials = numeric(1000)
> for(j in 1:1000) trials[j] = sum(runif(36,min=0,max=1))
> max(trials)
> count = 0
> for(j in 1:1000) if (trials[j]>22) count = count+1
> count
```

8. Let $p=0.88144$ and $q=1-0.88144$.
   a) If f is the p.f. of Binom$(100,p)$, then the requested probability is $f(80)$, or $\binom{100}{80}p^{80}\,q^{20}$, or 0.006657873364. In R, `dbinom(80,100,0.88144)` does the trick.
   b) The normal approximation of Binom$(100,p)$ is $\mathrm{N}\left(100\,p,\sqrt{100\,p\,q}\right)$. In R, the resulting approximation is

`pnorm(80.5,100*0.88144,sqrt(100*0.88144*(1-0.88144)))`
`- pnorm(79.5,100*0.88144,sqrt(100*0.88144*(1-0.88144)))`.
To use a Phi table, we convert to z-scores first. We let X_j be 1 if the j^{th} subject lives to 75 and 0 otherwise. We calculate

$$\begin{aligned}
\mathrm{P}\left(\mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{100}=80\right) &= \mathrm{P}\left(79.5\le \mathrm{X}_1+\mathrm{X}_2+\cdots+\mathrm{X}_{100}\le 80.5\right)\\
&\approx \mathrm{P}\left(79.5\le \mathrm{N}\left(100\,p,\sqrt{100\,p\,q}\right)\le 80.5\right)\\
&= \mathrm{P}\left(\frac{79.5-100\,p}{\sqrt{100\,p\,q}}\le \frac{\mathrm{N}\left(100\,p,\sqrt{100\,p\,q}\right)-100\,p}{\sqrt{100\,p\,q}}\le \frac{80.5-100\,p}{\sqrt{100\,p\,q}}\right)\\
&= \mathrm{P}\left(-2.673925\le \mathrm{Z}\le -2.364586\right)\\
&= \Phi(-2.364586)-\Phi(-2.673925)\\
&= 0.009025118-0.003748461\\
&= 0.005276657,
\end{aligned}$$

which is off by about 0.001.
c) The answer is $f(80)+f(81)+\cdots+f(100)$, or $\sum_{k=80}^{100}\binom{100}{k}p^{80}\,q^{100-k}$, or 0.9936516955. In R, `1 - pbinom(79,100,0.88144)` does the job, as does `pbinom(79,100,0.88144,lower.tail=FALSE)`, as does `sum(dbinom(80:100,100,0.88144))`.
d) Without the correction for continuity, the approximation is

$$1-\Phi\left(\frac{80-100\times 0.88144}{\sqrt{100\times 0.88144\times(1-0.88144)}}\right),$$

or 0.99411984.
e) With the correction for continuity, the approximation is

$$1-\Phi\left(\frac{79.5-100\times 0.88144}{\sqrt{100\times 0.88144\times(1-0.88144)}}\right),$$

or 0.99625154.

9. One might guess that with a population variance of $(15)^2$, or 225, it would be 50-50 for an observed sample variance to be greater than 225. That would lead to the conjecture that the requested probability must be greater than 0.5. But that isn't necessarily so. Let's do the calculation. For that, we use $(n-1)\,\mathrm{S}^2/\sigma^2\sim\chi^2_{n-1}$. In the first case, for $\sigma=15$ and $n=11$, we have $10\,\mathrm{S}^2/225\sim\chi^2_{10}$. Thus,

$$\begin{aligned}
\mathrm{P}\left(\mathrm{S}^2>220\right) &= \mathrm{P}\left(\frac{10\,\mathrm{S}^2}{225}>\frac{10\times 220}{225}\right)\\
&= \mathrm{P}\left(\chi^2_{10}>9.777778\right)\\
&= 0.4602013.
\end{aligned}$$

For $n=51$, the computation is

$$\begin{aligned}
\mathrm{P}\left(\mathrm{S}^2>220\right) &= \mathrm{P}\left(\frac{50\,\mathrm{S}^2}{225}>\frac{50\times 220}{225}\right)\\
&= \mathrm{P}\left(\chi^2_{50}>48.88889\right)\\
&= 0.5179379.
\end{aligned}$$

10. We use $\left(\overline{\mathrm{X}}-\mu\right)/\left(\mathrm{S}/\sqrt{n}\right)\sim \mathrm{t}_{n-1}$. In the first case, for $\sigma=5$, $\overline{\mathrm{X}}=11$, and $n=6$, we have

$$\begin{aligned}\mathrm{P}\left(\mu>10\right) &= \mathrm{P}\left(-\mu<-10\right)\\ &= \mathrm{P}\left(\overline{\mathrm{X}}-\mu<11-10\right)\\ &= \mathrm{P}\left(\frac{\overline{\mathrm{X}}-\mu}{S/\sqrt{6}}<\frac{11-10}{5/\sqrt{6}}\right)\\ &= \mathrm{P}\left(\mathrm{t}_5<0.4898979\right)\\ &= 0.6775284.\end{aligned}$$

For $n=51$, the first few lines are identical. The change begins in the third line, where we use 51 rather than 6:

$$\begin{aligned}\mathrm{P}\left(\mu>10\right) &= \mathrm{P}\left(-\mu<-10\right)\\ &= \mathrm{P}\left(\overline{\mathrm{X}}-\mu<11-10\right)\\ &= \mathrm{P}\left(\frac{\overline{\mathrm{X}}-\mu}{S/\sqrt{51}}<\frac{11-10}{5/\sqrt{51}}\right)\\ &= \mathrm{P}\left(\mathrm{t}_{50}<1.428286\right)\\ &= 0.9202867.\end{aligned}$$

The population mean is 11 so one would expect that a sample mean would exceed 10 with a fairly high probability, provided that the sample size was not too small. When $n=6$ the probability is only 0.68 because the sample size 6 does not greatly reduce the variance. But $n=51$ is large enough to greatly reduce the variability.

Chapter 9.

1. a) $\left[17.75-1.96(4.199)/\sqrt{16}, 17.75+1.96(4.199)/\sqrt{16}\right]$, or [15.69,19.81], is a 95% confidence interval.
$\left[17.75-2.5758(4.199)/\sqrt{16}, 17.75+2.5758(4.199)/\sqrt{16}\right]$, or [15.05,20.45], is a 99% confidence interval.
b) $\left[17.75-1.96(5.651)/\sqrt{16}, 17.75+1.96(5.651)/\sqrt{16}\right]$, or [14.98,20.52], is a 95% confidence interval.
$\left[17.75-2.5758(5.651)/\sqrt{16}, 17.75+2.5758(5.651)/\sqrt{16}\right]$, or [14.11,21.39], is a 99% confidence interval.
c) Because $\mathrm{t}_{0.025,15}=2.1314$ and $\mathrm{t}_{0.005,15}=2.9467$, we have
$\left[17.75-2.1314(5.651)/\sqrt{16}, 17.75+2.1314(5.651)/\sqrt{16}\right]$, or [14.74,20.76], is a 95% confidence interval, and
$\left[17.75-2.9467(5.651)/\sqrt{16}, 17.75+2.9467(5.651)/\sqrt{16}\right]$, or [13.59,21.91], is a 99% confidence interval.

2. The Agresti-Coull adjustment, necessary because there were fewer than 10 failures, results in 57 successes in 64 trials. The required confidence interval is

$$\frac{57}{64}\pm 1.96\sqrt{\frac{(57/64)(7/64)}{64}}$$

or $[0.8142, 0.9671]$.

3. Formula (9.1.8) provides the answer. With $z_{0.025}=1.996$ and $\mathrm{ME}_0=0.03$, we have

$$n=\left\lceil\left(\frac{1.96}{2\times 0.03}\right)^2\right\rceil=\lceil 1067.1\rceil=1068.$$

4. First, $z_{0.005}=2.576$ and $\mathrm{SE}\left(\overline{p}\right)=\sqrt{(0.09)(0.91)/1234}=0.008147$. It follows that $\mathrm{ME}=2.576\times 0.008147=0.020986$, or 2.0986%.

5. We use formula (9.4.5). The length of the interval is

$$2\,z_{\alpha/2}\sqrt{\frac{\widehat{p}_1\,(1-\widehat{p}_1)}{n}+\frac{\widehat{p}_2\,(1-\widehat{p}_2)}{m}} = 2\times 1.96\sqrt{\frac{(0.428)\,(1-0.428))}{550}+\frac{0.527\,(1-0.527)}{550}} = 0.11749.$$

6. We look up $\chi^2_{0.05,11} = 19.6751$ and $\chi^2_{0.95,11} = 4.5748$. A 90% confidence interval for the variance is

$$\left[\frac{11}{19.6751}(4.23)^2, \frac{11}{4.57481}(4.23)^2\right],$$

or [10, 43.02]. A 90% confidence interval for the standard deviation is $[\sqrt{10}, \sqrt{43.02}]$, or [3.16,6.56].

7. Give me a place to swim,
And a place to float,
And call this Lake, Ontario.
A place to swim,
A place to float,
Ontari-ari-ari-o![1]

The exam solution *assumes* that the population variances are the same. (The rule of thumb stated in the notes does lead to this assumption: the larger observed sample deviation does not exceed the smaller by more than a factor of 2.) With the assumption that the population variances are equal, the pooled variance is

$$\sqrt{\frac{19}{40}(261.111)^2+\frac{21}{40}(304.369)^2}\times\sqrt{\frac{1}{20}+\frac{1}{22}},$$

or 87.94. According to formula (9.4.4), the 95% confidence interval is

$$1196.75 - 1271.59 \pm \mathrm{t}_{0.025,20+22-2}\times 87.94$$

or

$$[1196.75 - 1271.59 - 2.0211\times 87.94, 1196.75 - 1271.59 + 2.0211\times 87.94],$$

or $[-252.58, 102.9]$. This is very close to the exam answer, E. The author does not know why there is a discrepancy in the fourth significant figure, but has not lost sleep over it.

Without pooling, the answer is obtained by formula (9.4.3). The standard deviation of $T_{\mathrm{M}} - T_{\mathrm{F}}$ is $\sqrt{(261.111)^2/22 + (304.369)^2/20}$, or 87.926. The number of degrees of freedom we use is 19, the smaller of $20-1$ and $22-1$. The Student-t value we use is $\mathrm{t}_{0.025,19} = 2.0930$. According to formula (9.4.3), the 95% confidence interval is

$$1196.75-1271.59\pm\mathrm{t}_{0.025,19}\times 87.926 \text{ or } [1196.75-1271.59-2.0930\times 87.926, 1196.75-1271.59+2.0930\times 87.926],$$

or $[-258.87, 109.19]$. As mentioned in the notes, this confidence interval is conservative (wider than it might need be). On the other hand, it does not rely on an assumption that is not fully justified.

8. If z is the z-value used, then

$$0.05 = z\times\sqrt{\frac{(0.6)(0.4)}{300}},$$

or $z = 1.767766953$. If the confidence interval was at level α, then

$$\frac{\alpha}{2} = 1 - \Phi(1.767766953) = 1 - 0.96145 = 0.03855.$$

The interval has $100(1-\alpha)\%$, or $100(1-0.07710)\%$, or 92.29% confidence level.

[1] Sung to the tune of *A Place to Stand, A Place to Grow*. `http://www.youtube.com/watch?v=rt-5tAWJxvU`

9. The t-tables give $t_{0.025,30} = 2.0423$ and $t_{0.025,40} = 2.0211$. Interpolating, we approximate $t_{0.025,39} = 2.0211 + (2.0423 - 2.0211)/10$, or $t_{0.025,39} = 2.02322$. The 95% confidence interval is $105 \pm 2.02322 \times 16/\sqrt{40}$, or $[99.88, 110.12]$.

10. Let p_M (respectively p_W) be the percentage of men (women) who exercise regularly. The standard deviation of $p_M - p_W$ is

$$\sqrt{\frac{(88/150)(1-88/150)}{150} + \frac{(130/200)(1-130/200)}{200}}, \qquad \text{or} \qquad 0.05248.$$

The requested confidence interval is

$$\frac{88}{150} - \frac{130}{200} \pm 1.96 \times 0.05248$$

or $[-0.166, 0.03951]$.

11. The required confidence interval is $\overline{W} \pm t_{0.95,10}\, S/\sqrt{11}$. We calculate $\overline{W} = 71.909091, S = 148.013820$, and $t_{0.95,10} = 1.812461$. The resulting interval is $[-8.977145, 152.795327]$. Because 0 is in this interval, we decline to conclude that X > Y on average.

12. We use the Range Rule of Thumb for a rough estimate S_{RR} of the unknown standard deviation: $\sigma \approx S_{RR} = (51008 - 870)/4 = 12534.5$. We look up $z_{0.05} = 1.644854$. Then

$$n = \left\lceil \left(\frac{z_{0.05}\, S_{RR}}{ME_0} \right)^2 \right\rceil = \left\lceil \left(\frac{1.644854 \times 12534.5}{2000} \right)^2 \right\rceil = \lceil 106.2695 \rceil = 107.$$

13. We use a paired confidence interval. The differences P - R are 3, 2, 3, 2, 2, -1, 2, 3. Then $\mu_{P-R} = 2$, $S_{P-R} = 1.309307$, and $t_{0.975,7} = 2.364624$. The requested confidence interval is $2 \pm 2.364624 \times 1.309307/\sqrt{8}$, or [0.9053921 , 3.094608].

14. We haven't been asked for the confidence interval, but it is little more work and we will calculate it for practice. First, $\widehat{p}_{ZNM} = 40/50 = 0.8$, $\widehat{p}_{BP} = 10/50 = 0.2$, and $\widehat{p}_{ZNM} - \widehat{p}_{BP} = 0.8 - 0.2 = 0.6$. Next, We look up $z_{0.025} = 1.959964$. The standard error is

$$SE(\widehat{p}_{ZNM} - \widehat{p}_{BP}) = \sqrt{\frac{(0.8)(0.2)}{50} + \frac{(0.2)(0.8)}{50}} = 0.08.$$

That is what is requested and the problem is over. Let's continue with the confidence interval, nevertheless. The margin of error is

$$ME\,(\widehat{p}_{ZNM} - \widehat{p}_{BP}) = 1.959964 \sqrt{\frac{(0.8)(0.2)}{50} + \frac{(0.2)(0.8)}{50}} = 0.1567971.$$

The confidence interval is 0.6 ± 0.1567971, or $[0.4432029, 0.7567971]$.

15. The requested critical value is $t_{0.015,3}$ In R, it is obtained from the call `qt(0.985,3)`, which returns 3.896046. Interpolation between the tabulated values $t_{0.025,3} = 3.1824$ and $t_{0.010,3} = 4.5407$ is not that accurate. It leads to the approximation

$$t_{0.015,3} \approx t_{0.010,3} + \frac{t_{0.025,3} - t_{0.010,3}}{0.025 - 0.010}\,(0.015 - 0.010) = 4.5407 + \frac{3.1824 - 4.5407}{0.025 - 0.010}\,(0.015 - 0.010) = 4.0879.$$

If the author had made this problem, he would have used answer choices with greater gaps between them in order to eliminate the problem of inaccuracy due to the use of tables.

16. The improvements are 16, 6, -2, 6, 10. The sample mean of this dataset is 7.2nd the sample standard deviation is 6.572671. The requested length is $2 \times t_{0.05,4} \times S/\sqrt{5}$, or $2 \times 2.131847 \times 6.572671/\sqrt{5}$, or 12.53265.

17. We are to compare two proportions. The sample proportion that is based on a sample of size 361 is 67/361, or 0.1855956. The sample proportion that is based on a sample of size 89 is 26/89, or 0.2921348. The standard deviation of the difference of sample proportions is

$$\sqrt{\frac{0.1855956(1-0.1855956)}{361} + \frac{0.2921348(1-0.2921348)}{89}},$$

or 0.05236606. We look up $z_{0.005} = 2.575829$. The margin of error is $2.575829 \times 0.05236606$, or 0.134886. The length of the confidence interval is twice the margin of error: 2×0.134886, or 0.269772.

18. The margin of error for the original confidence interval is (31.844 - 29.202)/2, or 1.321. Using the formula

$$\mathrm{ME}(\widehat{\mu}) = t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

for the margin of error for a confidence interval of level $100(1-\alpha)\%$ based on a sample size n and a Student-t distribution, we set the left side equal to 1.321, we substitute $\alpha = 0.05, n = 20$, and $t_{0.025,19} = 2.093024$ in the right side, and we solve for the observed sample standard deviation S, obtaining S = 2.822563. It follows that the length of the 90% confidence interval computed from the same sample data is

$$2 \times \mathrm{ME}(\widehat{\mu}) = 2 \times t_{0.05,19} \frac{2.822563}{\sqrt{20}} = 2 \times 1.729133 \times \frac{2.822563}{\sqrt{20}} = 2.182665.$$

19. The observed sample variance is $S^2 = 0.1601111$. We look up $\chi^2_{0.975,9} = 2.700389$ and $\chi^2_{0.025,9} = 19.02277$. A 95% confidence interval for σ^2 is $[9 \times 0.1601111/19.02277, 9 \times 0.1601111/2.700389]$, or $[0.07575133, 0.5336267]$. We square-root to obtain a 95% confidence interval of the standard deviation of the voltage: $\left[\sqrt{0.07575133}, \sqrt{0.5336267}\right]$, or [0.2752296,0.7304976].

20. We look up $t_{0.025,25} = 2.0595$, $\chi^2_{0.975,25} = 13.1197$, and $\chi^2_{0.025,25} = 40.6465$. For the mean, the confidence interval is $9.26 \pm 2.0595 \times 1.73/\sqrt{26}$, or [8.561251,9.958749]. For the *variance*, the confidence interval is $[25 \times (1.73)^2/40.6465, 25 \times (1.73)^2/13.1197]$, or $[1.840810, 5.703065]$. The interval for the standard deviation is obtained by square-rooting the endpoints: [1.3567652.388109].

21. We must treat this as a one-sample problem: the data will be paired (a first observation and a second observation for each mouse). The differences 129-113, 89-97, 136-139, 163-85, 118-75 give rise to the dataset 16 -8 -3 78 43 of observed improvement times. (The negative numbers are for the slow learner mice who took more time on their second attempts. Every species has a few.) The observed mean improvement time is 25.2. The observed standard deviation is 35.6609. With $n = 5$, we are clearly dealing with a small sample, hence the Student-t distribution. We look up the critical value $t_{0.95,4} = 2.131847$. The confidence interval is $25.2 \pm 2.131847 \times 35.6609/\sqrt{5}$, or $[-8.8, 59.2]$. Evidently the sample size was too small to glean anything useful. For all we know, the true mean improvement time might even be negative. Perhaps some mice found the maze more boring the second time through and decided to explore the dead ends.

22. We use a two-sample confidence interval for the difference of proportions. The sample proportion of at-risk women who develop eclampsia is $\widehat{p} = 96/4993 = 0.019227$ and for treated women $\widehat{p}_{\mathrm{MS}} = 40/4999 = 0.0080016$. Using $z_{0.025} = 1.959964$, the confidence interval for $p - p_{\mathrm{MS}}$ is

$$0.01922692 - 0.0080016 \pm 1.959964 \sqrt{\frac{0.019227(1-0.019227)}{4999} + \frac{0.008002(1-0.008002)}{4993}},$$

which evaluates to $0.01122532 \pm 1.959964 \times 0.002315587$, or 0.004538467. The length of the confidence interval is twice the margin of error, or 0.009076934.

23. The sample standard deviation of $\overline{\mathrm{X}} - \overline{\mathrm{Y}}$ is

$$\mathrm{S} = \sqrt{\frac{\mathrm{S_X^2}}{n_\mathrm{X}} + \frac{\mathrm{S_X^2}}{n_\mathrm{Y}}} = \sqrt{\frac{9}{15} + \frac{4}{36}} = 0.843274.$$

The conservative assignment of the critical value is $\mathrm{t}_{0.025,\min(15-1,36-1)}$, or $\mathrm{t}_{0.025,14}$, or 2.144787. The length of the confidence interval is $2 \times 2.144787 \times 0.843274$, or 3.617286. This is slightly longer than the intended exam answer, 3.525. That length is somewhat smaller because for the degrees of freedom on the exam is not the conservative choice. The remainder of this discussion concerns a complicated formula that is not covered in either these notes or the course on which they are based. But, for the record, the intended answer is obtained by using the following value for the number df of degrees of freedom:

$$\mathrm{df} = \frac{\mathrm{S}^4}{\frac{(\mathrm{S_X}/n_\mathrm{X})^2}{n_\mathrm{X}-1} + \frac{(\mathrm{S_Y}/n_\mathrm{Y})^2}{n_\mathrm{Y}-1}} = \frac{(0.4127004)^4}{\frac{(9/15)^2}{14} + \frac{(4/36)^2}{35}} = 19.39919.$$

Because $\mathrm{t}_{0.025,19.39919} = 2.090113$, the length of the confidence interval that was expected on the exam was $2 \times 2.090113 \times 0.843274$, or 3.525076. Not only is the formula for df complicated and unmotivated, it results in a noninteger value. The calculation of $\mathrm{t}_{0.025,19.39919} = 2.090113$ was by technology, but it could have been approximated by interpolating the df parameter.

Chapter 10.

A preliminary word about interpolations. The author was asked the following question by a First President University student:

> How do you interpolate for a t or chi-square value x to find a p-value? I ask because in Chapter 10, Exercise 16, the solution interpolates essentially as:
>
> $$p = \frac{\text{(bigger significance level - smaller significance level)}}{\text{(value corresp to bigger alpha - value corresp to smaller alpha)}}\ (x - \text{value corresp to smaller alpha}) + \text{smaller alpha}$$
>
> However, in the final exam of Fall 2014, Question 16, the solution interpolates essentially
>
> $$p = \frac{\text{(smaller significance level - bigger significance level)}}{\text{(value corresp to smaller alpha - value corresp to bigger alpha)}}\ (x - \text{value corresp to bigger alpha}) + \text{bigger alpha}$$
>
> So essentially the order is switched between the two, and I was wondering which one we are supposed to use in which situation?

The answer is, both equations lead to the *same* value. Let's see by way of an example using these values:

bigger confidence level = 0.050
corresponding value to bigger alpha = 2.3
smaller confidence level = 0.025
corresponding value to smaller alpha = 2.7

with x = 2.6.

The first equation becomes

$$\frac{(0.050 - 0.025)}{(2.3 - 2.7)}\,(2.6 - 2.7) + 0.025, \qquad \text{or } 0.03125.$$

The second equation becomes

$$\frac{(0.025 - 0.050)}{(2.7 - 2.3)}\,(2.6 - 2.3) + 0.050, \qquad \text{or } 0.03125.$$

Now, on to the solutions.

1. Let p be the true proportion of Hispanics sentenced to jury duty. Let $p_0 = 0.19$. We will test
   $H_0 : p = p_0$
   $H_1 : p \neq p_0$
   The observed sample proportion is 125/720, or 0.1736. Let $\delta = |0.1736 - p_0| = |0.1736 - 0.1900| = 0.0164$. We must calculate of $|\widehat{p} - p_0| \geq \delta$ assuming that the null hypothesis is true.

$$\begin{aligned}
&\mathrm{P}\left(|\widehat{p} - p_0| \geq \delta \,|\, p = p_0\right) \\
&\qquad = \mathrm{P}\left(\left|\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/720}}\right| \geq \frac{\delta}{\sqrt{p_0(1-p_0)/720}} \,\middle|\, p = p_0\right) \\
&\qquad = \mathrm{P}\left(\left|\frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/720}}\right| \geq 1.1217 \,\middle|\, p = p_0\right) \\
&\qquad = \mathrm{P}\left(|Z| \geq 1.1217\right) \\
&\qquad = 2\,\mathrm{P}\left(Z \geq 1.1217\right) \\
&\qquad = 2\left(1 - \mathrm{P}\left(Z \leq 1.1217\right)\right. \\
&\qquad = 2\left(1 - \Phi(1.1217)\right. \\
&\qquad = 0.262.
\end{aligned}$$

2. Answered without pooling:
   Let p (and p_m) be the proportion of women who have not been screened (respectively, who have been screened) who die from breast cancer. We will test
   $H_0 : p_m = p$
   $H_1 : p_m < p$
   The observed values of $\widehat{p}$ and $\widehat{p_m}$ are 200/32000, or 0.00625, 150/30000, or 0.005. The observed difference in proportions is 0.00625 - 0.005, or 0.00125. We calculate

$$\mathrm{SE}\left(\widehat{p} - \widehat{p_m}\right) = \sqrt{\frac{0.00625(1 - 0.00625)}{32000} + \frac{0.005(1 - 0.005)}{30000}} = 0.0006.$$

   We calculate the probability that the observed value of $\widehat{p} - \widehat{p_m}$ is greater than or equal to 0.00625 assuming that the expectation of $\widehat{p} - \widehat{p_m}$ is 0. We have

$$\begin{aligned}
\mathrm{P}\left(\widehat{p} - \widehat{p_m} \geq 0.00125 \,|\, p = p_m\right) &= \mathrm{P}\left(\frac{\widehat{p} - \widehat{p_m}}{0.0006} \geq \frac{0.00125}{0.0006} \,\middle|\, p = p_m\right) \\
&= \mathrm{P}\left(Z \geq 2.0833\right) \\
&= 1 - \mathrm{P}\left(Z \leq 2.0833\right) \\
&= 0.01861.
\end{aligned}$$

Answered with pooling (as intended on the exam):
The pooled sample proportion is (200+150)/(32000+30000), or 0.005645. Using this value,

$$\mathrm{SE}\,(\widehat{p}-\widehat{p_m}) = \sqrt{\frac{0.005645(1-0.005645)}{32000}+\frac{0.005645(1-0.005645)}{30000}} = 0.0006021,$$

and

$$\begin{aligned}\mathrm{P}\,(\widehat{p}-\widehat{p_m}\geq 0.00125\,|\,p=p_m) &= \mathrm{P}\left(\frac{\widehat{p}-\widehat{p_m}}{0.0006021}\geq\frac{0.00125}{0.0006021}\;\middle|\;p=p_m\right)\\ &= \mathrm{P}\,(Z\geq 2.0761)\\ &= 1-\mathrm{P}\,(Z\leq 2.0761)\\ &= 0.01894.\end{aligned}$$

3. Let $p_0 = 0.03$. Let p be the proportion of twin births among teenage mothers. Is $\mathrm{H}_0 : p = p_0$ or is $\mathrm{H}_a : p \neq p_0$? We calculate

$$\begin{aligned}\mathrm{P}\,(|\widehat{p}-p_0|\geq 15/1000-0.03\,|\,p=p_0) &= \mathrm{P}\left(\frac{|\widehat{p}-p_0|}{\sqrt{(p_0)\,(1-p_0)\,/1000}}\geq\frac{0.03-15/1000}{\sqrt{(0.03)(0.97)/1000}}\,|\,p=p_0\right)\\ &= \mathrm{P}\left(\frac{|\widehat{p}-p_0|}{\sqrt{(p_0)\,(1-p_0)\,/1000}}\geq 2.78064\,|\,p=p_0\right)\\ &= 2\,\mathrm{P}\left(\frac{\widehat{p}-p_0}{\sqrt{(p_0)\,(1-p_0)\,/1000}}\geq 2.78064\,|\,p=p_0\right)\\ &= 2\,\mathrm{P}\,(Z\geq 2.78064)\\ &= 2\,(1-\Phi(2.78064))\\ &= 0.005425\\ &\approx \frac{5}{1000}.\end{aligned}$$

4. We are testing $\mathrm{H}_0 : p_{\mathrm{NY}} = p_{\mathrm{StL}}$ versus $\mathrm{H}_a : p_{\mathrm{NY}} < p_{\mathrm{StL}}$. In the mechanics of the hypothesis test, either classical or modern, we assume the null hypothesis. The model for the difference therefore has mean 0. The standard deviation is

$$\sqrt{\frac{(53/100)(47/100)}{100}+\frac{(70/135)(65/135)}{135}},\ \text{or } 0.06588.$$

5. The emphasized words VERY LARGE tell us that the t-values that were used in the two tests were slightly greater than the corresponding z-values. The t-score calculated, the same for each student, would have been between the two t-values (in order for one to fail to reject and for the other to reject). Because $z_{0.05} = 1.645$ and $z_{0.025} = 1.96$, we see that 1.88 is the only t-score on offer that fits the bill.

6. We calculate

$$\begin{aligned}\mathrm{P}\,(\widehat{p}\geq 0.5503\,|\,p=0.5000) &= \mathrm{P}\left(\frac{\widehat{p}-0.5000}{\sqrt{(0.5000)(0.5000)/1003}}\geq\frac{0.5503-0.5000}{\sqrt{(0.5000)(0.5000)/1003}}\,|\,p=0.5000\right)\\ &= \mathrm{P}\,(Z\geq 3.186019629)\\ &= 1-\mathrm{P}\,(Z\leq 3.186019629)\\ &= 1-\Phi\,(3.186019629)\\ &= 0.0007212.\end{aligned}$$

7. Perhaps the wording with which the problem is stated might have been more clear, but only 8 cars in total are tested. Each of the eight cars is tested with regular and premium gasolene. The data sets on the two lines are *not* independent. The appropriate test is therefore a one-sided one-sample t-test of the differences. Thus, let X be the difference obtained by subtracting each entry on the first line from the entry below it: X : 3, 2, 3, 2, 2, -1, 2, 3. Let μ be the true mean of X. Our hypothesis test is

   $H_0 : \mu = 0$
   $H_0 : \mu > 0$

   We calculate $\overline{X} = 2, S^2 = 1.7143, S = 1.3093$. Then

$$\begin{aligned} P(\widehat{\mu} > 2 \,|\, \mu = 0) &= P(\widehat{\mu} - \mu > 2 - 0 \,|\, \mu = 0) \\ &= P\left(\frac{\widehat{\mu} - \mu}{S/\sqrt{8}} > \frac{2}{1.3093/\sqrt{8}} \,\middle|\, \mu = 0 \right) \\ &= P(t_7 > 4.3205) \\ &= 0.0017. \end{aligned}$$

8. First, we calculate the number of members of the sample that come from the "Other" class: $1000 - (716 + 214 + 41 + 9)$, or 20.

   a) 2013. The expected numbers are: White, not Hispanic or Latino: 0.680×1000, or 680; Black or African American: 0.237×1000, or 237; Asian: 0.380×1000, or 38; White, Hispanic or Latino: 0.023×1000, or 23; Other: 0.022×1000, or 22. The test statistic is $(716 - 680)^2/680 + (214 - 237)^2/237 + (41 - 38)^2/38 + (9 - 23)^2/23 + (20 - 22)^2/22$, or 13.08. The p-value is $P\left(\chi^2_4 \geq 13.08\right)$, or 0.011. We reject the null hypothesis: the composition of the sample does *not* match that of the county population.

   b) 2010. The expected numbers are: White, not Hispanic or Latino: 0.689×1000, or 689; Black or African American: 0.233×1000, or 233; Asian: 0.350×1000, or 35; White, Hispanic or Latino: 0.014×1000, or 14; Other: 0.029×1000, or 29. The test statistic is $(716 - 689)^2/689 + (214 - 233)^2/233 + (41 - 35)^2/35 + (9 - 14)^2/14 + (20 - 29)^2/29$, or 8.215. The p-value is $P\left(\chi^2_4 \geq 8.215\right)$, or 0.084. The evidence for rejecting the null hypothesis is insufficient: the composition of the sample *does* match that of the county population.

9. The expected values are:

		Age	Group
Leaning	18-35	36-50	Over 50
Conservative	20	25	15
Moderate	70	87.5	52.5
Liberal	30	37.5	22.5

   The test statistic comes to 29.35. Because $\chi^2_{0.05,(3-1)\times(3-1)} = 9.4877$, we reject the null hypothesis (and it isn't even close): political leaning and age are related.

10. The expected values are:

	Smoker	Non-smoker
Drinker	166.38	187.62
Non-drinker	115.62	130.38

The test statistic comes to 12.93. Because $\chi^2_{0.01,(2-1)\times(2-1)} = 6.6349$, we reject the null hypothesis (and it isn't even close): smoking and drinking were related in the 1980s.

11. We set $d_0 = 0$. By looking at the sample variances, it seems obvious that the assumption of equal variances (and hence pooling) is unwise. We calculate

$$\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = \sqrt{\frac{15291.7}{5} + \frac{2484.62}{10}} = 57.505$$

and

$$t_{\alpha/2,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = t_{\alpha/2,\min(5,10)-1} \times 57.505 = t_{0.025,4} \times 57.505 = 2.7764 \times 57.505 = 159.66.$$

Because $\left|\overline{X} - \overline{Y} - d_0\right| = |300.8 - 153.2| = 147.6$ is not greater than 159.66, we retain the null hypothesis that the population means are equal, a conclusion that might not have been evident from the values of the sample means.

12. We calculate

$$\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = \sqrt{\frac{(65.3)^2}{32} + \frac{(89.6)^2}{32}} = 19.6$$

and

$$d_0 - t_{\alpha,\mathrm{df}} \sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}} = 0 - t_{0.05,31} \times 19.6 = -1.6955 \times 19.6 = -33.2318.$$

Because $\overline{X} - \overline{Y} - d_0 = 533.7 - 585.2 - 0 = -51.5$, which is less than -33.2318 (i.e., farther away from 0), we reject the null hypothesis that the population means are equal. There is sufficient evidence to conclude that cell phone usage while driving increases reaction time. So does old age.

13. The standard deviation S_D of $\widehat{\mu_X} - \widehat{\mu_Y}$ is given by

$$S_D = \sqrt{\frac{S_X^2}{32} + \frac{S_Y^2}{32}} = \sqrt{\frac{(65.3)^2}{32} + \frac{(89.6)^2}{32}} = 19.6.$$

Let $\mu_D = \mu_X - \mu_Y$ be the mean of $\widehat{\mu_X} - \widehat{\mu_Y}$. If $\mu_X = \mu_Y$, then $\mu_D = 0$ and

$$\frac{\widehat{\mu_X} - \widehat{\mu_Y}}{S_D} = \frac{\widehat{\mu_X} - \widehat{\mu_Y} - \mu_D}{S_D} \sim t_{31}.$$

The observed value of $\widehat{\mu_X} - \widehat{\mu_Y}$ is $533.7 - 585.2$, or -51.5. It follows that the p-value of the stated one-sided hypothesis test is

$$\begin{aligned}
P\left(\widehat{\mu_X} - \widehat{\mu_Y} \le -51.5 \,\middle|\, \mu_X - \mu_Y = 0\right) &= P\left(\frac{\widehat{\mu_X} - \widehat{\mu_Y}}{S_D} \le \frac{-51.5}{19.6} \,\middle|\, \mu_X - \mu_Y = 0\right) \\
&= P\left(t_{31} \le -2.6276\right) \\
&= \texttt{pt}(-2.6276, 31) \qquad \text{(This is R code. See below for tables.)} \\
&= 0.006626.
\end{aligned}$$

As indicated, the value 0.00626 was obtained using software. An approximation by interpolation is somewhat involved because there is no line in the table for df = 31. First, convert the left tail expressed mathematically in the third to last line to a right tail:

$$P\left(t_{31} \le -2.6276\right) = P\left(t_{31} \ge 2.6276\right).$$

Next, we will calculate $\mathrm{P}\left(\mathrm{t}_{30} \geq 2.6276\right)$ and $\mathrm{P}\left(\mathrm{t}_{40} \geq 2.6276\right)$. For each of df = 30 and df = 40, we observe that 2.6276 is between the values tabulated for $\alpha = 0.010$ and $\alpha = 0.005$. The two lines we seek can be written as

$$\alpha = \frac{0.010 - 0.005}{\mathrm{t}_{0.010,\mathrm{df}} - \mathrm{t}_{0.005,\mathrm{df}}} \left(t - \mathrm{t}_{0.005,\mathrm{df}}\right) + 0.005.$$

For df = 30 and $t = 2.6276$ we obtain

$$\alpha = \frac{0.010 - 0.005}{2.4573 - 2.75} \left(2.6276 - 2.75\right) + 0.005 = 0.00709087803.$$

For df = 40 and $t = 2.6276$ we obtain

$$\alpha = \frac{0.010 - 0.005}{2.4233 - 2.7045} \left(2.6276 - 2.7045\right) + 0.005 = 0.00636735420.$$

Two down, one interpolation to go! For the final interpolation, we find the line segment in the line with df measured along the horizontal axis and α along the vertical axis. We interpolate between the points $(30, 0.00709087803)$ and $(40, 0.00636735420)$. The line is given by

$$\alpha = \frac{0.00636735420 - 0.00709087803}{40 - 30} \left(\mathrm{df} - 30\right) + 0.00709087803.$$

For df = 31 we obtain $\alpha = 0.007018525647$. Compare this value with the value 0.006626 obtained using software: the error is less than 0.0004. But it *was* quite a bit of work. On a multiple choice exam, make sure the answer choices require you to do the work before you do it. (In any event, the author does not make up exam questions that require interpolating both α and df. In this case, a little white lie—saying that the study involved 31 subjects instead of 32—would have eliminated two interpolations.)

The likelihood 0.006626 of the reported sample means being observed is less than a snowball's chance in hell. We reject the null hypothesis and conclude that cell phone usage while driving increases reaction time. So does old age.

14. The sample standard deviation is S = 21.37 and the sample variance is $\mathrm{S}^2 = 456.68$. Also, $\chi^2_{1-\alpha,5} = \chi^2_{0.90,5} = 1.6103$, $\sigma_0 = 30$, and $\sigma_0^2 = 900$. We calculate $\sigma_0^2 \times \chi^2_{0.90,5}/5 = 289.854$. Because 456.68 is not less than 289.854, we retain the null hypothesis.

15. We calculate $\mathrm{t}_{0.05,24} \sqrt{\frac{400}{25} + \frac{625}{25}} = 1.7109 \times 6.403 = 10.955$. Because the observed value of the test statistic, namely $\widehat{\mu}_{CM} - \widehat{\mu}_{TD} = -10$, is not less than the critical value, namely -10.955, our test statistic falls outside the ciritical region and we retain the null hypothesis.

16. The p-value is

$$\begin{aligned}
\mathrm{P}\left(\mathrm{S}^2 \geq (5.651)^2 \,\middle|\, \sigma = 4.199\right) &= \mathrm{P}\left(15\,\frac{\mathrm{S}^2}{(4.199)^2} \geq 15\frac{(5.651)^2}{(4.199)^2} \,\middle|\, \sigma = 4.199\right) \\
&= \mathrm{P}\left(15\,\frac{\mathrm{S}^2}{\sigma^2} \geq 27.1675 \,\middle|\, \sigma = 4.199\right) \\
&= \mathrm{P}\left(\chi^2_{15} \geq 27.1675\right) \\
&= 0.0274.
\end{aligned}$$

The value in the last line was obtained using software. An interpolation between the two tabulated values $\chi^2_{0.025,15} = 27.4884$ and $\chi^2_{0.050,15} = 24.9958$ leads to the line segment

$$p = \frac{(0.050 - 0.025)}{(24.9958 - 27.4884)} \left(x - 27.4884\right) + 0.025, \qquad \text{or} \qquad p = -0.01003\,x + 0.3007.$$

Substituting $x = 27.1675$ results in 0.02822 as the approximate p-value.

17. Assuming the null hypothesis, 0.19 is the fraction of Hispanics and 0.81 is the fraction of Others. The expected numbers of Hispanics and Others are, respectively, 0.19×720, or 136.8, and 0.81×720, or 583.2. We calculate

$$\frac{(125-136.8)^2}{136.8} + \frac{\big((720-125)-583.2\big)^2}{583.2} = 1.2566.$$

The p-value is

$$\mathrm{P}\left(\chi_1^2 \geq 1.2566\right) = 0.2623,$$

which is very close to the value obtained in Exercise 1 using a different test.

18. We calculate the Poisson probability density function $f(k) = \exp(u)\dfrac{u^k}{k!}$ for $u = 1$ and $k = 0, 1, 2, 3$. For $k = 4$, we will use $1 - \big(f(0)+f(1)+f(2)+f(3)\big)$ so that the probabilities sum to 1. In R, the command `c( dpois(0:3,1), 1 - sum(dpois(0:3,1)) )` is all that is needed. The result is 0.36787944, 0.36787944, 0.18393972, 0.06131324, 0.01898816. These represent the probability of one observation from the Poisson distribution being 0, 1, 2, 3, or 4, respectively. If we multiply these probabilities by 50, then we obtain the expected number of observations, 18.3939721, 18.3939721, 9.1969860, 3.0656620, 0.9494078, of 0, 1, 2, 3, 4 (or more), respectively. The test statistic is 1.504599 and the critical value is $\chi^2_{0.05,4}$, or 9.487729. The test statistic is less than the critical value so we retain the null hypothesis: R's simulation is a good fit.

The p-value is $\mathrm{P}\left(\chi_3^2 > 1.504599\right)$, or 0.6812093. This is a whopping probability. If the null hypothesis were true (i.e., R's simulation is a good one), then it would not at all be unlikely to observe a test statistic greater than or equal to 1.504599. We retain the null hypothesis.

19. The expected numbers are 50/8, 50/8, 50/8, 50/8, 50/8, 50/8, 50/8, 50/8. The observed value of the test statistic is 3.76. The critical value is $\chi^2_{0.5,7}$, or 6.345811. We retain the null hypothesis even though we have accepted a critical region so large that there is a 50-50 chance of rejecting a true null hypothesis.

20. The null hypothesis expected frquencies would be 50/4, or 12.5, for each orientation. The observed test statistic is therefore

$$\frac{(18-12.5)^2}{12.5} + \frac{(7-12.5)^2}{12.5} + \frac{(17-12.5)^2}{12.5} + \frac{(8-12.5)^2}{12.5}, \quad \text{or } 8.08.$$

The p-value is $\mathrm{P}\left(\chi_3^2 \geq 8.08\right)$, or 0.044387. This value can be approximated by interpolating using $\chi^2_{0.025,3} = 9.3484$ and $\chi^2_{0.050,3} = 7.8147$:

$$\text{p-value} = \frac{0.050-0.025}{7.8147-9.3484}(8.08-7.8147) + 0.050 = 0.04567549064.$$

The error that results from the interpolation is fairly small: about 0.001288. In any event, the p-value is small enough so that we may reject the hypothesis that the paintings were oriented entirely at random.

21. The introvert total is 60, the extrovert total is 140, the color preference totals are: red 100, yellow 20, green 36, and blue 44. The products of the row totals times the column totals divided by the table total results in the following expectations for the eight cells:

	Red	Yellow	Green	Blue
Introvert	30	6	10.8	13.2
Extrovert	70	14	25.2	30.8

The test statistic is $(16-30)^2/30 + (8-6)^2/6 + (12-10.8)^2/10.8 + \cdots + (20-30.8)^2/30.8$, or 23.09956710. The critical value is $\chi^2_{(2-1)\times(4-1),0.100}$, or 12.8382. Because the test statistic exceeds the critical value, we reject the null hypothesis: Personality type and color preference are *dependent*.

Chapter 11.

1. a) First, we obtain $t_{0.05/2,6-2} = 2.7764$. Next, we calculate $S_e = \sqrt{SSE/(n-2)} = \sqrt{22.4190/(6-2)} = 2.3674$ and $SE(b_1) = \frac{1}{\sqrt{n-1}} \times (S_e/S_X) = \frac{1}{\sqrt{6-1}} \times (2.3674/18.7083) = 0.05659$. The confidence interval for β_1 is $0.3771 \pm 2.7764 \times 0.05659$, or 0.3771 ± 0.1571, or $[0.2200, 0.5342]$.
   b) We continue by calculating

$$SE(\widehat{y}_*) = \sqrt{(0.05659)^2 \cdot (45-55)^2 + \frac{1}{6}(2.3674)^2} = 1.1200.$$

   The 95% confidence interval for the mean predicted yield at fertilizer level 45 is $31.5905 + 0.3771 \times 45 \pm 2.7764 \times 1.1200$, or 48.56 ± 3.1096 or $[45.4, 51.67]$.
   c) We continue by calculating

$$SE(\widehat{y}_*) = \sqrt{(0.05659)^2 \cdot (45-55)^2 + \frac{7}{6}(2.3674)^2} = 2.6190.$$

   The 95% confidence interval for Old MacDonald's grain crop yield at fertilizer level 45 is $31.5905 + 0.3771 \times 45 \pm 2.7764 \times 2.6190$, or 48.56 ± 7.2714 or $[41.29, 55.83]$.

2. The p-value is

$$\begin{aligned} P\left(t_{n-2} \geq \frac{b_1}{SE(b_1)}\right) &= P\left(t_4 \geq \frac{0.3771}{0.05659}\right) \\ &= P(t_4 \geq 6.6637) \\ &= 0.0013. \end{aligned}$$

   This very small p-value presents very strong evidence for rejecting the null hypothesis and concluding that $\beta_1 > 0$.

3. The correlation r is 0.6025357614. The observed value of the test statistic $r\sqrt{n-2}/\sqrt{1-r^2}$ is 2.135379.
   a) The relevant Student-t value, which is the critical value, is $t_{0.05,8}$, or 1.859548. (The R command is `qt(0.95,8)`.) Because the test statistic exceeds the critical value, the alternative hypothesis, $\rho > 0$, is accepted.
   b) The p-value is $P(t_8 \geq 2.135379)$, or 0.03262211. (The R command that obtains this value is `1-pt(2.135379,8)`.)

4. a) The values of b_0 and b_1 are 26.99033416 and 0.9025131179 respectively.
   b) The value of r^2 is 0.3630493438. The value of SST is $(n-1)\,S_Y^2$, or $9 \times (9.5008772)^2$, or 812.4. The value of SSR is r^2 SST, or $0.3630493438 \times 812.4$, or 294.9412869. The value of SSE is SST - SSR, or $812.4 - 294.9412869$, or 517.4587130. The value of S_e is $\sqrt{SSE/(n-2)}$, or $\sqrt{517.4587130/8}$, or 8.042533128. The value of $SE(b_1)$ is $(n-1)^{-1/2}\,S_e/S_X$, or $(1/3) \times 8.042533128/6.3429751$, or 0.4226477849.
   c) We look up $t_{0.05,8}$ and find that it is 1.859548038. The crtical value for the test statistic b_1 is $t_{0.05,8} \times SE(b_1)$, or $1.859548038 \times 0.4226477849$, or 0.7859338592. Because b_1 exceeds this value, the null hypothesis is rejected in favor of $\beta_1 > 0$.
   d) The p-value is

$$P\left(t_8 \geq \frac{b_1}{SE(b_1)}\right) = P\left(t_8 \geq \frac{0.9025131179}{0.4226477849}\right) = P(t_8 \geq 2.135378796) = 0.0326221225.$$

   The conventional conclusion based on this p-value and $\alpha = 0.05$ would be to reject the null hypothesis in favor of $\beta_1 > 0$.
   e) This p-value is double the one of part (d): about 0.0652. The conventional conclusion based on this

p-value and $\alpha = 0.05$ would be to retain the null hypothesis.
f) We look up $t_{0.005,8}$ and find 3.355387331. The 99% confidence interval is $b_1 \pm t_{0.005,8} \times \mathrm{SE}(b_1)$, or $0.9025131179 \pm 3.355387331 \times 0.4226477849$, or $0.9025131179 \pm 1.418147023$, or $[-.5156339051, 2.320660141]$. This is not very useful! Because 0 is included in the interval, we cannot be 99% sure that $\beta_1 > 0$. However, this revelation is not inconsistent with the results of parts (c) and (d), which were tests performed at significance level 0.05. With $\alpha = 0.05$, we find $t_{0.025,8} = 2.306004135$, and the confidence interval is $b_1 \pm 2.306004135 \times \mathrm{SE}(b_1)$, or [-0.0721144192, 1.877140655]. Notice that 0 is still in this interval. However, the presence of 0 in this 95% confidence interval is not inconsistent with parts (c) and (d). Those hypothesis test were one-sided and this confidence interval is not. On the other hand, the presence of 0 in the interval is entirely consistent with the conclusion of the two-sided test of part (e).
g) The appropriate SE for a specific individual is 14.08882298. The appropriate Student-t value is $t_{0.05,8} = 1.859548038$. The confidence interval is $26.99033416 + 5 \times 0.9025131179 \pm 1.859548038 \times 14.08882298$, or $[5.30405662, 57.70174288]$.
h) The appropriate SE for the average individual with a specified explanatory value is 6.169789128. The appropriate Student-t value is $t_{0.05,8} = 1.859548038$. The confidence interval is $26.99033416 + 45 \times 0.9025131179 \pm 1.859548038 \times 6.169789128$, or $[54.42858756, 80.77826138]$.

5. a) $b_1 = -361.6720588$, $b_0 = 734888.5118$
b) First,

$$\begin{aligned} \mathrm{SSE} &= \left(1 - r^2\right) \times \mathrm{SST} \\ &= \left(1 - r^2\right) \times (n-1)\,\mathrm{Sd}(\mathrm{Y})^2 \\ &= \left(1 - (-0.7658216844)^2\right) \times 15 \times (2248.439095)^2 \\ &= 31357904.9. \end{aligned}$$

Therefore,

$$\mathrm{S}_e = \sqrt{\frac{1}{n-2}\,\mathrm{SSE}} = \sqrt{\frac{1}{14}\,31357904.9} = 1496.612959$$

and

$$\mathrm{SE}(b_1) = \frac{1}{\sqrt{n-1}}\,\frac{\mathrm{S}_e}{\mathrm{S}_\mathrm{X}} = \frac{1}{\sqrt{15}}\,\frac{1496.612959}{4.760952286} = 81.16523361.$$

c) We look up $t_{0.05,14}$ and find that it is 1.761310136. The crtical value for the test statistic b_1 is $-t_{0.05,14} \times \mathrm{SE}(b_1)$, or $-1.761310136 \times 81.16523361$, or -142.9571486. Because b_1 is less than this value, the null hypothesis is rejected in favor of $\beta_1 < 0$.
d) The p-value is

$$\mathrm{P}\left(t_{14} \le \frac{b_1}{\mathrm{SE}(b_1)}\right) = \mathrm{P}\left(t_{14} \le \frac{-361.6720588}{81.16523361}\right) = \mathrm{P}\left(t_{14} \le -4.455997263\right) = 0.00027.$$

The conventional conclusion based on this p-value and $\alpha = 0.05$ would be to reject the null hypothesis in favor of $\beta_1 < 0$.
e) This p-value is double the one of part (d): about 0.00054. The conventional conclusion based on this p-value and $\alpha = 0.05$ (or, indeed a much smaller value) would be to retain the null hypothesis.
f) We look up $t_{0.005,14}$ and find 2.976842734. The 99% confidence interval is $b_1 \pm t_{0.005,14} \times \mathrm{SE}(b_1)$, or $-361.6720588 \pm 2.976842734 \times 81.16523361$, or $-361.6720588 \pm 241.6161359$, or $[-603.2881947, -120.0559229]$. Because 0 is not included in the interval, we can be 99% sure that $\beta_1 > 0$.
g) The appropriate SE for a specific year is 1724.641849. The appropriate Student-t value is $t_{0.05,14} = 1.761310136$. The confidence interval is $734888.5118 - 361.6720588 \times 1986 \pm 1.761310136 \times 1724.641849$, or $[13570, 19646]$.

6. The least squares line is $y = 78.15450644 - 0.6480686695\,x$. Also $\text{SST} = (n-1)\text{Sd(Y)}^2 = 7 \times (6.75991335)^2 = 319.875$ and $\text{SSE} = \left(1 - r^2\right) \times \text{SST} = \left(1 - (-0.5173838448)^2\right) \times 319.875 = 234.2489270$. It follows that $\text{S}_e = \sqrt{\text{SSE}/(8-2)} = \sqrt{234.2489270/6} = 6.2483188$ and $\text{SE}\,(b_1) = \text{S}_e / \left(\text{Sd(X)} \times \sqrt{n-1}\right) = 6.2483188 / \left(5.396758286 \times \sqrt{7}\right) = 0.437603909$.

   a) We calculate

$$\text{p-value} = 2\,\text{P}\left(\text{t}_{n-2} > \frac{|\,r\,|\sqrt{n-2}}{\sqrt{1-r^2}}\right) = 2\,\text{P}\,(\text{t}_6 > 1.480948081) = 0.1891232477.$$

   This probability is quite large. We retain the null hypothesis: $r = 0$.
   b) We calculate $-\text{t}_{0.025,6} \times \text{SE}\,(b_1) = -2.446911851 \times 0.437603909$, or -1.070778191. Because $b_1 = -0.6480686695$ is greater than this value, that is, because the point estimate b_1 of the slope lies to the right of the negative critical value, we retain the null hypothesis: $\beta_1 = 0$.
   c) We look up $\text{t}_{0.025,6} = 2.446911851$ and calculate $\text{t}_{0.025,10} \times \text{SE}\,(b_1) = 2.446911851 \times 0.437603909 = 1.070778191$. The confidence interval is $b_1 \pm \text{t}_{0.025,10} \times \text{SE}\,(b_1)$, or $-0.6480686695 \pm 1.070778191$, or $[-1.718846860, 0.4227095215]$. We have retained the null hypothesis that $\beta_1 = 0$. Therefore we are not surprised that 0 is in this interval.
   d) We have $y_* = b_0 + b_1\,x_* = 32.78969958$. The confidence interval is

$$32.78969958 \pm \text{t}_{0.025,n-2} \times \sqrt{\text{SE}\,(b_1)^2 \cdot (x_* - \bar{x})^2 + \frac{1}{n}\text{S}_e^2},$$

   for $x_* = 70$. We look up $\text{t}_{0.025,6} = 2.446911851$ and calculate its factor, $\text{SE}\left(\widehat{\bar{y}}_*\right)$, to be 2.923279274. The product $2.446911851 \times 2.923279274$ is 7.153006699. The confidence interval is $[32.78969958 - 7.153006699, 32.78969958 + 7.153006699]$, or $[25.63669288, 39.94270628]$.