r 1
Data and Distributions

Section 1.2

7.

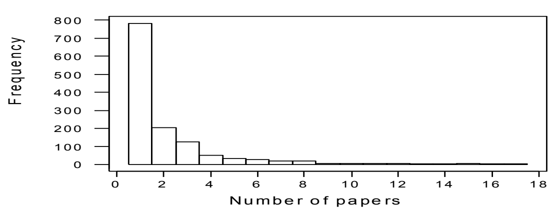
(a)

Number Nonconforming	Frequency	RelativeFrequency(Freq/60)
0	7	0.117
1	12	0.200
2	13	0.217
3	14	0.233
4	6	0.100
5	3	0.050
6	3	0.050
7	1	0.017
8	1	0.017
doesn't add exactly to 1 because relative frequencies have been rounded →		1.001

(b) The number of batches with at most 5 nonconforming items is $7+12+13+14+6+3=55$, which is a proportion of 55/60 = .917. The proportion of batches with (strictly) fewer than 5 nonconforming items is 52/60 = .867. Notice that these proportions could also have been computed by using the relative frequencies: e.g., proportion of batches with 5 or fewer nonconforming items = $1-(.05+.017+.107)=.916$; proportion of batches with fewer than 5 nonconforming items = $1-(.05+.05+.017+.107)=.866$.

8.

(a) The following histogram was constructed using MINITAB:



The most interesting feature of the histogram is the heavy positive skewness of the data.

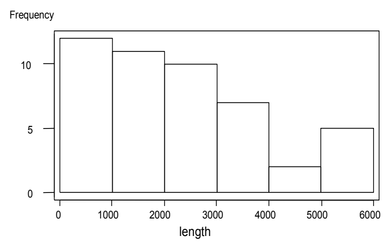
Note: One way to have MINITAB automatically construct a histogram from grouped data such as this is to use MINITAB's ability to enter multiple copies of the same number by typing, for example, 784(1) to enter 784 copies of the number 1. The frequency data in this exercise was entered using the following MINITAB commands:

```
MTB > set c1
DATA> 784(1) 204(2) 127(3) 50(4) 33(5) 28(6) 19(7) 19(8)
DATA> 6(9) 7(10) 6(11) 7(12) 4(13) 4(14) 5(15) 3(16) 3(17)
DATA> end
```

(b) From the frequency distribution (or from the histogram), the number of authors who published at least 5 papers is 33+28+19+…+5+3+3 = 144, so the proportion who published 5 or more papers is 144/1309 = .11, or 11%. Similarly, by adding frequencies and dividing by n = 1309, the proportion who published 10 or more papers is 39/1309 = .0298, or about 3%. The proportion who published more than 10 papers (i.e., 11 or more) is 32/1309 = .0245, or about 2.5%.

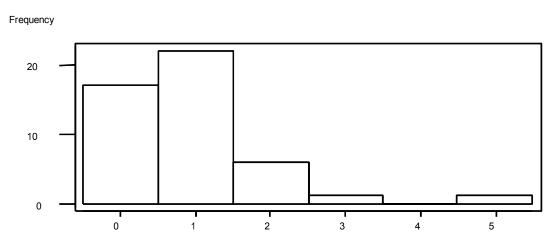
11.

(b) A histogram of this data, using classes of width 1000 separated at 0, 1000, 2000, and 6000 is shown below. The proportion of subdivisions with total length less than 2000 is (12+11)/47 = .489, or 48.9%. Between 2000 and 4000, the proportion is (10 + 7)/47 = .362, or 36.21%. The histogram shows the same general shape as depicted by the stem-and-leaf display in part (a).

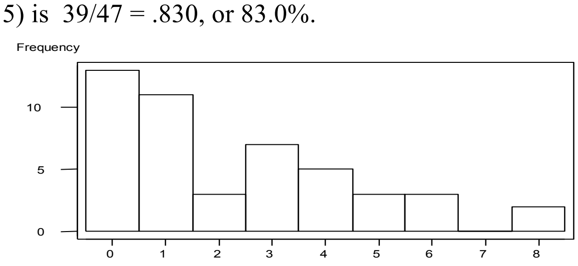


12.

(a) A histogram of the y data appears below. From this histogram, the number of subdivisions having no cul-de-sacs (i.e., y = 0) is 17/47 = .362, or 36.2%. The proportion having at least one cul-de-sac (y ≥ 1) is (47-17)/47 = 30/47 = .638, or 63.8%. Note that subtracting the number of cul-de-sacs with y = 0 from the total, 47, is an easy way to find the number of subdivisions with y ≥ 1.



(b) A histogram of the z data appears below. From this histogram, the number of subdivisions with at most 5 intersections (i.e., z ≤ 5) is 42/47 = .894, or 89.4%. The proportion having fewer than 5 intersections (z < 5) is 39/47 = .830, or 83.0%.



13.

(a) The frequency distribution is:

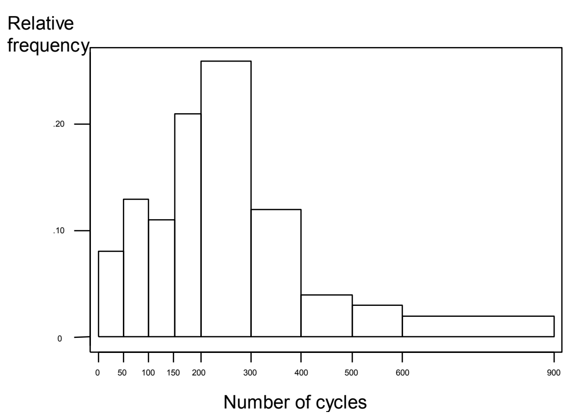
Class	Relative Frequency	Class	Relative Frequency
0–< 150	.193	900–<1050	.019
150–< 300	.183	1050–<1200	.029
300–< 450	.251	1200–<1350	.005
450–< 600	.148	1350–<1500	.004
600–< 750	.097	1500–<1650	.001
750–< 900	.066	1650–<1800	.002
		1800–<1950	.002

The relative frequency distribution is almost unimodal and exhibits a large positive skew. The typical middle value is somewhere between 400 and 450, although the skewness makes it difficult to pinpoint more exactly than this.

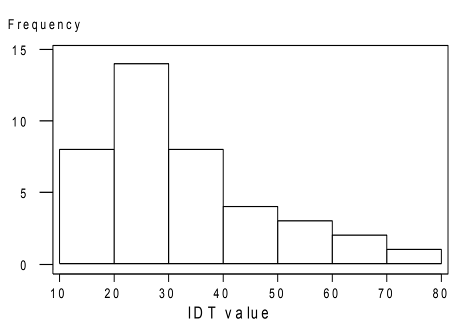
(b) The proportion of the fire loads less than 600 is .193+.183+.251+.148 = .775. The proportion of loads that are at least 1200 is .005+.004+.001+.002+.002 = .014.

15.

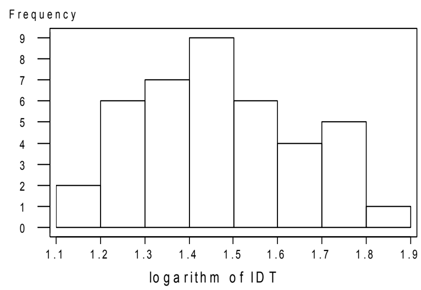
(b) Using class widths of different sizes, as specified in the exercise, the relative frequency histogram would appear as follows:



16. A histogram of the raw data appears below.



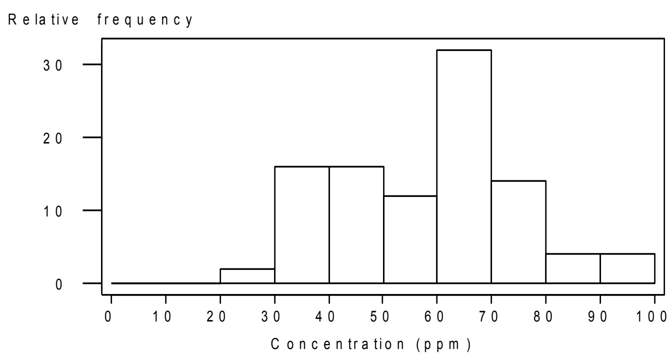
After transforming the data by taking logarithms (base 10), a histogram of the $\log_{10}$ data is shown below. The shape of this histogram is much less skewed than the histogram of the original data.



18.

(a) The classes overlap. For example, the classes 20-30 and 30-40 both contain the number 30, which happens to coincide with one of the data values, so it would not be clear which class to put this observation in.

(b) The histogram appears below:

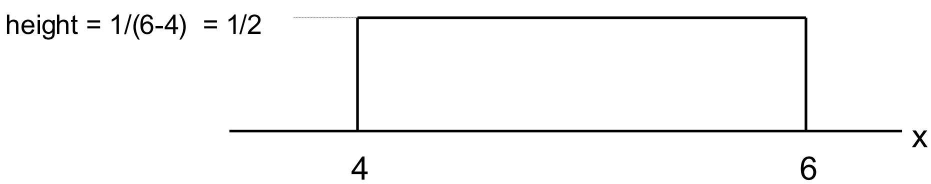


(c) The proportion of the concentration values that are less than 50 is 17/50 = .34, or 34%. The proportion of concentrations that are at least 60 is 27/50 = .54, or 54%.

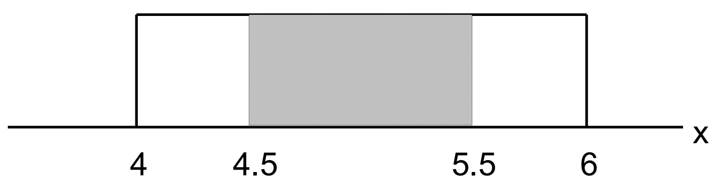
Section 1.3

19.

(a) The density curve forms a rectangle over the interval [4, 6]. For this reason, uniform densities are also called **rectangular densities** by some authors. Areas under uniform densities are easy to find (i.e., no calculus is needed) since they are just areas of rectangles. For example, the total area under this density curve is $\frac{1}{2}(6-4)$ = 1.



(b) The proportion of values between 4.5 and 5.5 is depicted (shaded) in the diagram below. The area of this rectangle is $\frac{1}{2}(5.5-4.5)$= .5. Similarly, the proportion of x values that exceed 4.5 would be $\frac{1}{2}(6-4.5)$= .75.



(c) The median of this distribution is 5 because exactly half the area under this density sits over the interval [4,5].

(d) Since 'good' processing times are short ones, we need to find the particular value x_0 for which the proportion of the data less than x_0 equals .10. That is, the area under the density to the left of x_0 must equal .10. Therefore, the area $=.10 = \frac{1}{2}(x_0-4)$, and so $x_0 - 4 = .20$. Thus, $x_0 = 4.20$.

20.

(a) The density function is $f(x) = 1/[5-(-5)] = 1/10$ over the interval [-5, 5] and $f(x) = 0$ elsewhere.

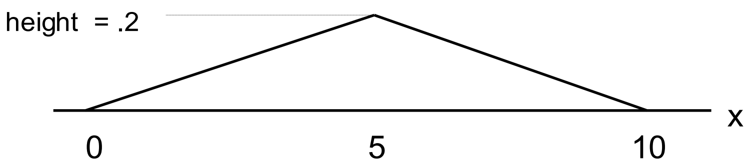
The proportion of x values that are negative is exactly .5 since the value $x = 0$ sits precisely in the middle of the interval [-5, 5].

(b) The proportion of values between -2 and 2 is $\frac{1}{10}[2-(-2)]$ = .4. The proportion of the x values falling between -2 and 3 is $\frac{1}{10}[3-(-2)]$ = .5.

(c) The proportion of the x values that lie between k and $k+4$ is $\frac{1}{10}[(k+4)-k]$ = .4.

22.

(a) The density curve forms an isosceles triangle over the interval [0, 10]. For this reason, such densities are often called **triangular densities**. The total area under this density curve is simply the area of the triangle, which is $\frac{1}{2}$ (base)(height) = $\frac{1}{2}$ (10)(.2) = 1. The height of the triangle is the value of f(x) at x = 5; i.e., f(5) = .4 -.04(5) = .2.



(b) Proportion $(x \le 3) = \frac{1}{2}(3-0)f(3) = \frac{1}{2}(3)(.12) = .18.$

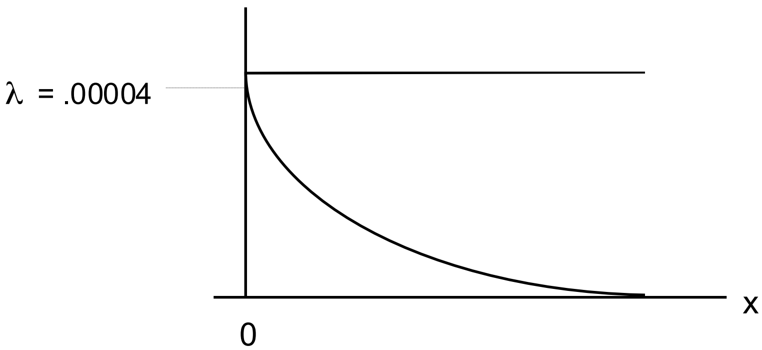
Proportion $(x \ge 7) = \frac{1}{2}(10-7)f(7) = \frac{1}{2}(3)(.12) = .18.$

Proportion $(x \ge 4) = 1 -$ Proportion $(x < 4) = 1 - \frac{1}{2}(4-0)f(4) = 1 - \frac{1}{2}(4)(.16) = .68.$

Proportion $(4 < x < 7) = 1 - [\text{Proportion } (x \le 4) + \text{Proportion } (x \ge 7)] = 1 - [.32+.18] = .50.$

23.

(a)



(b) $$\int_{20,0000}^{\infty} .00004e^{-.00004x}dx = \left[\frac{-(.00004)e^{-.00004x}}{(.00004)}\right]_{20,000}^{\infty} = -e^{-.00004x}\Big|_{20,000}^{\infty}$$
$= 0 - (-e^{-.00004(20,000)}) = e^{-.8} = .449.$

Note: for any exponential density curve, the area to the right of some fixed constant always equals $e^{-\lambda c}$, as our integration above shows. That is,

Proportion $(x > c) = \int_c^{\infty} \lambda e^{-\lambda x}dx = e^{-\lambda c}$.

We will use this fact in the remainder of the chapter instead of repeating the same type of integration as in part (a)

Proportion $(x \le 30,000) = 1 -$ Proportion $(x > 30,000) = 1 - e^{-\lambda c} =$
$1 - e^{-.00004(30,000)} = 1 - e^{-1.2} = .699.$

Proportion $(20,000 \le x \le 30,000) =$ Proportion $(x > 30,000) -$ Proportion$(x \le 20,000) = .699 - (1-.449)$
$= .148.$

(c) For the best 1%, the lifetimes must be at least x_0, where Proportion$(x \ge x_0) = .01$, which becomes $e^{-\lambda x_0} = .01$. Taking natural logarithms of both sides, $-\lambda x_0 = \ln(.01)$, so $x_0 = -\ln(.01)/\lambda =$ $4.60517/.00004 = 115,129.25$. For the worst 1%, we have Proportion$(x \le x_0) = .01$, which is equivalent to saying that Proportion$(x \ge x_0) = .99$, so $e^{-\lambda x_0} = .99$. Taking logarithms, $-\lambda x_0 = \ln(.99)$, so $x_0 = -\ln(.99)/\lambda = 251.26$.

26.

(a) (ii) qualifies as a distribution because the probabilities add exactly to 1. On the other hand, the probabilities in (i) add to .7, which disqualifies (i). In (iii), notice that the probability associated with x = 4 is negative, which is impossible for a valid distribution.

(b) Using (ii), the proportion of cars having at most 2 (i.e., 1 or less) under-inflated tires is p(0)+p(1)+p(2) = .4 + .1 +.1 = .6. Similarly, Proportion($x \geq 1$) = 1- Proportion(x =0) = 1-.4 = .6.

27.

(a) Proportion($x \leq 3$) = .10 + .15 + .20 + .25 = .70.
Proportion($x < 3$) = Proportion($x \leq 2$) = .10 + .15 + .20 = .45.

(b) Proportion($x \geq 5$) = 1 - Proportion($x < 5$) = 1 - (.10+.15+.20+.25+.20) = .10.

(c) Proportion($2 \leq x \leq 4$) = .20 + .25 + .20 = .65

(d) At least 4 lines will *not* be in use whenever 2 or fewer lines *are* in use. At most 2 lines are in use .45, or 45%, of the time from part (a) of this exercise.

28.

The sum of all the proportions must equal 1, so $1 = \sum_{y=1}^{5} cy$ = c[1+2+3+4+5] =15c and c=1/15.

Proportion($y \leq 3$) = p(1) + p(2) + p(3) = 1/15 +2/15 +3/15 = 6/15 = .4.

Proportion($2 \leq y \leq 4$) = 2/15 +3/15 +4/15 = 9/15 = .60.

Section 1.4

30.

(a) Proportion($z \leq 2.15$) = .9842 (Table I). Proportion($z < 2.15$) will also equal .9842 because the z distribution is continuous.

(b) Using the symmetry of the z density, Proportion($z > 1.50$) = Proportion($z < -1.50$) = .0668.
Proportion($z > -2.00$) = 1- Proportion($z \leq -2.00$) = 1 - .0228 = .9772.

(c) Proportion($-1.23 \leq z \leq 2.85$) = Proportion($z \leq 2.85$) - Proportion($z \leq -1.23$) = .9978 - .1093 =.8885.

(d) In Table I, z values range from -3.8 to +3.8. For $z < -3.8$, the left tail areas (i.e., table values) are .0000 (to 4 decimal places); for $z > +3.8$, left tail areas equal 1.0000 (to 4 places). Therefore, Proportion($z > 5$) = 1 - Proportion($z \leq 5$) - 1 - 1.0000 = .0000. Similarly, Proportion($z > -5$) = 1 - Proportion($z \leq -5$) = 1 - .0000 = 1.0000.

(e) Proportion($z < |2.50|$) = Proportion($-2.50 < z < 2.50$)
= Proportion($z < 2.50$) - Proportion($z < -2.50$) = .9938 - .0062 = .9876.

31.

Proportion($z \leq 1.78$) = .9625 (Table I)

(b) Proportion($z > .55$) = 1 - Proportion($z \leq .55$) = 1 - .7088 = .2912.

(c) Proportion($z > -.80$) = 1 - Proportion($z \leq -.80$) = 1 - .2119 = .7881.

(d) Proportion($.21 \le z \le 1.21$) = Proportion($z \le 1.21$) - Proportion($z \le .21$)
= .8869 - .5832 = .3037.

(e) Proportion($z \le -2.00$ or $z \ge 2.00$) = Proportion ($z \le -2.00$) + [1- Proportion($z < 2.00$)]
= .0228 + [1 - .9772] = .0456. Alternatively, using the fact that the z density is symmetric around z = 0, Proportion($z \le -2.00$) = Proportion($z \ge 2.00$), so the answer is simply
2 Proportion($z \le -2.00$) = 2(.0228) = .0456.

(f) Proportion($z \le -4.2$) = .0000

(g) Proportion($z > 4.33$) = .0000

32.

(a) Proportion($z \le z^*$) = .9082 when $z^* = 1.33$ (Table I).

(b) Proportion($z \le 1.33$) = .9082 and Proportion($z \le 1.32$) = .9066; the z value for 0.9082 is just under 1.33, and it is sufficient to approximate it with 1.329.

(c) Proportion($z > z^*$) = .1210 is a right-tail area. Converting to a left-tail area (so that we can use table I), Proportion($z \le z^*$) = 1 - .1210 = .8790. From Table I, $z^* = 1.17$ has a left tail area of .8790 and, therefore, has a right-tail area of .1210.

(b) Proportion($-z^* \le z \le z^*$) = .754. Because the z density is symmetric the two tail areas associated with $z < -z^*$ and $z > z^*$ must be equal and they also account for all the remaining area under the z density curve. That is, 2×Proportion($z \le -z^*$) = 1 - .754 = .246, which means that Proportion($z \le -z^*$) = .1230. From Table I, Proportion($z \le -1.16$) = .1230, so $-z^* = -1.16$ and $z^* = 1.16$.

(c) Proportion($z > z^*$) = .002 is equivalent to saying that Proportion($z \le z^*$) = .998. From Table I, Proportion($z \le 2.88$) = .9980, so $z^* = 2.88$. Similarly, you would have to go a distance of -2.88 (i.e., 2.88 units to the left of 0) to capture a left-tail area of .002.

34.)

(a) Let z^* denote the 91st percentile, so Proportion($z \le z^*$) = .9100. From Table I, Proportion($z \le 1.34$) = .9099 and Proportion($z \le 1.35$) = .9115 and, so the z^* is just over 1.34, and it is sufficient to approximate it with 1.341.

(b) Let z^* denote the 9th percentile, Proportion($z \le z^*$) = .0900. From Table I, $z^* \approx -1.34$. Note that the 9th percentile should be the same distance to the *left* of 0 that the 91st percentile is to the right of 0, so we could have simply used the answer to part (a), after attaching a minus sign.

(c) The 22nd percentile occurs at $z \approx -.77$.

36.

(a) Let x = current flow. Then, flows for which $18.5 \le x \le 22$, the upper value (x = 22) is 2 σ's above μ; the lower value (x = 18.5) is 1.5 σ's below μ. The corresponding z curve region is then $1.5 \le z \le 2$. Using Table I, Proportion($-1.5 \le z \le 2$) = Proportion($z \le 2$) - Proportion($z \le -1.5$) = .9772 - .0668 = .9104.

(b) The value x = 15 describes current flows that at 5 σ's below the mean. Since virtually all the area (to 4 decimal places anyway) under a normal curve lies within ±3.8 σ's (see Table I) from the mean, we can say that Proportion($z > 5$) should be approximately 1.0000.

(c) We want to find the current flow x* that satisfies Proportion($x \geq x^*$) = .05. The equivalent statement about a z curve is Proportion($z \geq z^*$) = .05, or, in terms of left-tail areas, Proportion($z \leq z^*$) = .95. From Table I, $z^* \approx 1.645$, which is a value that is 1.645 σ's above the mean of the z distribution. Therefore, x* must lie 1.645 σ's above the mean of the x data: x* = 20 + 1.645(1) = 21.645.

37.

(a) Let x = yield strength. Then, $x \leq 40$, when standardized, becomes $z \leq (40-43)/4.5 = -.67$. From Table I, Proportion($z \leq -.67$) = .2514. Similarly, standardizing $x \geq 60$ yields $z \geq (60-43)/4.5 = 3.78$. From Table I, Proportion($z > 3.78$) = 1 - Proportion($z \leq 3.78$) = 1- .9999 = .0001.

(b) Standardizing $40 \leq x \leq 50$ gives $(40-43)/4.5 \leq z \leq (50-43)/4.5$ or, $-.67 \leq z \leq 1.56$. From Table I, this proportion equals Proportion($z \leq 1.56$) - Proportion($z \leq -.67$) = .9406 - .2514 = .6892. **[The answer at the back of the text is still incorrect.]**

(c) We need to find the value of x* for which Proportion($x > x^*$) = .75, or equivalently, Proportion($x \leq x^*$) = .25. The corresponding statement about a z curve is Proportion($z \leq z^*$) = .25. From Table I, $z^* \approx -.675$; i.e., z^* is .675 σ's below the mean. So, x* must be .675 σ's below the mean of the x data: x* = 43 - .675(4.5) = 39.963. **[The answer at the back of the text is still incorrect.]**

38.

(a) When standardized, $x > 100$ becomes $z > (100-96)/14$ or, $z > .29$. From Table I, Proportion($z > .29$) = 1 - Proportion($z \leq .29$) = 1 - .6141 = .3859.

(b) When standardized, $x \geq 75$ becomes $z \geq (75-96)/14$ or, $z > -1.5$. From Table I, Proportion($z \geq -1.5$) = 1 - Proportion($z < -1.5$) = 1 - .0668 = .9332.

(c) When standardized, $50 < x < 75$ becomes $(50-96)/14 < z < (75-96)/14$ or, $-3.29 < z < -1.5$. From Table I, Proportion($-3.29 < z < -1.5$) = Proportion($z < -1.5$) - Proportion($z < -3.29$) = .0668 - .0005 = .0663.

(d) Proportion($x \leq a$) = .05, when standardized, becomes Proportion($z \leq \frac{a-96}{14}$) = .05. From Table I, 5% of the area under the z curve lies to the left of z = -1.645. Therefore, $\frac{a-96}{14} = -1.645$, so a = 96 -1.645(14) = 72.97. For the right-tail area, $\frac{b-96}{14} = +1.645$, so b = 96+1.645(14) = 119.03.

40.

(d) Proportion($8.8-c < x < 8.8+c$) = .98; standardizing gives Proportion($(8.8-c-8.8)/2.8 < z < (8.8+c-8.8)/2.8$) = Proportion($-c/2.8 < z < c/2.8$) = .98. This means that the right and left tail areas Proportion($z < -c/2.8$) and Proportion($z > c/2.8$) both equal .01. From Table I, z = -2.33 has (approximately) a left tail area of .01, so -c/2.8 = -2.33, or, c = 6.524.

Section 1.6

53.

(a) Let x = number of red lights encountered. Then x has a binomial distribution with n = 10, $\pi = .40$. Proportion($x \leq 2$) = .006 + .040 + .121 = .167 (using Table II). Similarly, Proportion($x \geq 5$) = .201 + .111 +.042 + .011 + .002 + .000 = .367.

(b) Proportion($2 \le x \le 5$) = .121 + .215 + .251 + .201 = .788.

54.

(a) Let x = number of bits erroneously transmitted. Then, x is binomial with n = 20, π = .10, so Proportion($x \le 2$) = .122 + .270 + .285 = .677 (from Table II).

(b) Proportion($x \ge 5$) = .032 + .009 + .002 + .000 + ….+ .000 = .043.

(c) 'More than half' means 11 or more, so Proportion($x \ge 11$) = .000 + … + .000 = .000.

56.

(c) Proportion($10 \le x \le 20$) = .006 + .011 + … + .089 + .089 = .556.
Proportion($10 < x < 20$) = .011 +… + .089 = .461.

57.

(a) Proportion(x = 1) = .164 (Table III, λ = .2)

(b) Proportion($x \ge 2$) = 1 - Proportion($x \le 1$) = 1 - (.819 + .164) = 1 - .983 = .017.

(c) Let y denote the number of missing pulses on 10 disks. Then, Proportion($y \ge 2$) = 1 - Proportion($y \le 1$) = 1 - (.135 + .271) = .594 (Table III, with λ = 2).

58.

Using Table III (λ = 20), Proportion($x \ge 15$) = 1 - Proportion($x \le 14$) = 1 - (.000 + .000+ … + .001+ .001 + .003 + .006 + .011 + .018 + .-27 + .039) = 1 - .106 = .894. Similarly, Proportion($x \le 25$) = 1 - Proportion($x \ge 26$) = 1 - (.034 + .025 + .018 + .013 + .008) = 1 - .098 = .902.

59.

Although x has a binomial distribution with n = 1000 and π = 1/200 = .005, its distribution can be approximated by a Poisson distribution with $\lambda = n\pi$ = 1000/200 = 5. Therefore, using Table III, Proportion($x \ge 8$) = .065 + .036 + .018 + .008 + .003 + .001 = .131. Note: because the Table entries are rounded to 3 places, you would get a slightly different answer (of .135) if you worked the problem by first adding the proportions for $x \le 7$, then subtracting from 1.

Proportion($5 \le x \le 10$) = .175 + .146 + .104 + .065 + .036 + .018 = .544.

Supplementary Exercises

61.

(a) Let x = fracture strength. Then, Proportion(x < 90) = .50 because 90 is the mean of the (assumed) normal distribution of x. Table I must be used for the other proportions: Proportion(x < 95) = Proportion(z < (95-90)/3.75) = Proportion(z < 1.33) = .9082; therefore, Proportion(x ≥ 95) = 1 - Proportion(x < 95) = 1 - .9082 = .0918.

(b) Proportion(85 ≤ x ≤ 95) = Proportion((85-90)/3.75 ≤ z ≤ (95-90)/3.75) = Proportion(-1.33 ≤ z ≤ 1.33) = Proportion(z ≤ 1.33) - Proportion(z < -1.33) = .9082 - .0918 = .8164. Likewise, Proportion(80 ≤ x ≤ 100) = Proportion(-2.67 ≤ z ≤ 2.67) = .9962 -.0038 = .9924.

(c) Let x* denote the value exceeded by 90% of the x data. From Table I, the corresponding value z* for the z distribution is z* = -1.28, which is 1.28 σ's below the mean of the z distribution. Therefore, x* must be 1.28 σ's below the mean of the x data: x* = 90 - 1.28(3.75) = 85.20.

(b) The corresponding interval for the z distribution is .99 = Proportion(-z* ≤ z ≤ z*), which means that the left-tail proportion Proportion(z ≤ z*) = .99 + .005 = .9950. From Table I, z* ≈ 2.58; i.e., 2.58 σ's from the mean. Therefore, the two values for the x data must lie 2.59 σ's on either side of the mean: 90 ± 2.58(3.75) = 80.325 and 99.675.

62.

(a) Use the formula for right-tail areas given in the answer to Exercise 23(b):

Proportion(x ≤ 100) = 1- Proportion(x ≥ 100) = 1 - $e^{-\lambda(100)}$ = 1 - $e^{-.01386(100)}$ = 1 - .25007 = .7499, or, .75.

Proportion(x ≤ 200) = 1 - $e^{-\lambda(200)}$ = 1 - $e^{-.01386(200)}$ = 1 - .06253 = .9375.

Proportion(100 ≤ x ≤ 200) = Proportion(x > 100) - Proportion(x > 200) = $e^{-.01386(100)}$ - $e^{-.01386(200)}$ = .25007 - .06253 = .1875.

(b) Proportion(x ≥ 50) = $e^{-.01386(50)}$ = .50007, or, *almost* exactly 50%.

(c) Let $\tilde{x}$ denote the median. Then, .50 = Proportion(x ≥ $\tilde{x}$) = $e^{-\lambda(\tilde{x})}$. Taking logarithms of both sides, ln(.50) = -λ$\tilde{x}$, so $\tilde{x}$ = -.6931472/-.01386 = 50.01. Note that you could have guessed from the answer to (b) that $\tilde{x}$ is very close to 50.

64.

(c) Proportion(5 ≤ x ≤ 10) = .196 + .163 + .111 + .062 + .030 + .011 = .573.

(d) Proportion(5< x < 10) = .163 + .111 + .062 + .030 = .366

65.

(a) Accommodating everyone who shows up means that x ≤ 100, so Proportion(x ≤ 100) =.05 + .10 + … + .24 + .17 = .82.

(b) Proportion(x > 100) = 1 - Proportion(x ≤ 100) = 1 - .82 = .18.

(c) The first standby passenger will be able to fly as long as the number who show up is x ≤ 99, which leaves one free seat. This proportion is .65. The third person on the standby list will get a seat as long as x ≤ 97, where Proportion(x ≤ 97) = .05 + .10 + .12 = .27.

70.

Let X = the number of components that function. X is binomial with n = 5, π = .9.
(Proportion of 3 out of 5 systems that will function.)

$$= \text{Proportion}(x \geq 3) = p(3) + p(4) + p(5) = \binom{5}{3}\pi^3(1-\pi)^2 + \binom{5}{4}\pi^4(1-\pi)^1$$

$$+\binom{5}{5}\pi^5(1-\pi)^0 = 10\pi^3(1-\pi)^2 + 5\pi^4(1-\pi) + \pi^5$$

$$= (10)(.9)^4(.1)^2 + (5)(.9)^4(.1) + (.9)^5 = .99144$$

Alternatively, use the Binomial Table II.

73.

Letting X = "bursting strength", we first find the proportion of all bottles having bursting strength exceeding 300 PSI. Proportion(X > 300) = Proportion(z > ((300-250)/30)) = Proportion(z > 1.67) = .0475 (from Table I). Then, Y = "the number of bottles in a carton of 12 with bursting strength over 300 PSI" is a binomial variable with n = 12 and π = .0475. So the proportion of all cartons with at least one bottle with a bursting strength over 300 PSI is Proportion(Y $\geq$ 1) = 1 – Proportion(Y = 0) = 1 – $(1\text{-p})^{1/2}$ = 1 – $(1 - .0475)^{12}$ = .4423

Chapter 2
Numerical Summary Measures

Section 2.1

4.

The three quantities of interest are: $\bar{x}_n, x_{n+1}$, and $\bar{x}_{n+1}$. Their relationship is as follows:

$$\bar{x}_{n+1} = \left(\frac{1}{n+1}\right)\sum_{i=1}^{n+1} x_i = \left(\frac{1}{n+1}\right)\left[\sum_{i=1}^{n} x_i + x_{n+1}\right] = \left(\frac{1}{n+1}\right)\left[n\bar{x}_n + x_{n+1}\right] = \left(\frac{n\bar{x}_n + x_{n+1}}{n+1}\right)$$

For the strength observations, $\bar{x}_{10} = 43.34$ and $x_{11} = 81.5$. Therefore, $\bar{x}_{11} = \left(\frac{(10)(43.34)+81.5}{10+1}\right) = 46.81$

6.

(a) The *reported* blood pressure values would have been: 120 125 140 130 115 120 110 130 135. When ordered these values are: 110 115 120 120 125 130 130 130 135 140. The median is 125.

(b) 127.6 would have been reported to be 130 instead of 125. Since this value is the middle value, the median would change from 125 to 130. This example illustrates that the median is sensitive to rounding in the data.

8.

$$\mu = \int_{-1}^{1} (x)(.75)(1-x^2)\,dx = \left[\frac{.75x^2}{2} - \frac{.75x^4}{4}\right]_{-1}^{1} = 0$$

The mean value of x equals 0.

9.

(a) $\mu = \int_0^2 x(.5x)\,dx = .5\left[\frac{x^3}{3}\right]_0^2 = 4/3$. The mean μ does not equal 1 because the density curve is not symmetric around x = 1.

(b) Half the area under the density curve to the left (or right) of the median $\tilde{\mu}$, so, $.50 = \int_0^{\tilde{\mu}} .5x\,dx = .5\left[\frac{x^2}{2}\right]_0^{\tilde{\mu}} = (\tilde{\mu}^2)/4$, or, $\tilde{\mu}^2 = 4(.5) = 2$ and $\tilde{\mu} = 1.414$. $\mu < \tilde{\mu}$ because the density curve is negatively skewed.

(c) $\mu \pm \frac{1}{2} = \frac{4}{3} \pm \frac{1}{2} = \frac{5}{6}$ and $\frac{11}{6}$. The area under the curve between these two values is $\int_{5/6}^{11/6} .5x\,dx$ =

$\left[\frac{x^2}{4}\right]_{5/6}^{11/6}$ = .667. Similarly, the proportion of the times that are within one-half hour of $\tilde{\mu}$ is:

$\int_{.914}^{1.914} .5x\,dx = \left[\frac{x^2}{4}\right]_{.914}^{1.914} = .707.$

10.

$$\mu = \int_a^b x\left(\frac{1}{b-a}\right)dx = \left[\frac{x^2}{2}\left(\frac{1}{b-a}\right)\right]_a^b = \left(\frac{b+a}{2}\right)$$

12.

$$\mu = \sum_{x=0}^{6} xp(x) = \begin{bmatrix}(0)(.10)+(1)(.15)+(2)(.20)+(3)(.25)+ \\ (4)(.20)+(5)(p(5))+(6)(p(6))\end{bmatrix} = 2.64$$

$\Rightarrow 2.1+5(p(5))+6(p(6))=2.64$
$\Rightarrow 5(p(5))+6(p(6))=.54$
Also, we know that $\sum p(x)=1$
$\Rightarrow p(5)+p(6)=1-[(.10)+(.15)+(.20)+(.25)+(.20)]$
$\Rightarrow p(5)+p(6)=1-.90=.10$
Therefore, p(5) = .10 – p(6)

Returning to our first equation:
$\Rightarrow 5(.10-p(6))+6(p(6))=.54$
$\Rightarrow .5-5(p(6)+6(p(6))=.54$
$\Rightarrow p(6)=.04$
Therefore, p(5) = .1 - .04 = .06

Finally, p(5) = .06 and p(6) = .04

13.

$$\mu = \sum_{x=0}^{4} x \cdot p(x) = 0(.4) + 1(.1) + 2(.1) + 3(.1) + 4(.3) = 1.8$$

Section 2.2

15.

(a) $\bar{x} = \frac{1}{n}\sum_i x_i$ = 577.9/5 = 115.58. Deviations from the mean:

116.4 - 115.58 = .82, 115.9 - 115.58 = .32, 114.6 -115.58 = -.98,
115.2 - 115.58 = -.38, and 115.8-115.58 = .22.

(b) $s^2 = [(.82)^2 + (.32)^2 + (-.98)^2 + (-.38)^2 + (.22)^2]/(5-1) = 1.928/4 = .482$, so $s = .694$.

(c) $\sum_i x_i^2 = 66{,}795.61$, so $s^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - \frac{1}{n}\left(\sum_i x_i\right)^2\right] =$

$[66{,}795.61 - (577.9)^2/5]/4 = 1.928/4 = .482$.

16.

(a) The transformed data are: 16.4 15.9 14.6 15.2 15.8.

Relevant quantities are: $\sum_{i=1}^{n} x_i^2 = 1{,}215.61 \quad \sum_{i=1}^{n} x_i = 77.9$

$$\text{So, } s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{1{,}215.61 - \frac{(77.9)^2}{5}}{4} \quad \text{So, } s^2 = .482$$

Notice that s^2 for the transformed data equals s^2 for the original data.

(b) $\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{n}\sum(x_i - c) = \frac{1}{n}\sum x_i - \frac{1}{n}(nc) = \bar{x} - c$

Since $\bar{y} = \bar{x} - c$ we know that $s_y^2 = \frac{\sum(y_i - \bar{y})^2}{n-1} = \frac{\sum((x_i - c) - (\bar{x} - c))^2}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1} = s_x^2$

18.

The sample mean, $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{10}(1{,}162) = 116.2$.

The sample standard deviation, $s = \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{140{,}992 - \frac{(1{,}162)^2}{10}}{9}} = 25.75$

On average, we would expect a fracture strength of 116.2. In general, the size of a typical deviation from the sample mean (116.2) is about 25.75. Some observations may deviate from 116.2 by more than this and some by less.

22.

Let X = the number of drivers who travel between a particular origin and destination during a designated time period. X has a Poisson distribution with $\lambda = 20$.

(a) $\mu_x = \lambda = 20$

Find $P(\mu - 5 \le x \le \mu + 5) = P(15 \le x \le 25)$

Using Table III with $\lambda = 20$, we obtain:

$P(15 \le x \le 25) = .052 + .065 + .076 + .084 + .089 + .089 + .085 + .077 +$

$.067 + .056 + + .045 = .785$

(b) $\sigma_x = \sqrt{\lambda} = \sqrt{20} = 4.47$

Find $P(\mu - \sigma \le x \le \mu + \sigma)$

$P(20-4.47 \le x \le 20+4.47) = P(15.53 \le x \le 24.47)$

But, since X is an integer–valued random variable, only the integers between 16 and 24 satisfy this requirement. So, we find $P(16 \le x \le 24)$ using Table III with λ = 20, we obtain:

$P(16 \le x \le 24) = .065 + .076 + .084 + .089 + + .089 + .085 + .077 + .067 + .056$

$= .688$

23.

(a) $\sigma = \sqrt{\lambda} = \sqrt{5} = 2.236$, so x values that lie between 5 - 2.236 = 2.764 and 5 + 2.236 = 7.236 are within one standard deviation from the mean. Because x is integer-valued, only the integers between 3 and 7 satisfy this requirement, i.e., Proportion(2.764 ≤ x ≤ 7.236) = Proportion(3 ≤ x ≤ 7) = p(3) + p(4) + p(5) +p(6) + p(7) = .140 + .175 + .175 + .146 + .104 = .740 (using Table III with λ = 5).

(b) To exceed the mean by *more* than 2 standard deviations, x values must be greater than 5 + 2(2.236) = 9.472. The integer values of x than satisfy this requirement are x = 10 and greater. From Table III, Proportion(x > 9.472) = Proportion(x ≥ 10) = .018 +.008 + .003 +.001 = .030. Note: because table entries are rounded to 3 places, a slightly different answer results if you calculate the proportion as 1 - Proportion(x ≤ 9) = 1 - .966 = .032.

24.

(a) $\sigma = \sqrt{\sum (x-\mu)^2 p(x)}$

$$\sigma = \sqrt{\begin{gathered}(0-1.8)^2(.4) + (1-1.8)^2(.1) + (2-1.8)^2(.1) + \\ (3-1.8)^2(.1) + (4-1.8)^2(.3)\end{gathered}}$$

$\sigma = \sqrt{2.96} = 1.72$

(b) $P(\mu - \sigma \le x \le \mu + \sigma) = P(.08 \le x \le 3.52) = P(1 \le x \le 3) = .3$

$P(x \rangle \mu + 3\sigma) + P(x \langle \mu - 3\sigma) = P(x \rangle 6.96) + P(x \langle -3.36) = 0 + 0 = 0$

25.

$\sigma^2 = \sum (x-\mu)^2 p(x) = \sum (x^2 - 2\mu x + \mu^2) p(x) =$

$\sum x^2 p(x) - 2\mu \sum x p(x) + \mu^2 \sum p(x) = \sum x^2 p(x) - 2\mu^2 + \mu^2 = \sum x^2 p(x) - \mu^2$.

For the mass function given in Exercise 24, $\sigma^2 = \sum x^2 p(x) - \mu^2 = (0)^2(.4) + (1)^2(.1) + (2)^2(.1) + (3)^2(.1) + (4)^2(.3) - (1.8)^2 = 6.2 - 3.24 = 2.96$.

26.

Extending the result from exercise 25, we know:

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \left[\int_{-\infty}^{\infty} x^2 f(x)dx\right] - \mu^2$$

So, in the case of a uniform from a to b,

$$\sigma^2 = \left[\int_a^b x^2 \left(\frac{1}{b-a}\right) dx\right] - \mu^2 = \left(\frac{1}{b-a}\right)\left[\frac{x^3}{3}\right]_a^b - \mu^2$$

$$= \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) - \left(\frac{a+b}{2}\right)^2 = \frac{1}{b-a}\left(\frac{1}{3}\right)(b^3 - a^3) - \left(\frac{1}{4}\right)(a+b)^2$$

$$=\frac{1}{3}\left(a^2+ab+b^2\right)-\frac{1}{4}\left(a+b\right)^2=\frac{1}{12}\left(b-a\right)^2$$

For the task completion time application, a = 4 and b = 6.

$$\mu=\left(\frac{a+b}{2}\right)=\left(\frac{4+6}{2}\right)=5 \qquad \sigma^2=\frac{\left(b-a\right)^2}{12}=\frac{\left(6-4\right)^2}{12}=\frac{1}{3}$$

$$P\left(x>\mu+\sigma\right)+P\left(x<\mu-\sigma\right)=P\left(x>5+\sqrt{1/3}\right)+P\left(x<5-\sqrt{1/3}\right)$$

$$=P\left(x>5.577\right)+P\left(x<4.423\right)=\frac{6-5.577}{2}+\frac{4.423-4}{2}=.2115+2115=.423$$

27.

The proportion of x values between μ - 1.5σ and μ + 1.5σ is the same as the proportion of z values between -1.5 and +1.5: Proportion(-1.5 ≤ z ≤ 1.5) = Proportion(z ≤ 1.5) - Proportion(z ≤ -1.5) = .9332-.0668 = .8664. The proportion of x value that exceed μ by more than 2.5 σ's equals Proportion(z > 2.5) = 1 - Proportion(z ≤ 2.5) = 1 - .9938 = .0062.

28.

X is binomial with $\pi=.2$ and $n=25$.

$$\sigma^2=n\pi\left(1-\pi\right)=\left(25\right)\left(.20\right)\left(.80\right)=4 \quad \sigma=2$$

Also, $\mu=n\pi=\left(25\right)\left(.20\right)=5$

$$P\left(x>\mu+2\sigma\right)=P\left(x>5+2\left(2\right)\right)=P\left(x>9\right)$$

Using Table II:

$$P\left(x>9\right)=P\left(x\geq 10\right)=.011+.004+.002=.017$$

(Notice that $P\left(x\geq 13\right)\approx 0$)

Section 2.3

34.

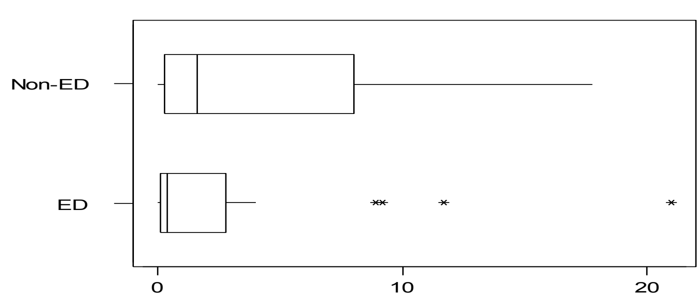
Since 5% of all lengths exceed 3.75 mm, then 3.75 is the 95th percentile of the distribution. Because the circuits are normally distributed, 3.75 is 1.645 standard deviations above the mean; that is, 3.75 = $\mu+1.645\sigma$. Furthermore, 3.85 is the 99th percentile of the distribution, and so it is 2.33 standard deviations above the mean: 3.85 = $\mu+2.33\sigma$. We then have two equations and two unknowns:

$$\mu + 1.645\sigma = 3.75$$
$$\mu + 2.33\sigma = 3.85$$

Subtracting the top equation from the bottom equation yields $.685\sigma = .10$, and so $\sigma = .10/.685 = .146$. Then substituting $\sigma = .146$ into either of the equations gives $\mu = 3.51$.

39.

(c) A comparative boxplot appears below. The outliers in the ED data are clearly visible. There is noticeable positive skewness in both samples; the Non-Ed data has more variability then the Ed data; the typical values of the ED data tend to be smaller than those for the Non-ED data.



Section 2.4

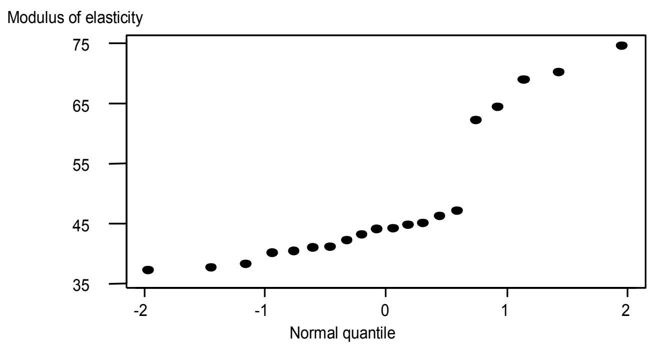
45.

The normal quantiles are easy to generate using Minitab or Excel. For example, in Minitab, typing the following commands will generate the normal quantiles:

```
MTB> set c2
MTB> 1:20
MTB> let c3 = (c2-.5)/20
MTB> invcdf c3 c4;
SUBC> norm 0 1.
```

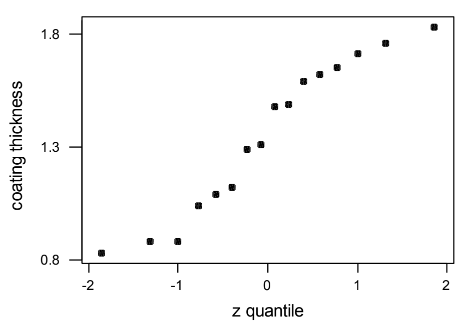
(Note: you don't have to type 'END' to end the input of data into column C2; typing any valid Minitab command will automatically end data input and then will execute the command)

Although it isn't necessary in this exercise, remember to sort (from smallest to largest) the data before plotting it versus the normal quantiles. The quantile plot for this data is shown below. The pattern is obviously nonlinear, so a normal distribution is implausible for this data. The apparent break that appears in the data in the right side of the graph is indicative of data that contains outliers.



46.

The quantile plot (created using the same technique as outlined in the textbook) is shown below:



The z quantiles were computed by first computing the sample quantiles using the equation:
$\left[{(i-.5)}/{16}\right]$ where i = 1, 2, …, 16

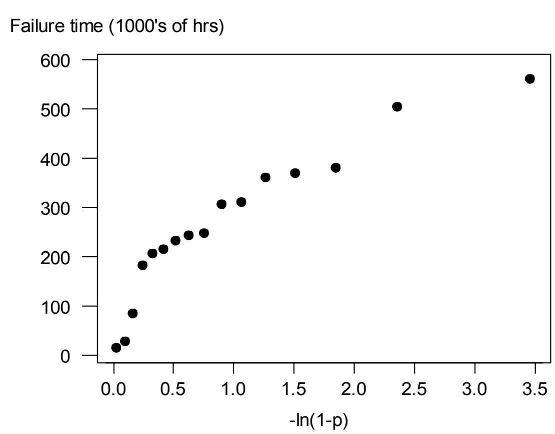
Then the z quantile was found for each sample quantile.

coating thickness	order	sample quantile	z quantile
0.83000	1	0.03125	-1.86273
0.88000	2	0.09375	-1.31801
0.88000	3	0.15625	-1.00999
1.04000	4	0.21875	-0.77642
1.09000	5	0.28125	-0.57913
1.12000	6	0.34375	-0.40225
1.29000	7	0.40625	-0.23720
1.31000	8	0.46875	-0.07841
1.48000	9	0.53125	0.07841
1.49000	10	0.59375	0.23720
1.59000	11	0.65625	0.40225
1.62000	12	0.71875	0.57913
1.65000	13	0.78125	0.77642
1.71000	14	0.84375	1.00999
1.76000	15	0.90625	1.31801
1.83000	16	0.96875	1.86273

Regarding the normality of the coating thickness data, the plot seems to exhibit some curvature. We should have some concerns about the normality assumption.

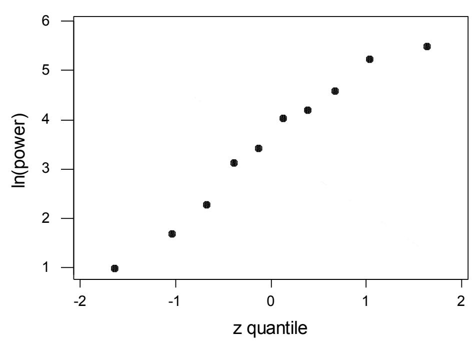
51.

Let η_p denote the pth quantile of an exponential distribution with parameter λ. Then, the area to the *right* of η_p is 1-p. Recall from Exercise 23 of Chapter 1, that the right tail area (i.e., the area past x = c) for an exponential distribution is simply $e^{-\lambda c}$. Therefore, $e^{-\eta_p \lambda}$ = 1-p. Taking logarithms of both sides, $-\eta_p\lambda$ = ln(1-p), so $\eta_p = \lambda$(-ln(1-p). That is, the quantiles η_p are linearly related to the quantitites -ln(1-p), so a plot of the sample quantiles $x_{(i)}$ versus -ln(1-p_i) is a straight line. In this exercise, n = 16, so the values of p_i equal (i-.5)/16 for i = 1,2,…16. Use Minitab or Excel to compute the plotting values -ln(1-p_i). The quantile plot for this data appears below. Because the plot exhibits curvature, an exponential distribution would not be appropriate for this data.



52.

(a) A normal quantile plot of x follows.
Clearly the variable, hourly median power, is not normally distributed, as the normal quantile plot is curvilinear.



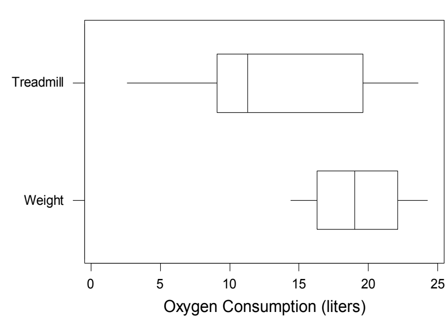
(b) By taking the natural logarithm of the variable and constructing a normal quantile plot of it, we obtain:

This plot looks quite linear indicating that it is plausible that these observations were sampled from a lognormal distribution.

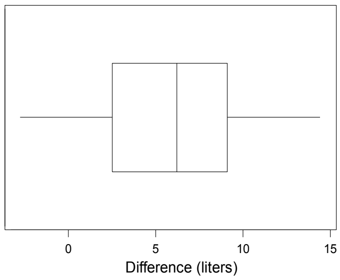
Supplementary Exercises

58.

(a) It appears that the oxygen consumption for weight training is greater on average than that for the treadmill exercise. The distribution of the weight training consumption appears to be symmetric, whereas the treadmill measurements appear to be somewhat positively skewed.



(b) The boxplot of the sample differences is fairly symmetric. From the boxplot, we note that most of the differences are positive, and an average difference value is about 6. The boxplot does suggest that the oxygen consumption is greater than the oxygen consumption for the treadmill exercise.



60. (a) Chebyshev's inequality is more conservative than is the empirical rule. This is because Chebyshev's rule makes no assumptions about the shape of the distribution of the values while the empirical rule assumes the shape of the distribution in approximately normal.

	Chebyshev's Rule	**Empirical Rule**
Percentage within 1 standard deviation	No statement	Approximately 68%
Percentage within 2 standard deviations	At least 75%	Approximately 95%
Percentage within 3 standard deviations	At least 89%	Approximately 99.7%

(b) $\mu = 100 \Rightarrow \lambda = .01$

$\sigma = 100$

$$\mu \pm \sigma \Rightarrow 100 \pm 100 \Rightarrow (0,\ 200)\ ;\ \int_0^{200} .01e^{-.01x}\,dx = 1 - e^{-.01(200)} = .8647$$

$$\mu \pm 2\sigma \Rightarrow 100 \pm 200 \Rightarrow (0,\ 300)\ ;\ \int_0^{300} .01e^{-.01x}\,dx = 1 - e^{-.01(300)} = .9502$$

$$\mu \pm 3\sigma \Rightarrow 100 \pm 300 \Rightarrow (0,\ 400)\ ;\ \int_0^{400} .01e^{-.01x}\,dx = 1 - e^{-.01(400)} = .9817$$

So:

Percentage within	**Chebyshev's Rule**	**Exponential**
1σ	No statement	86.47%
2σ	At least 75%	95.02%
3σ	At least 89%	98.17%

(c) In part (a) a normal distribution is assumed, since the empirical rule is applied.

In part (b) an exponential distribution with $\lambda = .01$ is specified, a positively skewed distribution.

In either case Chebyshev's inequality may not accurately estimate any particular distribution as it must accommodate all distributions.

61.

The transformation of the x_i's into the z_i's is analogous to 'standardizing' a normal distribution. The purpose of standardizing is to reduce a distribution (or, in this exercise, a set of data) to one that has a mean of 0 and a standard deviation of 1. To show that this transformation achieves this goal, note that:

$\sum z_i = \sum \frac{1}{s}(x_i - \bar{x}) = \frac{1}{s}\sum (x_i - \bar{x}) = \frac{1}{s}(0) = 0$, so, dividing this sum by n, $\bar{z}$ = 0. Next,
$\frac{1}{n-1}\sum (z_i - \bar{z})^2 = \frac{1}{n-1}\left(\frac{1}{s^2}\right)\sum (x_i - \bar{x})^2 = \left(\frac{1}{s^2}\right)s^2 = 1$.

65.

(a) Let y denote the capacitance of a capacitor. Capacitors will conform to specification when y is in the interval from 95 nf to 105 nf. Therefore, Proportion(95 ≤ y ≤ 105) = Proportion((95-98)/2 ≤ z ≤ (105-98)/2) = Proportion(-1.5 ≤ z ≤ 3.5). Using Table I, this proportion is equivalent to Proportion(z ≤ 3.5) - Proportion(z ≤ -1.5) = .9998 - .0068 = .9930, or, about 93.3%.

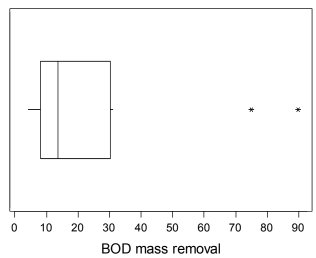
(b) The number of capacitors in a batch of 20 that conform to specifications will have a binomial distribution with n = 20 and π = .9330. Therefore, the proportion of batches containing at least 19 conforming capacitors is Proportion(x ≥ 20) = Proportion(x = 19) + Proportion(x = 20). Using the formula for the binomial mass function:

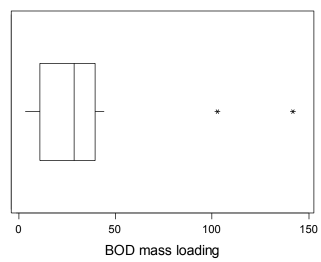
$$\frac{20!}{19!1!}(.9930)^{19}(1-.9930)^{1} + \frac{20!}{20!0!}(.9930)^{20}(1-.9930)^{0} = .9914$$

Section 3.1

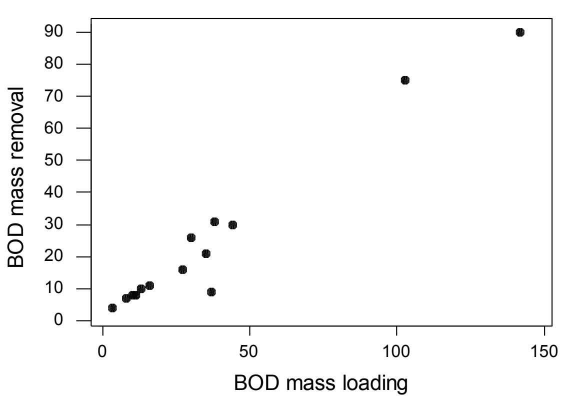
4.

(a)





On both the BOD mass loading boxplot and the BOD mass removal boxplot there are 2 outliers. Both variables are positively skewed.

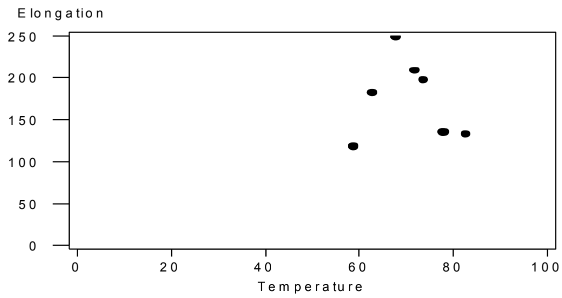


(b)

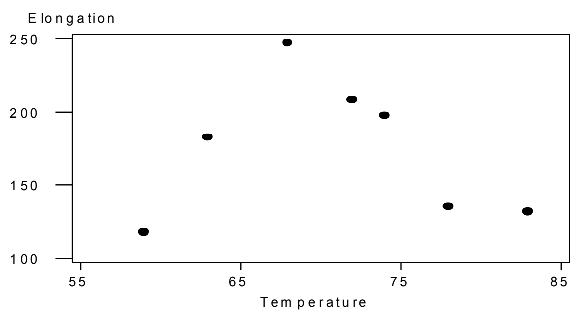
There is a strong linear relationship between BOD mass loading and BOD mass removal. As the BOD mass loading increases so does the BOD mass removal. The two outliers seen on each of the boxplots are seen to be correlated here. There is one observation that appears not to match the linear pattern. This value is (37, 9). One might have expected a larger value for BOD mass removal.

5.

(a) The scatter plot with axes intersecting at (0,0) is shown below.

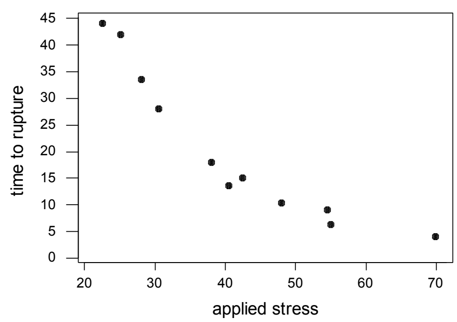


(b) The scatter plot with axes intersecting at (55,100) appears below. The plot in (b) makes it somewhat easier to see the nature of the relationship between the two variables.

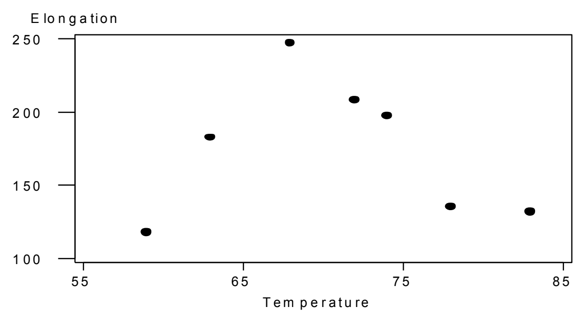


(c) A parabola appears to provide a good fit to both graphs.

8.

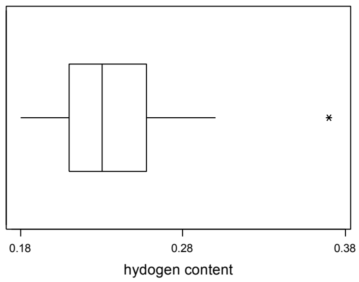


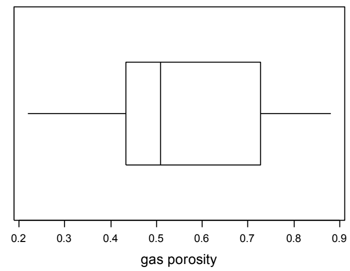
There appears to be a quadratic relationship between applied stress and time to rupture. The data has very little variation around the curvilinear pattern.



Section 3.2

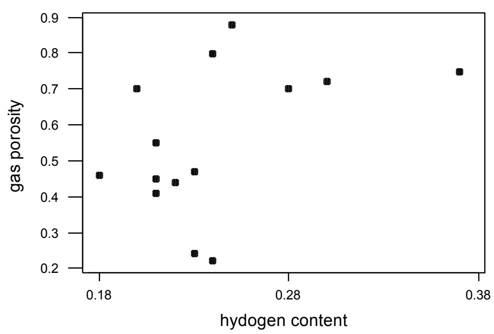
10. (a) The hydrogen content boxplot shows an outlier value of .37.





The gas porosity boxplot does not show any remarkable features.

(b)



The nature of the relationship between hydrogen content and gas porosity is unclear. There are two data pairs that do not seem to fit with the other 12 pairs, ((.23, .24) and (.24, .22)). One might have expected the gas porosity values to be larger than they were. Without these two pairs, a positive linear relationship between hydrogen content and gas porosity exists, although there is a large amount of variation.

10.

(c) First compute relevant quantities:

$$\sum x = 3.37 \qquad \sum y = 7.79$$
$$\sum x^2 = .84190 \qquad \sum y^2 = 4.8805$$
$$\sum xy = 1.9333 \qquad n = 14$$

Then, using the equations provided in Section 3.2, compute Pearson's sample correlation, r.

$$S_{xy} = \left(1.9333 - \frac{(3.37)(7.79)}{14}\right) = .05814$$

$$S_{xx} = \left(.84190 - \frac{(3.37)^2}{14}\right) = .03069$$

$$S_{yy} = \left(4.8805 - \frac{(7.79)^2}{14}\right) = .54592$$

$$\text{So, } r = \left(\frac{(.05814)}{\sqrt{.03069}\sqrt{.54592}}\right) = .44917$$

The computation of r = .449 does confirm the impression that there is a positive linear relationship, although it is a weak relationship.

11. (a) SS_{xy} = 5530.92 - (1950)(47.92)/18 = 339.586667, SS_{xx} = 251,970 - $(1950)^2$/18 = 40,720,and SS_{yy} = 130.6074 - $(47.92)^2$/18 = 3.033711, so

$$r = \frac{339.586667}{\sqrt{40720}\sqrt{3.033711}} = .9662$$

There is a very strong positive correlation between the two variables.

(b) Because the association between the variables is positive, the specimen with the larger shear force will tend to have a larger percent dry fiber weight.

(c) Changing the units of measurement on either (or both) variables *will have no effect on the calculated value of r*, because any change in units will affect both the numerator and denominator of r by exactly the same multiplicative constant.

12. (a) First compute relevant quantities:

$$\sum x = 44{,}615 \qquad \sum y = 3{,}860$$
$$\sum x^2 = 170{,}355{,}425 \qquad \sum y^2 = 1{,}284{,}450$$
$$\sum xy = 14{,}755{,}500 \qquad n = 12$$

Then, using the equations provided in Section 3.2, compute Pearson's sample correlation, r.

$$S_{xy} = 404{,}341.67$$
$$S_{xx} = 4{,}480{,}572.92$$
$$S_{yy} = 42{,}816.67$$
$$So,\ r = \left(\frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}\right) = .9233$$

A value of r = .9233 means that there is a strong positive linear relationship between TOST time and RBOT time.

(b) The value of r does not depend on which of the two variables is labeled as the x variable. Thus, had we let x = RBOT time and y = TOST time, the value of r would have remained the same.

(c) The value of r does not depend on the unit of measure for either variable. Thus, had we expressed RBOT time in hours instead of minutes, the value of r would have remained the same.

14. Using a correlation coefficient to summarize the relationship between the artist (x) and the sales price (y) is not appropriate. To compute and interpret a correlation coefficient both x and y variables must be quantitative variables. While the y variable, sale price, is quantitative, the x variable, artist, is not.

15. Let d_0 denote the (fixed) length of the stretch of highway. Then, d_0 = distance = (rate)(time) = xy. Dividing both sides by x, gives the equation $y = d_0/x$ which means the relationship between x and y is curvilinear (in particular, the curve is a hyperbola). However, for values of x that are fairly close to one another, sections of this hyperbola can be approximated very well by a straight line with a negative slope (to see this, draw a picture of the function d_0/x for a particular value of d_0). This means that r should be closer to -.9 than to any of the other choices.

16. The value of the sample correlation coefficient using the squared y values would not necessarily be approximately 1. If the y-values are greater than 1, then the squared y-values would differ from each other by more than the y-values differ from one another. Hence, the relationship between x and y^2 would be less like a straight line, and the resulting value of the correlation coefficient would decrease. [Note: I have yet to find an example where r is less than about .96 for (x, y^2), however.]

17. (a) $SS_{xx} = 37695 - (561)^2/9 = 2726$, $SS_{yy} = 40223 - (589)^2/9 = 1676.222$, and $SS_{xy} = 38281 - (561)(589)/9 = 1566.666$, so

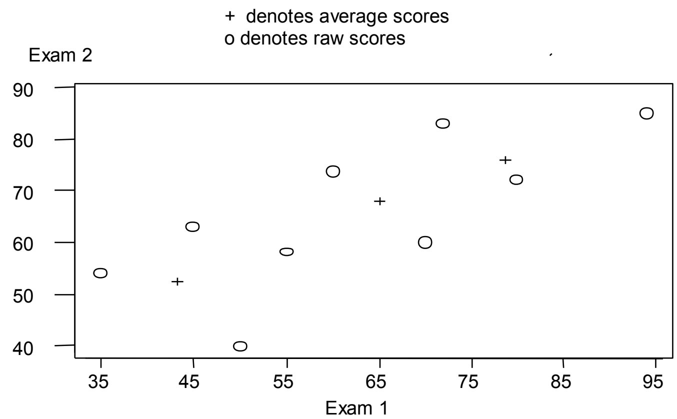
$$r = \frac{1566.667}{\sqrt{2726}\sqrt{1676.222}} = .733.$$

(b) $\bar{x}_1 = (70+72+94)/3 = 78.667$, $\bar{y}_1 = (60+83+85)/3 = 76$.
$\bar{x}_2 = (80+60+55)/3 = 65$, $\bar{y}_2 = (72+74+58)/3 = 68$.
$\bar{x}_3 = (45+50+35)/3 = 43.333$, $\bar{y}_3 = (63+40+54)/3 = 52.333$.

$S_{xx} = [(78.667)^2+(65)^2+(43.333)^2 - (78.667+65+43.333)^2/3] = 634.913$,
$S_{yy} = [(76)^2+(68)^2+(52.333)^2-(76+68+52.333)^2/3] = 289.923$,
$S_{xy} = [(78.667)(76)+(65)(68)+(43.333)(52.333)-(187)(196.333)/3] = 428.348$, so

$$r = \frac{428.348}{\sqrt{634.913}\sqrt{289.923}} = .9984.$$

(c) The correlation among the averages is noticeably higher than the correlation among the raw scores, so these points fall much closer to a straight line than do the unaveraged scores. The reason for this is that averaging tends to reduce the variation in data, making it more likely that the averages will fall close to a straight line than the more variable raw data.



Section 3.3

18. (a) To obtain the least squares regression equation, first we will compute the slope and the vertical intercept using the equations provided in Section 3.3.

$$S_{xy} = \left[25{,}825 - \frac{(517)(346)}{14}\right] = 13{,}047.71$$

$$S_{xx} = \left[39{,}095 - \frac{(517)^2}{14}\right] = 20{,}002.93$$

$$So,\ b = \left(\frac{S_{xy}}{S_{xx}}\right) = \left(\frac{13{,}047.71}{20{,}002.93}\right) = .652289939$$

$$\text{Also}:\ a = \bar{y} - b\bar{x} = \left(\frac{346}{14}\right) - (.652289939)\left(\frac{517}{14}\right) = .62615$$

Thus, the equation for the least squares line is:

$$\hat{y} = 0.626 + 0.6523x$$

(b) When $x = 35$, $\hat{y} = .62615 + .65228939(35)$ = 23.46

The corresponding residual is:
$residual = (y - \hat{y}) = (21 - 23.46) = -2.46$

(c) Yes, there is a very large residual corresponding to the (37, 9) pair. The predicted value $\hat{y}$ equals 24.76 and since the observed value is 9, the residual equals (9 – 24.76) = -15.76.

(d) We need to compute r^2.

$$SSResid = 392.0$$

$$SSTo = \left(17{,}454 - \frac{(346)^2}{14}\right) = 8{,}902.86$$

$$So,\ r^2 = \left(1 - \frac{SSResid}{SSTo}\right) = 1 - \left(\frac{392.0}{8{,}902.86}\right) = .956$$

Thus, 95.6% of the observed variation in removal can be explained by the approximate linear relationship between removal and loading, a very impressive result.

(e) After deleting these two pairs of observations, the new least squares line becomes:

$$\hat{y} = 2.29 + 0.564x$$

The new $r^2 = 68.8\%$.

So, you can see the large effect these two pairs of observations had on the analysis. The estimate of the slope was decreased and the fit of the least squares line, as measured by r^2, is not nearly as good.

21. (a) The following stem-and-leaf display shows that: a typical value for this data is a number in the low 40s, there is some positive skew in the data, there are some potential outliers (79.5 and 80.0), and there is a reasonably large amount of variation in the data (e.g., the spread 80.0 - 29.8 = 50.2 is large compared with the typical values in the low 40s.)

```
Stem-and-leaf of MoE       N  = 27
Leaf Unit = 1.0
    1    2 9
    3    3 33
   13    3 5566677889
  (4)    4 1223
   10    4 56689
    5    5 1
    4    5
    4    6 2
    3    6 9
    2    7
    2    7 9
    1    8 0
```

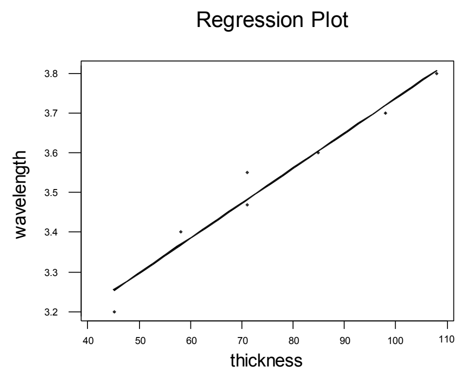
(b) No, the strength values are not uniquely determined by the MoE values. For example, note that the two pairs of observations having strength values of 42.8 have different MoE values.

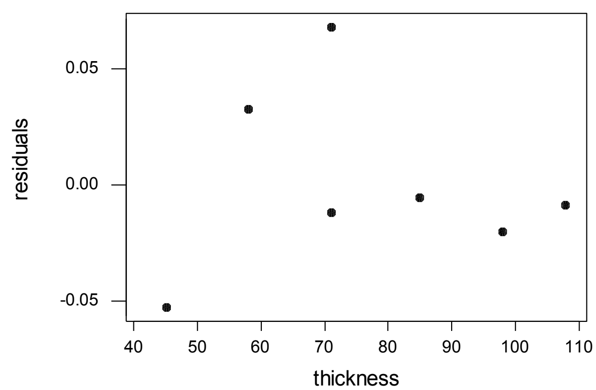
(c) The least squares line is $\hat{y}$ = 3.2925 + .10748x. For a beam whose modulus of elasticity is x = 40, the predicted strength would be $\hat{y}$ = 3.2925 + .10748(40) = 7.59. The value x = 100 is far beyond the range of the x values in the data, so it would be dangerous (i.e., potentially misleading) to extrapolate the linear relationship that far.

(d) From the printout, SSResid = 18.736, SSTo = 71.605, and the coefficient of determination is r^2 = .738 (or, 73.8%) . The r^2 value is large, which suggests that the linear relationship is a useful approximation to the true relationship between these two variables.

(e) There is no obvious pattern in the residuals that might suggest anything other than an approximately linear relationship between the variables.

22. Using Minitab to run this regression, we obtain the following least squares regression line: $\hat{y} = 2.86 + 0.00882x$





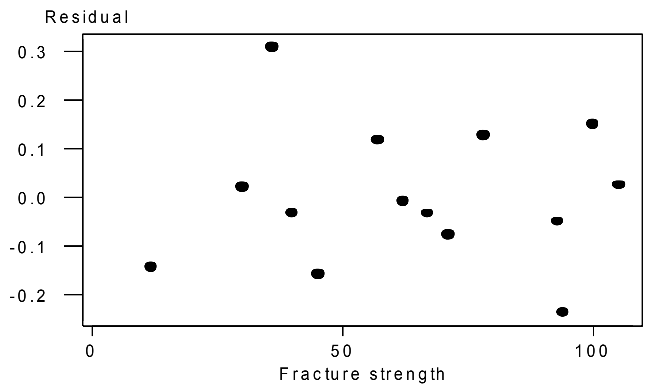
96.1% of the observed variation in PL wavelength can be attributed to the approximate linear relationship between the PL wavelength and layer thickness. This means that the linear fit is very good.

Looking at the residual plot, there is no particular pattern to the plot and there are no surprising residuals. Thus, there does not seem to be any problem with our linear fit.

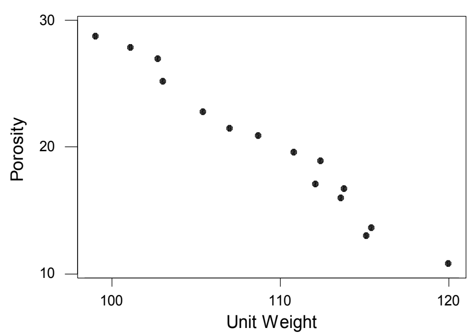
23. (a) From the 'Parameter Estimate' column of the printout, the least squares line is $\hat{y}$ = 3.620906 - 0.014711x, where x = fracture strength. Substituting the value x = 50 into the equation gives a predicted attenuation of $\hat{y}$ = 3.620906 - 0.014711(50) = 2.8854, or, about 2.89.

(b) From the 'Sum of Squares' column of the printout, SSResid = .26246 and SSTo = 2.55714. The r^2 value is .8974, or 89.74%. s_e is called the 'Root MSE' (where MSE stands for 'mean square error') in the printout, so s_e = .14789. The high vale of r^2 and the small value of s_e (compared to the typical size of the y data values) indicate that the least squares line effectively summarizes the relationship between the variables.

(c) The plot of the residuals versus x (below) shows no discernible patterns, suggesting that no modification to the straight-line model is needed.



24. (a)



The scatter plot above suggests a strong, negative linear relationship between porosity and unit weight.

(b) $b = S_{xy} / S_{xx} = -471.542 / 521.196 = -.9047 \approx -.905$

$a = \bar{y} - b\bar{x} = (1/15)(299.8) - (-.905)[(1/15)(1640.1)] = 19.987 + .905(109.34) \approx 118.9$

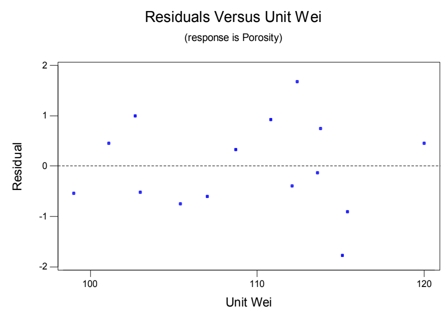
So the estimate of the least squares line is $\hat{y} = a + bx = 118.9 - .905x$.

The predicted porosity value for the weight at $x = 110$ is $\hat{y} = 118.9 - .905(110) = 19.35$.

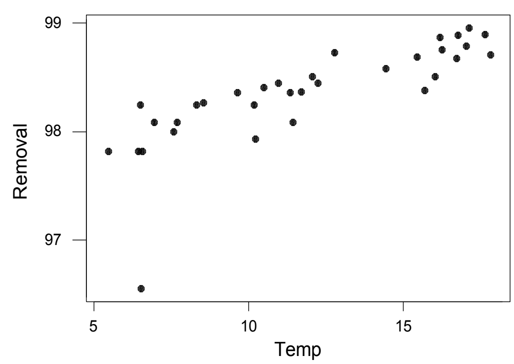
(c) By substituting $x = 135$ into the least squares equation, the predicted porosity value is $\hat{y} = 118.9 - .905(135) = -3.275$. Thus, we get a negative weight value, which is not physically possible. We get this ridiculous value because the point $x = 135$ is far beyond the range of the data; that is, we extrapolated.

(d) We can assess the fit of the least squares line by considering the coefficient of determination r^2. We can calculate r^2 in either of two ways. First of all, r^2 by definition is 1 – SSResid/SSTo. We can compute SSTo by the formula SSTo = $S_{yy} = \sum y_i^2 - (1/n)(\sum y_i)^2 = 6430.06 - (1/15)(299.8)^2 =$ 438.06. We can now calculate $\text{SSResid} = \text{SSTo} - bS_{xy} = 438.06 - (-.905)(-471.542) = 11.31$. So, $r^2 = 1 - 11.31/438.06 = 1 - .026 = .974$. A second way to calculate r^2 in simple linear regression is to take the square of the sample correlation coefficient :
$r^2 = S_{xy}^2/(S_{xx}S_{yy}) = (-471.542)^2/[(521.196)(438.06)] = .97388 \approx .974$. Therefore, 97.4% of the porosity values can be explained by a linear relationship with the unit weight.

(e) A plot of the residuals versus the x values (unit weight) follows.



25. (a) Yes, the following plot does suggest the aptness of a simple linear regression model.



(b) We need to calculate S_{xx} and S_{xy}:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - (1/n)(\sum x_i)^2 = 5099.2412 - (1/32)(384.26)^2 \approx 485.00$$

$$S_{xy} = \left(\sum (x_i - \bar{x})\right)\left(\sum (y_i - \bar{y})\right) = \sum x_i y_i - (1/n)(\sum x_i)(\sum y_i)$$
$$= 37{,}850.7762 - (1/32)(384.26)(3149.04) = 36.71025$$

Therefore, $b = S_{xy}/S_{xx} = 36.71025/485 \approx .0756$, and
$a = \bar{y} - b\bar{x} = (1/32)(3149.04) - .0756[(1/32)(384.26)] = 98.4075 - 13.212(12.008125) \approx 97.5$.
Hence, the least squares line is given by $\hat{y} = 97.5 + .0756x$.
The following MINITAB output summarizes the least squares regression. The output gives $\hat{y} = 97.5 + .0757x$.

```
The regression equation is
Removal = 97.5 + 0.0757 Temp

Predictor        Coef      SE Coef          T        P
Constant      97.4986       0.0889    1096.17    0.000
Temp         0.075691     0.007046      10.74    0.000

S = 0.1552      R-Sq = 79.4%     R-Sq(adj) = 78.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2.7786      2.7786    115.40    0.000
Residual Error    30      0.7224      0.0241
Total             31      3.5010
```

Now, the point prediction of removal efficiency at the temperature value of 10.50 is $\hat{y} = 97.5 + .0756(10.5) = 98.2938$. The residual is given by $y - \hat{y} = 98.41 - 98.2938 = 0.1162$.

25.

(c) The size of a typical of deviation of points in the scatter plot from the least squares line is given by the standard deviation about the least squares line: $s_e = \sqrt{\dfrac{\text{SSResid}}{n-2}}$.

$$\text{SSResid} = \text{SSTo} - bS_{xy}$$
$$\text{SSTo} = S_{yy} = \sum y_i^2 - (1/n)(\sum y_i)^2 = 309{,}892.6548 - (1/32)(3149.04)^2 = 3.501$$
$$\text{SSResid} = \text{SSTo} - bS_{xy} = 3.501 - (.0756)(36.71025) = .7257$$
$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}} = \sqrt{\frac{.7257}{32-2}} = \sqrt{.02419} \approx .1552.$$

NOTE: From the MINITAB output above, we could see that the output value of `S = 0.1552` gives s_e, and we could have calculated s_e by calculating the square root of the mean square residual error of .0241.

(d) The proportion of observed variation in removal efficiency that can be attributed to the approximate linear relationship is given by $r^2 = 1 - \dfrac{\text{SSResid}}{\text{SSTo}} = 1 - \dfrac{.7257}{3.501} = .7927$, which can also be found from the MINITAB output.

25.

(e) First compute the new summary statistics:

$$\sum x_i = 384.26 + 6.53 = 390.79$$
$$\sum x_i^2 = 5099.241 + 6.53^2 = 5141.8819$$
$$\sum y_i = 3149.04 + 96.55 = 3245.59$$
$$\sum y_i^2 = 309{,}892.6548 + 96.55^2 = 319{,}214.5573$$
$$\sum x_i y_i = 37{,}850.78 + (6.53)(96.55) = 38{,}481.25$$

Then compute the least squares line:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - (1/n)(\sum x_i)^2 = 5141.8819 - (1/33)(390.79)^2 \approx 514.10$$
$$S_{xy} = \left(\sum (x_i - \bar{x})\right)\left(\sum (y_i - \bar{y})\right) = \sum x_i y_i - (1/n)(\sum x_i)(\sum y_i)$$
$$= 38{,}481.25 - (1/33)(390.79)(3245.59) = 46.58$$

Therefore, $b = S_{xy} / S_{xx} = 46.58 / 514.1 \approx .0906$, and
$a = \bar{y} - b\bar{x} = (3245.59/33) - (.0906)(390.79/33) = 97.28$
So our new equation is $\hat{y} = 97.28 + .0906$.

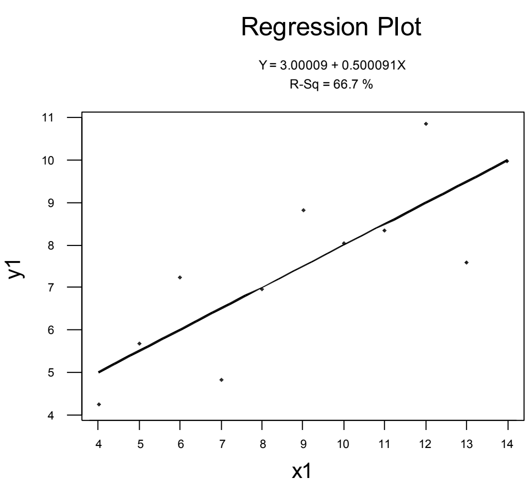
Next, compute s_e and r^2:

$$\text{SSTo} = S_{yy} = \sum y_i^2 - (1/n)(\sum y_i)^2 = 319{,}214.5573 - (1/33)(3245.59)^2 = 6.85$$
$$\text{SSResid} = \text{SSTo} - bS_{xy} = 6.85 - (.0906)(46.58) = 2.63$$
$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}} = \sqrt{\frac{2.63}{33-2}} = \sqrt{.0848} \approx .2913.$$
$$r^2 = 1 - \frac{\text{SSResid}}{\text{SSTo}} = 1 - \frac{2.63}{6.85} = .616$$

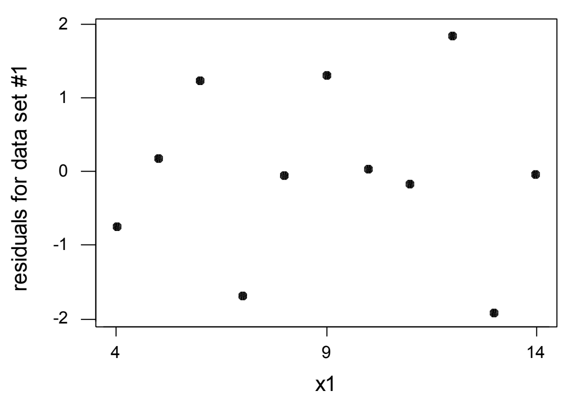
From above, we see that the intercept of our least squares equation decreased from 97.5 to about 97.28. The change in the slope is noticeable: the new slope increased from .0756 to .0906. The addition of this new results in a much larger s_e of .2913, which is about twice the original s_e value of .1552.

Consequently, the r^2 valued decreased; the new r^2 of .616 is down from the original r^2 value of .7257. Moreover, s_e increases.

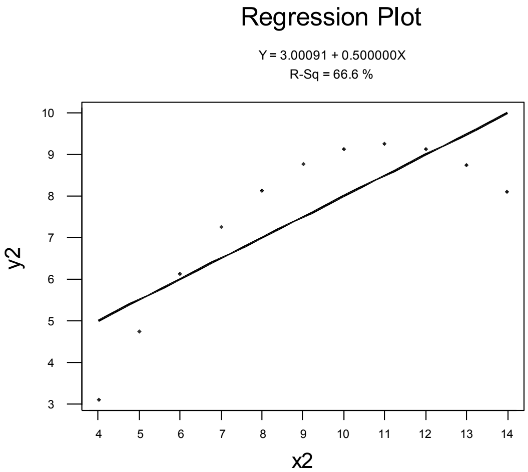
26. Data set #1

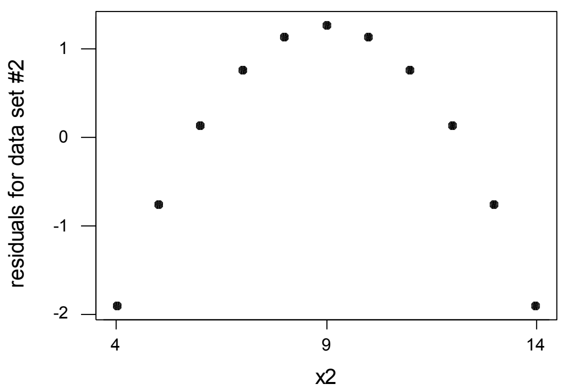


Fitting a straight line seems appropriate here. There is no indication of a problem. There is a fair amount of scatter around the least squares line, however. This fact is quantified by the r^2 value of about 67%.



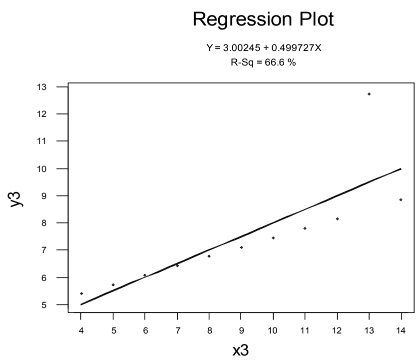
26. Data set #2

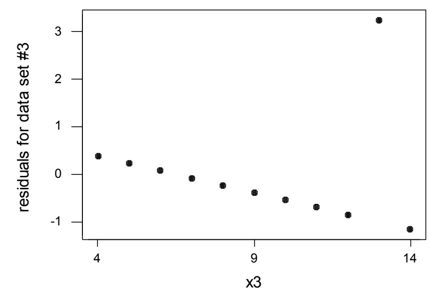




It is inappropriate to fit a straight line through this data. There is a quadratic relationship between these two variables. So, a model that incorporates this quadratic relationship should be fit instead of a linear fit.

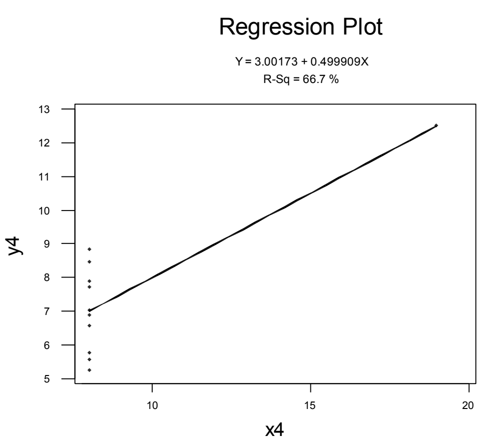
26. Data set #3

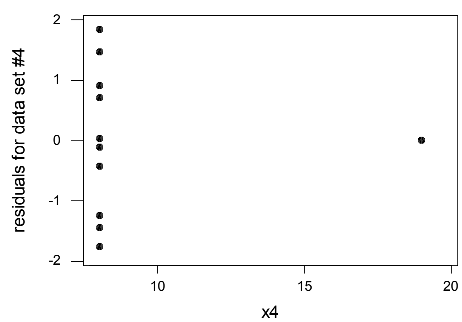




In this data set there is one pair of values that is a clear outlier (13.0, 12.74). The y-value of 12.74 is much larger than one would expect based on the other data. This pair of values exerts a lot of influence on the linear fit. It should be investigated. Perhaps it was a recording error or some other similar problem. With the outlier included, it is inappropriate to fit a straight line. With it excluded, an excellent linear fit is achieved.

Data set #4

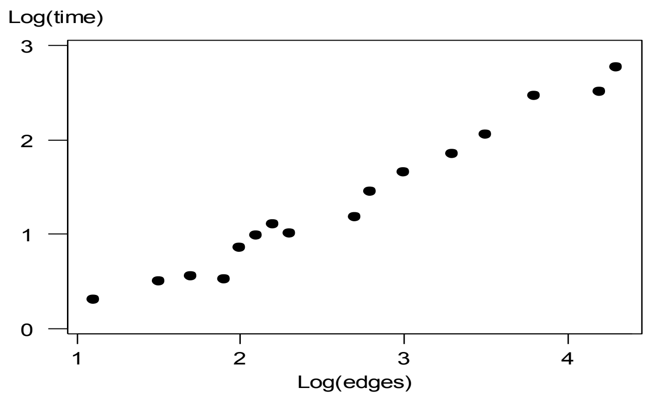




It is inappropriate to fit a straight line through this data. There is no linear relationship between the variables. There is one pair of observations that is a clear outlier (19.0, 12.50). It should be investigated.

Section 3.4

27. (a) The plot is shown below. The plot suggests that a straight-line relationship log(time) $\approx$ a + b(log(edges)) provides a good fit to the data. Letting y denote the recognition time and letting x denote the number of edges on a part, exponentiating both sides of this equation yields the approximate equation $y \approx kx^b$ (where $k = 10^a$, since logarithms base 10 were used). This is a 'power relationship' between x and y.

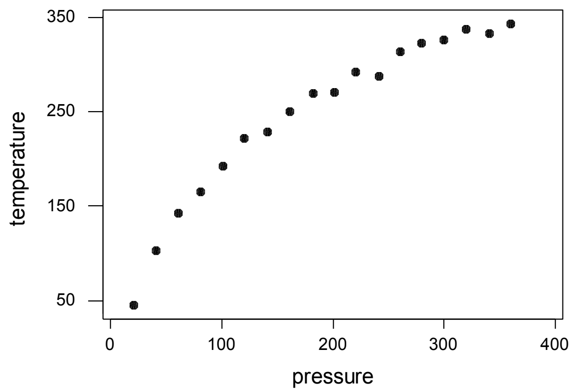


(b) The following printout (from Minitab) shows a least squares fit for this data. The large value of r^2 (97.5%) and the small value of s_e (.1241, which is small compared to the typical y values) indicate that the least squares line provides an excellent fit to this data.

```
The regression equation is
logtime = - 0.760 + 0.798 loedg

Predictor       Coef       StDev          T        P
Constant    -0.76011     0.09330      -8.15    0.000
loedg        0.79839     0.03320      24.05    0.000

S = 0.1241      R-Sq = 97.6%     R-Sq(adj) = 97.5%

Analysis of Variance

Source       DF          SS          MS          F        P
Regression    1      8.9112      8.9112     578.23    0.000
Error        14      0.2158      0.0154
Total        15      9.1270
```

(c) Substituting the values of a = -.76011 an b = .79839 from the printout in (b) into the power relationship in part (a), the nonlinear relationship between y and x can be described by the equation $y \approx .17374x^{.79839}$ (where $.17374 = 10^{-.76011}$).

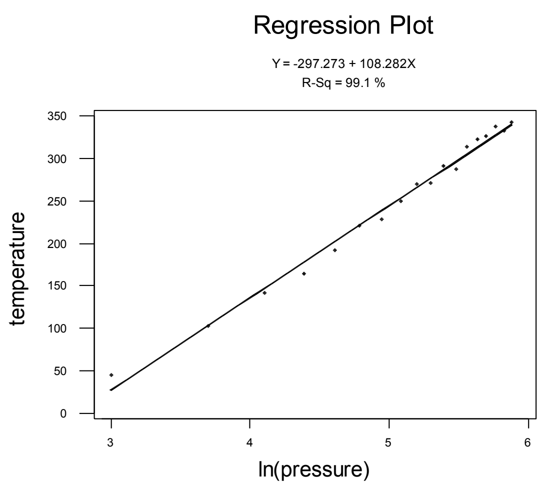
28. (a)



It is not appropriate to fit a straight line to this data. There is clear curvature to the scatterplot. Given that the pattern of curvature seen in the scatterplot is consistent with segment 4 of Figure 3.15, a transformation “down the ladder” for x or y would be appropriate. A scatterplot of (ln(x), y) is quite straight. A regression of y on ln(x) produced the following least squares regression equation:

$$\hat{y} = -297.27 + 108.28\ell n(x)$$

where y = temperature and x = pressure.

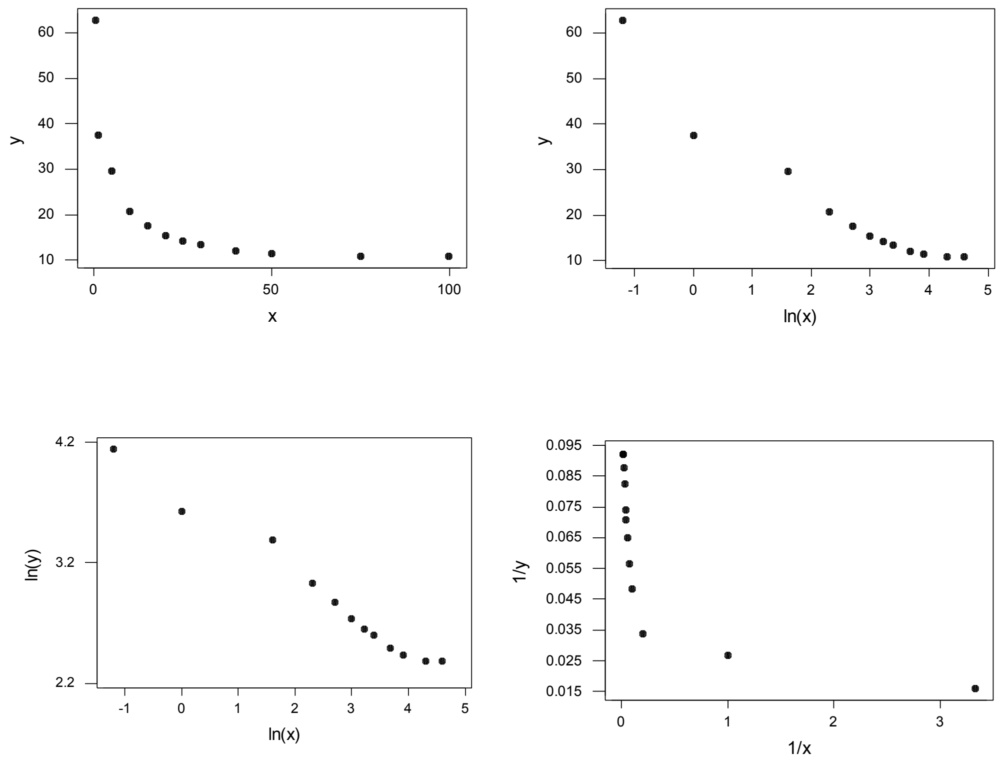


(Note: $r^2 = 99\%$)

To predict the value of temperature that would result from a pressure of 200:

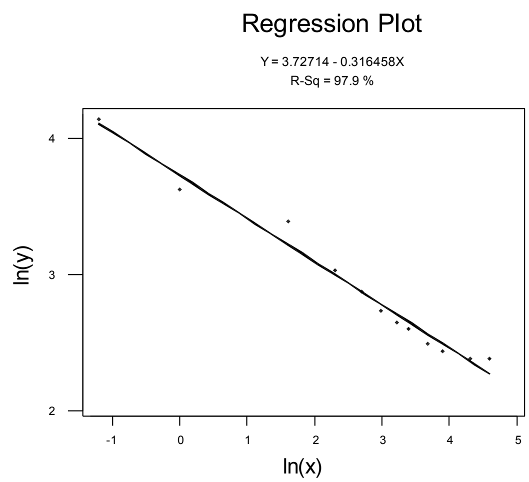
$$\hat{y} = -297.27 + 108.28\ell n(200) = 276.43\ ^\circ F$$

30. (a)



30.

(b) The ln(x) versus ln(y) transformation seems to do the best job of producing an approximate linear relationship.



A regression of ln(y) on ln(x) produced the following least squares regression equation:

$$\ell\hat{n}(y) = 3.73 - 0.316\ell n(x)$$

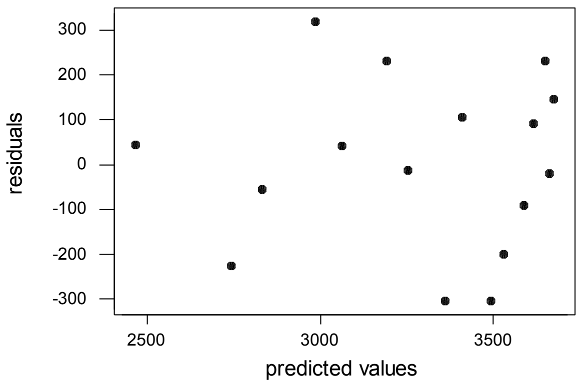
(Note: $r^2 = 98\%$)

To predict lead content, y, when distance, x, equals 45m:

$$\ell\hat{n}(y) = 3.73 - 0.316\ell n(45) = 2.527$$
$$\Rightarrow y = e^{2.527} = 12.516$$

We would predict lead content, y = 12.516 ppm.

32. (a)



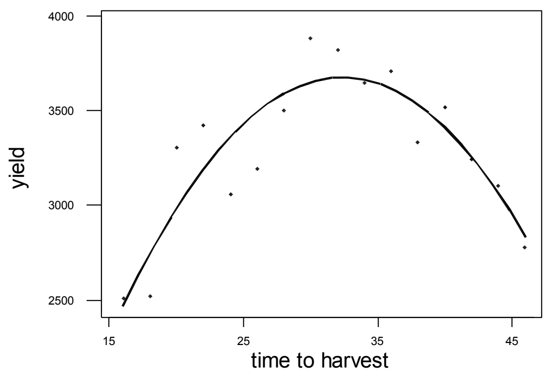
It is not possible to transform this data as described in Section 3.4 so as to approximate a linear relationship between the transformed data. There is a quadratic relationship between yield and time to harvest, so a quadratic model should be fit.

32.

(b) A quadratic fit to this data using Minitab produced:

$$\hat{y} = -1070.4 + 293.483x - 4.5358x^2$$

9%)



To predict yield when time to harvest is 25 days:

$$\hat{y} = -1070.40 + 293.483(25) - 4.5358(25)^2 = 3431.8$$

We would predict 3,432 kg/ha when time to harvest is 25 days.

Finally, the residual plot for this regression does not show us any unusual pattern or unusual residuals. We believe our quadratic fit is quite good.

33. (a) From the 'Coef' column in the printout, the least squares equation is $\hat{y}$ = -251.6 + 1000.1x -135.44x^2. For a water supply value of x = 3.0, the predicted wheat yeild would be $\hat{y}$ = -251.6 + 1000.1(3.0) - 135.44$(3.0)^2$ = 1529.74.

(b) From the printout, SSResid = 202,227 and SSTo = 1,372,436. These values yield an r^2 value of 1 - SSResid/SSTo = 1 - 202,227/1,372,436 = .8526, or, about 85.3%. The large value of r^2 suggests that the quadratic fit is good.

Section 3.5

36. (a) The coefficient of multiple determination is R^2. R^2 = 78% in this regression.

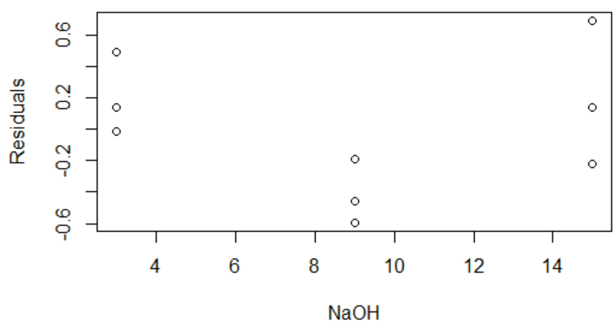
So, 78% of the observed variation in surface area can be attributed to the stated approximate relationship between surface area and the predictors.

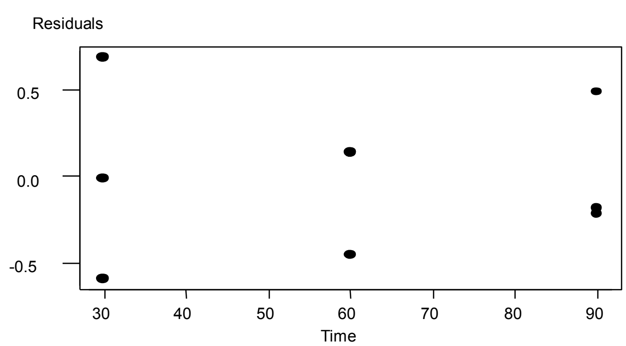
(b) $x_1 = 2.6$ $x_2 = 250$ $x_1x_2 = (2.6)(250) = 650$

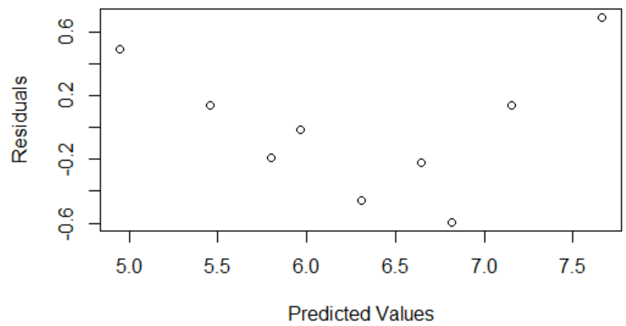
$$\hat{y} = 185.49 - 45.97(2.6) - 0.3015(250) + (0.0888)(650) = 48.313$$

(c) No, it is not legitimate to interpret b_1 in this way. It is not possible to increase by 1 unit the cobalt content, x_1, while keeping the interaction predictor, x_3, fixed. When x_1 changes, so does x_3, since $x_3 = x_1 x_2$.

39. All three residual plots (below) show evidence of curvature, indicating that higher-order terms (e.g., x_1^2, x_2^2) should be considered as additions to the model.







40. (a) $f(500, 500) = .30$

(b) $f(200, 0) + f(200, 250) = .20 + .10 = .30$

(c)

x	200	500
$f_1(x)$	.50	.50

y	0	250	500
$f_2(y)$	.25	.25	.50

42. Given:

x	0	1	2
$f_1(x)$	.80	.15	.05

where x = machinery defects.

Given:

y	0	1	2	3	4
$f_2(y)$	.50	.25	.15	.08	.02

where y = cosmetic flaws.

(a) Since x and y are independent,
$f(x, y) = f_1(x)\, f_2(y)$

	f(x,y)	y = **0**	**1**	**2**	**3**	**4**
x	**0**	.400	.2000	.1200	.064	.016
	1	.075	.0375	.0225	.012	.003
	2	.025	.0125	.0075	.004	.001

(b) The proportion of machines with no major defect or cosmetic flaw = $f(0, 0) = .40$.

The proportion of machines with at least one defect or flaw = $1 - f(0, 0) = .60$.

(c) The proportion of machines with the number of cosmetic flaws exceeding the number of major defects

= f(0, 1) + f(0, 2) + f(0, 3) + f(0, 4) + f(1, 2) + f(1, 3) + f(1, 4) + f(2, 3) + f(2, 4) = .20 + .12 + .064 + .016 + .0225 + .012 + .003 + .004 + .001 = .4425

43. Summing across the rows (for x) and the columns (for y), the marginal distributions of x and y are:

x:	200	500
	.50	.50

y:	0	250	500
	.25	.25	.50

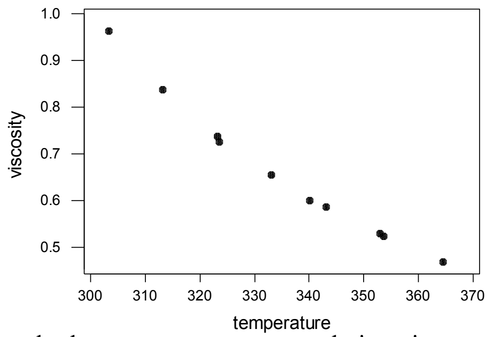
So, the means of the two distributions are: μ_x = 200(.50) + 500(.50) = 350 and μ_y = 0(.25) + 250(.25) + 500(.50) = 312.5. The covariance between the two variables is then: covariance(x,y) = $\sum\sum(x-\mu_x)(y-\mu_y)f(x,y)$, where f(x,y) is the joint mass function given in exercise 38. That is, covariance(x,y) = (200-350)(0-312.5)(.20) + (200-350)(250-312.5)(.10)+…+ (500-350)(500-312.4)(.30) = 9375.0.

Using the marginal distributions again, the standard deviations of each variable can be computed:

$\sigma_x^2 = \sum(x-\mu_x)^2 p(x)$ = $(200\text{-}350)^2(.50) + (500\text{-}350)^2(.50)$ = 22500, so σ_x = 150 and, $\sigma_y^2 = \sum(y-\mu_y)^2 p(y)$ = $(0\text{-}312.5)^2(.25) + (250\text{-}312.5)^2(.25) + (500\text{-}312.5)^2(.50)$ = 42,968.75, so σ_y = 207.289. The population correlation coefficient is then ρ = covariance/($\sigma_x\sigma_y$) = 9375/[(150)(207.289)] = .302. Because it is closer to 0 than to 1, this value of ρ does not suggest that there is a strong relationship between x and y.

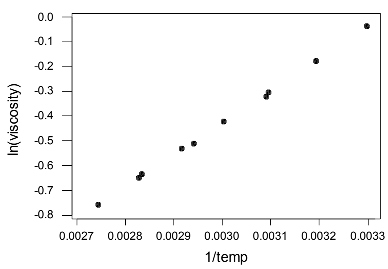
Supplementary Exercises

44. (a)



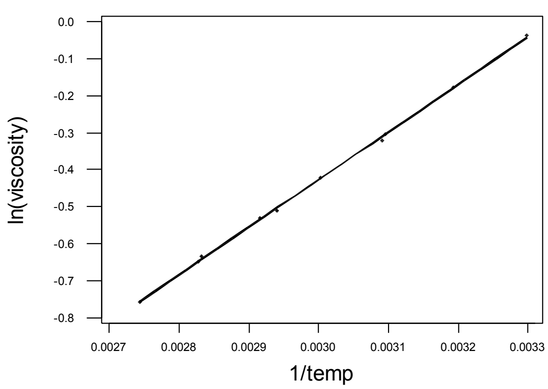
The scatterplot between temperature and viscosity reveals a slight curvature. So, a transformation to and/or y would be appropriate prior to fitting a straight line.

(b)



Once both variables are transformed, a linear relationship exists between them.

44. (c)



The least squares regression equation is:

$$\ell\hat{n}(y) = -4.275 + 1{,}282.8\left(\frac{1}{x}\right)$$

When x = 325

$$\ell\hat{n}(y) = -4.275 + 1{,}282.8\left(\frac{1}{325}\right) = -0.3279$$

$$\Rightarrow y = e^{-.3279} = .7204$$

When the temperature is 325 degrees Kelvin, we would predict the viscosity to be .7204 mPas.

(d) A straight line does an excellent job of summarizing the relationship between ln(y) and $\frac{1}{x}$. One way to measure this is by using r^2. In this regression r^2 = 99.9%, meaning that 99.9% of the observed variation in ln(viscosity) can be explained by the linear relationship between ln(viscosity) and the reciprocal of temperature.

47 (a) Since stride rate is being predicted, y = stride rate and x = speed. Therefore, $SS_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 3880.08 - (205.4)^2/11 = 44.7018$, $SS_{yy} = \sum y_i^2 - (\sum y_i)^2/n = 112.681 - (35.16)^2/11 = .2969$, and $SS_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 660.130 - (205.4)(35.16)/11 = 3.5969$. Therefore, $b = SS_{xy}/SS_{xx} = 3.5969/44.7018 = .0805$ and $a = (35.16/11) - (.0805)(205.4/11) = 1.6932$. The least squares line is then $\hat{y} = 1.6932 + .0805x$.

(b) Predicting speed from stride rate means that y = speed and x = stride rate. Therefore,interchanging the x and y subscripts in the sums of squares computed in part (a), we now have $SS_{xx} = .2969$ and $SS_{xy} = 3.5969$ (note that SS_{xy} does not change when the roles of x and y are reversed). The new regression line has a slope of $b = SS_{xy}/SS_{xx} = 3.5969/.2969 = 12.1149$ and an intercept of $a = (205.4/11) - (12.1149)(35.16/11) = -20.0514$; that is, $\hat{y} = -20.0514 + 12.1149x$.

(c) For the regression in part (a), $r = 3.5969/[\sqrt{44.7018}\sqrt{.2969}] = .9873$, so $r^2 = (.9873)^2 = .975$. For the regression in part (b), r is also equal to .9873 (since reversing x and y has no effect on the formula for r). So, both regressions have the same coefficient of determination. For the regression in part (a), we conclude that about 97.5% of the observed variation in rate can be attributed to the approximate linear relationship between speed and rate. In part (b), we conclude that about 97.5% of the variation in speed can be attributed to the approximate linear relationship between rate and speed.

48. Values for the vertical intercept and the slope will be changed using the same constant that was used to change y. For example, suppose that the least squares equation used to predict speed (in ft/sec) from stride rate was:

$$\hat{y} = -20.0514 + 12.1149x$$

Then, since 1 m is about 3.2808 ft, if speed was expressed in m/sec instead of ft/sec, the new least squares equation would be:

$$\hat{y} = \left(\frac{-20.0514}{3.2808}\right) + \left(\frac{12.1149}{3.2808}\right)x$$

$$\Rightarrow \hat{y} = -6.1117 + 3.6927x$$

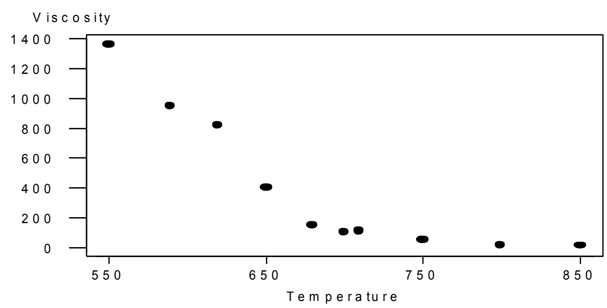
More generally, if each y value in the sample is multiplied by the same number c, the slope and vertical intercept of the least squares line will also change by multiplying each of them by c.

49. (a) Substituting x = .005 into the least squares equation, a stress value of $\hat{y}$ = 88.791 + 5697.0(.005) - 328,161$(.005)^2$ = 109.07 would be predicted.

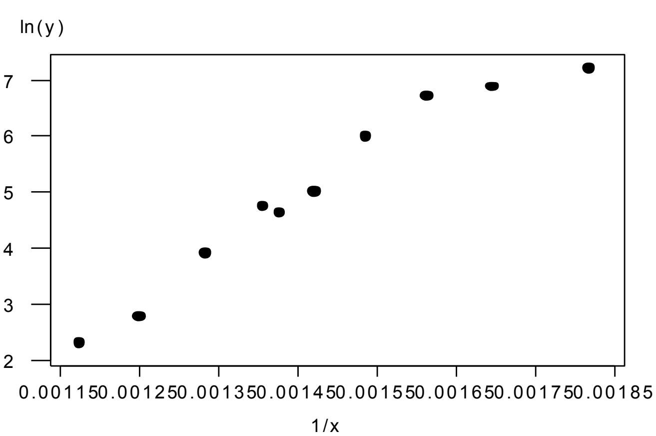
(b) From the stress values (i.e, y values given in the exercise,) SSTo = SS_{yy} = $\sum y_i^2 - (\sum y_i)^2/n$ = 107,604 - $(1034)^2$/10 = 688.40. Subtracting predicted values from actual values gives the residuals: -3.16, -1.87, 5.07, 1.93, 2.84, -3.36, -1.48, -2.22, 2.07, and 0.20. The sum of squares of these residuals is SSResid = 73.711. The coefficient of determination r^2 = 1 - SSResid/SSTo = 73.711/688.40 = .893. Therefore, about 89.3% of the observed variation in stress can be attributed to the approximate linear relationship between stress sand strain.

(c) Substituting x = .03 into the least squares equation yields a predicted stress value of $\hat{y}$ = 88.791 + 5697.0(.03) - 328,161$(.03)^2$ = -35.64, which is not at all realistic (since stress values can not be negative). The problem is that the value x = .03 lies far outside the region of x values that were used to fit the regression equation (the maximum x value used was x = .017). Extrapolation such as this is often unreliable (at best) and sometimes leads to ridiculous (i.e., impossible) predictions as in this case.

51. (a) The curvature that is apparent in the plot of y versus x (see below) indicates that merely fitting a straight-line to the data would not be the best strategy. Instead, one should search for some transformation of the x or y data (or both) that would give a more linear plot.



51.
(b) The plot below shows the graph of ln(y) versus 1/x. Because it appears to be approximately linear, a straight-line fit to such data should provide a reasonable approximation to the relationship between the two variables.



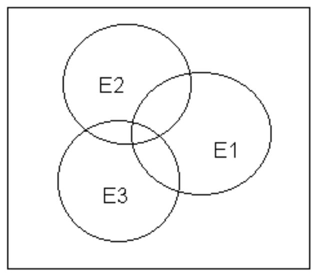
The following Minitab printout shows the results of fitting a regression line to the transformed data. From the printout, the prediction equation is $\ln(y) \approx -7.2557 + 8328.4(1/x)$. The r^2 value of 95.3% indicates that the fit is quite good. When temperature is 720 (i.e., x = 720), the equation gives a predicted value of $\ln(y) \approx -7.2557 + 8328.4(1/720) = 4.31152$. Exponentiating both sides gives a predicted y value of $y \approx e^{4.31152} = 74.6$.

```
The regression equation is
logey = - 7.26 + 8328 recipx

Predictor       Coef       StDev          T        P
Constant     -7.2557      0.9670      -7.50    0.000
recipx        8328.4       651.1      12.79    0.000

S = 0.3882      R-Sq = 95.3%     R-Sq(adj) = 94.8%

Analysis of Variance

Source       DF          SS          MS         F        P
Regression    1      24.663      24.663    163.63    0.000
Error         8       1.206       0.151
Total         9      25.869
```

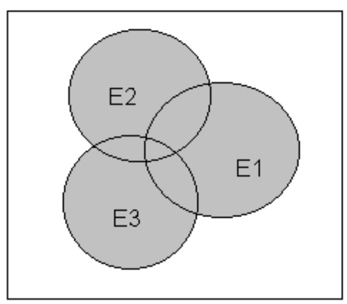
Chapter 5
Probability and Sampling Distributions

Section 5.1

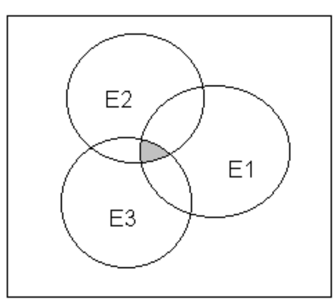
5.2. A Venn diagram of these three events is



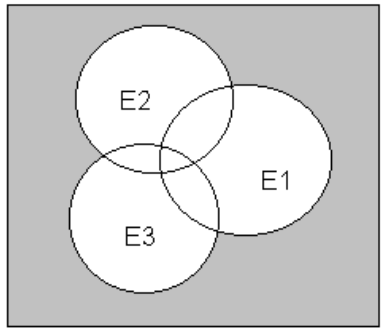
(a) The event 'at least one plant is completed by the contract date' is represented by the shaded area covered by all three circles:



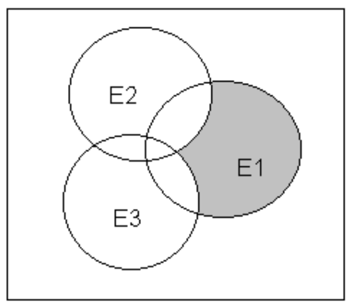
(b) The event 'all plants are completed by the contract date' is the shaded area where all three circles overlap:



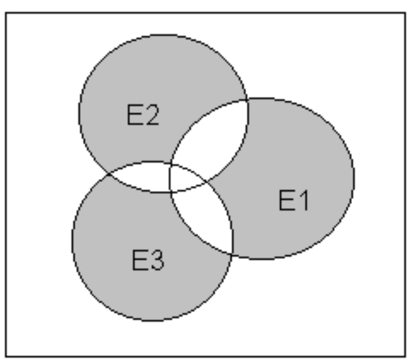
(c) The event ‘none of the plants is completed by the contract date’ is the complement of the shaded area in (a):



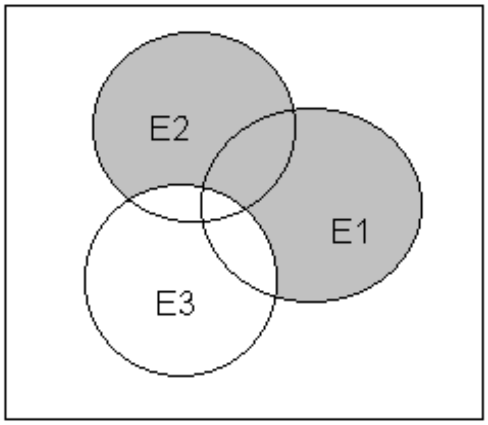
(d) The event ‘only the plant at site 1 is completed by the contract date’ is shown shaded:



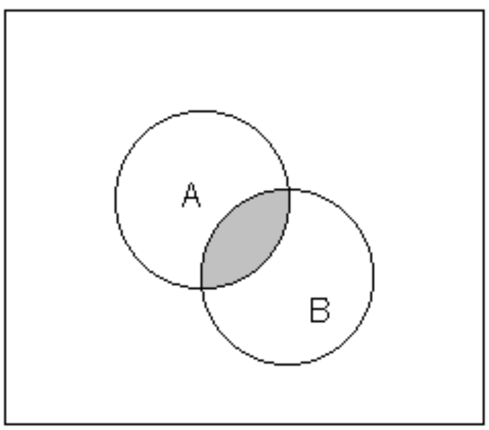
(e) The event ‘exactly one of the three plants is completed by the contract date is:



(f) The event ‘either the plant at Site 1 or Site 2 or both plants are completed by the contract date’ is:



5.7. The event *A and B* is the shaded area where A and B overlap in the following Venn diagram. Its complement consists of all events that are either not in A or not in B (or not in both). That is, the complement can be expressed as *A′ or B′*.



Section 5.2

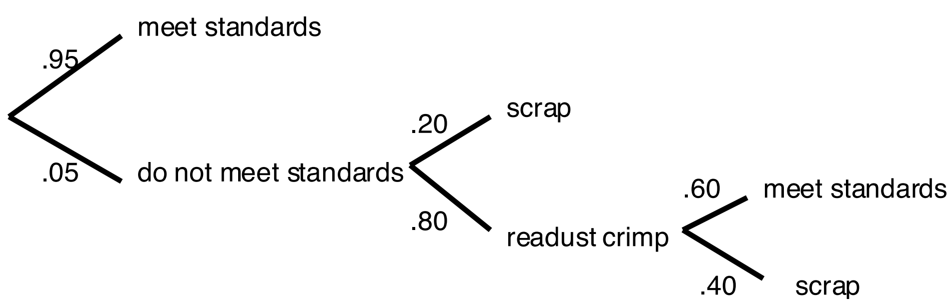
5.10. For i = 1, 2, 3, ... , 10, let A_i denote the event 'component i functions correctly'. The problem indicates that $P(A_i) = .999$ for each component which, by the law of complementary events, means that each $P(A_i') = 1-.999 = .001$. For a series system built from these 10 components to function correctly, *all ten* must function, so P(system functions correctly) = $P(A_1$ and A_2 and A_3 and ... and $A_k)$. According to the problem, this probability exceeds $1 - [P(A_1') + P(A_2') + P(A_3') + ... + P(A_k')] = 1 - [(.001) + (.001) + ... + (.001)] =$

1 - 10(.001) = 1 - .01 = .99. That is, there is *at least* a 99% probability that the system will function correctly.

5.11. Letting A_i denote the event that the i^{th} component fails (note that this is different from the definition of A_i used in problem 5.10), the probability that the entire series system fails is denoted by P(A_1 or A_2 or A_3 or ... or A_k). Given that each P(A_i) = .01, the problem states that P(A_1 or A_2 or A_3 or ... or A_k) $\leq$ P(A_1) + P(A_2) + P(A_3) + ... + P(A_k) = 5(.01) = .05. That is, there is *at most* a 5% chance of system failure.

Section 5.3

5.14. It is helpful to first put the probabilities in this exercise on the tree diagram of problem 5.5.

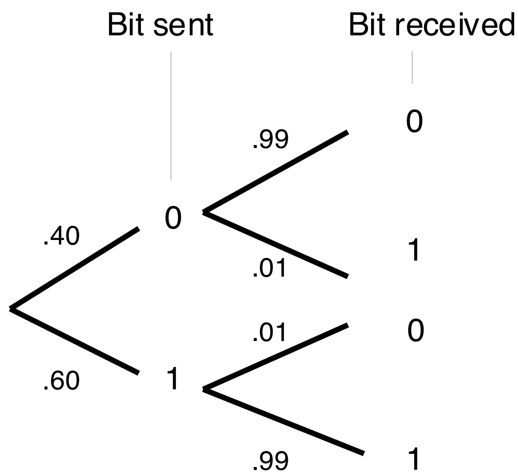


(a) The branches associated with passing the second inspection have probabilities .80 (those that require recrimping) and .60 (the recrimped fasteners that pass inspection). Recall that probabilities of following any branch on a tree diagram is simply the product of the probabilities of the sub-branches. Therefore, (.80)(.60) = .48, or, 48% of the fasteners that fail the initial inspection will go on to pass the second inspection. Be careful, this problem asks for the percentage of *failed* fasteners that eventually pass inspection, not the percentage of *all* fasteners initially submitted for inspection.

(b) From the tree diagram, the fasteners that pass inspection are those that either pass on first inspection (95%), or pass on second inspection after recrimping. These two events are mutually exclusive, so their probabilities can be added. Following the branches through the tree, the proportion of *all* fasteners that pass the second inspection is (.05)(.80)(.60) = .024. Therefore, 95% + 2.4% = 97.4% of all fasteners eventually pass inspection.

(c) Let I denote the event that a fastener passes inspection and let F denote the event that a fastener passes the first inspection. Then the problem asks for the conditional probability P(F | I). Using the conditional probability formula, P(F | I) = P(F and I)/P(I). Note that P(I) = .024 was calculated in part (b) of this question. Also, the event *F and I* can be simplified to F; i.e., *F and I = F* , so P(*F and I*) = P(F) = .95. The desired probability is then P(F | I) = .95/.974 = .9754, or 97.54%.

5.15. The probabilities are shown on the tree diagram of problem 5.6:



(a) The proportion/percentage of 1s received is the sum of the proportion of 0s that were sent and mistakenly received as 1s, (.40)(.01), plus the proportion of 1s that were sent and correctly received as 1s, (.60)(.99); i.e., (.40)(.01) + (.60)(.99) = .004 + .594 = .598, or 59.8%.

(b) Let S denote the event that a 1 is sent and let R be the event that a 1 is received. Then the problem asks for the conditional probability $P(S \mid R)$. The conditional probability formula gives $P(S \mid R) = P(S \text{ and } R)/P(R)$. Part (a) gives P(R) = .634. From the tree diagram P(S and R) is the product of the probabilities (.60)(.99) = .594. Therefore, $P(S \mid R)$ = .594/.634 = .937, or 93.7%.

5.16. The probabilities of independent events A and B must satisfy the equation P(*A and B*) = P(A)·P(B). If A and B were also mutually exclusive, then P(*A and B*) would equal 0, which would mean that P(A)·P(B) = P(*A and B*) = 0. This would require that at least one of A or B have zero probability of occurring. Although this is technically possible, most events of interest have non-zero probabilities, making P(A)·P(B) non-zero. It is therefore impossible for independent events with non-zero probabilities to be mutually exclusive.

5.18. (a) Since people's blood type is independent, $P(1_A \text{ and } 2_A) = P(1_A)P(2_A) = (.42)(.42) = .1764$

(b) $P(1_B \text{ and } 2_B) = (.10)^2 = .01$
$P(1_{AB} \text{ and } 2_{AB}) = (.04)^2 = .0016$
$P(1_0 \text{ and } 2_0) = (.44)^2 = .1936$

(c) P(matching blood types) = P(both have A) + P(both have B) + P(both have AB) + P(both have 0) = .1764 + .01 +.0016 + .1936 = .3816

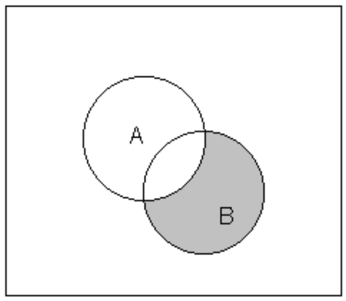
(c) Discriminating Power = P(the 2 people do not have matching blood types)
= 1 – P(the 2 people do have matching blood types)
= 1 – (.3816) = .6184

5.20. Let A denote the event that components 3 and 4 *both* work correctly and let B denote the event that *at least one* of components 1 or 2 works correctly. Then P(systems works) = P(A or B). From the general addition law, P(A or B) = P(A) +P(B) - P(A and B). Because all components act independently of one another, P(A) = P(3 and 4 work) = P(3 works)·P(4 works) = (.9)(.9) = .81. P(B) = P(1 or 2 works) = P(1 works) + P(2 works) - P(1 and 2 work) = .9 + .9 - (.9)(.9) = .99. Finally, events A and B are independent

5.23 (a) Let S_i denote the event that the i^{th} point signals a problem with the manufacturing process. Then, the probability that *none* of the 10 points give such a signal is P(no signals) = P(S_1' and S_2' and ... S_{10}') = P(S_1')·P(S_2')· ...P(S_{10}') = $(1-.01)^{10}$ = .90438. Therefore, the probability of having *at least one* point signal a problem is P(at least one signal) = 1 - P(no signals) = 1 - .90438 = .0956.

(b) P(at least one in 25 signals a problem) = 1 - $(1-.01)^{25}$ = .2222.

5.24 The event *A′ and B* is shaded in the following Venn diagram:



From the diagram you can also see that the mutually exclusive events *A and B* and *A′ and B* comprise event B; i.e., B = (*A and B*) or (*A′ and B*). Using the addition law for exclusive events, P(B) = P(*A and B*) + P(*A′ and B*), which can be rearranged as P(*A′ and B*) = P(B) - P(*A and B*). Using the fact that A and B are independent, P(*A and B*) = P(A)· P(B), so, P(*A′ and B*) = P(B) - P(*A and B*) = P(B) - P(A)· P(B) = [1-P(A)]·P(B) = P(A′)·P(B), which shows that A′ and B are independent.

Section 5.4

34. (a) x = measurements of the length of an object
σ = 1 mm

$$P(-2 \le x-\mu \le 2) = P\left(\frac{-2}{\sigma} \le \frac{x-\mu}{\sigma} \le \frac{2}{\sigma}\right) = P(-2 \le z \le 2)$$
$$= P(z \le 2) - P(z \le -2) = .9772 - .0228 = .9544$$

(b) When σ = .5 mm, $P(-2 \le x-\mu \le 2) = P(-4 \le z \le 4) \approx 1$

Section 5.5

5.43. (a) x has a binomial distribution with n = 5 and π = .05. Writing P(.05-.01 ≤ p ≤ .05+.01) in terms of x, we find P(.05-.01 ≤ x/5 ≤ .05+.01) = P((.04)5 ≤ x ≤ (.06)5) = P(.2 ≤ x ≤ .3). Because x can only have integer values, there is no x between .2 and .3, so the probability of this event is 0.

(b) For n= 25, P(.04 ≤ p ≤ .06) = P((.04)25 ≤ x ≤ (.06)25) = P(1 ≤ x ≤ 1.5) = P(x = 1) = $\binom{25}{1}(.05)^1(.95)^{24}$ = 0.36498.

5.45. (a) x = 'disconnect force' has a uniform distribution on the interval [2,4]. M is the maximum of a sample of size n = 2 from the uniform density on [2,4]. The larger of two items randomly selected from the interval [2,4] should, on average, tend to be closer to the upper end of the interval.

(b) Using the same reasoning as in part (a), the largest value in a sample of n = 100 will, most likely, be even closer to the upper end of the interval [2,4] than is the largest value in a sample of size n =2. So the average of all M's based on n=100 ought to be larger than the average value of all M's based on n = 2.

(c) For larger samples (e.g., n = 100), the maximum values will usually be fairly close to the upper endpoint of 4, which means that the variability amongst such values will tend to be small. For smaller samples (e.g., n = 2), it is easier for the value of M to wander over the interval [2,4], which means that the variability among these values will be larger than for n = 100.

Section 5.6

5.46

x = diameter of a piston ring $\mu = 12\ cm \quad \sigma = .04\ cm$

(a) $\mu_{\bar{x}} = \mu = 12\ cm$

$$\sigma_{\bar{x}} = \left(\frac{\sigma}{\sqrt{n}}\right) = \left(\frac{.04}{\sqrt{n}}\right)$$

(b) When n = 64 $\mu_{\bar{x}} = 12\ cm \quad \sigma_{\bar{x}} = \left(\frac{.04}{\sqrt{64}}\right) = .005$

(c) The mean of a random sample of size 64 is more likely to lie within .01 cm of μ, since $\sigma_{\bar{x}}$ is smaller.

5.47.

(a) $\mu_p = \pi = .80.$ $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.80(1-.80)}{25}} = .08$

(b) Since 20% do <u>not</u> favor the proposed changes (so $\pi = .20$), the mean & standard deviation of the sampling distribution of this proportion are $\mu_p = \pi = .20$ and $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.20(1-.20)}{25}} = .08.$

(c.) For n = 100 and $\pi = .80$, $\mu_p = \pi = .80.$ $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{.80(1-.80)}{100}} = .04.$ Notice that it was necessary to <u>quadruple</u> the sample size (from n=25 to n=100) in order to cut σ_p in half (from $\sigma_p = .08$ to $\sigma_p = .04$).

5.50. (a) x = the weight of a bag of fertilizer

$\mu = 50$ lbs $\quad \sigma = 1$ lb $\quad$ n = 100

$$P(49.75 \le \bar{x} \le 50.25) \approx P\left(\frac{49.75-50}{1/\sqrt{100}} \le z \le \frac{50.25-50}{1/\sqrt{100}}\right)$$

$$= P(-2.5 \le z \le 2.5) = P(z \le 2.5) - P(z \le -2.5)$$

$$= .9938 - .0062 = .9876$$

(b) μ = 49.8 lbs σ = 1 lb n = 100

$$P(49.75 \le x \le 50.25) \approx P(-0.5 \le z \le 4.5) = P(z \le 4.5) - P(z \le -0.5)$$

$$= 1 - .3085 = .6915$$

5.51

(a)

x = 'lifetime of battery" has a normal density with μ = 8 hours and σ = 1 hour. Therefore, P(average of 4 exceeds 9 hours) = P($\bar{x}$ > 9) = P(z > $\frac{9-8}{1/\sqrt{4}}$) =

P(z > 2) = 1 - P(z<2) = 1 - .9772 = .0228.

(b) Having the total lifetime of 4 batteries exceeds 36 hours is the same thing as having their average exceed 9, so the probability of this event is .0228, the same as in part (a).

(c) .95 = P(T > T_0) = P(T/4 > T_0/4) = P($\bar{x}$ > T_0/4) = P(z > $\frac{T_0/4-8}{1/\sqrt{4}}$) =

P(z > T_0/2-16). For a standard normal distribution, P(z > -1.645) ≈ .95, so we must have T_0/2-16 = -1.645, which gives T_0 = 28.71 hours.

5.54. (a) $\mu = \sum xp(x) = (0)(.8) + (1)(.1) + (2)(.05) + (3)(.05) = .35$

$$\sigma = \sqrt{\sum (x-\mu)^2 p(x)}$$

$$= \sqrt{(0-.35)^2(.8) + (1-.35)^2(.1) + (2-.35)^2(.05) + (3-.35)^2(.05)}$$

$$= .7921$$

(b) $n = 64 \quad \mu_{\bar{x}} = \mu = .35$

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = (.7921/\sqrt{64}) = .099$$

(c) $P(\bar{x} > 1) \approx P\left(z > \frac{1-.35}{.099}\right) = P(z > 6.57) \approx 0$

5.55

(a) Let p = 'proportion of resistors exceeding 105 Ω'. Then the sampling distribution of p is approximately normal with $\mu_p = \pi = 0.02$ and $\sigma_p = \sqrt{\dfrac{\pi(1-\pi)}{n}} = \sqrt{\dfrac{.02(1-.02)}{100}} = 0.014$.

(b) $P(p < .03) = P(z < \dfrac{.03-.02}{.014}) = P(z < .71) = 0.7611$.

5.56

x = length of an object σ = 1 mm n = 2

$$P(-2 \le x-\mu \le 2) = P\left(\frac{-2}{\sigma/\sqrt{n}} \le z \le \frac{2}{\sigma/\sqrt{n}}\right)$$

$$= P(-2.83 \le z \le 2.83) = P(z \le 2.83) - P(z \le -2.83)$$

$$= .9977 - .0023 = .9954$$

Supplementary Problems – Chapter 5

5.59

(a) The sum of the parcel areas is 15+20+…+20 = 90, so $P(B_1 \text{ or } B_2 \text{ or } B_3) = P(B_1) + P(B_2) + P(B_3) = 15/90 + 20/90 + 25/90 = 60/90 = 2/3$.

(b) $P(B_5') = 1 - P(B_5) = 1 - 20/90 = 7/9$.

5.60 A = {assembly fails}
B = {C_1 fails}

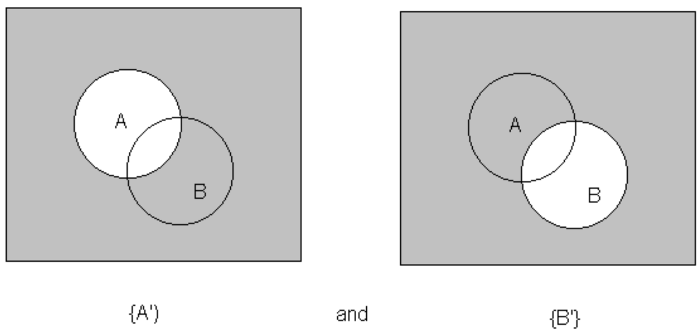
(a) $A \mid B$ means that: given component C_1 fails, the entire assembly fails.

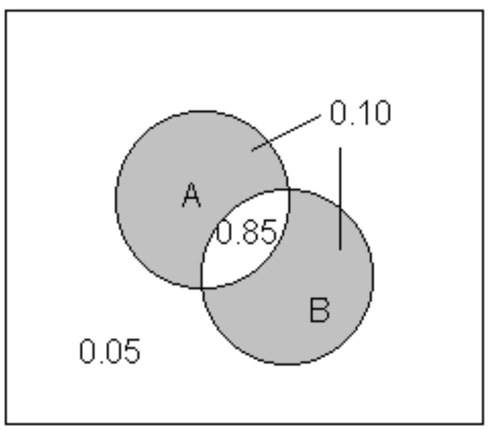
$B|A$ means that: given the entire assembly fails, component C_1 fails.

(b) We know that if a single component fails, the entire assembly will fail; thus, $P(A|B)=1$. We also know that if the entire assembly fails, it may or may not be due to the failure of component C_1, thus $P(B|A)<1$. So: $P(A|B) \neq P(B|A)$.

5.61. Let B_i denote the event that the i^{th} battery operates correctly. Therefore, $P(B_i) = .95$ for each i = 1,2,3,4. Then, P(tool fails) = P(at least one battery fails) = 1 - P(all batteries operate correctly) = 1 - P(B_1 and B_2 and B_3 and B_4) = $1 - P(B_1)P(B_2)P(B_3)P(B_4) = 1 - (.95)^4 = .1855$.

5.62.





The probability that both pumps fail on a given day is .85

5.63

P(at least one event occurs) = P(A or B) = 1 - P(neither event occurs) = 1 - P(A' and B') = 1-P(A')P(B') = (1-P(A)(1-P(B)). We have used the fact that A' and B' are independent to simplify P(A' and B').

5.69. (a) x = 'battery voltage' has a mean value of $\mu = 1.5$ and a standard deviation of $\sigma = .2$ volts. The sampling distribution of $\bar{x}$ (based on n = 4) has a mean value of $\mu_{\bar{x}} = \mu = 1.5$ and a standard error of $\sigma_{\bar{x}} = \sigma/\sqrt{n} = .2/\sqrt{4} = .1$

(a) T is related to $\bar{x}$ by the equation $T = 4\bar{x}$, so the mean of T should be 4 times the mean of $\bar{x}$. That is, $\mu_T = 4\mu_{\bar{x}} = 4(1.5) = 6$. Similarly, the standard deviation of T should be 4 times that of $\bar{x}$, so $\sigma_T = 4\sigma_{\bar{x}} = 4(.1) = .4$

5.70

x = resistance n = 5 $\mu = 100\ ohms \quad \sigma = 1.7\,ohms$

First, we must assume that the population density for x is symmetric. Then, we can use the Central Limit Theorem to proceed.

(a) Note: The question should read: "What is the probability that the *average* resistance in the circuit exceeds 105 ohms?"

$$P(\bar{x} > 105) \approx P\left(\bar{x} > \frac{105-100}{1.7/\sqrt{5}}\right) = P(z > 6.58) \approx 0$$

(b) $P(T > 511) + P(T < 489) = P\left(\bar{x} > \frac{511}{5}\right) + P\left(\bar{x} < \frac{489}{5}\right)$

$$P(\bar{x} > 102.2) + P(\bar{x} < 97.8) = P\left(z > \frac{102.2-100}{1.7/\sqrt{5}}\right) + P\left(z < \frac{97.8-100}{1.7/\sqrt{5}}\right)$$

$= P(z > 2.89) + P(z < -2.89) = 1 - P(z < 2.89) + P(z < -2.89)$

$= 1 - .9981 + .0019 = .0038$

(c) We know that: $P(-1.96 \le z \le 1.96) = .95$

So, $\dfrac{\bar{x}-100}{1.7/\sqrt{n}} = 1.96 \Rightarrow \bar{x} = 100 + 1.96\left(\dfrac{1.7}{\sqrt{n}}\right)$

Also, $T = 510 = n\bar{x} \Rightarrow n = \dfrac{T}{\bar{x}} = \dfrac{510}{\bar{x}}$

Thus: $n = \dfrac{510}{100 + 1.96\left(\dfrac{1.7}{\sqrt{n}}\right)}$

Solving for n produces a value just over 5.

5.73. (a) $P(\text{line}\,1) = \dfrac{500}{1500} = \dfrac{1}{3} \approx .333$

$P(\text{crack}) = .50(500/1500) + (.44)(400/1500) + (.40)(600/1500) = .444$

(b) By reading down the column titled Line 1, it follows that $P(\text{blemish}\,|\,1) = .15$.

(c) $P(1\,|\,\text{surface defect}) = \dfrac{P(1\text{ and surface defect})}{P(\text{surface defect})}$,

where $P(\text{surface defect}) = (.10)(500/1500) + (.08)(400/1500) + (.15)(600/1500) = .11467$.

So $P(1\,|\,\text{surface defect}) = \dfrac{P(1\text{ and surface defect})}{P(\text{surface defect})} = \dfrac{(.10)(500/1500)}{.11467} \approx .2907$

Chapter 7
Estimation and Statistical Intervals

Section 7.1

7.4. (a) Recall: $\mu_p = \pi$ and $\sigma_p = \sqrt{\dfrac{\pi(1-\pi)}{n}}$

Even though π is unknown, we can still set an upper bound on σ_p.

When $\pi = .5$, σ_p is maximized. So, $\sigma_p = \dfrac{1}{2\sqrt{n}}$.

In our case, $n = 10 \Rightarrow \sigma_p = .15811$

So, $P(-.10 < (p - \pi) < .10) = P\left(\dfrac{-.10}{.15811} < z < \dfrac{.10}{.15811}\right)$

$= P(-.63 < z < .63) = .4714$

Section 7.2

7.7. (a) The entry in the 3.0 row and .09 column of the z table is .9990. Similarly, the entry for -3.09 is .0010. Therefore, the area under the z curve between -3.09 and +3.09 is .9990 - .0010 = .9980. The confidence level is then 99.8%.

(b) Following the example in part (a), the z-table entries corresponding to z = -2.81 and z = +2.81 are .9975 and .0025, respectively. Therefore the area between these two z values is .9975 - .0025 = .9950. The confidence level is then 99.5%.

(c) The z-table entries corresponding to z = -1.4 and z = +1.44 are .9251 and .0749, respectively. Therefore, the area between these two z values is .9251 - .0749 = .8502. The confidence level is then 85.02%.

(d) The coefficient of s/ $\sqrt{n}$ is not written, but is understood to be 1.00. The z-table entries corresponding to z = -1.001 and z = +1.00are .8413 and .1587, respectively. Therefore, the area between these two z values is .8413 - .1587 = .6826. The confidence level is then 68.26%.

7.8. (a) 98% of the standard normal curve area must be captured. This requires that 1% of the area is to be captured in each tail of the distribution. So, P(Z < -z critical value) = .01 and P(Z > z critical value) = .01. Thus, the z critical value = 2.33.

(b) 85% of the standard normal curve area must be captured. This requires that 7.5% of the area is to be captured in each tail of the distribution. So, P(Z < -z critical value) = .075 and P(Z > z critical value) = .075. Thus, the z critical value = 1.44.

(c)75% of the standard normal curve area must be captured. This requires that 12.5% of the area is to be captured in each tail of the distribution. So, P(Z < -z critical value) = .125 and P(Z > z critical value) = .125. Thus, the z critical value = 1.15.

(d) 99.9% of the standard normal curve area must be captured. This requires that .05% of the area is to be captured in each tail of the distribution. So, P(Z < -z critical value) = .0005 and P(Z > z critical value) = .0005. Thus, the z critical value is conservatively 3.32.

(Notice that, in this case, there are several possible z critical values listed in the standard normal table.)

7.10. (a) The sample mean is the mid-point of each of the confidence intervals. So,

$$\left(\frac{115.6+114.4}{2}\right)=115$$

To confirm: $\left(\frac{115.9+114.1}{2}\right)=115$

The sample mean resonance frequency is 115 Hz.

(b) The first interval has the 90% confidence level, (114.4, 115.6). We know this because it is the more narrow of the two intervals and a 90% confidence interval will be more narrow than a 99% confidence interval, given the same sample data.

7.11. (a) Decreasing the confidence level from 95% to 90% will decrease the associated z value and therefore make the 90% interval narrower than the 95% interval. *(Note: see the answer to Exercise 9 above)*

The statement is not correct. Once a particular confidence interval has been created/calculated, then the true mean is either in the interval or not. The 95% refers to the process of creating confidence intervals; i.e., it means that 95% of all the possible confidence intervals you could create (each based on a new random sample of size n) will contain the population mean (and 5% will not).

The statement is not correct. A confidence interval states where plausible values of the population mean are, not where the individual data values lie. In statistical inference, there are three types of intervals: **confidence intervals** (which estimate where a population mean is), **prediction intervals** (which estimate where a single value in a population is likely to be), and **tolerance intervals** (which estimate the likely range of values of the items in a population. The statement in this exercise refers to the likely range of all the values in the population, so it is referring to a tolerance interval, not a confidence interval.

(d) No, the statement is not exactly correct, but it is close. We *expect* 95% of the intervals we construct to contain μ, but we also expect a little variation. That is, in any group of 100 samples, it is possible to find only, say, 92 that contain μ. In another group of 100 samples, we might find 97 that contain μ, and so forth. So, the 95% refers to the *long run* percentage of intervals that will contain the mean. 100 samples/intervals is <u>not</u> the long run.

7.14. (a) A 95% two-sided confidence interval for the true average dye-layer density for all such trees is:

$$\bar{x} \pm (1.96)\left(\frac{s}{\sqrt{n}}\right)$$

$$1.28 \pm (1.96)\left(\frac{.163}{\sqrt{69}}\right)$$

$$1.028 \pm 0.03846$$

$$(.9895, 1.0665)$$

Interpretation 1: We are 95% confident that the (true) average is between 0.9895 and1.0665.

Interpretation 2: There is a 95% probability that a random 95% CI, computed using our formula, will yield an interval that covers the (true) average.

(b) $n = \left[\frac{1.96s}{B}\right]^2 = \left[\frac{1.96(.16)}{.025}\right]^2 \approx 158$

A sample size of 158 trees would be required.
(Note: The researchers wanted an interval width of .05. So, the bound on the error of estimation, B, is half of the width. B = .025)

7.17

(a) The entry in the z table corresponding to z = .84 is .7995, so the confidence level is 79.95% or, approximately, 80%.

(b) The entry in the z table corresponding to z = 2.05 is .9798, so the confidence level is 97.98% or, approximately, 98%.

(c) The entry in the z table corresponding to z = .67 is .7486, so the confidence level is 74.86% or, approximately, 75%.

7.18
A 95% upper confidence bound for the true average charge-to-tap time is:

$$\bar{x} + (1.645)\left(\frac{s}{\sqrt{n}}\right)$$

$$382.1 \pm (1.645)\left(\frac{31.5}{\sqrt{36}}\right)$$

$$(382.1 + 8.64) = 390.74 \text{ min}$$

That is, with 95% confidence, the value of μ lies in the interval (0 min, 390.74 min).

7.20
A 90% lower confidence bound for the true average shear strength is:

$$\bar{x} - (1.28)\left(\frac{s}{\sqrt{n}}\right)$$

$$4.25 - (1.28)\left(\frac{1.30}{\sqrt{78}}\right)$$

$$(4.25 - .188) = 4.062\, kip$$

That is, with 90% confidence, the value of μ lies in the interval (4.062, ∞).

Section 7.3

7.22. In this problem, we have $n = 507$ and $p = 142/507 \approx .28008$. For a two-sided 99% confidence interval, we use $z^* = 2.576$:

$$\frac{p+\frac{(z^*)^2}{2n} \pm z^*\sqrt{\frac{p(1-p)}{n}+\frac{(z^*)^2}{4n^2}}}{1+\frac{(z^*)^2}{n}} = \frac{.28008+\frac{2.576^2}{2(507)} \pm 2.576\sqrt{\frac{.28008(1-.28008)}{507}+\frac{2.576^2}{4(507^2)}}}{1+\frac{2.576^2}{507}}$$

$$= \frac{.28662 \pm 2.576(.0201)}{1.013078} \Rightarrow (.231, .334)$$

7.23. Here, we have $n = 539$ and $p = 133/539 \approx .24675$. For a 95% lower confidence bound, we use the value of $z^* = 1.645$:

$$\frac{p+\frac{(z^*)^2}{2n} - z^*\sqrt{\frac{p(1-p)}{n}+\frac{(z^*)^2}{4n^2}}}{1+\frac{(z^*)^2}{n}} = \frac{.24675+\frac{1.645^2}{2(539)} - 1.645\sqrt{\frac{.24675(1-.24675)}{539}+\frac{1.645^2}{4(539^2)}}}{1+\frac{1.645^2}{539}}$$

$$= \frac{.2493 - 1.645(.0186)}{1.0051} = .217 \Rightarrow \text{lower bound is .217.}$$

We are 95% confident that at least 21.7% of all households in this Midwestern city own firearms.

7.24. Here, we have $n = 487$ and $p = 7.2\% = .072$. For a 99% upper confidence bound, we use the value $z^* = 2.33$:

$$\frac{p+\frac{(z^*)^2}{2n} + z^*\sqrt{\frac{p(1-p)}{n}+\frac{(z^*)^2}{4n^2}}}{1+\frac{(z^*)^2}{n}} = \frac{.072+\frac{2.33^2}{2(487)} + 2.33\sqrt{\frac{.072(1-.072)}{487}+\frac{2.33^2}{4(487^2)}}}{1+\frac{2.33^2}{487}}$$

$$= \frac{.0776 + 2.33(.012)}{1.011} = .104 \Rightarrow \text{upper bound is .104.}$$

We are 99% confident that at most 10.4% of all births by nonsmoking women in this metropolitan area result in low birth weight.

7.25. (a) Following the same format used for most confidence intervals, i.e., **statistic ± (critical value) (standard error),** an interval estimate for $\pi_1 - \pi_2$ is:

$$(p_1 - p_2) \pm z\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

(b) The response rate for no-incentive sample: $p_1 = 75/110 = .6818$, while the return rate for the incentive sample is $p_2 = 66/98 = .6735$. Using $z = 1.96$ (for a confidence level of 95%), a two-sided confidence interval for the true (i.e., population) difference in response rates $\pi_1 - \pi_2$ is:

$$(.6818 - .6735) \pm (1.96)\sqrt{\frac{.6818(1-.6818)}{110} + \frac{(.6735)(1-.6735)}{98}}$$

$= .0083 \pm .1273 = (-.119, .1356)$.

The fact that this interval contains 0 implies that we cannot say anything about the equality of the proportions. In other words, including incentives in the questionnaires may or may not have a significant effect on the response rate.

(c) Let $\tilde{p}_i$ denote the sample proportion by adding 1 success and 1 failure to the i^{th} sample. We calculate $\tilde{p}_i = (x+1)/(n+2)$, where x is the number of successes (or failures, whichever is desired) in the sample. Then: $\tilde{p}_1 = (75+1)/(110+2) = .67857$ and $\tilde{p}_2 = (66+1)/(98+2) = .67$. Using the format of the equation given in part (a) above, we have the following 95% confidence interval:

$$(\tilde{p}_1 - \tilde{p}_2) \pm z\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

$$(.67857 - .67) \pm 1.96\sqrt{\frac{.67857(1-.67857)}{110+2} + \frac{.67(1-.67)}{98+2}}$$

$$(.00857) \pm 1.96\sqrt{.004158} \;\Rightarrow\; .00857 \pm .1264 \;\Rightarrow\; (-.1178,\ .1350)$$

This interval contains 0; so, it is not clear if including incentives in the questionnaire has an effect on the response rate. This is the same conclusion as in part (b).

7.27. (a) As in Exercise 25, the usual confidence interval format *statistic ± (critical value)(standard error)* gives a confidence interval for:

$\ln(\pi_1/\pi_2)$ of: $\ln(p_1/p_2) \pm z\sqrt{\frac{n_1-u}{n_1 u} + \frac{n_2-v}{n_2 v}}$.

(b) Since we want to estimate the ratio of return rates for incentive group to the non-incentive group, we will call group 1 the incentive group (to match the subscripts in the formula above). The number of returns for the non-incentive group is v = 75 out of n_2 = 110, so p_2 = .6818. For the incentive group, u = 78 out of n_1 = 100, so p_1 = .78. The 95% confidence interval for $\ln(\pi_1/\pi_2)$ is:

$$\ln(.7800/.6818) \pm (1.96)\sqrt{\frac{100-78}{100(78)} + \frac{110-75}{110(75)}} = .1346 \pm .1647$$

= [-.0301, .2993]. To find a 95% interval for π_1/π_2, we exponentiate both end points of this interval, $[e^{-.0301}, e^{.2993}]$ = [.9702, 1.3489], or, about [.970, 1.349]. Because the 95% confidence interval includes 1, then at 95% confidence level, we cannot say anything about the equality of pi1 and pi2. As in problem 25, there is insufficient evidence from this data to suggest that incentive has an effect on the questionnaire return rate. In other words, we don't know if incentive has an effect or not.

7.28

$$n = \pi(1-\pi)\left[\frac{2.575}{B}\right]^2$$

Since an estimate of π is not provided, a conservative estimate is .50. (However, it seems improbable that as many as 50% of the coffeepot handles will be cracked.)

$$n = (.50)(.50)\left[\frac{2.575}{.1}\right]^2$$

$$n \approx 166$$

A sample of 166 coffeepot handles from the shipment should be inspected.

7.29

For 90% confidence, the associated z value is 1.645. Since nothing is known about the likely values of π we use .25, the largest possible value of π(1-π), in the sample size formula: n = $(.25)\left(\frac{1.645}{.05}\right)^2$ = 270.6. To be conservative, we round this value up to the next highest integer and use n = 271.

7.33

Let μ_1 denote the average toughness for the high-purity steel and let μ_2 denote the average toughness for the commercial purity steel. Then, a lower 95% confidence bound for μ_1-μ_2 is given by: ($\bar{x}_1$ - $\bar{x}_2$) - $z\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ = (65.6-59.2) - (1.645) $\sqrt{\frac{(1.4)^2}{32}+\frac{(1.1)^2}{32}}$ = 6.4 - .518 = 5.882. Because this lower interval bound exceeds 5, it gives a reliable indication that the difference between the population toughness levels does exceed 5.

7.34

We need the following equation to be true:

$$B=(\text{z critical value})\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}}$$

So, $.5=(1.96)\sqrt{\frac{2^2}{n}+\frac{2^2}{n}}$

Solving for n: $n\approx 123$

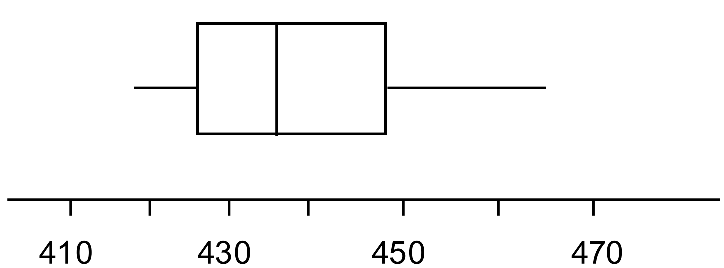
A sample of 123 batteries of each type should be taken.

Section 7.4

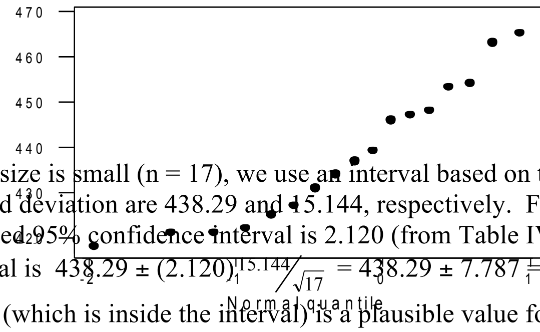
7.39

Recall that this text's definition of upper and lower quartiles may differ a little from the definitions used by some other authors. <u>We define the upper and lower quartiles to be the medians of the lower and upper halves of the data</u> (for n odd, the median is included in both halves). This usually gives results that are very close to the actual (or estimated) quartiles used by other authors, but there are often small differences. For example, in this exercise the medians of the lower and upper halves of the data are: lower quartile = median of lower half (including the median, 437) = 425; upper quartile = median of the upper half (including median) = 448. The software package Minitab, however, estimates the lower quartile as the $.25(n+1)^{th} = .25(17+1)^{th} = 4.5^{th}$ item in the sorted list (here, 422 and 425 are the 4^{th} and 5^{th} items in the sorted list, and the 4.5^{th} value is defined to be their average, which is 423.5). So, be careful when comparing your answers to those in various software packages.

(a) Using 425 and 448 as the lower and upper quartiles, the IQR = 448 - 425 = 23. To check for outliers, we calculate the values 425 + 1.5(IQR) = 425 - 1.5(23) = 390.5 and 448 + 1.5(IQR) = 448 - 1.5(23) = 482.5. Since the maximum and minimum in the data are 465 and 418, which are inside the limits just calculated, there are no outliers in the data. The median of the data is 437 and a boxplot of the data appears below:



(b) A quantile plot can be used to check for normality. Refer to the answer to Exercise 43 of Chapter 2 for an easy method of creating a quantile plot in Minitab. The quantilile plot below shows a fairly strong linear pattern, supporting the assumption that this data came from a normal population:



(c) Since the sample size is small (n = 17), we use an interval based on the t distribution. The sample mean and standard deviation are 438.29 and 15.144, respectively. For n -1 = 17 -1 = 16 df, the critical t value for a 2-sided 95% confidence interval is 2.120 (from Table IV). Therefore, the desired confidence interval is $438.29 \pm (2.120)\,{}^{15.144}\!/\!\sqrt{17} = 438.29 \pm 7.787 =$ [430.5, 446.1]. This interval suggests that 440 (which is inside the interval) is a plausible value for the mean polymerization. The value of 450, however, is not plausible, since it lies outside the interval.

7.40.

Given: n = 14

$\bar{x} = 8.48$

$s = .79$

(a) A 95% lower confidence bound for the true average proportional limit stress of all such joints is:

$$\bar{x} - (\text{t critical value})\left(\frac{s}{\sqrt{n}}\right)$$

The t critical value is obtained from Table IV with df = (n – 1) = 13 for a one-sided interval.

$$8.48 - (1.771)\left(\frac{.79}{\sqrt{14}}\right)$$

$$8.48 - .37 = 8.11$$

That is, with a confidence level of 95%, the value of μ lies in the interval $(8.11, \infty)$.

Our above confidence limit is valid only if the distribution of proportional limit stress is normal.

7.42. Given: n = 26

$\bar{x} = 370.69$

$s = 24.36$

(a) A 95% upper confidence bound for the population mean escape time is:

$$\bar{x} + (\text{t critical value})\left(\frac{s}{\sqrt{n}}\right)$$

The t critical value is obtained from Table IV with df = (n – 1) = 25 for a one-sided interval.

$$370.69 - (1.708)\left(\frac{24.36}{\sqrt{26}}\right)$$
$$370.69 + 8.16 = 378.85$$

That is, with a confidence level of 95%, the value of μ, the population mean escape time, lies in the interval $(0,\ 378.85)$.

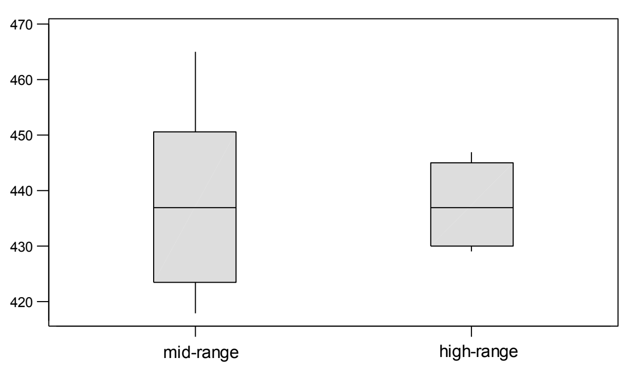
7.45. (a) Confidence level = area between -.687 and 1.725 = .95 - .25 = .70, or, 70%. Note that the value 1.725 is in Table IV and is associated with a cumulative area of .95 (for df = 20).

(b) Confidence level = area between -.860 and 1.325 = .90 - .20 = .70, or, 70%. Note that the value 1.325 is in Table IV and is associated with a cumulative area of .90 (for df = 20).

(c) Confidence level = area between -1.064 and 1.064 = (1-2(.15)) = .70, or, 70%. Note that the symmetry of the t distribution allows us to say there is evidence that the area to the left of 1.064 is also .15.

All three intervals have 70% confidence, but they have different widths. The interval in part (c) has the smallest width of 2(1.064) $s/\sqrt{n}$, and is therefore the best choice among the three.

Section 7.5
7.48
(a)



The most notable feature of these boxplots is the larger amount of variation present in the mid-range data as compared to the high-range data. Otherwise, both boxplots look reasonably symmetric and there are no outliers present.

(b) Minitab output:

	n	sample mean	sample standard deviation
Mid-range	17	438.3	15.1

High-range	11	437.45	6.83

A 95% Confidence Interval for (μ mid range - μ high range) is

(-7.9 , 9.6)

(Note: df = 23.)

The above analysis was performed by Minitab. The confidence interval was computed as follows:

$$(438.3-437.45)\pm(2.069)\sqrt{\frac{(15.1)^2}{17}+\frac{(6.83)^2}{11}}$$

using df = 23, resulting in:

$$.85\pm 8.69$$
$$(-7.84, 9.54)$$

Since plausible values for $(\mu_1 - \mu_2)$ are both positive and negative (i.e., the interval spans zero) we would say that there is not sufficient evidence from data to suggest that μ_1 and μ_2 differ.

7.49. The approximate degrees of freedom for this estimate are: df $\approx \dfrac{\left[\frac{11.3^2}{6}+\frac{8.3^2}{8}\right]^2}{\left[\frac{11.3^2}{6}\right]^2\Big/5+\left[\frac{8.3^2}{8}\right]^2\Big/7}$ = 893.586/101.175 = 8.83, so we round down and use df $\approx$ 8. For a 95% 2-sided confidence interval with 8 df, the critical t value is 2.306, so the desired interval is: ($\bar{x}_1$ - $\bar{x}_2$) - t $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ = (40.3-21.4) ± (2.306) $\sqrt{\frac{11.3^2}{6}+\frac{8.3^2}{8}}$ = 18.9 ± (2.306)(5.4674) = 18.9 ± 12.607 = [6.3, 31.5]. Because 0 is not contained in this interval, there is strong evidence that μ_1-$\mu_2 \neq 0$; i.e., we can say that there is evidence from data in favor of the claim that the population means are not equal. Calculating a confidence interval for μ_2-μ_1 would change only the order of subtraction of the sample means (i.e., $\bar{x}_2$ - $\bar{x}_1$ = -18.9), but the standard error calculation would give the same result as before. Therefore, the 95% interval estimate of μ_2-μ_1 would be [-31.5, -6.3], just the negatives of the endpoints of the interval for μ_1-μ_2. Since 0 is not in this interval, we reach exactly the same conclusion as before: the population means are not equal.

7.50. (a) A 99% confidence interval for the difference between the true average load for the fiberglass beams and that for the carbon beams is:

$$(33.4-42.8)\pm(\text{t critical value})\sqrt{\frac{(2.2)^2}{26}+\frac{(4.3)^2}{26}}$$

where df = 37.

Since Table IV does not contain a row for 37 degrees of freedom, one can either interpolate or obtain the exact t critical value. The t critical value obtained from Minitab is 2.7154.

Thus, the 99% confidence interval for $(\mu_1 - \mu_2)$ is:

$$-9.4 \pm (2.7154)(.94726)$$
$$-9.4 \pm 2.6$$
$$(-12, -6.8)$$

We are very confident that the true average load for carbon beams exceeds that for fiberglass beams by between 6.8 and 12 kN.

(b) The upper limit of the interval calculated in part (a) does not give a 99% upper confidence bound for $(\mu_1 - \mu_2)$. A 99% upper confidence bound is:

$$-9.4 + (2.4314)(.94726)$$
$$-9.4 + 2.3 = -7.1$$

where the t critical value was obtained from Minitab using the t-distribution with 37 degrees of freedom and a cumulative probability of .99.

So, with a confidence level of 99%, the value of $(\mu_1 - \mu_2)$ lies below - 7.1. Meaning that the true average load for carbon beams exceeds that for fiberglass beams by at least 7.1kN.

7.52. (a) This exercise calls for a paired analysis. First, compute the difference between indoor and outdoor concentrations of hexavalent chromium for each of the 33 houses. These 33 differences are summarized as follows:

$n = 33 \quad \bar{d} = -.4239 \quad s_d = .3868$, where d = (indoor value – outdoor value)

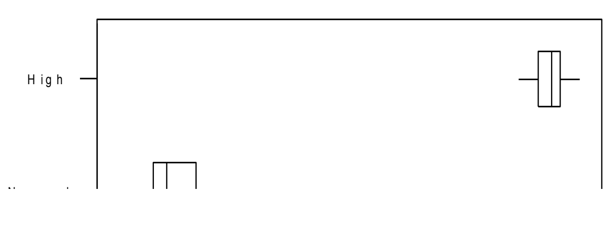
Then the t critical value based on df = (n – 1) = 32 must be determined. Table IV does not contain a value for 32 degrees of freedom, so one can either interpolate or obtain the exact t critical value. The t critical value obtained from Minitab is 2.0369.

Thus, a 95% confidence interval for the population mean difference between indoor and outdoor concentrations is:

$$-.4239 \pm (2.0369)\left(\frac{.3868}{\sqrt{33}}\right)$$
$$-.4239 \pm .13715$$
$$(-.5611, -.2868)$$

We can be highly confident, at the 95% confidence level, that the true average concentration of hexavalent chromium outdoors exceeds the true average concentration indoors by between .2868 and .5611 nanograms/m^3.

7.53. (a) The median of the 'Normal' data is 46.80 and the upper and lower quartiles are 45.55 and 49.55, which yields an IQR of 49.55 - 45.55 = 4.00. The median of the 'High' data is 90.1 and the upper and lower quartiles are 88.55 and 90.95, which yields an IQR of 90.95 - 88.55 = 2.40.



The most significant feature of these boxplots is the fact that their locations (medians) are far apart.

(b) This data is paired because the two measurements are taken for each of 15 test conditions. Therefore, we have to work with the differences of the two samples. A quantile plot of the 15 differences shows that the data follows (approximately) a straight line, indicating that it is reasonable to assume that the differences follow a normal distribution. Taking differences in the order 'Normal' - 'High', we find $\bar{d}$ = -42.23 and s_d = 4.34. A 95% confidence interval for the difference between the population means is given by $\bar{d} \pm t\, {}^{s_d}\!/\!_{\sqrt{n}}$, where n = 15 is the number of pairs and the critical t value is t = 2.145, based on n-1 = 14 df. The desired interval is: -42.23 ± (2.145) ${}^{4.34}\!/\!_{\sqrt{15}}$ = -42.23 ± 2.404 = [-44.63, -39.83]. Because 0 is not contained in this interval, we can say that there is evidence in data of the claim that the difference between the population means is not 0; i.e., we can say there is evidence from data of the claim that the two population means are not equal.

7.54. (a) Yes, there is doubt about the normality of the population distribution of the differences. In about the middle of the plot there is a "jump" that would not be consistent with an assumption of normality.

(b) A 95% lower confidence bound for the population mean difference is:

$$\bar{d} - (\text{t critical value})\left(\frac{s_d}{\sqrt{n}}\right)$$

$$-38.60 - (1.761)\left(\frac{23.18}{\sqrt{15}}\right)$$

where the t critical value was found in Table IV, with df = (n – 1) = 14 for a one-sided interval.

$$(-38.60 - 10.54) = -49.14$$

Therefore, with a confidence level of 95%, the population mean difference is in the interval $(-49.14, \infty)$.

(c) A 95% upper confidence bound for the corresponding population mean difference is:

$$38.60 + 10.54 = 49.14$$
$$(-\infty, 49.14).$$

7.55

This is paired data with n = 60, $\bar{d}$ = 4.47, and s_d = 8.77. The critical t value associated with df = n-1 = 60-1 = 59 and 99% confidence is approximately 2.66 (Table IV). Alternatively, since df = 59 is large, we could simply use the 99% z value of 2.58, which we do in the following calculation: $\bar{d} \pm t\, {}^{s_d}\!/\!\sqrt{n}$ = 4.47 ± (2.58) ${}^{8.77}\!/\!\sqrt{60}$ = 4.47 ± 2.92 = [1.55, 7.39]. Therefore, we estimate that the average blood pressure in a dental setting exceeds the average blood pressure in a medical setting by between 1.55 and 7.39. The interval does not contain 0, which suggests that the true average pressures are indeed different in the two settings.

Supplementary Exercises

7.73

The center of any confidence interval for μ_1-μ_2 is always $\bar{x}_1 - \bar{x}_2$, so $\bar{x}_1 - \bar{x}_2$ = (-473.3 + 1691.9)/2 = 609.3. Furthermore, the half-width of this interval is [1691.9-(-473.3)]/2 = 1082.6. Equating this value to the expression for the half-width of a 95% interval, 1082.6 = (1.96) $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$, we find $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ = 1082.6/1.96 = 552.35. For a 90% interval, the associated z value is 1.645, so the 90% confidence interval is then $\bar{x}_1 - \bar{x}_2 \pm (1.645)\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ = 609.3 ± (1.645)(552.35) = 609.3 ± 908.6 = [-299.3, 1517.9].

7.76. A 95% upper confidence bound for the true proportion of all such screws that would strip out during assembly is:

$$p = \left(\frac{44}{2{,}000}\right) = .022$$

$$p + \left(1.645\sqrt{\frac{p(1-p)}{n}}\right)$$

$$.022 + (1.645)\sqrt{\frac{(.022)(.978)}{2{,}000}} \;=\; .022 + .0054 \;=\; .0274$$

That is, with 95% confidence, the value of π lies in the interval (0%, 2.74%).

We cannot be sure that at most 2% of these screws suffer from this defect, since there are plausible values for π that exceed 2%.

7.86. (a) We shall construct a 95% confidence interval for the true proportion of all American adults who are obese. Here, n = 4115, and $p = 1276/4115 \approx .310085$. With 95% confidence, we use the value $z^* = 1.96$ for a two-sided confidence interval:

$$\frac{p + \frac{(z^*)^2}{2n} \pm z^*\sqrt{\frac{p(1-p)}{n}+\frac{(z^*)^2}{4n^2}}}{1+\frac{(z^*)^2}{n}} = \frac{.310085 + \frac{1.96^2}{2(4115)} \pm 1.96\sqrt{\frac{.310085(1-.310085)}{4115}+\frac{1.96^2}{4(4115^2)}}}{1+\frac{1.96^2}{4115}}$$

$$\Rightarrow \quad \frac{.31055 \pm 1.96(.007214)}{1.0009} \Rightarrow (.296, .324)$$

We are therefore 95% confident that between 29.6% and 32.4% of all American adults are obese.

(b) We shall construct a one-sided 95% confidence interval for the true proportion of all American adults who are obese. Here, $n = 4115$, and $p = 1276/4115 \approx .310085.$ With 95% confidence, we use the value $z^* = 1.645$ for a one-sided confidence interval:

$$\frac{p + \frac{(z^*)^2}{2n} - z^* \sqrt{\frac{p(1-p)}{n} + \frac{(z^*)^2}{4n^2}}}{1 + \frac{(z^*)^2}{n}} = \frac{.310085 + \frac{1.645^2}{2(4115)} - 1.645\sqrt{\frac{.310085(1-.310085)}{4115} + \frac{1.645^2}{4(4115^2)}}}{1 + \frac{1.645^2}{4115}}$$

$$\Rightarrow \quad \frac{.310412 - 1.645(.007213)}{1.0007} \Rightarrow .2983$$

Since the interval value dips below 30%, there is no evidence from data in favor of the claim that the 2002 percentage is more than 1.5 times the 1998 percentage.

Chapter 8
Testing Statistical Hypotheses

Section 8.1

8.1

(a) Yes, $\sigma > 100$ is a statement about a population standard deviation, i.e., a statement about a population parameter.

(b) No, this is a statement about the statistic $\bar{x}$, not a statement about a population parameter.

(c) Yes, this is a statement about the population median $\tilde{\mu}$.

(d) No, this is a statement about the statistic s (s is the sample, not population, standard deviation.

(e) Yes, the parameter here is the ratio of two other parameters; i.e, σ_1/σ_2 describes some aspect of the populations being sampled, so it is a parameter, not a statistic.

(f) No, saying that the difference between two samples means is -5.0 is a statement about sample results, not about population parameters.

(g) Yes, this is a statement about the parameter λ of an exponential population.

(h) Yes, this is a statement about the proportion π of successes in a population.

(i) Yes, this is a legitimate hypothesis because we can make a hypothesis about the population distribution [see (4) at the beginning of this section].

(j) Yes, this is a legitimate hypothesis. We can make a hypothesis about the population parameters [see (3) at the beginning of this section].

8.2

The purpose of inspecting pipe welds in nuclear power plants is to determine if the welds are defective (i.e., do not conform to specifications). The benefit of the experimental design where H_0: $\mu = 100$ and H_a: $\mu > 100$ over the design H_0: $\mu = 100$ and H_a: $\mu < 100$ can be understood in terms of Type 1 and Type II error.

In the first design, a type I error corresponds to saying there is evidence from the data that supports the claim that the mean weld population strength is greater than 100 when in fact the mean population strength is less than or equal to 100. Under this design, a type II error corresponds to saying there is not evidence from the data that supports the claim that the mean weld population strength is greater the 100 when in fact such evidence does exist.

In the second design, a type I error corresponds to saying there is evidence from the data that supports the claim that the mean weld population strength is less than 100 when in fact the mean population strength is greater than or equal to 100. Under this design, a type II error corresponds to saying there is not evidence from the data that supports the claim that the mean weld population strength is less the 100 when in fact such evidence does exist.

Thus, under the first design, setting a suitably small significance level will allow you to minimize the chance of claiming that the weld population conforms to specification based on the data, when in fact such a claim is unsupported. Conversely, a small significance level in the second design will minimize the chance of claiming that the weld population fails to conform to specification based on the data, when in fact such a claim is unsupported. The first instance of Type I error is a danger to public safety; the second only to reputation. Thus, the first design is superior.

8.3 Let μ denote the average amperage in the population of all such fuses. Then the two relevant hypotheses are H_0:μ = 40 (the fuses conform to specifications) and H_a:μ ≠ 40 (the average amperage either exceeds 40 is less than 40).

8.4

Let σ denote the population standard deviation of sheath thickness. The relevant hypotheses are:

$$H_0: \sigma = .05 \textit{ versus } H_a: \sigma < .05$$

This is because the company is interested in obtaining conclusive evidence that $\sigma < .05$. A Type I error would be: concluding that the true standard deviation of sheath thickness is less than .05mm when, in fact, it is not.

A Type II error would be: concluding that the true standard deviation of sheath thickness is equal to .05mm when, in fact, it is really less than .05mm.

8.5

Let μ denote the average breaking distance for the new system. The relevant hypotheses are H_0:μ = 120 versus H_a:μ < 120, so implicitly H_0 really says that μ ≥ 120. A Type I error would be: *concluding that the new system really does reduce the average breaking distance (i.e., rejecting H_0) when, in fact (i.e., when H_0 is true) it doesn't.* A Type II error would be: *concluding that the new system does not achieve a reduction in average breaking distance (i.e., not rejecting H_0) when, in fact (i.e, when H_0 is false) it actually does.*

8.6

Let μ denote the true average compressive strength of the mixture. The relevant hypotheses are:

$$H_0: \mu = 1{,}300 \textit{ versus } H_a: \mu > 1{,}300$$

A Type I error would be: concluding that the mixture meets the strength specifications when, in fact, it does not.

A Type II error would be: concluding that the mixture does not meet the strength specifications when, in fact, it does.

8.13

(a) This is a test about the population average μ = average silicon content in iron. The null hypothesis value of interest is μ = .85, so the test statistic is of the form z = $\dfrac{\bar{x} - .85}{s/\sqrt{n}}$. From the wording of the exercise it seems that a 2-sided test is appropriate (since the silicon content is supposed to average .85 and not be substantially larger or smaller than that number), so the relevant hypotheses are H_0:μ = .85 versus H_a:μ ≠ .85 . We can verify that a 2-sided test was done by calculating the P-value associated with the z value of -.81 given in the printout: the area to the left of -.81 is .2090 (from Table I), so the 2-sided P-value associated with z = -.81 is 2(.2090) = .418 ≈ .42.

(b) The P-value of .42 is quite large, so we don't expect it to lead to rejecting H_0 for any of the usual values of α used in hypothesis testing. Indeed, P = .42 exceeds both α = .05 and α = .10, so in neither case would this data lead to rejecting H_0. It appears to be quite likely that the average silicon content does not differ from .85

8.14. (a) If H_a had been $\mu > 750$, the p-value would have been $P(z > -2.14)$, which is clearly not equal to .016. In fact, the p-value would have been .9838.

(b) At a significance level of .05, since .016 < .05, we would reject H_0 and say there is evidence from data in favor of the claim that the true average lifetime is smaller than what is advertised. Therefore, the customer should not purchase the light bulbs. Whereas, at a significance level of .01, since .016 > .01, we would fail to reject H_0. That is, there is insufficient evidence to claim that the true average lifetime is smaller than what is advertised. Therefore, the customer should purchase the light bulbs.

8.16

Let μ denote the true average penetration. Since we are concerned about the specifications not being met, the relevant hypotheses are:

H_0: the specifications are met (μ = 50) , versus
H_a: the specifications are not met ($\mu > 50$).

The test statistic is: $z = \left[\dfrac{52.7-50}{4.8/\sqrt{35}} \right] = 3.33$

The corresponding p-value = $P(z > 3.33) = .0004$

Since p-value = .0004 < α = .05, we reject H_0 and say that there is evidence from data in favor of the claim that the true average penetration exceeds 50 mils. Thus, the specifications have not been met.

Section 8.2

19. (a) Let μ denote the true average writing lifetime. The wording in this exercise indicates that the investigators believe, a priori, that μ can't be less than 10 (i.e., $\mu \geq 10$), so the relevant hypotheses are H_0: $\mu = 10$ versus H_a: $\mu < 10$.

(b) The degrees of freedom are d.f. = n-1 = 18-1 = 17, so the P-value = $P(t < -2.3) = P(t > 2.3) = .017$. Since P-value = .017 < α = .05, we should reject H_0 and say that there is evidence from data that the design specification has not been satisfied.

(c) For t = -1.8, P-value = $P(t<-1.8) = P(t>1.8) = .045$, which exceeds α = .01. Therefore, in this case H_0 would not be rejected. That is, there is not sufficient evidence from data to claim that the design specification is not satisfied.

(d) For t = -3.6, P-value = $P(t<-3.6) = P(t>3.6) = .001$. This P-value is smaller than either α = .01 or α = .05, so in either case H_0 would be rejected. In fact, H_0 would be rejected for any value of α that exceeds .001. For such values of α, we would say that there is no evidence from data to support the claim that the design specification has been satisfied.

22. Let μ denote the true average escape time. The relevant hypotheses are H_0: μ = 360 versus H_a: μ > 360. We choose these hypotheses because the wording in this exercise indicates that the investigators believed, a priori, that μ would be at most 6 minutes (i.e., $\mu \le 360$). This belief should represent the null hypothesis.

The test statistic is: $t = \left[\dfrac{370.69 - 360}{24.36/\sqrt{26}}\right] = 2.24$

The df = (n – 1) = 25.

The corresponding p-value = P(t > 2.24) $\approx$ .017

Since the p-value = .017 < α = .05, we reject H_0 and say that there is evidence from data that contradicts the investigators a priori belief. That is, we have sufficient evidence to say that there is evidence from data that supports the claim that the true average escape time exceeds 6 minutes.

23. From the wording in this exercise the value 3000 is a target value for viscosity, so we don't want to have the true average viscosity μ smaller or larger than 3000, which means that a 2-sided test is needed. To test H_0:μ = 3000 versus H_a:$\mu \neq 3000$, the test statistic is t = $\dfrac{\bar{x} - \mu_0}{s/\sqrt{n}} = \dfrac{2887.6 - 3000}{84.026/\sqrt{5}}$ = -2.99 $\approx$ - 3.0. Using d.f. =n-1 = 5-1 = 4, Table VI gives the 2-sided P-value of 2P(t < -3.0) = 2P(t > 3.0) = 2(.020) = .040. For any value of α greater than .040 (such as α = .05) we would reject H_0, but we would not reject H_0 for values of α smaller than .040

26. Let μ_1 denote the true average gap detection threshold for normal subjects and let μ_2 denote the true average gap detection threshold for CTS subjects. Since we are interested in whether the gap detection threshold for CTS subjects exceeds that for normal subjects, a lower tailed test is appropriate. So, we test:

$$H_0 : (\mu_1 - \mu_2) = 0 \quad versus \quad H_a : (\mu_1 - \mu_2) < 0$$

Using the sample statistics provided, the test statistic is:

$$t = \left[\frac{1.71 - 2.53}{\sqrt{\frac{(.53)^2}{8} + \frac{(.87)^2}{10}}} \right]$$

$$t = -\,2.46 \approx -\,2.5$$

Using the equation for df provided in the section, the approximate df = 15.1, which we round down to 15.

The corresponding p-value = P(t < -2.5) = .012.

Since the p-value = .012 > α = .01, we fail to reject H_0. We have insufficient evidence to claim that the true average gap detection threshold for CTS subjects exceeds that for normal subjects.

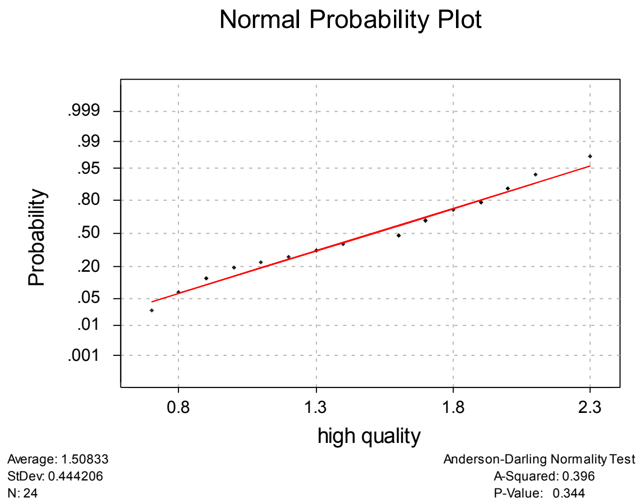
27. Let μ_1 denote the true average strength for wire-brushing preparation and let μ_2 denote the average strength for hand-chisel preparation. Since we are concerned about any possible difference between the two means, a 2-sided test is appropriate, so we test H_0: μ_1-μ_2 = 0 versus H_a: μ_1-$\mu_2 \neq 0$ using the results given in the exercise: n_1 = 12, $\bar{x}_1$ = 19.20, s_1 = 1.58, n_2 = 12, $\bar{x}_2$ = 23.13, s_2 = 4.01 . The test statistic is:

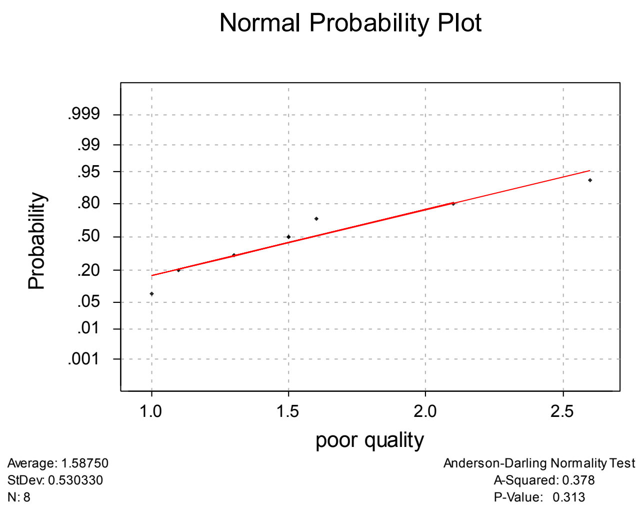
$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} = \frac{19.20 - 23.13}{\sqrt{\frac{1.58^2}{12} + \frac{4.01^2}{12}}} = -3.93/1.244 = -3.16 \approx -3.2.$$

The approximate d.f. $\approx \dfrac{\left[\frac{1.58^2}{12} + \frac{4.01^2}{12}\right]^2}{\left[\frac{1.58^2}{12}\right]^2 \Big/ 11 + \left[\frac{4.01^2}{12}\right]^2 \Big/ 11}$ = 2.396/.1672 = 14.34, which we round <u>down</u> to d.f. = 14.

For a 2-sided test, we use Table VI to find the P-value = 2P(t < -3.2) = 2(.006) = .006. Since P-value = .006 is less than α = .05, we reject H_0 and say there is evidence from data in favor of the claim that there is a difference between the two population average strengths.

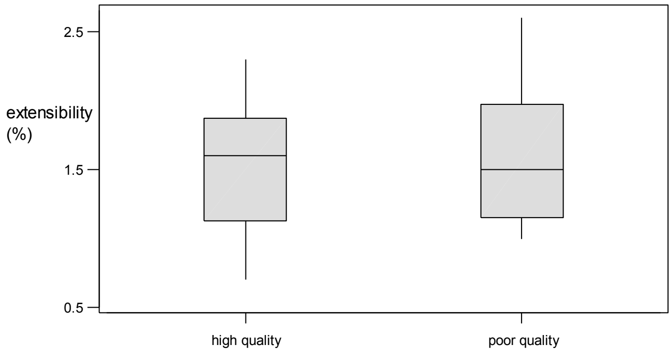
28. (a)





Using Minitab to generate normal probability plots, we see that both plots illustrate sufficient linearity. Therefore, it is plausible that both samples have been selected from normal population distributions.

(b)



The comparative boxplot does not suggest a difference between average extensibility for the two types of fabrics.

(c) Minitab produced the following analysis:

```
Two sample T for high quality vs poor quality

                N      Mean     StDev   SE Mean
high quality  24      1.508     0.444     0.091
poor quality   8      1.587     0.530      0.19

95% CI for mu high quality - mu poor quality:
( -0.543,  0.38)

T-Test mu high quality = mu poor quality (vs not =):
T = -0.38   p-value = 0.71   DF = 10
```

The hypotheses tested were:
$H_0 : (\mu_H - \mu_P) = 0 \quad versus\, H_a : (\mu_H - \mu_P) \neq 0$

A two-tailed test was used because we are interested in whether a "difference" exists between the two types of fabrics.

Using the Minitab output, the test statistic, t = -.38. The corresponding df is 10 and the p-value = 2(P(t < -.38)) = .71.

Since the p-value = .71 is extremely large, we fail to reject H_0. That is, there is insufficient evidence to claim that the true average extensibility differs for the two types of fabrics.

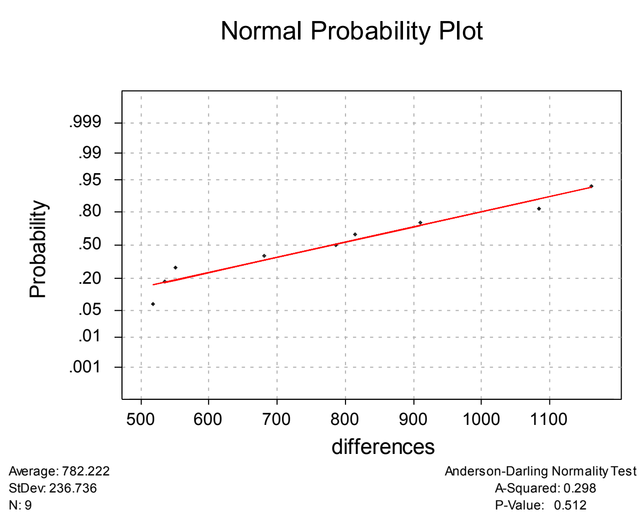
29. Let μ_1 denote the true average proportional stress limit for red oak and let μ_2 denote the average stress limit for Douglas fir. The wording of the exercise suggests that we are interested in detecting any differences between the two averages, which means a 2-sided test is appropriate, so we test H_0: $\mu_1-\mu_2 = 0$ versus H_a: $\mu_1-\mu_2 \neq 0$. The test statistic is:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}} = \frac{8.48 - 6.65}{\sqrt{\frac{.79^2}{14} + \frac{1.28^2}{10}}} = 1.83/.4565 = 4.01 \approx 4.0.$$

The approximate d.f. $\approx \dfrac{\left[\frac{.79^2}{14} + \frac{1.28^2}{10}\right]^2}{\left[\frac{.79^2}{14}\right]^2\Big/13 + \left[\frac{1.28^2}{10}\right]^2\Big/9} = .04344/.003135 = 13.85$, which we round down to d.f. = 13. For a 2-sided test, we use Table VI to find the P-value = 2P(t > 4.0) = 2(.001) = .002. Since P-value = .002 is very small (smaller than the usual significance like .05 or .01), we reject H_0 and say that there is evidence from data that supports the claim that there is a difference between the two average stress limits.

36. Since the experimental design involved growing each of two types of wheat in different locations, the locations serve as a blocking variable and the appropriate analysis is a paired t-test. The analysis requires we compute the difference between Sundance winter and Manitou spring wheat yields for each of the 9 plot locations.

(a) In order to proceed, we must check the assumption that the 9 sample differences have been randomly sampled from a population that is normally distributed. A normal probability plot of the 9 differences produced in Minitab suggests that the normality assumption is reasonable, since the pattern in the plot is reasonably linear.



(b) Let μ_d denote the true average difference in yield for the winter wheat versus the spring wheat. Since we are interested in determining if the average yield for winter wheat is more than 500 kg/ha higher than for spring wheat, the relevant hypotheses are:

H_0: $\mu_d = 500$ *versus* H_a : $\mu_d > 500$

The descriptive statistic for the nine differences are:
$n = 9 \quad \bar{d} = 782.2 \quad s_d = 236.7$

The corresponding test statistic is:

$$t = \left[\frac{782.2 - 500}{236.7/\sqrt{9}} \right] = 3.58$$

With df = (n - 1) = 8, the p-value = P(t > 3.6) = .004

Since, p-value = .004 < most choices for $\alpha\,(e.g., \alpha = .05 \text{ or } \alpha = .01)$, we reject H_0. That is, we have sufficient evidence to claim that the true average yield for the winter wheat is more than 500 kg/ha higher than for the spring wheat.

37. Each sample is measured by two different methods so the data is paired. The observed differences (MSI minus SIB) are: 0.03, -0.51, -0.80, -0.57, -0.66, -0.63, -0.18, 0.01. The sample mean and standard deviation of these observations are $\bar{d}$ = -.4138 and s_d = .3210. The wording in the exercise indicates that we are interested in detecting *any* difference between the two methods, so a 2-sided test is required. To test H_0: $\mu_d = 0$ versus H_a: $\mu_d \neq 0$, the test statistic is:

$$t = \frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{-.4138 - 0}{.3210/\sqrt{8}} = -3.64 \approx -3.6.$$

For d.f. = n-1 = 8-1 = 7, we use Table VI to find the P-value for this 2-sided test: P-value = 2P(t < -3.6) = 2(.004) = .008. At significance levels of either $\alpha = .05$ or $\alpha = .01$, H_0 would be rejected and we would say that there is evidence from data in favor of the claim that there is a difference between the two methods. However, at $\alpha = .001$ we would not reject H_0 since the P-value of .008 exceeds .001.

38. Because the data in Exercise 58 of Chapter 2 is paired, we proceed to a paired *t*-test. (We use a *t* procedure since the sample size is relatively small. For this test, we require that the differences are normally distributed.) We test $H_0 : \mu_d = 5$ versus $H_a : \mu_d > 5$.

Relevant sample statistics include $n = 15$, $\sum d_i = 91$, $\sum d_i^2 = 869.12$, $\bar{d} = 91/15 = 6.07$, $S_{dd} = 869.12 - (1/15)(91^2) = 317.053$, $s^2 = 317.053/(15-1) = 22.65$, $s = 4.76$. The test-statistic takes the form of a one-sample *t*-test:

$$t = \frac{\bar{d} - \Delta_0}{s/\sqrt{n}} = \frac{6.07 - 5}{4.76/\sqrt{15}} = .8706$$

The df for a paired *t*-test are $n - 1 = 15 - 1 = 14$. The critical *t*-value for a significance level of .05 is $t = 1.761$. Clearly, the test-statistic value of .8706 is small, so we fail to reject the null hypothesis. (Even without looking at the *t*-table, it is clear that we fail to reject H_0, since the test-statistic is so small.) Thus, there is insufficient evidence to say that there is evidence in favor of the claim that the true average consumption after weight training exceeds that for the treadmill exercise by more than 5.

As mentioned above, the normality assumption about the difference distribution is required for this *t*-test. To check the plausibility of normality, we could construct a quantile plot of the differences, or we could look at a normal probability plot of the data.

40. Let d_i denote the difference in energy intake and expenditure for player *i*. We test $H_0: \mu_d = 0$ versus $H_a: \mu_d \neq 0$. Relevant summary statistics are $n = 7$, $\sum d_i = 12.3$, $\sum d_i^2 = 30.21$, $\bar{d} = 12.3/7 = 1.757$, $S_{dd} = 30.21 - (1/7)(12.3^2) = 8.597$, $s^2 = 8.597/(7-1) = 1.4329$, $s = \sqrt{1.4329} = 1.197$. The *t* test statistic is given by:

$$t = \frac{\bar{d} - \Delta_0}{s/\sqrt{n}} = \frac{1.757 - 0}{1.197/\sqrt{7}} = 3.88$$

Because this is a two-tailed test, the critical *t*-values (with 7 – 1 = 6 df) for the significance levels of .05, .01, and .001 are 2.447, 3.707, and 5.959. So yes, the conclusion does depend on whether the level of .05, .01, or .001 is used. At the .05 and .01 significance levels, H_0 would be rejected in favor of H_a, but we would fail to reject H_0 if the significance level were .001.

Section 8.3

42. Using the number 1 (for business), 2 (for engineering), 3 (for social science), and 4 (for agriculture), let π_i = the true proportion of all clients from discipline i. If the Statistics Department's expectations are correct, then the relevant null hypothesis is:

$H_0: \pi_1 = .40 \quad \pi_2 = .30 \quad \pi_3 = .20 \quad \pi_4 = .10$ *versus*

H_a : The Statistics Department's expectations are not correct

Using the proportions in H_0, the expected number of clients are:

Client's discipline	Expected number of clients
Business	(120)(.40) = 48
Engineering	(120)(.30) = 36
Social science	(120)(.20) = 24
Agriculture	(120)(.10) = 12

Since all expected counts are at least 5, the chi-squared test can be used. The value of the χ^2 test statistic is:

$$\chi^2 = \left[\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right]$$

$$= \left[\frac{(52-48)^2}{48} + \frac{(38-36)^2}{36} + \frac{(21-24)^2}{24} + \frac{(9-12)^2}{12}\right] = 1.57$$

For df = (k – 1) = (4 – 1) = 3, the corresponding p-value = $P(\chi^2 > 1.57)$. Using Table VII, we find that the p-value > .10. Since the p-value is larger than $\alpha = .05$, we fail to reject H_0. We have no evidence to suggest that the Statistics Department's expectations are incorrect.

43. Using the numbering 1 (for Winter), 2 (Spring), 3(Summer), and 4 (Fall), let π_i = the true proportion of all homicides committed in the i^{th} season. If the homicide rate doesn't depend upon the season, then we would expect the rates to be equal; i.e., H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4 = .25$. The alternative hypothesis in this case would be H_a: *At least one of the proportions does not equal .25*. Using the proportions in H_0, the expected numbers of homicides (out of n = 1361) are shown below the actual numbers from the problem:

Season	**Winter**	**Spring**	**Summer**	**Fall**	**total #**
Observed	328	334	372	327	1361
Expected	340.25	340.25	340.25	340.25	1361
χ^2 contribution	.4410	.1148	2.9627	.5160	

The χ^2 test statistic value is $\chi^2 =$ .4410 +.1148 +2.9627 +.5160 = 4.0345. For d.f = k-1 = 4-1 = 3, the value 4.0345 is smaller than any of the entries in Table VII (column df=3), so the P-value associated with 4.0345 must be larger than .10. Therefore, since P-value > .10 > .05 = α, H_0 is not rejected and we say that this data does not support the belief that there are different homicide rates in the different seasons.

44. Using the equation given for π_i, the relevant hypotheses to test are:

$$H_0 : \pi_1 = .033, \pi_2 = .067, \pi_3 = .100, \pi_4 = .133, \pi_5 = .167, \pi_6 = .167$$
$$\pi_7 = .133, \pi_8 = .100, \pi_9 = .067, \pi_{10} = .033$$
$$\textit{versus}$$
$$H_a : \textit{The a priori proportions are incorrect}$$

Using the proportions in H_0, the expected number of retrieval requests are:

Storage location	Expected number of retrieval requests
1	(200)(..0333...) = 6.667
2	(200)(.0666…) = 13.333
3	(200)(.1000…) = 20.000
4	(200)(.1333…) = 26.667
5	(200)(.1666…) = 33.333
6	(200)(.1666…) = 33.333
7	(200)(.1333…) = 26.667
8	(200)(.1000…) = 20.000
9	(200)(.0666…) = 13.333
10	(200)(.0333…) = 6.667

Since all expected counts are at least 5, the chi-squared test can be used.

location	1	2	3	4	5
observed	4	15	23	25	38
expected	6.667	13.333	20.000	26.667	33.333

location	6	7	8	9	10
observed	31	32	14	10	8
expected	33.333	26.667	20.000	13.333	6.667

The value of the χ^2 test statistic is:

$$\chi^2 = \left[\sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}\right] = 6.6125$$

For df = (k – 1) = (10 – 1) = 9, the corresponding p-value = $P(\chi^2 > 6.6125)$. Using Table VIII, we find that the p-value > .10. Since the p-value, which is larger than .10 is also larger than $\alpha = .10$, we fail to reject H_0. We have no evidence to suggest that the a priori proportions are incorrect.

45. This is a χ^2 test of the homogeneity of several proportions. The hypotheses to test are H_0: *germination rate is homogenous across the different seed types* and H_a: *germination rate depends on the seed type*. The data (with expected values underneath) are shown below:

	Seed Type 1	2	3	4	5	Total
Germinated	31 22.52	57 53.33	87 87.10	52 56.88	10 17.18	237
Failed to germinate	7 15.49	33 36.67	60 59.90	44 39.12	19 11.82	163
Total	38	90	147	96	29	400

The χ^2 test statistic is χ^2 = 3.198 + 0.253 + 0.000 + 0.419 + 3.002 +4.649 + 0.368 + 0.000 + 0.609 + 4.365 = 16.864.

For d.f. = (5-1)(2-1) = 4, the value of χ^2 = 16.864 falls between entries 14.86 and 18.46 of Table VII, so we can say that .001 < P-value < .005. Thus, the P-value is smaller that the specified significance level of α = .01, so H_0 is rejected and we say that there is evidence from data in favor of the claim that germination rates do depend upon seed type. More specifically, note that the largest χ^2 contributions are associated with seed types 1 and 5 (which account for most of the χ^2 value), so we conclude that seed types 1 and 5 are most different (in terms of germination rate) from what we would expect if germination rate did not vary with seed type.

46. This is a χ^2 test of the homogeneity of several proportions. The hypotheses to test are:

H_0: the design configurations are homogeneous with respect to type of failure
versus
H_a: the design configurations are not homogeneous with respect to type of failure

Using the same procedure outlined in the text, the χ^2 test analysis produced by Minitab is shown below:

Chi-Square Test

```
(Expected counts are printed below observed counts)

                           failure mode
              1         2         3         4    Total

config 1     20        44        17         9       90
          16.11     43.58     18.00     12.32

config 2      4        17         7        12       40
           7.16     19.37      8.00      5.47

config 3     10        31        14         5       60
          10.74     29.05     12.00      8.21

Total        34        92        38        26      190

Chi-Sq =   0.942 +  0.004 +  0.056 +  0.893 +
           1.393 +  0.290 +  0.125 +  7.781 +
           0.051 +  0.131 +  0.333 +  1.255 =   13.253

DF = 6, p-value = 0.039
```

As reported by Minitab, the χ^2 test statistic = 13.253. There are (r – 1)(k – 1) = (3 – 1)(4 – 1) = 6 degrees of freedom. The p-value = P(χ^2 > 13.253) = .039. To compute the p-value using Table VIII, we can say that .035 < p-value <.040. Since the p-value = .039 < α .05, we reject H_0. We conclude that the configuration appears to have an effect on type of failure. More specifically, note the large χ^2 contribution associated with configuration 2, failure mode 4. Interpreting the data for configuration 2 only, it appears that the engineer found many more failures in mode 4 than one would have expected. Also, the engineer found few failures in mode 1.

Supplementary Exercises

61. (a) Let μ denote the true average soil heat flux covered with coal dust. The sample mean and standard deviation of this data are $\bar{x}$ = 30.79 and s = 6.530. The relevant hypotheses are $H_0{:}\mu$ =29.0 versus $H_a{:}\mu$ > 29.0 and the calculated test statistic value is t = $\frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{30.79-29.0}{6.530/\sqrt{8}}$ = .775. For d.f. = n-1 = 8-1 = 7, the P-value ≈ .225 (from Table VI), so we would not reject H_0 at significance level α = .05. Therefore is no evidence from data in favor of the claim that coal dust increases the true average heat flux.

(b) A P-value > .10 for the Ryan-Joiner test indicates that the hypothesis H_0:*population is normal* can not be rejected. This is good, since normality of the population is one of the requirements for the validity of the t-test in part (a). So this result does support the use of the t-test in part (a).

62. (a) No, it does not appear plausible that the distribution is normal. Notice that the mean value, $\bar{x} = 215$, is not nearly in the middle of the range of values, 5 to 1176. The midrange would be about 585. Since the mean is so much lower than this, one would suspect the distribution is positively skewed.

However, it is not necessary to assume normality if the sample size is "large enough", due to the central limit theorem. Since this problem has a sample size which is "large enough" (i.e., 47 > 30), we can proceed with a test of hypothesis about the true mean consumption.

(b) Let μ denote the true mean consumption. Since we are interested in determining if there is evidence to contradict the prior belief that μ was at most 200 mg, the following hypotheses should be tested.

H_0: $\mu = 200$ versus H_a: $\mu > 200$.

The value of the test statistic is:

$$z = \left(\frac{\bar{x} - 200}{s/\sqrt{n}} \right) = \left(\frac{215 - 200}{235/\sqrt{47}} \right) = .44$$

The corresponding p-value = $P(z > .44) = .33$.

Since p-value = .33 > most any choice of α, we fail to reject H_0. There is insufficient evidence to suggest that the true mean caffeine consumption of adult women exceeds 200 mg per day.

65. (a) The uniformity specification is that σ not exceed .5, so the relevant hypotheses are $H_0{:}\sigma = .5$ versus $H_a{:}\sigma > .5$ (i.e, we want to see if the data shows that the specified uniformity has been exceeded). The test statistic is $\chi^2 = (n-1)s^2/\sigma^2 = (10-1)(.58)^2/(.5)^2 = 12.1104$. From the χ^2 table (Table VII) with d.f. = n-1 = 10-1 = 9, we note that 12.1104 is smaller than the smallest entry (14.68) in the d.f. = 9 column, so the P-value > .10. Therefore, H_0 should not be rejected at any of the usual significance levels (e.g., .05, .01) and we say that there is no evidence from data that contradicts the uniformity specification.

(b) No. The calculated test statistic is $\chi^2 = (n-1)s^2/\sigma^2 = (10-1)(.58)^2/(.7)^2 = 6.1788$ under the new null hypothesis. Since 6.1788 is smaller than 12.1104, the smallest entry in the d.f. = 9 column, so the right-tail area > .100. All we know is that the left-tail area < .9. Since we are having the alternative hypothesis $H_a{:}\sigma < .7$, we have to do a left-tail test and with the given information we cannot decide whether to reject the null at the usual levels (e.g., .05, .01).

66. Let π denote the true proportion of front-seat occupants involved in head-on collisions, in a certain region, who sustain no injuries. Given the wording of the exercise, the relevant hypotheses are:

H_0: $\pi = \left(\frac{1}{3}\right)$ versus H_a: $\pi < \left(\frac{1}{3}\right)$

[Note: $(319)\left(\frac{1}{3}\right) = 106.3$ and $(319)\left(\frac{2}{3}\right) = 212.7$ are each ≥ 5].

So, the test statistic is:

$$z=\left[\frac{p-\pi_0}{\sqrt{\frac{(\pi_0)(1-\pi_0)}{n}}}\right]=\left[\frac{\left(\frac{95}{319}\right)-\left(\frac{1}{3}\right)}{\sqrt{\frac{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)}{319}}}\right]=-1.35$$

The corresponding p-value = P(z < -1.35) = .0885

Since p-value = .0885 > α=.05, we fail to reject H_0. We have insufficient evidence to claim that less than one-third of all such accidents result in no injuries.

67. Let π denote the proportion of students in the population who have cheated in this manner. The relevant hypotheses are then $H_0:\pi = .20$ versus $H_a:\pi > .20$ (i.e., we want to know if the data shows that the proportion is even greater than 20%). The sample proportion is p = x/n = 124/480 = .2583, so the test statistic is:

$z = \dfrac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \dfrac{.2583-.20}{\sqrt{\frac{(.20)(.80)}{480}}} \approx 3.19$. From Table I, the upper-tailed P-value associated with 3.19 is P-value = P(z > 3.19) = P(z < -3.19) = .0007. With a P-value this small, H_0 would certainly be rejected at the usual significance levels (e.g., .05, .01), so we say that there is evidence from data in favor of the claim that more than 20% of all students have cheated in the manner described in the article.

In other words, to test H1: sigma < .7, we need left-areas, especially small (e.g., .1, .05, .01) left-areas, but the Table in the book does not give small left areas.

70. (a) Let μ_1 denote the true mean strength for males and μ_2 denote the true mean strength for females. The hypotheses tested here were:

$$H_0:(\mu_1-\mu_2)=0 \quad versus \quad H_a:(\mu_1-\mu_2)\neq 0$$

If one assumes equal population variances, and uses the pooled sample variance you will obtain: s_p = 31.77, t = 2.47, and df = 24. The corresponding p-value for this test is: 2P(t > 2.5) = 2(.010) = .02. These values are quite close to those reported in the exercise.

Notice, however, that the assumption of equal population variances and the t-test statistic that accompanies this assumption is not described in the body of the chapter.

If one uses the method described in the body of the chapter, then t = 2.84 and df = 18. This results in a p-value = 2(P(t > 2.8) = 2(.006) = .012.

(b) Revise the hypotheses:

$$H_0:(\mu_1-\mu_2)=25 \quad versus \quad H_a:(\mu_1-\mu_2)>25$$

The test statistic (without assuming equal population variances) is:

$$t = \left[\frac{(129.2 - 98.1) - 25}{\sqrt{\frac{(39.1)^2}{15} + \frac{(14.2)^2}{11}}} \right] = .556$$

With df = 18, the p-value = P(t > .6) = .278

Since the p-value is greater than any sensible choice of α, we fail to reject H_0. There is insufficient evidence that the true average strength for males exceeds that for females by more than 25N.

75. The relevant hypotheses are H_0:*the leader's winning % is homogeneous across all 4 sports* versus H_a:*the leader's winning % differs among the 4 sports*. The appropriate test is a χ^2 test of homogeneity of several proportions. The following table shows the actual observations along with the expected values (underneath each observation) for the χ^2 test:

	Leader wins	Leader loses	totals
Basketball	150 155.28	39 33.72	189
Baseball	86 75.59	6 16.41	92
Hockey	65 65.73	15 14.27	80
Football	72 76.41	21 16.59	93
Totals	373	81	454

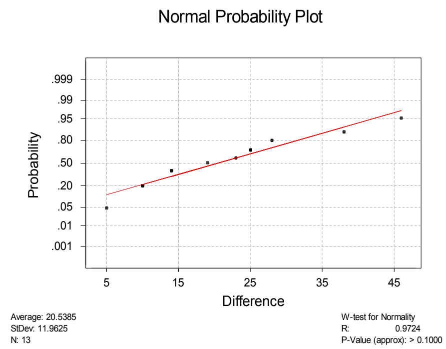
The test statistic value is:

$$\begin{aligned} \chi^2 = \ & 0.180 + 0.827 + \\ & 1.435 + 6.607 + \\ & 0.008 + 0.037 + \\ & 0.254 + 1.171 = 10.518 \end{aligned}$$

From Table VII, with d.f. = (2-1)(4-1) = 3, the P-value associated with χ^2 = 10.518 is P-value ≈ .015. Since the P-value of .015 is smaller than the specified significance level of α = .05, H_0 is rejected and we say that there is evidence from data in favor of the claim that the leader's winning % is not the same across all 4 sports. In particular, the win percentages are 79.4% (basketball), 93.5% (baseball), 81.3% (hockey), and 77.4% (football), so it appears that the leader's winning percentage is much higher for baseball than for the other three sports.

77. Let μ_d denote the true mean difference in retrieval time. We shall test $H_0 : \mu_d = 10$ versus $H_a : \mu_d > 10$ at the $\alpha = .05$ significance level. We will use a paired *t* test, which assumes that the paired differences are normally distributed. From the data, we have $\bar{d} = 20.538$ and $s_d = 11.9625$

By using the Ryan-Joiner test of normality, it appears plausible that this normality condition is satisfied. The following probability plot also appears fairly linear.



For the paired-t test, the appropriate test statistic is $t = \dfrac{\bar{d} - \Delta_0}{s_d / \sqrt{n}} = \dfrac{20.538 - 10}{11.9625 / \sqrt{13}} \approx 3.176$. With $n = 13$, we have $\text{df} = n - 1 = 13 - 1 = 12$. So the appropriate t critical value is 1.782. So we reject H_0 and say there is evidence from data in favor of the claim that the true mean difference in retrieval time does exceed 10 seconds.

Chapter 9
The Analysis of Variance

Section 9.1

8. (a) Let μ_i denote the true mean flow rate for nozzle type i where i = A, B, C, D. The relevant hypotheses are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus
H_a: at least two of the population means differ.

The F test statistic value = 3.68.

Using α = .01, df_1 = 3, and df_2 = 20, a 1% upper-tail area critical value of 4.94 is found. Since the test statistic, F = 3.68 is less than the critical value of 4.94, we fail to reject H_0. We have insufficient evidence to suggest a difference in mean flow rates between the four nozzle types.

(b) Since the p-value = .029 > α = .01, we fail to reject H_0. Thus, we reach the same conclusion here as was reached in part (a)

Section 9.2

11. (a) There are k = 4 populations (brands) and a total of n = (5)(4) = 20 sample observations. Therefore, SST has n-1 = 19 d.f., SSTr has k-1 = 4-1 = 3 d.f., and SSE has n-k = 20-4 = 16 d.f. Since 14,713.69 = MSE = SSE/(n-k) = SSE/16, we solve to find SSE = (16)(14,713.69) = 235,419.04. Next, SSTr = SST-SSE = 310,500.76 - 235,419.04 = 75,081.72. Finally, MSTr = SSTr/(k-1) = 75,081.72/3 = 25,027.24 and F = MSTr/MSE = 25,027.24/14,713.69 = 1.7009.

Source	**df**	**SS**	**MS**	**F**
Brand	3	75,081.72	25,027.24	1.7009
Error	16	235,419.04	14,713.69	
Total	19	310,500.76		

(b) H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$, where μ_i is the true (population) mean number of driving miles until plug failure for spark plug type i
H_1: at least two of the population means are not equal

(c) Using Appendix Table VIII for the F distribution with (k-1,n-k)=(3,16) degrees of freedom, we find that the p-value associated with the test statistic F=1.7009 is larger than 0.1. Since this p-value is larger than the prescribed α = .05, we cannot reject H_0.

12. Let μ_i denote the true mean yield caused by salinity level i where i = 1.6, 3.8, 6.0, 10.2.

The relevant hypotheses are:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ versus
H_a: at least two of the population means differ.

The ANOVA Table from a Minitab analysis appears below. (Note: we entered the data into columns C1, C2, C3, and C4 and used the one-way (unstacked) ANOVA command).

Source	df	SS	MS	F	p-value
Factor	3	456.50	152.17	17.11	0.000

Error	14	124.50	8.89
Total	17	581.00	

Although Minitab computes an exact p-value, we can also approximate it by using Table VIII with $df_1 = 3$ and $df_2 = 14$. The p-value = $P(F > 17.11)$, resulting in an approximated p-value < .001.

Since the p-value is less than $\alpha = .05$, we reject H_0. We have sufficient evidence to claim that there are differences in the true mean yield caused by the different salinity levels.

13. (a) Let μ_A, μ_B, and μ_C denote the true (i.e., population) moduli of elasticity of the three grades of lumber. Then the hypotheses of interest are $H_0: \mu_A = \mu_B = \mu_C$ versus H_a: *at least two of the population means are not equal*.

(b) Since the sample sizes are equal, we can simply average the three sample means to find the grand mean $\bar{\bar{x}} = (1.63+1.56+1.42)/3 = 1.5367$. Next, SSE = $(n_1\text{-}1)s_1^2 + (n_2\text{-}1)s_2^2 + (n_3\text{-}1)s_3^2 = (10\text{-}1)(.27)^2 + (10\text{-}1)(.24)^2 + (10\text{-}1)(.26)^2 = 1.7829$ and SSTr = $n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2 = 10(1.63\text{-}1.5367)^2 + 10(1.56\text{-}1.5367)^2 + 10(1.42\text{-}1.5367)^2 = .22867$. In addition, k = 3 and n = (3)(10) = 30, so SST has d.f. = n - 1 = 30-1 = 29, SSTR has d.f. = k-1 = 3- 1 = 2, and SSE has d.f. - n-k = 30-3 = 27. Therefore, MSTr = SSTr/(k-1) = .22867/2 = .11433 and MSE = SSE/(n-k) = 1.7829/27 = .066033. The test statistic is F = MSTr/MSE = .11433/.066033 = 1.731.

Since the value of F = 1.731 is smaller than those listed in Table VIII (for $df_1 = 2$, $df_2 = 27$), we can say that the P-value > .10. Since the P-value exceeds $\alpha = .01$, H_0 can not be rejected and there is no evidence from data in favor of the claim that there is any differences between the three population means.

14. (a)

Source	df	SS	MS	F
Factor	4	.929	.232	2.15
Error	25	2.69	.108	
Total	29	3.619		

The ANOVA Table values were obtained as follows:

df associated with SSTr = (k – 1) = (5 – 1) = 4
df associated with SSE = (n – k) = (30 – 5) = 25
df associated with SST = (n – 1) = (30 – 1) = 29

Next compute $\bar{\bar{x}}$. Since the sample sizes are equal, we can just average the five sample means to obtain the grand mean. $\bar{\bar{x}} = 2.448$

So, SSTr =$[(6)(2.58 – 2.448)^2 + \ldots + (6)(2.49 – 2.448)^2] = .929$

And, SSE = (SST – SSTr) = (3.619 - .929) = 2.69

$$\text{MSTr} = \left(\frac{.929}{4}\right) = .232\ ;\ \text{MSE} = \left(\frac{2.69}{25}\right) = .108\ ;\ F = \left(\frac{MST}{MSE}\right) = \left(\frac{.232}{.108}\right) = 2.15$$

(b) Let μ_i denote the true mean DNA contents for carbohydrate source group i where i = 1. 2, 3, 4, 5.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ versus
H_a: at least 2 of the population means differ

The value of the test statistic is F = 2.15. The corresponding p-value when $df_1 = 4$ and $df_2 = 25$ is:

p-value = P(F > 2.15) > .10.

Since the p-value is greater than $\alpha = .05$, we fail to reject H_0. We have insufficient evidence to claim that different carbohydrate sources lead to different average DNA contents.

15. The relevant hypotheses are H_0: $\mu_{L1} = \mu_{L2} = \mu_{L3} = \mu_L$ versus H_a: *at least two of the population means differ*. The ANOVA table from a Minitab printout appears below (*note: we entered the data into columns c1, c1, c3, and c4 and used the 'unstack' ANOVA command*).

```
Analysis of Variance
Source     DF        SS         MS        F        P
Factor      3    1.0559     0.3520     3.96    0.053
Error       8    0.7114     0.0889
Total      11    1.7673
```

Although Minitab computes the exact p-value, we can still approximate it by using $df_1 = 3$, and $df_2 = 8$ in Table VIII. From Table VIII we find the p-value associated with F = 3.96 lies somewhere between .05 and .10; i.e., .05 < p-value < .10. At a significance level of $\alpha = .05$, this p-value is not small enough to reject H_0, so there is no evidence from data in favor of the claim that there is a difference between the mean percentage of methyl alcohol reported by the four laboratories.

17. (a) Changing units of measurement amounts to simply multiplying each observation by an appropriate conversion constant, c. In this exercise, c = 2.54. Next, note that replacing each x_i by cx_i causes any sample mean to change from $\bar{\bar{x}}$ to c$\bar{\bar{x}}$ while the grand mean also changes from $\bar{\bar{x}}$ to c$\bar{\bar{x}}$. Therefore, in the formulas for SSTr and SSE, replacing each x_i by cx_i will introduce a factor of c^2. That is, SSTR(for the cx_i data) = $n_1(c\bar{x}_1 - c\bar{\bar{x}})^2 + \ldots + n_k(c\bar{x}_k - c\bar{\bar{x}})^2 = n_1c^2(\bar{x}_1 - \bar{\bar{x}})^2 + \ldots + n_kc^2(\bar{x}_k - \bar{\bar{x}})^2$ = (c^2)SSTr(for the original x_i data). The same thing happened for SSE; i.e., SSE(for the cx_i data) = (c^2)SSE(for the original x_i data). Using these facts, we also see that SST(for the cx_i data) = SSTR(for the cx_i data) + SSE(for the cx_i data) = (c^2)SSTr(for the original x_i data)+ (c^2)SSE(for the original x_i data) = (c^2)[SSTR(for the original x_i data)+SSE(for the original x_i data)] = (c^2)SST(for the original x_i data). The net effect of the conversion, then, is to multiply all the sums of square in the ANOVA table by a factor of $c^2 = (2.54)^2$. Because neither the number of treatments nor the number of observations is altered, the entries in the degrees of freedom column of the ANOVA table is not changed. Notice also that the F-ratio remains unchanged: F(for the cx_i data) = MSTR(for the cx_i data)/MSE(for the cx_i data) = (c^2)MSTR(for the original x_i data)/[$(c)^2$MSE(for the original x_i data)] = MSTR(for the original x_i data)/MSE(for the original x_i data) = F-ratio(for the original x_i data). This makes sense, for otherwise we could change the significance of an ANOVA test by merely changing the units of measurement.

(b) The argument in (a) holds for *any* conversion factor c, not just for c = 2.54. We can say then, that *any* change in the units of measurement will change the 'Sum of Squares' column in the ANOVA table, but the degrees of freedom and F ratio will remained unchanged.

21. (a) The relevant hypotheses are H_0: $\mu_0 = \mu_{45} = \mu_{90}$ versus H_a: *at least two of the population means differ*. The ANOVA table from a Minitab printout appears below (*note: we entered the data into columns c1, c1, and c3, and used the 'unstack' ANOVA command*).

```
Analysis of Variance
Source      DF        SS        MS        F        P
Factor       2     25.80     12.90     2.35    0.120
Error       21    115.48      5.50
Total       23    141.28
```

Minitab computes the exact P-value, but we can still approximate it by using $df_1 = 2$ and $df_2 = 21$ in Table VIII. In Table VIII, F = 2.35 is smaller than any of the entries in the table (for $df_1 = 2$, $df_2 = 21$), so we know that the P-value > .10. Since P-value > .10 > .05 = α, H_0 is not rejected and we say that there is no evidence from data in favor of the claim that there is a difference between the mean strengths at the three different orientations.

(b) The results in (a) are favorable for the practice of using wooden pegs because the pegs do not appear to be sensitive to the angle of application of the force. Since the angle at which force is applied could vary widely at construction sites, it is good to know that peg strength is not seriously affected by different construction practices (i.e., possibly different force application angles).

23. (a) Let x_i denote the true value of an observation. Then x_i +c is the measured value reported by an instrument which is consistently off (i.e., out of calibration) by c units. Therefore, if $\bar{x}$ denotes the mean of the true measurements, then $\bar{x}$+c is the mean of the measured values. Similarly, the grand mean of the measured values equals $\bar{\bar{x}}$ + c, where $\bar{\bar{x}}$ is the grand mean of the true values. Putting these results in the formula for SSTR, we find SSTr(measured values) = $n_1(\bar{x}_1 + c - (\bar{\bar{x}} + c))^2 + \ldots +$ $n_k(\bar{x}_k + c - (\bar{\bar{x}} + c))^2 = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2$ = SSTR (true values). Furthermore, note that any sample variance is unchanged by the calibration problem since the deviations from the mean for the measured data are identical to the deviations from the mean for the true values; i.e., $(x_i + 2.5) - (\bar{x} + 2.5) = (x_i - \bar{x})$. Therefore, SSE (measured values) = $(n_1-1)s_1^2 + \ldots + (n_k-1)s_k^2$ = SSE (true values). Finally, because SSTR and SSE are unaffected, so too will SST be unaffected by the calibration error since SST = SSTR + SSE. Thus, *none* of the sums of squares are changed by the calibration error. Obviously, the degrees of freedom are unchanged too, so the net result is that there will be no change in the entire ANOVA table.

(b) Calibration error will not change any of the ANOVA table entries and therefore will not affect the results of an ANOVA test. That is, if all data points are shifted (up or down) by the same amount c, the ANOVA entries will not be affected. However, the mean of each sample *will* shift by an amount equal to c.

Supplementary Exercises

41. (a) From Table VIII, $F_{.05}(1,10) = 4.96$ and $t_{.025}(10) = 2.228$ and $(2.228)^2 \approx 4.96$. The equality is approximate because the entries in the F and t table entries are rounded.

(b) $F_\alpha(df_1{=}1, df_2) = (t_{\alpha/2})^2$, so for $\alpha = .05$: $F_{.05}(1, df_2) = (t_{05/2})^2 = (t_{.025})^2$, which approaches $(z_{.025})^2 = (1.96)^2 = 3.8416$.

45. (a) $n_1 = 9$ and $n_2 = 4$, so k = 2 and $n = n_1+n_2 = 9+4 = 13$. The grand mean is the weighted average of the sample means, $\bar{\bar{x}} = [9(-.83)+4(-.70)]/[9+4] = -.79$. SSE = $(n_1-1)s_1^2 + (n_2-1)s_2^2 = (9-1)(.172)^2 + (4-$

1)$(.184)^2$ = .33824 and SSTr = $n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2$ = $9(-.83-(-.79))^2 + 4(-.70-(-.79))^2$ = .0468. Therefore, MSTR = SSTR/(k-1) =.0468/(2-1) = .0468, MSE = SSE/(n-k) = .33824/(13-2) = .03075, and F = MSTR/MSE = .0468/.03075 = 1.522. For $df_1 = 1$ and $df_2 = 11$, the P-value associated with F = 1.522 exceeds .10 (Table VIII). Therefore, since the P-value is larger than $\alpha = .01$, we can not reject H_0: *no difference between average diopter measurements for the symptom 'present' and 'absent'*. The data does not show a significant difference between the two groups of pilots.

(b) The equivalent test from Chapter 8 is the independent samples t-test. The test statistic is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$= \dfrac{-.83-(-.70)}{\sqrt{\frac{0.172^2}{9} + \frac{0.184^2}{4}}}$ = -1.199. For df = $\dfrac{(se_1^2 + se_2^2)^2}{\dfrac{se_1^4}{n_1-1} + \dfrac{se_2^4}{n_2-1}} = \dfrac{(\frac{0.172^2}{9} + \frac{0.184^2}{4})^2}{\dfrac{\frac{0.172^4}{81}}{9-1} + \dfrac{\frac{0.184^4}{16}}{4-1}} = 5$ (where $se_1^2 = s_1^2/n_1$ and $se_2^2 = s_2^2/n_2$), the p-value associated with t = -1.199 (for a 2-sided test) is approximately 2(.142) = .284 (from Table VI). Since the p-value ≈ .284 exceeds $\alpha = .01$, we do not reject H_0: $\mu_1 - \mu_2 = 0$. That is, the test does not show a significant difference between the two groups of pilots. This is the same conclusion as in part (a).

Chapter 11
Inferential Methods in Regression and Correlation

Section 11.1

1. (a) The slope of the estimated regression line (β = .095) is the expected change of in the response variable y for each one-unit increase in the x variable. This, of course, is just the usual interpretation of the slope of a straight line. Since x is measured in inches, a one-unit increase in x corresponds to a one-inch increase in pressure drop. Therefore, the expected change in flow rate is .095 m^3/min.

   (b) When the pressure drop, x, changes from 10 inches to 15 inches, then a 5 unit increase in x has occurred. Therefore, using the definition of the slope from (a), we expect about a 5(.095) = .475 m^3/min. increase in flow rate (it is an *increase* since the sign of β = .095 is *positive*).

   (c) For x = 10, $\mu_{y.10}$ = -.12 + .095(10) = .830. For x = 15, $\mu_{y.15}$ = -.12 + .095(15) = 1.305.

   (d) When x = 10, the flow rate y is normally distributed with a mean value of $\mu_{y.10}$ = .830 and a standard deviation of $\sigma_{y.10} = \sigma$ = .025. Therefore, we standardize and use the z table to find: P(y > .835) = P(z > $\frac{.835-.830}{.025}$) = P(z > .20) = 1 - P(z ≤ .20) = 1- .5793 = .4207 (using Table I).

2. (a) The slope of the estimated regression line (β = -.01) is the expected change in reaction time for a one degree Fahrenheit increase in the temperature of the chamber.

   So, with a one degree Fahrenheit increase in temperature, the true average reaction time will decrease by .01 hours.

   With a 10 degree increase in temperature, the true average reaction time will decrease by (10)(.01) = 1 hour.

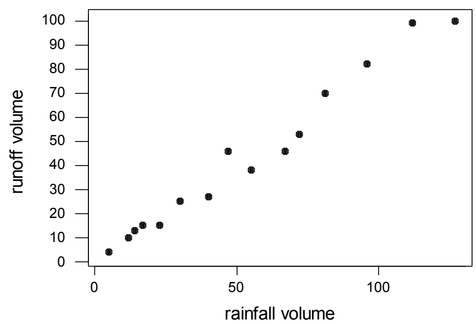
   (b) When x = 200, $\mu_{y\bullet 200} = 5 - .01(200) = 3$

   When x = 250, $\mu_{y\bullet 250} = 5 - .01(250) = 2.5$

   (c) $P(2.4 < y < 2.6 \text{ when } x = 250) =$
   $P\left(\frac{2.4-2.5}{.075} < z < \frac{2.6-2.5}{.075}\right) =$
   $P(-1.33 < z < 1.33) = P(z < 1.33) - P(z < -1.33) =$
   $.9082 - .0918 = .8164$

   Next, the probability that all five observed reaction times are between 2.4 and 2.6 is $(.8164)^5$ = .3627

4. (a)



The scatterplot appears linear, so a simple linear regression model seems reasonable.

(b) The following quantities are needed:

$$\bar{x} = 53.200$$
$$\bar{y} = 42.867$$
$$S_{xy} = \left[51232 - \left(\frac{(798)(643)}{15}\right)\right] = 17024.4$$
$$S_{xx} = \left[63040 - \left(\frac{(798)^2}{15}\right)\right] = 20586.4$$
$$S_{yy} = \left[41999 - \left(\frac{(643)^2}{15}\right)\right] = 14435.7$$
$$b = \left(\frac{S_{xy}}{S_{xx}}\right) = \left(\frac{17024.4}{20586.4}\right) = .82697$$
$$a = (\bar{y} - b\bar{x}) = (42.867 - (.82697)(53.2)) = -1.1278$$

(c) $\mu_{y\cdot 50} = -1.1278 + (.82697)(50) = 40.2207$

(d) SSResid = $S_{yy} - bS_{xy}$ = 14435.7 – (.82697)(17024.4) = 357.07

$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}} = \sqrt{\frac{357.07}{15.2}} = 5.24$$

(e) $r^2 = 1 - \left(\frac{\text{SSResid}}{\text{SSTo}}\right) = 1 - \left(\frac{\text{SSResid}}{S_{yy}}\right) =$

$$1 - \left(\frac{357.07}{14435.7}\right) = .9753$$

So, 97.53% of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall.

5. (a) Using the formulas for the various sums of squares, we find:

$$SS_{xy} = \sum x_i y_i - \tfrac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) = 40.968 - (12.6)(27.68)/9 = 2.216$$
$$SS_{xx} = \sum x_i^2 - \tfrac{1}{n}\left(\sum x_i\right)^2 = 18.24 - (12.6)^2/9 = .600.$$

Therefore, the estimated slope is: $b = SS_{xy}/SS_{xx} = 2.216/.600 = 3.6933$. The estimated intercept is $a = \bar{y} - b\bar{x} = (27.68)/9 - (3.6933)(12.6)/9 = -2.0951$. The estimated regression line is then: $\hat{y} = a + bx = -2.0951 + 3.6933x$.

(b) For x = 1.5, the point estimate of the *average* y value is: $\mu_{y.1.5} \approx -2.0951 + 3.6933(1.5) = 3.445$. For another measurement made when x = 1.5, the point estimate of the y value (for this x value) would be the same; i.e., $\hat{y} = 3.445$.

(c) $SSTo = SS_{yy} = \sum y_i^2 - \tfrac{1}{n}\left(\sum y_i\right)^2 = 93.3448 - (27.68)^2/9 = 8.2134$. Therefore, $SSResid = SSTo - b \cdot SS_{xy} = 8.2134 - (3.6933)(2.216) = .0290$. The estimate of σ is then $s_e \approx \sqrt{\frac{SSResid}{n-2}} = \sqrt{\frac{.0290}{9-2}} = .0644$.

(d) The coefficient of determination is: $r^2 = 1 - \frac{SSResid}{SSTo} = 1 - \frac{.0290}{8.2134} = .996$. Almost all (i.e., about 99.6%) of the observed variation in diffusivity can be attributed to the simple linear regression model between diffusivity and temperature.

Section 11.2

9. Let β denote the true average change in runoff (for each $1m^3$ increase in rainfall). To test the hypotheses $H_0{:}\beta = 0$ versus $H_a{:}\beta \neq 0$, the calculated t statistic is $t = b/s_b = .82697/.03652 = 22.64$ which (from the printout) has an associated P-value of P = 0.000. Therefore, since the P-value is so small, H_0 is rejected and we say that there is evidence from data in favor of the claim that there is a useful linear relationship between runoff and rainfall (no surprise here!).

A confidence interval for β is based on n-2 = 15-2 = 13 degrees of freedom. The t critical value for, say, 95% confidence is 2.160 (from Table IV), so the interval estimate is: $b \pm (\text{t critical}) \cdot s_b = .82697 \pm (2.160)(.03652) = .82697 \pm .07888 = [.748, .906]$ Therefore, we can be confident that the true average change in runoff, for each $1m^3$ increase in rainfall, is somewhere between $.748m^3$ and $.906m^3$.

10. H_0: $\alpha = 0$ versus H_a: $\alpha \neq 0$.

The calculated t statistic is:

$$t = \left(\frac{a}{s_a}\right) = \left(\frac{-1.128}{2.368}\right) = -.48$$

The corresponding p-value = .642. Therefore, since the p-value is large, we would not reject the null hypothesis. We cannot say that there is evidence from data in favor of the claim that the vertical intercept of the population line is nonzero.

12. (a) Method 1: Hypothesis Test

H_0: $\beta = 0$ versus H_a: $\beta \neq 0$

The calculated test statistic, t = -10.243, with a corresponding p-value of < .001.

Therefore, since the p-value is small, we reject the null hypothesis and say that there is evidence from data in favor of the claim that there is a useful linear relationship between these two variables.

Method 2: A confidence interval for β.

$$b \pm (t \text{ critical value})\, s_b$$

A 95% confidence interval for β is:

$$-.014711 \pm (2.179)(.00143620)$$
$$-.014711 \pm .003129$$
$$(-.01784, -.011582)$$

Note: the t critical value was found using df = (n – 2) = (14 – 2) = 12.

With a high degree of confidence, we estimate that the true mean change in attenuation, for each unit increase in fracture strength is somewhere between -.0116 and -.0178 units. Since the plausible values for β are all negative, we say that there is evidence from data in favor of the claim that there is a useful (and negative) linear relationship between these two variables.

(b) The t ratio for testing model utility would be the same value regardless of which of the two variables was defined to be the independent variable. This can be easily seen by looking at the t test statistic for testing if the population correlation coefficient is equal to zero. In that equation the only values required are the sample size (n) and the sample correlation coefficient (r). Both r and n are not dependent on which variable was the independent variable. So, the model utility test is not dependent on which variable was the independent variable.

13. (a) From the printout in Problem 21 (Chapter 3), the error d.f. = n-2 = 25, so the t-critical value for a 95% confidence interval is 2.060 (Table IV). The confidence interval is then: b ± (t critical)·s_b = .10748 ± (2.060)(.01280) = .10748 ± .02637 = [.081, .134]. Therefore, we estimate with a high degree of confidence that the true average change in strength associated with a 1 GPa increase in modulus of elasticity is between .081 MPa and .134 MPa.

(b) Letting β denote the true average change in strength (for each 1 GPa increase in modulus of elasticity), the relevant hypotheses to test are H_0:β = .1 versus H_a:β > .1. The test statistic is t = (b-.1)/s_b = (.10748-.1)/(.01280) = .58 ≈ .6, based on n-2 = 25 degrees of freedom. [*Caution: the t-value from the printout in Problem 21 tests the hypothesis H_0:β = 0, so we don't use that t-ratio*]. The P-value for t = .6 is (from Table VI) P = .277. A large P-value such as this would not lead to rejecting H_0, so there is not enough evidence to contradict the prior belief.

17. (a) A simple linear regression model is given by $\hat{y} = a + bx$, where $b = S_{xy} / S_{xx}$.

$$S_{xy} = \sum x_i y_i - (1/n)\left(\sum x_i\right)\left(\sum y_i\right) = 2759.6 - (1/19)(221.1)(193) = 249.4647$$

$$S_{xx} = \sum x_i^2 - (1/n)\left(\sum x_i\right)^2 = 3056.69 - (1/17)(221.1)^2 = 181.0894$$

So $b = S_{xy} / S_{xx} = 249.4647/181.0894 = 1.37758$

We can compute a 95% confidence interval for the true population slope coefficient β :
We compute $b \pm s_b \cdot t$, where *t* is the appropriate critical value with $n-2 = 17-2 = 15$ df. So the appropriate critical *t*-value for a 95% confidence interval is 2.131.

$s_b = s_e / \sqrt{S_{xx}}$, where $s_e = \sqrt{\text{SSResid}/(n-2)}$, and $\text{SSResid} = \text{SSTo} - b \cdot S_{xy}$

$$\text{SSTo} = S_{yy} = \sum y_i^2 - (1/n)\left(\sum y_i\right)^2 = 2975 - (1/17)(193)^2 = 783.88235$$

$$\text{SSResid} = \text{SSTo} - b \cdot S_{xy} = 783.88235 - (1.37758)(249.4647) = 440.22$$

$$s_e = \sqrt{\text{SSResid}/(n-2)} = \sqrt{440.22/(17-2)} = 5.417$$

$$s_b = s_e / \sqrt{S_{xx}} = 5.417/\sqrt{181.0894} \approx .4025$$

So a 95% confidence interval for β is given by $1.37758 \pm (.4025)(2.131)$, which gives the interval (.51985, 2.2531). Therefore, we are 95% confident that the true mean percentage increase in nausea is between .51985% and 2.23531% for a 1 unit increase in dose.

(b) Yes, there does appear to be a meaningful linear relationship between the two variables. Our 95% confidence interval for β is entirely positive, suggesting that a positive, linear relationship exits.

(c) No, we should not try to predict *y* at the *x* value of 5.0, since it is smaller than the smallest *x*-value used in the regression analysis. Remember, we should not extrapolate outside the range of *x*-values used for the regression analysis.

(d) By deleting the observation $(x, y) = (6.0, 2.50)$, our new summary statistics become:

$n = 17 - 1 = 16$; $\sum x_i = 221.1 - 6.0 = 215.1$; $\sum y_i = 193 - 2.50 = 190.50$;
$\sum x_i^2 = 3056.69 - 6.0^2 = 3020.69$; $\sum x_i y_i = 2759.6 - (6.0)(2.50) = 2744.6$;
$\sum y_i^2 = 2975 - 2.50^2 = 2968.75$

$$S_{xy} = \sum x_i y_i - (1/n)\left(\sum x_i\right)\left(\sum y_i\right) = 2744.6 - (1/16)(215.1)(190.50) = 183.5656;$$

$$S_{xx} = \sum x_i^2 - (1/n)\left(\sum x_i\right)^2 = 3020.69 - (1/16)(215.1)^2 = 128.9394$$

So $b = S_{xy} / S_{xx} = 183.5656/128.9894 \approx 1.42366$

By deleting $(x, y) = (6.0, 2.50)$, our new estimate for β is $b = 1.42366$, which is still fairly close to the value of $b = 1.37758$ that was computed in part (a) above. Moreover, this new estimate of $b = 1.42366$ falls well within the 95% CI that was obtained in part (a) above. Therefore, the point (6.0, 2.50) does not appear to exert any undue influence on our regression analysis.

Section 11.3

18.

(d) We will estimate β using the 95% confidence interval given by $b \pm s_b \cdot t$:

$$6.21075\times10^{-4} \pm (7.578\times10^{-5})(2.571) \quad \Rightarrow \quad (4.26\times10^{-4}, 8.159\times10^{-4})$$

We are 95% confident that the true average change in mist associated with a 1 cm/sec increase in velocity is between 4.26×10^{-4} and 8.159×10^{-4}.

19. (a) The mean of the x data in Problem 21 (Chapter 3) is $\bar{x}$ = 45.11. Since x = 40 is closer to 45.11 than is x = 60, the quantity $(40-\bar{x})^2$ must be smaller than $(60-\bar{x})^2$. Therefore, since these quantities are the only ones that are different in the two $s_{\hat{y}}$ values, the $s_{\hat{y}}$ value for x = 40 must necessarily be smaller than the $s_{\hat{y}}$ value for x = 60. Said briefly, the closer x is to $\bar{x}$, the smaller the value of $s_{\hat{y}}$.

(b) From the printout in Problem 21 (Chapter 3), the error degrees of freedom is d.f. = 25. Therefore, for a 95% confidence interval the t-critical value is t = 2.060 (Table IV), so the interval estimate when x = 40 is:

$\hat{y}$ ± (t-critical) $s_{\hat{y}}$ = 7.592 ± (2.060)(.179) = 7.592 ± .369 = [7.223, 7.961]

We estimate, with a high degree of confidence, that the true average strength for all beams whose MoE is 40 GPa is between 7.223 MPa and 7.961 MPa.

(c) From the printout in Problem 21 (Chapter 3), s_e = .8657, so the 95% prediction interval is:

$\hat{y}$ ± (t-critical) $\sqrt{s_e^2 + s_{\hat{y}}^2}$ = 7.592 ± (2.060) $\sqrt{(.8657)^2 + (.179)^2}$ = 7.592 ± 1.821 = [5.771, 9.413].

Note that the prediction interval is almost 5 times (1.821/.369 = 4.93≈ 5) as wide as the confidence interval.

21. n= 15, so the error d.f. = n -2 = 15-2 =13 and, therefore, the t-critical value for a 95% prediction interval is 2.160 (from Table IV). The prediction interval for x = 40 is centered at $\hat{y}$ = -1.128 + .82697(40) = 31.9508. The prediction interval is then:

$\hat{y}$ ± (t-critical) $\sqrt{s_e^2 + s_{\hat{y}}^2}$ = 31.9508 ± (2.160) $\sqrt{(5.24)^2 + (1.44)^2}$ = 31.95 ± 11.74

= [20.21, 43.69]. Even though the r^2 value is large (r^2 = 73.8%), the prediction interval is rather wide, so precise information about future runoff levels can not be obtained from this model.

23. (a) Let β denote the true average change in milk protein for each 1 kg/day increase in milk production. The relevant hypotheses to test are H_0:β = 0 versus H_a:β ≠ 0. The test statistic is t = b/s_b based on n-2 = 14-2 = 12 degrees of freedom. In order to find s_b, we first find s_e:

$s_e^2 = \frac{SS\,Resid}{n-2} = \frac{.02120}{12}$ = .001767, so s_e = .0420.

Then, $s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{.0420}{\sqrt{762.012}}$ = .00152, which then gives a calculated t value of t = b/s_b = .024576/.00152 ≈ 16.2. The 2-sided P-value associated with t = 16.2 is approximately 2(.000) = .000 (Table VI), so H_0 is rejected in favor of the conclusion that there is a useful linear relationship between protein and production. We should not be surprised by this result since the r^2 value for this data is .956.

(b) For a 99% confidence interval based on d.f. = 12, the t-critical value is 3.055 (from Table IV). The estimated regression line gives a value of $\hat{y}$ = .175576 + .024576(30) = .913 when x = 30. Therefore, $s_{\hat{y}} = (.0420)\sqrt{\frac{1}{14} + \frac{(30-29.56)^2}{762.012}}$ = .01124 and the 95% confidence interval is then: .913 ±(3.055)(.01124) = .913 ± .034 = [.879, .947].

(c) The 99% prediction interval for protein from a single cow is:
$\hat{y}$ ± (t-critical) $\sqrt{s_e^2 + s_{\hat{y}}^2}$ = .913 ± (3.055) $\sqrt{(.0420)^2 + (.01124)^2}$
= .913 ± .133 = [.780, 1.046].

24. $$s_a = s_{a+b(0)} = s_e\sqrt{\frac{1}{n} + \frac{(0-\bar{x})^2}{S_{xx}}}$$

$$= .0420\sqrt{\frac{1}{14} + \frac{(29.564)^2}{762.012}} = .0464$$

H_0: α = 0 versus H_a: α ≠ 0

The value of the test statistic is:

$$t = \left(\frac{.175576 - 0}{.0464}\right) = 3.78$$

With 12 degrees of freedom, the corresponding p-value obtained from Minitab is .0026.

Since the p-value is so small, we reject the null hypothesis and say that there is evidence from data in favor of the claim that the vertical intercept is a nonzero value.

Section 11.4

25. (a) The mean value of y, when x_1 = 50 and x_2 = 3 is -.800 + .060(50) + .900(3) = 4.9 hours.

(b) When the number of deliveries (x_2) is held fixed, then average change in travel time associated with a one-mile (i.e., one unit) increase in distance traveled (x_1) is .060 hours. Similarly, when the distance traveled (x_1) is held fixed, then the average change in travel time associated with one extra delivery (i.e., a one-unit increase in x_2) is .900 hours.

(c) Under the assumption that y follows a normal distribution, the mean and standard deviation of this distribution are 4.9 (because x_1 = 50 and x_2 = 3) and σ = .5 (since σ is assumed to be constant regardless of the values of x_1 and x_2). Therefore, $P(y \le 6) = P(z \le (6-4.9)/.5) = P(z \le 2.20) = .9861$ (from Table I). That is, in the long run, about 98.6% of all days will result in a travel time of at most 6 hours.

Section 11.5

31. (a) To test H_0: $\beta_1 = \beta_2 = 0$ versus H_a: *at least one of* β_1 *and* β_2 *is not zero*, the test statistic is:

$$F = \frac{MS\,\text{Re}\,gr}{MS\,\text{Re}\,sid} = 319.31 \text{ (from printout).}$$

Alternatively, you could use the formula

$$F = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{.991/2}{(1-.991)/6} = 330.33.$$

Discrepancy between the two methods is caused by the rounding in R^2 (e.g., as you can see from the sums of squares in the printout, R^2 = 715.50/722.22 = .9907, which was then rounded to .991). Regardless, the P-value associated with and F-value of 319 or so is zero, so, at any reasonable level of significance, H_0 should be rejected. There does appear to be a useful linear relationship between temperature difference and at least one of the two predictors.

(b) The degrees of freedom for SSResid = n-(k+1) = 9 - (2+1) = 6 (which you could simply read in the "DF" column of the printout). Therefore, the t-critical value for a 95% confidence interval is 2.447 (Table IV). The desired confidence interval is:

$b_2 \pm$ (t-critical) s_{b_2} = 3.0000 ± (2.447)(.4321) = 3.0000 ± 1.0573, or about [1.943, 4.057].

That is, holding furnace temperature fixed, we estimate that the average change in temperature difference on the die surface will be somewhere between 1.943 and 4.057.

(c) When x_1 = 1300 and x_2 = 7, the estimated average temperature difference is $\hat{y} = -199.56 + .2100x_1 + 3.000x_2 = -199.56 + .2100(1300) + 3.000(7) = 94.44$. The desired confidence interval is then 94.44 ± (2.447)(.353) = 94.44 ± .864, or, about [93.58, 95.30].

(d) From the printout, s_e = 1.058, so the prediction interval is:
$94.44 \pm (2.447)\sqrt{(1.058)^2 + (.353)^2} = 94.44 \pm 2.729$, or, about [91.71, 97.17].

32. (a) Yes, there does appear to be a useful linear relationship between y and the predictors. We determine this by observing that the p-value corresponding to the model utility test is < .0001 (F test statistic = 18.924).

(b) H_o: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$

The value of the t test statistic is 3.496. Its corresponding p-value is .0030. Since the p-value < α = .01, we reject the null hypothesis and say that there is evidence from data in favor of the claim that the interaction predictor does provide useful information about y.

(c) A 95 % confidence interval for the mean value of surface area under the stated circumstances requires the following quantities:

$$\hat{y} = 185.486 - 45.969(2) - .3015(500) + .0888(2)(500) = 31.598$$

Next, with (n – k – 1) = (20 – 3 – 1) = 16 df, the t critical value is 2.120.

So, the 95% confidence interval is:

$$31.598 \pm (2.120)(4.69)$$
$$31.598 \pm 9.9428$$
$$(21.6552, 41.5408)$$

33. (a) The appropriate hypotheses are H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$ versus H_a: *at least one of the β_i's is not zero.* The test statistic is F = $\dfrac{R^2/k}{(1-R^2)/(n-(k+1))} = \dfrac{.946/4}{(1-.946)/(25-(4+1))}$ = 87.6. The test is based on df_1 = 4, df_2 = 20. From Table XII, the P-value associated with F = 6.59 is .001, so the P-value associated with 87.6 is obviously .000. Therefore, H_0 can be rejected at any reasonable level of significance. We say that there is evidence from data in favor of the claim that at least one of the four predictor variables appears to provide useful information about tenacity.

(c) The estimated average tenacity when x_1 = 16.5, x_2 = 50, x_3 = 3, and x_4 = 5 is: $\hat{y}$ = 6.121 -.082x_1 +.113x_2 +.256x_3 -.219x_4 = 6.121 -.082(16.5) +.113(50) +.256(3) -.219(5) = 10.091. For a 99% confidence interval based on 20 d.f., the t-critical value is 2.845. The desired interval is: 10.091 ± (2.845)(.350) = 10.091 ± .996, or, about [9.095, 11.087]. Therefore, when the four predictors are as specified in this problem, the true average tenacity is estimated to be between 9.095 and 11.087.

34. (a) Yes, there does appear to be a useful linear relationship between repair time and the two model predictors. We determine this by conducting a model utility test.

H_0: $\beta_1 = \beta_2 = 0$ versus H_a: At least one of these two β's are not zero.

The test statistic requires the following quantities.

SSRegr = (SSTo – SSResid) = (12.72 – 2.09) = 10.63

MSRegr = (SSRegr/k) = (10.63/2) = 5.315

MSResid = (SSResid/n – k – 1) = (2.09/9) = .2322

So, F = (MSRegr/MSResid) = (5.315/.232) = 22.91

With 2 numerator degrees of freedom and 9 denominator degrees of freedom, the F critical value at α = .05 is 4.26. Since 22.91 > 4.26, we reject the null hypothesis and say that there is evidence from data in favor of the claim that at least one of the two predictor variables is useful.

(b) H_o: $\beta_2 = 0$ versus H_a: $\beta_2 \neq 0$

The t test statistic = $\left(\dfrac{1.250-0}{.312}\right)=4.01$

The corresponding p-value = 2P(t > 4.01). With df = 9 and using Minitab, the exact p-value is .003. Since the p-value < α =.01, we reject the null hypothesis and say that there is evidence from data in favor of the claim that the "type of repair" variable does provide useful information about repair time, given that the "elapsed time since the last service" variable remains in the model.

(c) A 95% confidence interval for β_2 is :

$$1.250 \pm (2.262)(.312)$$

$$1.250 \pm .7057$$

$$(.5443 \text{ , } 1.9557)$$

(Note: df = n – (k + 1) = 12 – (2 + 1) = 9)

We estimate, with a high degree of confidence, that when an electrical repair is required the repair time will be between .54 and 1.96 hours longer than when a mechanical repair is required, while the "elapsed time" predictor remains fixed.

(d) To compute the prediction interval we need the following quantities.

$$\hat{y} = .950 + .400(6) + 1.250(1) = 4.6$$

$$s_e^2 = \text{MSResid} = .23222$$

For a 99% prediction interval, the t critical value with df = 9 equals 3.250.
So, our 99% prediction interval is:

$$4.6 \pm (3.250)\sqrt{.23222 + (.192)^2}$$

$$4.6 \pm 1.69$$

$$(2.91, 6.29)$$

The prediction interval is quite wide, suggesting a variable estimate for repair time under these conditions.

35. (a) The negative value of b_2, which is the coefficient of x^2 in the model, indicates that the parabola $b_0 + b_1x + b_2x^2$ opens downward.

(b) R^2 = 1 - SSResid/Ssto = 1 - .29/202.87 = .9986, so about 99.86% of the variation in output power can be attributed to the relationship between power and frequency.

(c) With an R^2 this high, it is very likely that the test statistic will be significant. The relevant hypotheses are H_0:$\beta_1=\beta_2=0$ versus H_a: *at least one of the* β_i*'s is not zero*. The test statistic is: F = $\frac{SS\,Regr/k}{SS\,Resid/(n-(k+1))} = \frac{(202.87-2.9)/2}{.29/5}$ = 1746. Clearly, the P-value associated with F = 1746 is 0, so H_0 is rejected and we say that there is evidence from data in favor of the claim that the model is useful for predicting power.

(d) The relevant hypotheses are H_0:$\beta_2=0$ versus H_a:$\beta_2 \neq 0$. The test statistic is: t = b_2/s_{b_2} = -.00163141/.00003391 = 48. The P-value for this statistic is 0 and H_0 is rejected in favor of the conclusion that the quadratic predictor provides useful information.

(e) The estimated average power when x = 150 is $\hat{y}$ = -1.5127 + .391902x - .00163141x^2 = -1.5127 + .391902(150) - .00163141$(150)^2$ = 20.57. The t-critical value based on 5d.f. is 4.032, so the 99% confidence interval is: 20.57 ± (4.032)(.1410) = 20.57 ± .57 or, about [20.00, 21.14] To find the prediction interval, we must first find s_e. s_e^2 = SSResid/(n-3) = .29/5 = .058, so s_e = .241. Therefore, the 99% prediction interval is:

20.57 ± (4.032) $\sqrt{(.241)^2 + (.141)^2}$ = 20.57 ± 1.13, or, about [19.44, 21.70].

36. (a) For a 1% increase in the percentage plastics, we would expect a 28.9 kcal/kg increase in energy content.

Also, for a 1% increase in the moisture percentage, we would expect a 37.4 kcal/kg decrease in energy content.

(b) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ versus

H_a: At least one of the four predictors is useful.

The value of the F test statistic is 167.71, with a corresponding p-value that is extremely small. So, we reject the null hypothesis and say there is evidence from data in favor of the claim that at least one of the four predictors is useful in predicting energy content, using a linear model.

(c) H_o: $\beta_3 = 0$ versus H_a: $\beta_3 \neq 0$

The value of the test statistic is t = 2.24 with a corresponding p-value = .034. Since the p-value < α = .05, we reject the null hypothesis and say that there is evidence from data in favor of the claim that percent garbage provides useful information about energy consumption, given that the other three predictors remain in the model.

(d) To compute the 95% confidence interval, the following quantities are required:

$$\hat{y} = 2245 + 28.9(20) + 7.64(25) + 4.30(40) - 37.4(45) = 1503$$

The t critical value for a 95% interval and 25 degrees of freedom is 2.060.

So, a 95% confidence interval for the true average energy content under these circumstances is:

$1503 \pm (2.060)(12.47)$
1503 ± 25.69
$(1477, 1529)$

Because the interval is reasonably narrow, we would say that there evidence from data in favor of the claim that the mean energy content has been precisely estimated.

(e) A 95% prediction interval for the energy content of a waste sample having the specified characteristics is:

$1503 \pm (2.060)\sqrt{(31.48)^2 + (12.47)^2}$
1503 ± 69.75
$(1433, 1573)$

Supplementary Exercises

59. (a) We use a chi-square test with $(3-1)(2-1) = 2$ degrees of freedom. The chi-square test-statistic is $X^2 = 19.184$, and our P-value is about 0. From the following table, it is clear that soccer players tend to have higher concussion rates than the other groups.

```
Expected counts are printed below observed counts

            Yes       No    Total
    1        45       46       91
          30.71    60.29

    2        28       68       96
          32.40    63.60

    3         8       45       53
          17.89    35.11

Total        81      159      240

Chi-Sq =  6.647 +  3.386 +
          0.598 +  0.304 +
          5.465 +  2.784 = 19.184
DF = 2, P-Value = 0.000
```

(b) Since the sample correlation coefficient is negative, this suggests that the score on the recall test somewhat tends to decrease as the total number of competitive seasons played increases.

(c) To test whether $\mu_1 = \mu_2$, we use an independent two-sample t-test. The test-statistic is:

$t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \dfrac{37.50 - 39.63}{\sqrt{\dfrac{9.13^2}{26} + \dfrac{10.19^2}{56}}} = -.9769$. Because our test-statistic is small in magnitude, there does not appear to be a significant difference in the mean test scores between the soccer and non-soccer players.

(d) We test $\mu_S = \mu_{NS} = \mu_C$ using a one-way ANOVA at the $\alpha = .05$ level. First note that the grand mean of all these individuals is $\overline{\overline{x}} = \dfrac{91(.30)+96(.49)+53(.19)}{91+96+53} = \dfrac{84.41}{240} = .3517$. The relevant quantities include:

$$\text{SSTr} = \sum_i n_i(\bar{x}_i - \overline{\overline{x}})^2 = 91(.30-.3517)^2 + 96(.49-.3517)^2 + 53(.19-.3517)^2 = 3.4652$$

$$\text{SSE} = \sum_i (n_i - 1)\, s_i^2 = (91-1)(.67^2) + (96-1)(.87^2) + (53-1)(.48^2) = 124.2873$$

$$\text{MSTr} = \text{SSTr}/(k-1) = 3.4652/(3-1) = 1.7326$$

$$\text{MSE} = \text{SSE}/(n-k) = 124.2873/(240-3) = .524419$$

F = MSTr / MSE = 1.7326 / .524419 = 3.3038 , with (2, 237) degrees of freedom

By looking at the F table with 2 numerator d.f. and 240 d.f., we see that the critical value for the .05 level is 3.03. So the approximate P-value of this test is .05, and we say that there is evidence from data in favor of the claim that there are differences in the average non-soccer concussions.