INSTRUCTOR'S SOLUTIONS MANUAL

SONGFENG ZHENG
Missouri State University

PROBABILITY & STATISTICS WITH R FOR ENGINEERS AND SCIENTISTS

Michael Akritas
The Pennsylvania State University

PEARSON

Boston Columbus Hoboken Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montreal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Reproduced by Pearson from electronic files supplied by the author.

Publishing as Pearson, 501 Boylston Street, Boston, MA 02116.

ISBN-13: 978-0-321-85308-0
ISBN-10: 0-321-85308-3

www.pearsonhighered.com

PEARSON

Contents

Chapter 1
Basic Statistical Concepts

1.2 Populations and Samples

1. (a) The population consists of the customers who bought a car during the previous year.

   (b) The population is not hypothetical.

2. (a) There are three populations, one for each variety of corn. Each variety of corn that has been and will be planted on all kinds of plots make up the population.

   (b) The characteristic of interest is the yield of each variety of corn at the time of harvest.

   (c) There are three samples, one for each variety of corn. Each variety of corn that was planted on the 10 randomly selected plots make up the sample.

3. (a) There are two populations, one for each shift. The cars that have been and will be produced on each shift make up the population.

   (b) The populations are hypothetical.

   (c) The characteristic of interest is the number of nonconformances per car.

4. (a) The population consists of the all domestic flights, past or future.

   (b) The sample consists of the 175 domestic flights.

   (c) The characteristic of interest is the air quality, quantified by the degree of staleness.

5. (a) There are two populations, one for each teaching method.

   (b) The population consists of all students who took or will take a statistics course for engineering using one of each teaching methods.

   (c) The populations are hypothetical.

   (d) The samples consist of the students whose scores will be recorded at the end of the semester.

1.3 Some Sampling Concepts

1. The second choice provides a closer approximation to simple random sample.

2. (a) It is not a simple random sample.

   (b) In (a), each member of the population does not have equal chance to be selected, thus it is not a simple random sample. Instead, the method described in (a) is a stratified sampling.

3. (a) The population includes all the drivers in the university town.

   (b) The student's classmates do not constitute a simple random sample.

   (c) It is a convenient sample.

   (d) Young college students are not experienced drivers, thus they tend to use seat belts less. Consequently, the sample in this problem will underestimate the proportion.

4. We identify each person with a number from 1 to 70. Then we write each number from 1 to 70 on separate, identical slips of paper, put all 70 slips of paper in a box, and mix them thoroughly. Finally, we select 15 slips from the box, one at a time, without replacement. The 15 selected numbers specify the desired sample of size $n = 15$ from the 70 iPhones. The R command is

   $$y = \textit{sample(seq(1,70), size=15)}$$

   A sample set is 52 8 14 48 62 6 70 35 18 20 3 41 50 27 40.

5. We identify each pipe with a number from 1 to 90. Then we write each number from 1 to 90 on separate, identical slips of paper, put all 90 slips of paper in a box, and mix them thoroughly. Finally, we select 5 slips from the box, one at a time, without replacement. The 5 selected numbers specify the desired sample of size $n = 5$ from the 90 drain pipes. The R command is

   $$y = \textit{sample(seq(1,90), size=5)},$$

   A sample set is 7 38 65 71 57.

6. (a) We identify each client with a number from 1 to 1000. Then we write each number from 1 to 1000 on separate, identical slips of paper, put all 1000 slips of paper in a box, and mix them thoroughly. Finally, we select 100 slips from the box, one at a time, without replacement. The 100 selected numbers specify the desired sample of size $n = 100$ from the 1000 clients.

(b) Using stratified sampling: Get a simple random sample of size 80 from the sub-population of Caucasian-Americans, a simple random sample of size 15 from the sub-population of African-Americans, and a simple random sample of size 5 from the sub-population of Hispanic-Americans. Then combine the three subsamples together.

(c) The R command for part (a) is

$$y = sample(seq(1{,}1000),\ size{=}100)$$

and the R command for part (b) is

$$y1 = sample(seq(1{,}800),\ size{=}80)$$
$$y2 = sample(seq(801{,}950),\ size{=}15)$$
$$y3 = sample(seq(951{,}1000),\ size{=}5)$$
$$y = c(y1,\ y2,\ y3)$$

7. One method is to take a simple random sample of size n from the population of N customers (of all dealerships of that car manufacturer) who bought a car the previous year.

   The second method is to divide the population of the previous year's customers into three strata according to the type of car each customer bought and perform stratified sampling with proportional allocation of sample sizes. That is, if N_1, N_2, N_3 denote the sizes of the three strata, take simple random samples of approximate sizes (due to round-off) $n_1 = n(N_1/N)$, $n_2 = n(N_2/N)$, $n_3 = n(N_3/N)$, respectively, from each of the three strata. Stratified sampling assures that the sample representation of the three strata equals their population representation.

8. It is not a simple random sample because products from facility B have a smaller chance to be selected than products from facility A.

9. No, because the method excludes samples consisting of n_1 cars from the first shift and $n_2 = 9 - n_1$ from the second shift for any (n_1, n_2) different from $(6, 3)$.

1.4 Random Variables and Statistical Populations

1. (a) The variable of interest is the number of scratches in each plate. The statistical population consists of 500 numbers, 190 zeros, 160 ones, and 150 twos.

   (b) The variable of interest is quantitative.

   (c) The variable of interest is univariate.

2. (a) Statistical population: If there are N undergraduate students enrolled at PSU, the statistical population is a list of length N and the i-th element in the list is the major of the i-th student. The variable of interest is qualitative. Another possible variable: gender.

(b) Statistical population: If there are N restaurants on campus, the statistical population consists of a list of N numbers, and the i-th element is the capacity of the i-th restaurant. The variable of interest is quantitative. Another possible variable: food type.

(c) Statistical population: If there are N books in Penn State libraries, the statistical population consists of a list of N numbers, and the i-th element is the check-out frequency of the i-th book in the library. The variable of interest is quantitative. Another possible variable: pages of the book.

(d) Statistical population: If there are N steel cylinders made in the given month, the population consists of a list of N numbers, and the i-th element is the diameter of the i-th steel cylinder made in the given month. The variable of interest is quantitative. Another possible variable: weight.

3. (a) The variable of interest is univariate.

(b) The variable of interest is quantitative.

(c) If N is the number cars of available for inspection, the statistical population consists of N numbers, $\{v_1, \cdots, v_N\}$, where v_i is the total number of engine and transmission nonconformances of the ith car.

(d) If the number of nonconformances in the engine and transmission are recorded separately for each car, the new variable would be bivariate.

4. (a) The variable of interest is the degree of staleness. Statistical population consists of a list of 175 numbers, and the i-th number is the degree of staleness of the air in the i-th domestic flight.

(b) The variable of interest is quantitative.

(c) The variable of interest is univariate.

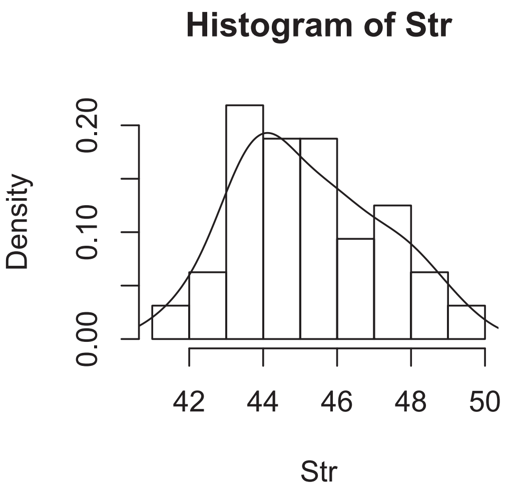
5. (a) The variable of interest is the type of car a customer bought and his/her satisfaction level. Statistical population: If there are N customers who bought a new car in the previous year, the statistical population is a list of N elements, and the i-th element is the car type the i-th customer bought along with his/her satisfaction level, which is a number between 1 to 6.

(b) The variable of interest is bivariate.

(c) The variable of interest has two components. The first is qualitative and the second is quantitative.

1.5 Basic Graphics for Data Visualization

1. The histogram produced by the commands is shown as following:



The stem and leaf plot is as following:

The decimal point is at the |

41 | 5

42 | 39

43 | 1445788

44 | 122357

45 | 1446

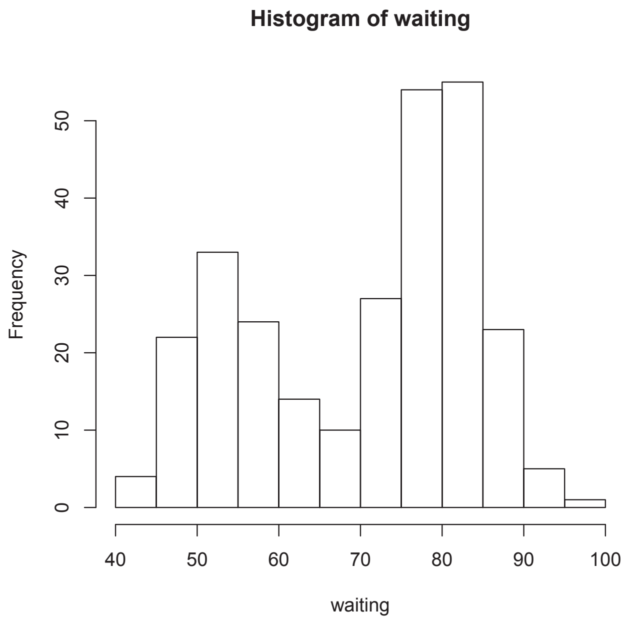
46 | 00246

47 | 3577

48 | 36

49 | 3

2. The histogram on the waiting time is as following



The corresponding stem and leaf plot is given below. It is clear that the shape of the stem and leaf plot is similar to that of the histogram.

The decimal point is 1 digit(s) to the right of the |

4 | 3

4 | 55566666777788899999

5 | 00000111111222223333333444444444

5 | 555555666677788889999999

6 | 00000022223334444

6 | 555667899

7 | 00001111123333333444444

7 | 555555556666666667777777777778888888888888889999999999

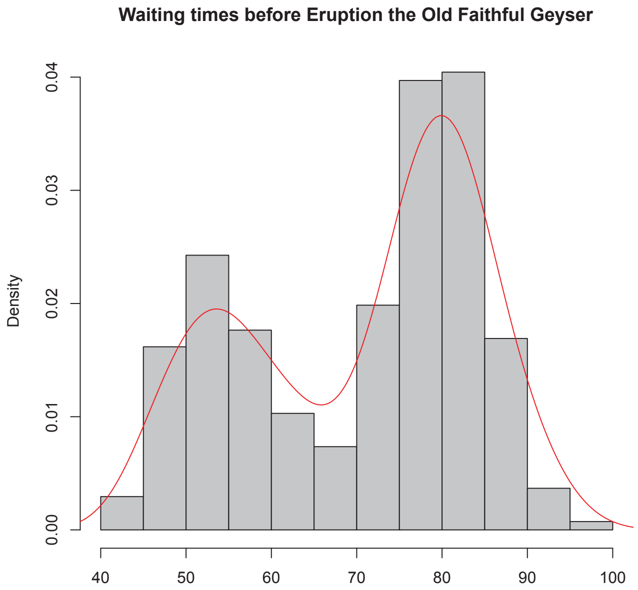
8 | 000000001111111111111222222222222333333333333334444444444

8 | 55555566666677888888999

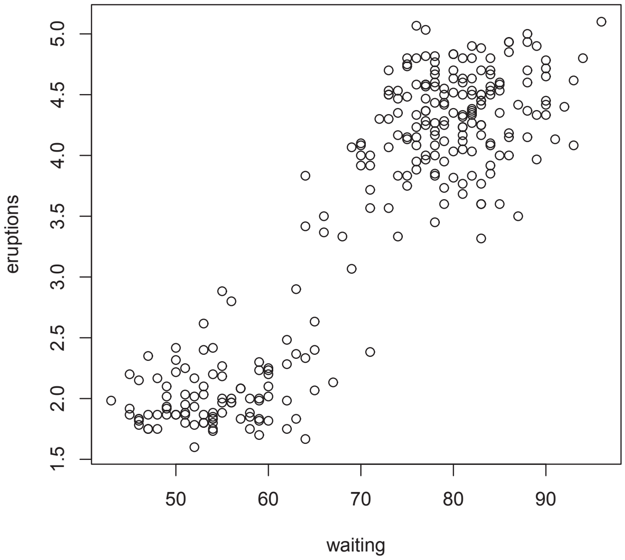
9 | 00000012334

9 | 6

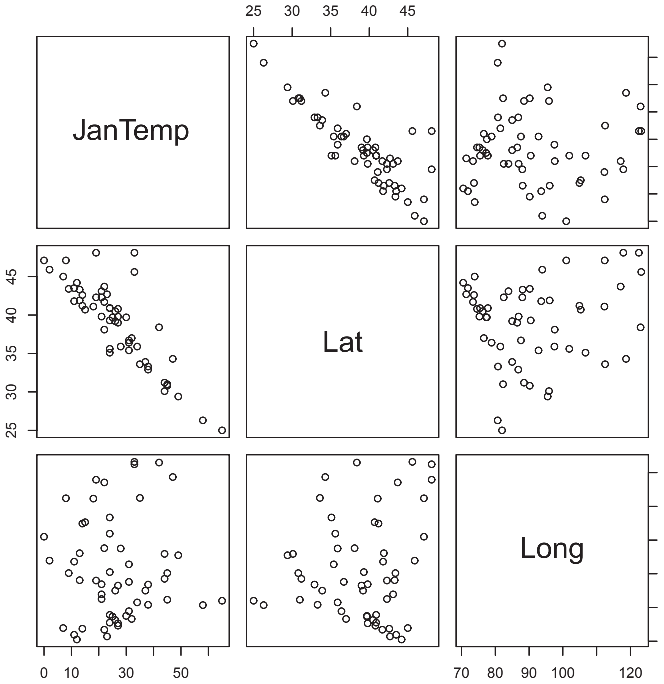
The histogram with title and the colored smooth curve superimposed is shown as



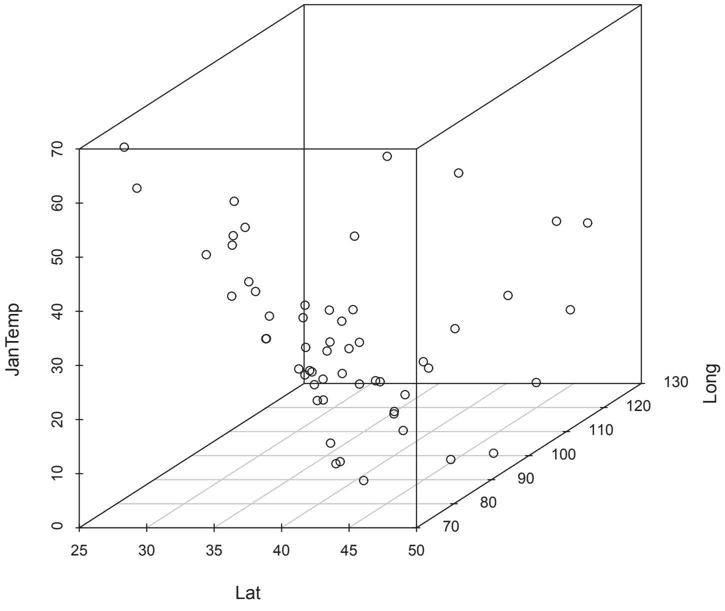
3. The scatterplot is shown below. From the scatter plot, it seems that if the waiting time before eruption is longer, the duration is also longer.



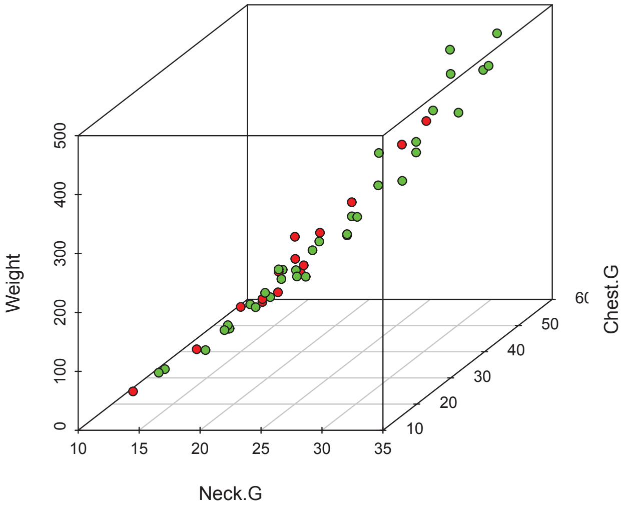
4. (a) The scatterplot matrix is given below. From the figure, it seems that the latitude is a better predictor of the temperature because as the latitude changes, the temperature shows a clear pattern, while there is no pattern as the longitude changes.



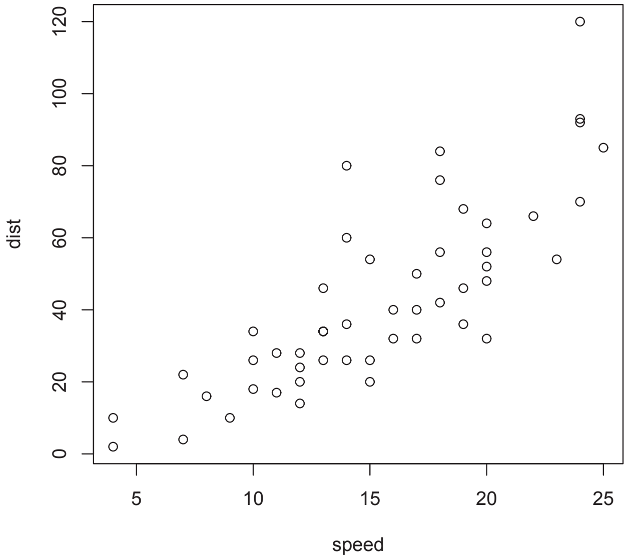
(b) The following figure gives the 3D scatter plot. The 3D scatter plot also shows that the latitude is a better predictor for the temperature.



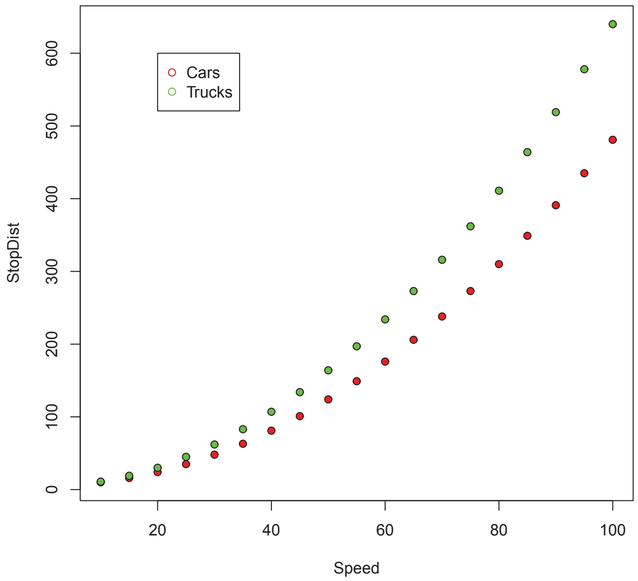
5. The 3D scatterplot is shown below



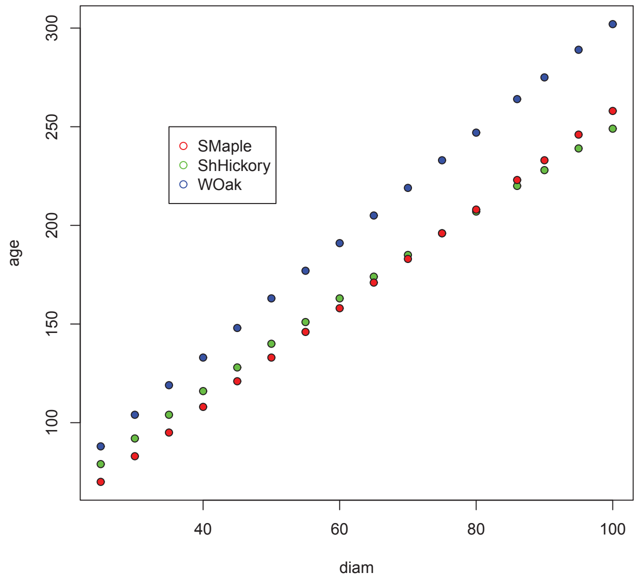
6. The scatterplot is shown below. From the scatterplot, it is clear that in general, if the speed is high, the breaking distance is larger.



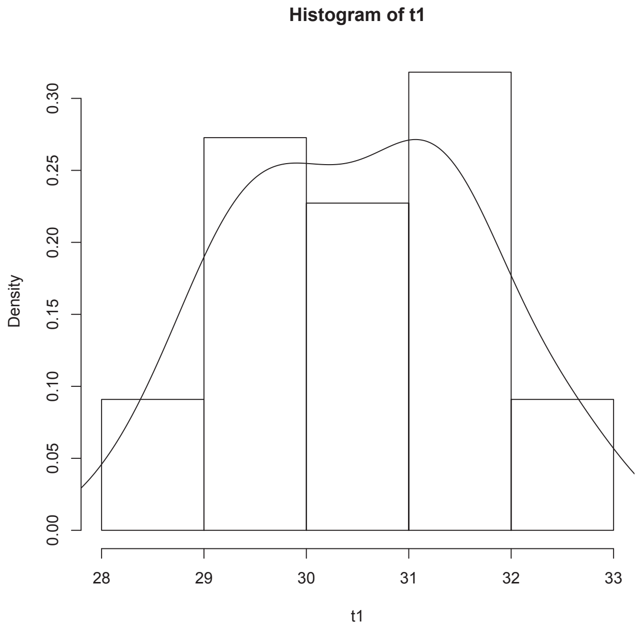
7. The required graph is given below:



8. The resulting graph is given below. The figure shows that for SMaple and WOak, the growing speed in terms of the diameter of the tree is constant, while for ShHickory, when the tree gets older, it grows faster.



9. (a) The basic histogram with smooth curve superimposed:



(b) The stem and leaf plot for the reaction time of Robot 1 is given below. The decimal point is at the |

28 | 4

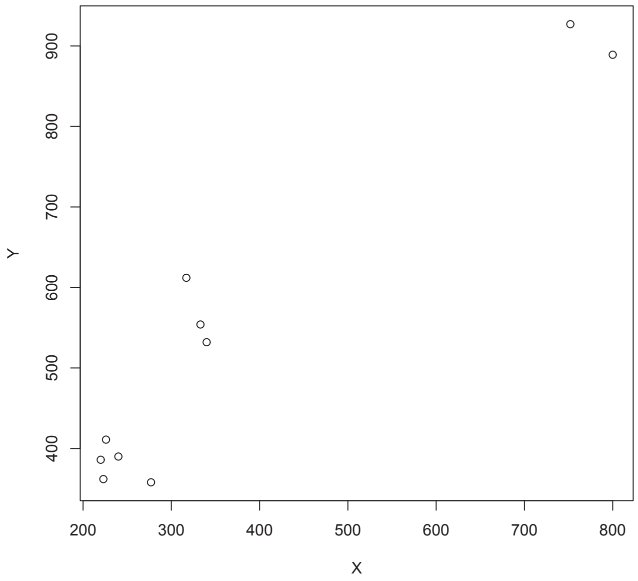
29 | 0133688

30 | 03388

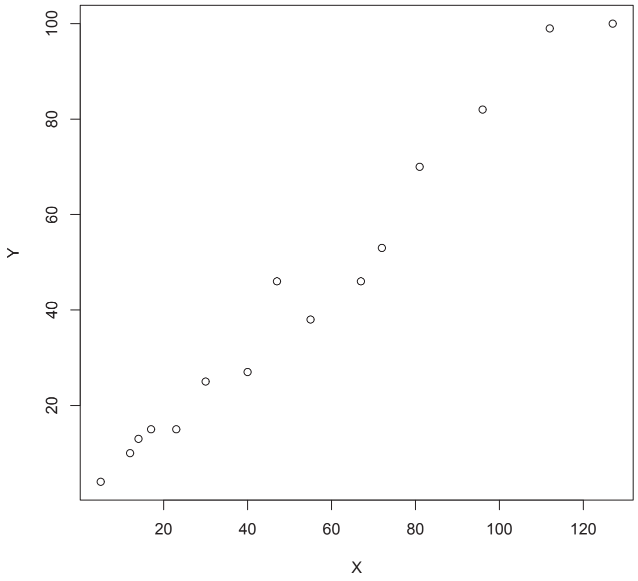
31 | 0234669

32 | 47

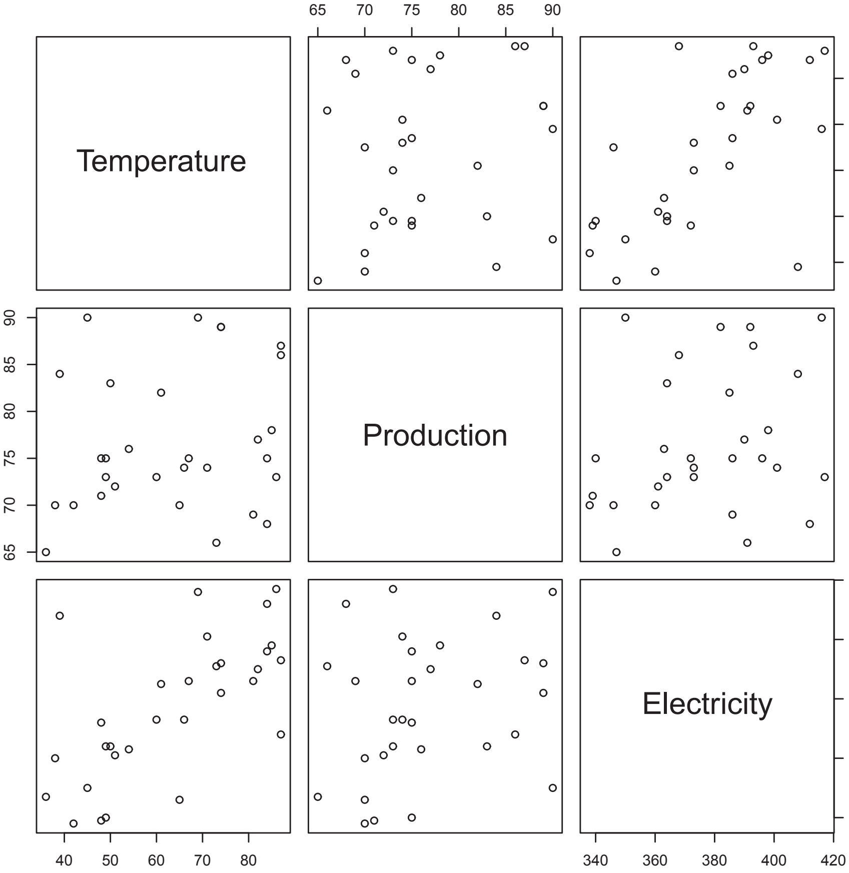
10. The produced basic scatter plot is given below. It seems that the surface conductivity can be used for predicting sediment conductivity.



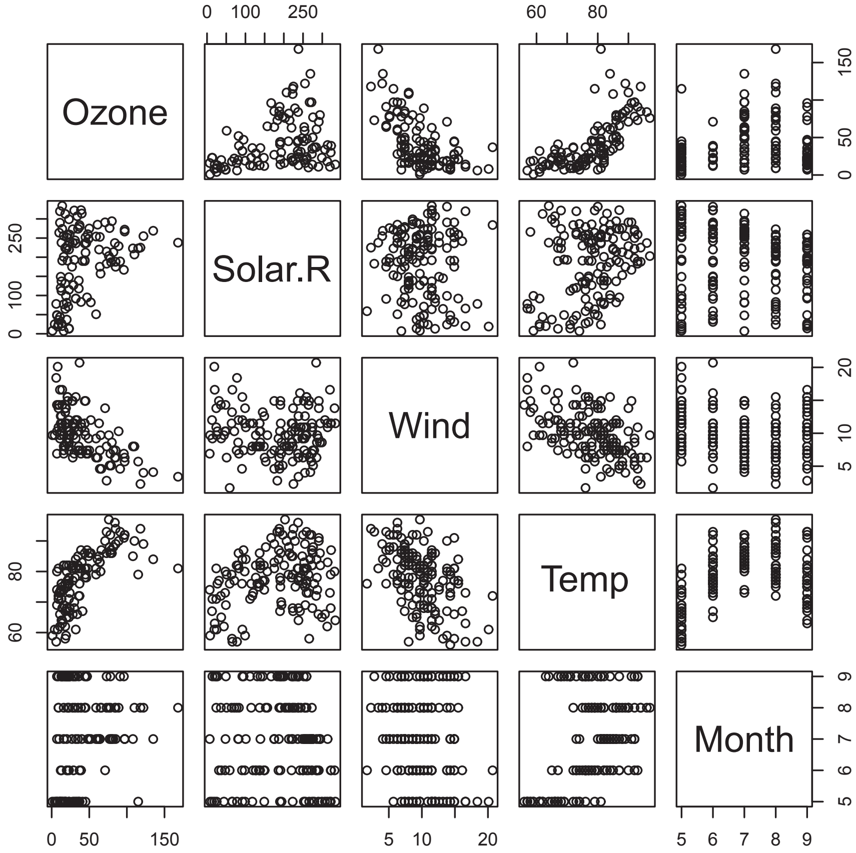
11. The produced basic scatter plot is given below. It seems that the rainfall volume is useful for predicting the runoff volume.



12. The produced scatterplot matrix is as following, and it seems that the variable temperature is a better single predictor for the amount of electricity consumed.


Temperature
Production
Electricity
65
70
75
80
85
90
40
50
60
70
80
340
360
380
400
420

13. The produced scatterplot matrix is as following



According to the scatterplot matrix, we can answer the questions as

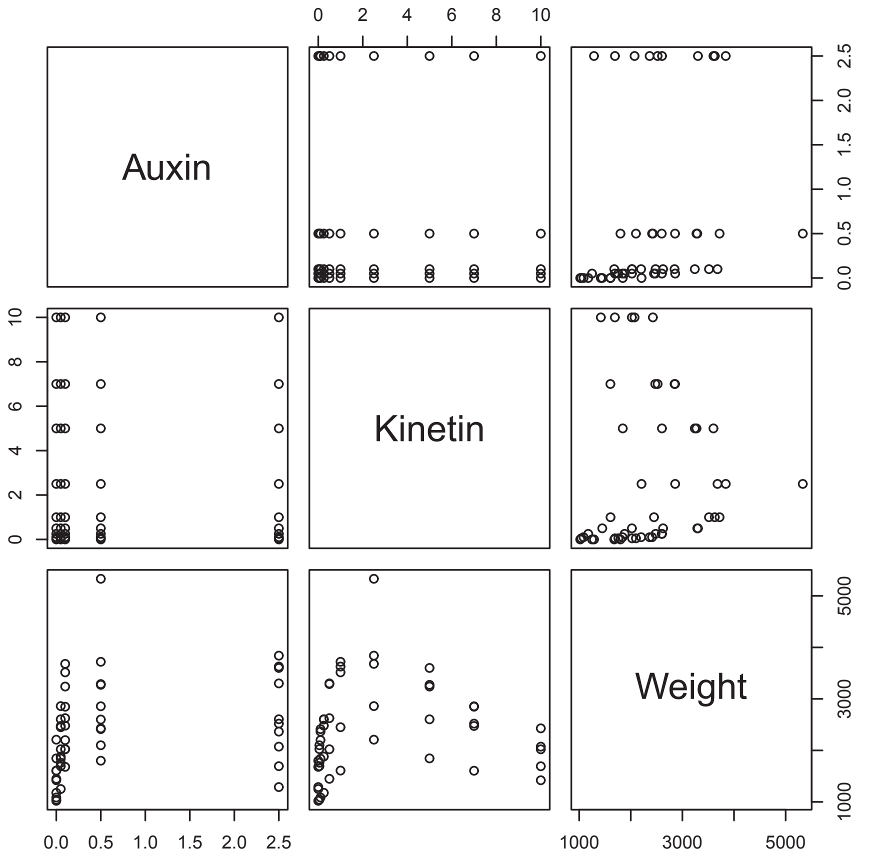
(a) Yes.

(b) No.

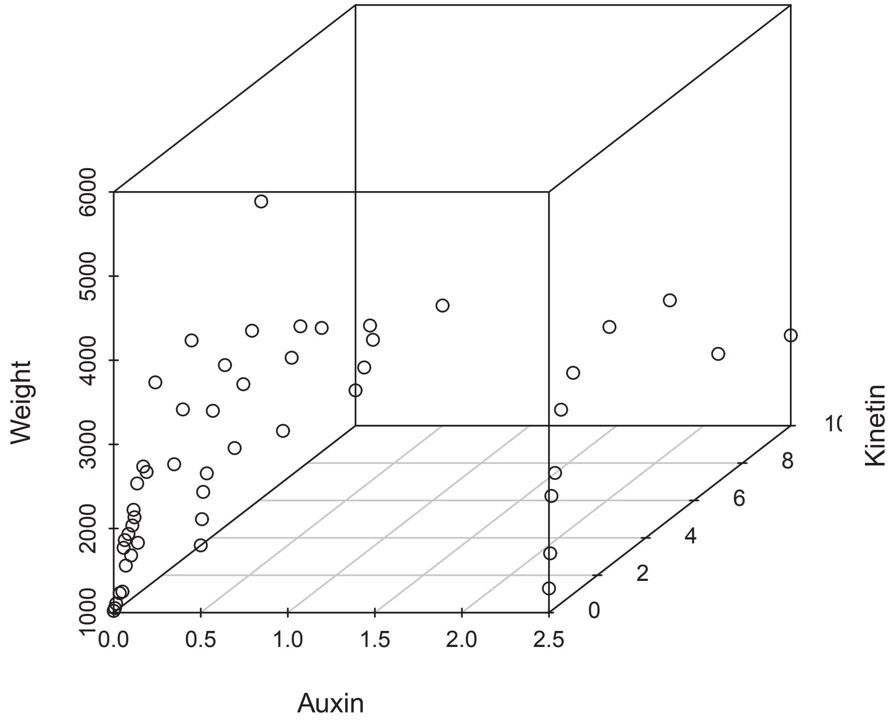
(c) When there is increased solar radiation, the ozone level is more likely to increase, but the variability also increases.

(d) August.

14. The produced scatterplot matrix is as following

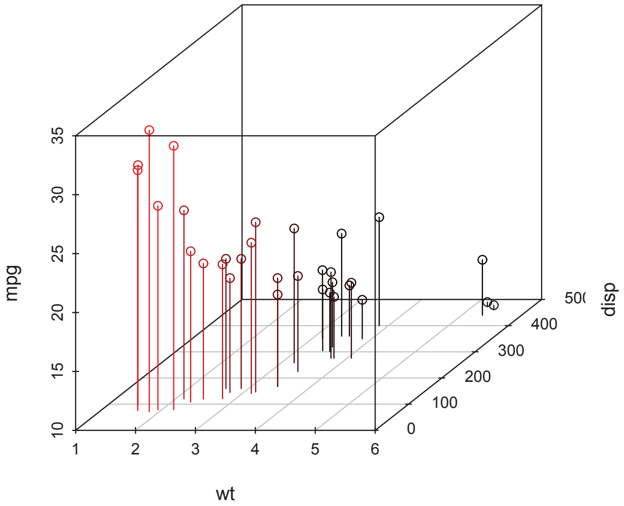


The produced scatterplot matrix is as following

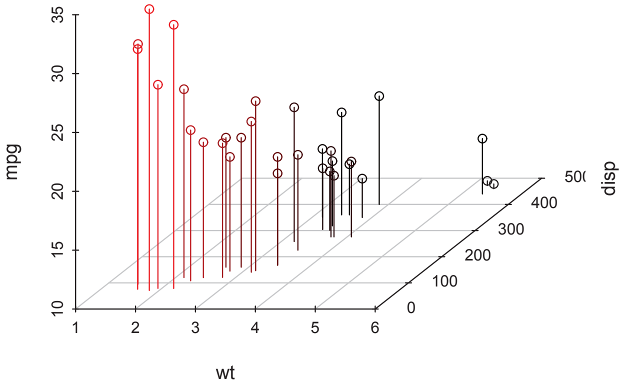


From these figures, it seems that the variables auxin and kinetin as not good predictors for the callus wight.

15. The produced 3D scatterplot is given below:

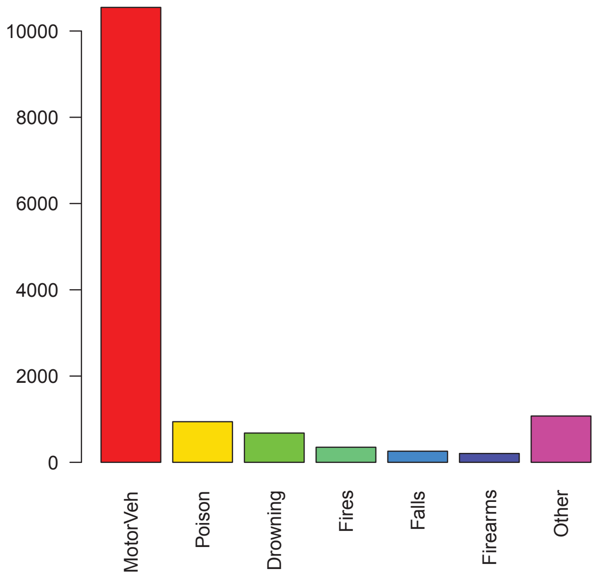


The resulting 3D scatterplot after replacing *box=T* by *box=F*:

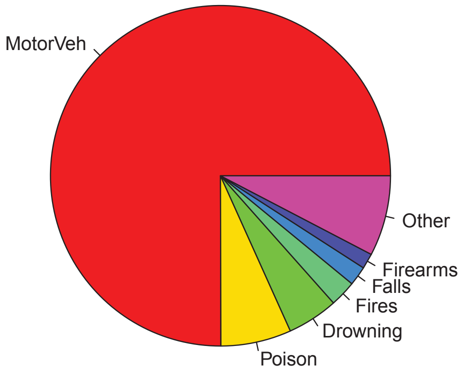


We can clearly see the difference.

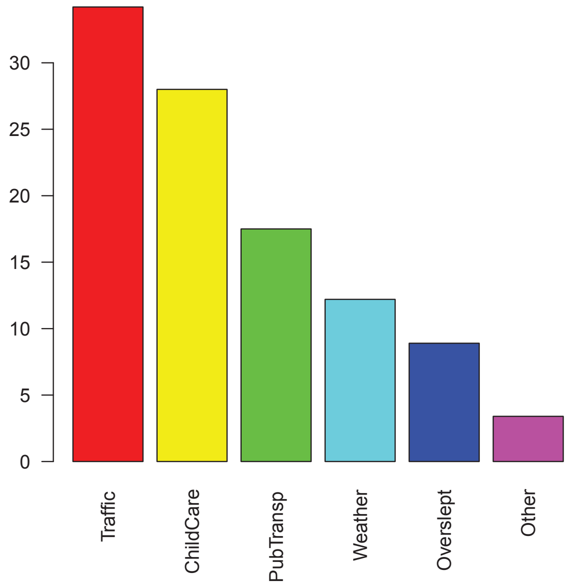
16. The produced bar graph is shown below



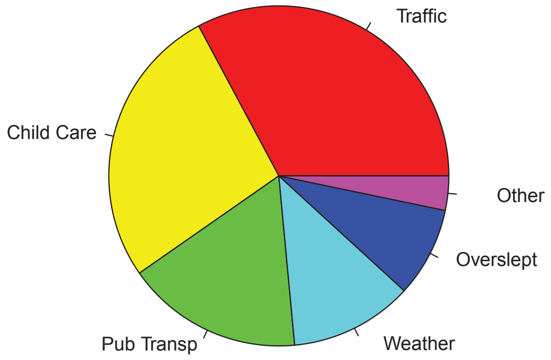
The produced pie graph follows



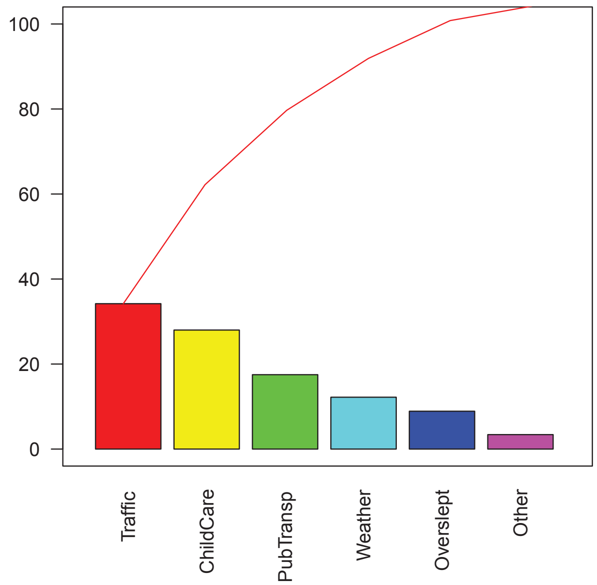
17. (a) The produced bar graph is shown below



The produced pie graph follows



(b) The produced figure is shown below



1.6 Proportions, Averages, and Variances

1. (a) $\bar{x}$.

   (b) S.

   (c) $\hat{p}$.

2. (a) μ.

   (b) σ.

   (c) p.

3. $\hat{p} = 4/14 = 0.286$. It estimates the proportion of time when the ozone level is below 250.

4. They estimate the proportion of all concrete cylinders, constructed by the specifications listed, whose 28-day compressive-strength measure is no more than 44, and at least 47, respectively.

5. (a) $\sigma = 0.877, \sigma^2 = 0.769$.

   (b) $S = 0.949, S^2 = 0.9$.

6. After repeating the commands five times, we obtain the five pairs of $(\bar{x}, S)$ as (3.28, 0.90), (3.34, 0.64), (3.52, 0.50), (3.38, 0.73), and (3.32, 0.83).

7. (a) $\mu = 0.92, \sigma = 0.8207, \sigma^2 = 0.6736$.

(b) $\bar{x} = 0.91, S = 0.8177, S^2 = 0.6686.$

8. (a) After running the commands five times, we obtain the results as (0.44, 0.31, 0.25), (0.27, 0.33, 0.40), (0.38, 0.33, 0.29), (0.34, 0.38, 0.28), and (0.39, 0.30, 0.31). Each of the results gives an estimation of the population proportions, for example, the first gives the estimated proportions of 0, 1 and 2 are 0.44, 0.31, and 0.25, respectively.

   (b) After running the commands five times, we obtain the results as (0.87, 0.62, 0.79), (0.94, 0.62, 0.79), (1.06, 0.66, 0.81), (1.09, 0.65, 0.81), and (0.94, 0.70, 0.84). Each of the above results gives the estimated values of μ, σ^2, and σ.

9. (a) $\mu_X = 3.5, \sigma_X^2 = 2.92.$

   (b) After running the commands five times, we obtain the following results (3.52, 2.64), (3.49, 2.70), (3.43, 3.03), (3.37, 3.10), (3.74, 3.00). We can observe that the sample mean and sample variance approximate the population mean and population variance reasonably well.

   (c) After running the commands five times, we obtain the following sample proportions: (0.12, 0.20, 0.13, 0.25, 0.12, 0.18), (0.19, 0.17, 0.17, 0.21, 0.17, 0.09), (0.20, 0.11, 0.17, 0.17, 0.17, 0.18), (0.14, 0.14, 0.19, 0.18, 0.19, 0.16), and (0.13, 0.28, 0.12, 0.18, 0.14, 0.15). They are reasonably close to 1/6.

10. (a) $\sigma_X^2 = 0.25.$

    (b) $S_1^2 = 0, S_2^2 = 0.5, S_3^2 = 0.5, S_4^2 = 0.$

    (c) $E(Y) = (0 + 0.5 + 0.5 + 0)/4 = 0.25.$

    (d) We can see that $\sigma_X^2 = E(Y)$. If the sample variances in part (b) were computed according to a formula that divides by n instead of $n-1$, $E(Y)$ would have been 0.125.

11. (a) $\bar{x}_1 = 30, \bar{x}_2 = 30.$

    (b) $S_1^2 = 0.465, S_2^2 = 46.5.$

    (c) There is more uniformity among cars of type A (smaller variability in achieved gas mileage), so type A cars are of better quality.

12. (a) For the mean value:

$$\mu_w = \frac{\sum_{i=1}^N w_i}{N} = \frac{\sum_{i=1}^N (c_1 + v_i)}{N} = \frac{Nc_1 + \sum_{i=1}^N v_i}{N} = c_1 + \mu_v.$$

For the variance:

$$\sigma_w^2 = \frac{\sum_{i=1}^N (w_i - \mu_w)^2}{N} = \frac{\sum_{i=1}^N (c_1 + v_i - (c_1 + \mu_v))^2}{N} = \frac{\sum_{i=1}^N (v_i - \mu_v)^2}{N} = \sigma_v^2.$$

Consequently, $\sigma_w = \sqrt{\sigma_w^2} = \sqrt{\sigma_v^2} = \sigma_v$.

(b) For the mean value:

$$\mu_w = \frac{\sum_{i=1}^N w_i}{N} = \frac{\sum_{i=1}^N c_2 v_i}{N} = \frac{c_2 \sum_{i=1}^N v_i}{N} = c_2 \mu_v.$$

For the variance:

$$\sigma_w^2 = \frac{\sum_{i=1}^N (w_i - \mu_w)^2}{N} = \frac{\sum_{i=1}^N (c_2 v_i - c_2 \mu_v))^2}{N} = \frac{c_2^2 \sum_{i=1}^N (v_i - \mu_v)^2}{N} = c_2^2 \sigma_v^2.$$

Consequently, $\sigma_w = \sqrt{\sigma_w^2} = \sqrt{c_2^2 \sigma_v^2} = |c_2| \sigma_v$.

(c) Let $u_i = c_2 v_i$ for $i = 1, 2, \cdots, N$. Then we have $w_i = c_1 + u_i$ for $i = 1, 2, \cdots, N$. From part (b), we have

$$\mu_u = c_2 \mu_v, \;\; \sigma_u^2 = c_2^2 \sigma_v^2, \;\; \sigma_u = |c_2| \sigma_v.$$

From part (a),

$$\mu_w = c_1 + \mu_u = c1 + c2\mu_v, \;\; \sigma_w^2 = \sigma_u^2 = c_2^2 \sigma_v^2, \;\; \sigma_w = \sigma_u = |c_2| \sigma_v.$$

13. (a) For the mean value:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n c_1 + x_i}{n} = \frac{nc_1 + \sum_{i=1}^n x_i}{n} = c_1 + \bar{x}.$$

For the variance:

$$S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (c_1 + x_i - (c_1 + \bar{x}))^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = S_x^2.$$

Consequently, $S_y = \sqrt{S_y^2} = \sqrt{S_x^2} = S_x$.

(b) For the mean value:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n c_2 x_i}{n} = \frac{c_2 \sum_{i=1}^n x_i}{n} = c_2 \bar{x}.$$

For the variance:

$$S_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^n (c_2 x_i - c_2 \bar{x}))^2}{n-1} = \frac{c_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}{n-1} = c_2^2 S_x^2.$$

Consequently, $S_y = \sqrt{S_y^2} = \sqrt{c_2^2 S_x^2} = |c_2| S_x$.

(c) Let $u_i = c_2 x_i$ for $i = 1, 2, \cdots, n$. Then we have $y_i = c_1 + u_i$ for $i = 1, 2, \cdots, n$. From part (b), we have

$$\bar{u} = c_2 \bar{x}, \;\; S_u^2 = c_2^2 S_x^2, \;\; S_u = |c_2| S_x.$$

From part (a),

$$\bar{y} = c_1 + \bar{u} = c1 + c2\bar{x}, \;\; S_y^2 = S_u^2 = c_2^2 S_x^2, \;\; S_y = S_u = |c_2| S_x.$$

14. Let x_i, $i = 1, \cdots, 7$, be the temperature expressed in Celsius scale, and let y_i, $i = 1, \cdots, 7$, be the temperature expressed in Fahrenheit scale. Then $y_i = 1.8x_i + 32$. From the given information, $\bar{x} = 31$ and $S_x = 1.5$. By the results in 13 (c), we have

$$\bar{y} = 1.8\bar{x} + 32 = 1.8 \times 31 + 32 = 87.8, \quad S_y = |1.8|S_x = 1.8 \times 1.5 = 2.7.$$

15. Let the coded data be $y_i = (x_i - 81.2997) \times 10000$, for $i = 1, \cdots, 7$. Thus, by the result of 13 (c), $S_y^2 = 10000^2 S_x^2 = 10^8 S_x^2$. Therefore, $S_x^2 = 10^{-8} S_y^2 = 6.833 \times 10^{-7}$.

16. (a) The estimated population mean is $\bar{x} = 192.8$ and the estimated population variance is $S_x^2 = 312.31$

    (b) Let y_i be the second-year salary, for $i = 1, \cdots, 15$.

    (i) Since $y_i = x_i + 5$, $\bar{y} = \bar{x} + 5 = 197.8$ and $S_y^2 = S_x^2 = 312.31$.

    (ii) Since $y_i = 1.05x_i$, $\bar{y} = 1.05\bar{x} = 202.44$ and $S_y^2 = 1.05^2 S_x^2 = 344.33$.

1.7 Medians, Percentiles, and Boxplots

1. (a) The sample median is $\tilde{x} = 717$, the 25th percentile is $q_1 = (691+699)/2 = 695$, and the 75th percentile is $q_3 = (734 + 734)/2 = 734$.

   (b) The sample interquartile range is $IQR = q_3 - q_1 = 734 - 695 = 39$.

   (c) The sample percentile is $100 \times (19 - 0.5)/40 = 46.25$.

2. (a) The sample median is $\tilde{x} = 30.55$, the 25th percentile is $q_1 = 29.59$, and the 75th percentile is $q_3 = 31.41$.

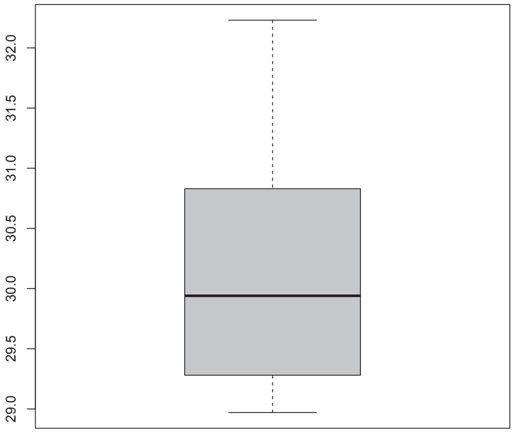
   (b) The sample interquartile range is $IQR = q_3 - q_1 = 31.41 - 29.59 = 1.82$.

   (c) The sample percentile is $100 \times (19 - 0.5)/22 = 84.09$.

3. (a) After running the code, we obtain the results as $x_{(1)} = 28.97$, $q_1 = 29.30$, $\tilde{x} = 29.94$, $q_3 = 30.82$, and $x_{(n)} = 32.23$.

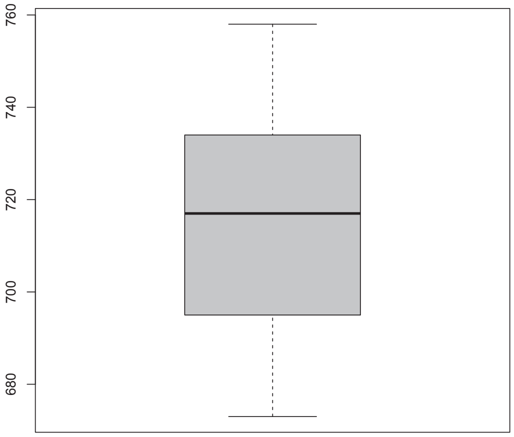
   (b) The 90th percentile is 31.068.

   (c) The boxplot is shown as follows



Clearly, there are no outliers.

4. (a) The boxplot is shown as follows



(b) The 30th, 60th, and 90th sample percentiles are 700.7, 720.8, and 746.0 , respectively.

1.8 Comparative Studies

1. (a) The experimental units are the batches of cake.

(b) The factors are baking time and temperature.

(c) The levels for baking time are 25 and 30 minutes, and the levels for temperature are 275°F, 300°F, and 325°F.

(d) All the treatments are (25, 275), (25, 300), (25, 325), (30, 275), (30,300), (30, 325).

(e) The response variable is qualitative.

2. (a) There are three populations involved in this study.

   (b) True

   (c) False

   (d) In this study, each of the three watering regimens is considered as a treatment.

   (e) With the changes in the study:

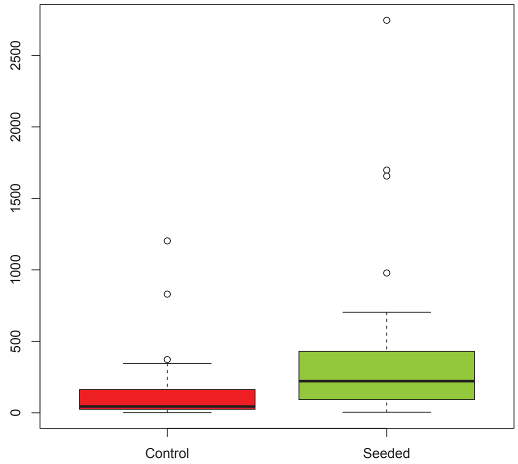
      (i) This will change the number of populations.

      (ii) Watering regimen with levels W_1, W_2, W_3, and location with levels L_1, L_2, L_3. The treatments are all (W_i, L_j) where $i = 1, 2, 3$ and $j = 1, 2, 3.$.

3. (a) Let $\mu = (\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5)/5$, then the contrasts that represent the effects of each area are $\alpha_i = \mu_i - \mu$, for $i = 1, \cdots, 5$.

   (b) The contrast is $(\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4 + \mu_5)/3$.

4. The comparative boxplot is shown as follows



The comparative boxplot shows that, in general, the seeded clouds could produce more rainfall than the unseeded clouds.

5. The three control versus treatment contrasts are $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_4 - \mu_1$.

6. (a) There are four populations involved in this study.

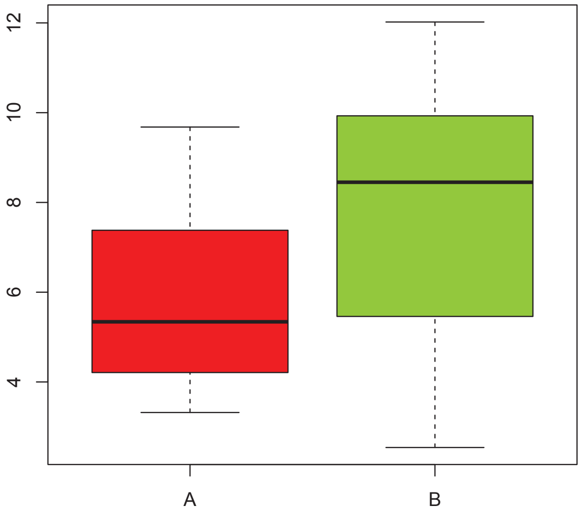
   (b) In this study, each of the four new types of paint is considered as a treatment.

   (c) The three control versus treatment contrasts are $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_4 - \mu_1$.

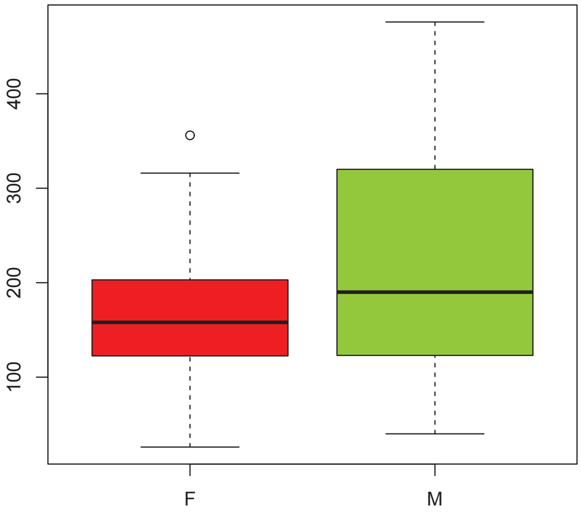
7. (a) This will change the number of populations.

   (b) Paint type with levels $T_1, \cdots, T_4$, and location with levels $L_1, \cdots, L_4$. The treatments are all (T_i, L_j) where $i = 1, \cdots, 4$ and $j = 1, \cdots, 4$.

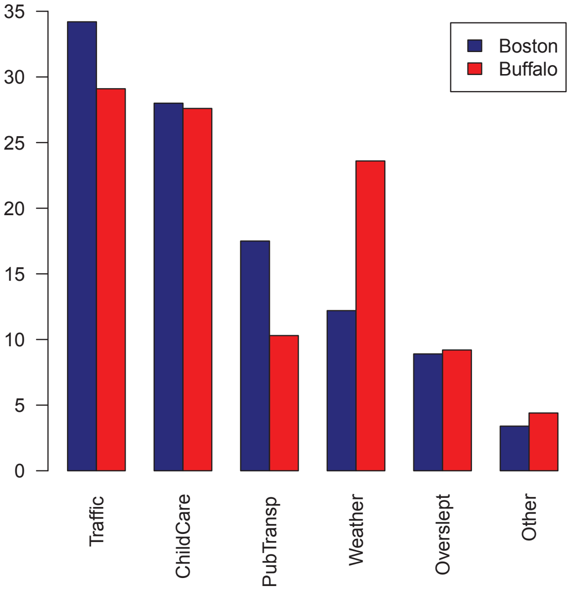
8. The comparative boxplot is given as follows, and it shows that the type B material, on average, has higher ignition time than type A material.



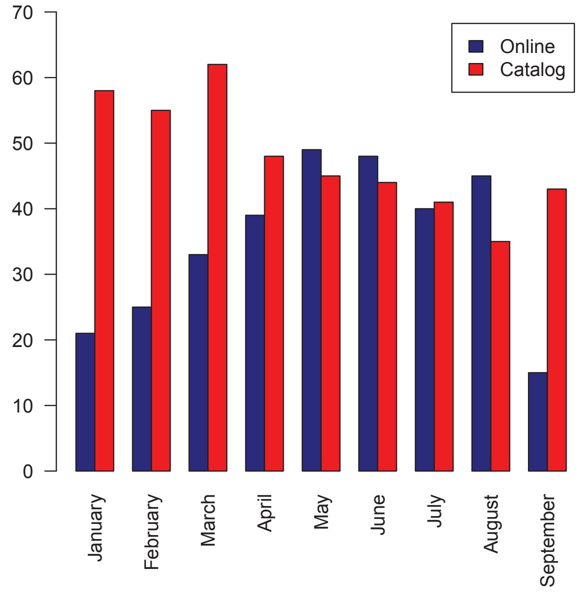
9. The comparative boxplot is given below, and it shows that the male bears, on average, are heavier than female bears, but the weights of male bears are spread in a wider range.



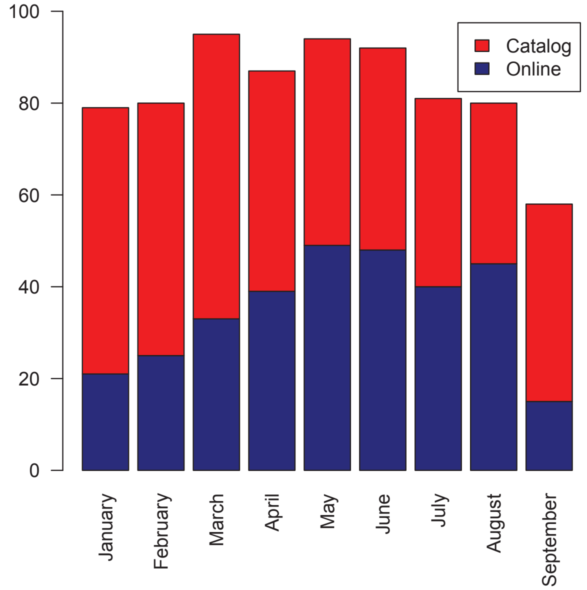
10. The comparative bar graph is shown in the following figure. The reason "weather" is the one with the biggest difference between the two cities for being late to work.



11. (a) The comparative bar graph for online and catalog volumes of sale is as follows



(a) The stacked bar graph for online and catalog volumes of sale is as follows



(c) The comparative bar graph is better for comparing the volume of sales for online and catalog, while the stacked bar graph is better for showing variation in the total volume of sales.

12. The watering and location effects will be confounded. The three watering regimens should be employed in each location. The root systems in each location should be

assigned randomly to a watering regimen.

13. The paints and location effects will be confounded. The four types of new paint should be used in each location. The road segments should be assigned randomly to a new type of paint.

14. (a) There are four populations in this study.

    (b) True

    (c) False

    (d) The factor fertilization has two levels, F_1 and F_2, and the factor watering has two levels, W_1 and W_2.

    (e) True

15. (a) Of 2590 male applicants, about 1192 were admitted. Similarly, of the 1835 female applicants, about 557 were admitted. Thus, the admission rates for men and women are 0.46 and 0.30, respectively.

    (b) Yes

    (c) No, because the major specific admission rates are higher for women for most majors.

16. (a) This is not an additive design because the Pygmalion effect is stronger for female recruits.

    (b) Here, $\bar{\mu}_{..} = (8 + 13 + 12 + 10)/4 = 10.75$. Therefore, the main gender effects are

$$\alpha_F = \bar{\mu}_{F.} - \bar{\mu}_{..} = (8 + 13)/2 - 10.75 = -0.25$$

and

$$\alpha_M = \bar{\mu}_{M.} - \bar{\mu}_{..} = (10 + 12)/2 - 10.75 = 0.25.$$

The main Pygmalion effects are

$$\beta_C = \bar{\mu}_{.C} - \bar{\mu}_{..} = (8 + 10)/2 - 10.75 = -1.75$$

and

$$\beta_P = \bar{\mu}_{.P} - \bar{\mu}_{..} = (13 + 12)/2 - 10.75 = 1.75.$$

    (c) The interaction effects are computed as following:

$$\gamma_{FC} = \mu_{FC} - (\bar{\mu}_{..} + \alpha_F + \beta_C) = 8 - (10.75 - 0.25 - 1.75) = -0.75,$$

$$\gamma_{FP} = \mu_{FP} - (\bar{\mu}_{..} + \alpha_F + \beta_P) = 13 - (10.75 - 0.25 + 1.75) = 0.75,$$

$$\gamma_{MC} = \mu_{MC} - (\bar{\mu}_{..} + \alpha_M + \beta_C) = 10 - (10.75 + 0.25 - 1.75) = 0.75,$$

$$\gamma_{MP} = \mu_{MP} - (\bar{\mu}_{..} + \alpha_M + \beta_P) = 12 - (10.75 + 0.25 + 1.75) = -0.75.$$

17. (a) Omitted

(b) The interaction plot shows that the traces are not parallel; therefore, there is interaction between pH and temperature.

(c) $\bar{\mu}_{..} = (108 + 103 + 101 + 100 + 111 + 104 + 100 + 98)/8 = 103.125$. Therefore, the main PH effects are

$$\alpha_I = \bar{\mu}_{I.} - \bar{\mu}_{..} = (108 + 103 + 101 + 100)/4 - 103.125 = -0.125$$

and

$$\alpha_{II} = \bar{\mu}_{II.} - \bar{\mu}_{..} = (111 + 104 + 100 + 98)/4 - 103.125 = 0.125.$$

The main temperature effects are

$$\beta_A = \bar{\mu}_{.A} - \bar{\mu}_{..} = (108 + 111)/2 - 103.125 = 6.375,$$

$$\beta_B = \bar{\mu}_{.B} - \bar{\mu}_{..} = (103 + 104)/2 - 103.125 = 0.375,$$

$$\beta_C = \bar{\mu}_{.C} - \bar{\mu}_{..} = (101 + 100)/2 - 103.125 = -2.625,$$

and

$$\beta_D = \bar{\mu}_{.D} - \bar{\mu}_{..} = (100 + 98)/2 - 103.125 = -4.125.$$

(d) The interaction effects are computed as following:

$$\gamma_{IA} = \mu_{IA} - (\bar{\mu}_{..} + \alpha_I + \beta_A) = 108 - (103.125 + (-0.125) + 6.375) = -1.375,$$

$$\gamma_{IB} = \mu_{IB} - (\bar{\mu}_{..} + \alpha_I + \beta_B) = 103 - (103.125 + (-0.125) + 0.375) = -0.375,$$

$$\gamma_{IC} = \mu_{IC} - (\bar{\mu}_{..} + \alpha_I + \beta_C) = 101 - (103.125 + (-0.125) + (-2.625)) = 0.625,$$

$$\gamma_{ID} = \mu_{ID} - (\bar{\mu}_{..} + \alpha_I + \beta_D) = 100 - (103.125 + (-0.125) + (-4.125)) = 1.125,$$

$$\gamma_{IIA} = \mu_{IIA} - (\bar{\mu}_{..} + \alpha_{II} + \beta_A) = 111 - (103.125 + 0.125 + 6.375) = 1.375,$$

$$\gamma_{IIB} = \mu_{IIB} - (\bar{\mu}_{..} + \alpha_{II} + \beta_B) = 104 - (103.125 + 0.125 + 0.375) = 0.375,$$

$$\gamma_{IIC} = \mu_{IIC} - (\bar{\mu}_{..} + \alpha_{II} + \beta_C) = 100 - (103.125 + 0.125 + (-2.625)) = -0.625,$$

and

$$\gamma_{IID} = \mu_{IID} - (\bar{\mu}_{..} + \alpha_{II} + \beta_D) = 98 - (103.125 + 0.125 + (-4.125)) = -1.125.$$

18. (a) The R codes are as following:

```
SMT=read.table("SpruceMothTrap.txt", header=T)
mcm=tapply(SMT$Moth, SMT[,c(1, 3)], mean)
alphas=rowMeans(mcm)-mean(mcm)
betas=colMeans(mcm)-mean(mcm)
gammas=t(t(mcm-mean(mcm)-alphas) -betas)
```

The computed matrix of cell means is

	Lure		
Location	Chemical	Scent	Sugar
ground	26.57143	24.28571	28.14286
lower	42.71429	38.57143	37.57143
middle	37.28571	34.28571	41.57143
top	30.42857	29.14286	32.57143

The computed main effects for ground, lower, middle, and top are -7.261905, 6.023810, 4.119048, and -2.880952, respectively, while the computed main effects for Chemical, Scent, and Sugar are 0.6547619, -2.0238095, and 1.3690476, respectively.

The computed interaction effects are

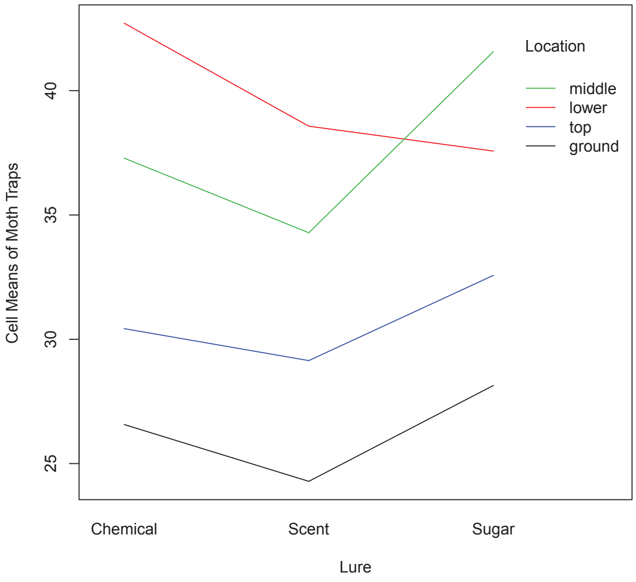
	Lure		
Location	Chemical	Scent	Sugar
ground	-0.4166667	-0.02380952	0.4404762
lower	2.4404762	0.97619048	-3.4166667
middle	-1.0833333	-1.40476190	2.4880952
top	-0.9404762	0.45238095	0.4880952

(b) The R commands for the interaction plot are shown as following

attach(SMT) # so variables can be referred to by name

interaction.plot(Lure,Location, Moth, col=c(1,2,3,4), lty = 1, xlab="Lure", ylab="Cell Means of Moth Traps", trace.label="Location")

The interaction plot is given in the following figure. According to this figure, there are interactive effects.



19. (a) The R codes are as following:

ALN=read.table("AdLocNews.txt", header=T)

mcm=tapply(ALN$Inquiries, ALN[,c(1, 3)], mean)

alphas=rowMeans(mcm)-mean(mcm)

betas=colMeans(mcm)-mean(mcm)

gammas=t(t(mcm-mean(mcm)-alphas) -betas)

The computed matrix of cell means is

	Section		
Day	Business	News	Sports
Friday	12.00	15.50	14.25
Monday	14.50	11.25	7.50
Thursday	10.75	7.25	9.00
Tuesday	11.75	13.25	9.50
Wednesday	11.50	12.25	9.75

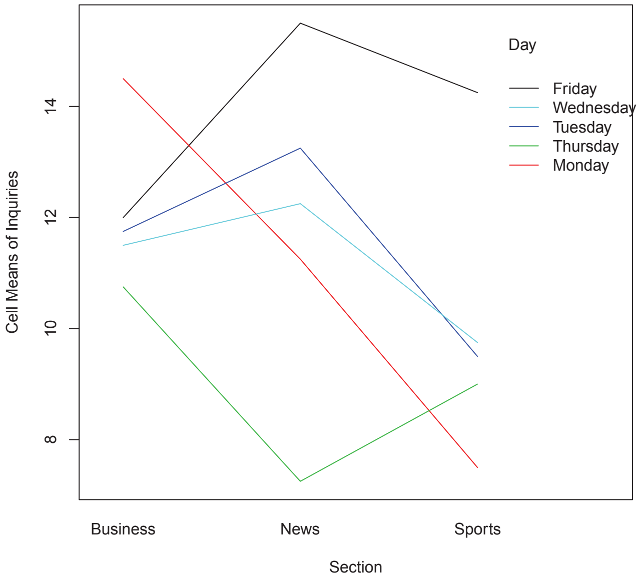
The computed main effects for Friday, Monday, Thursday, Tuesday, and Wednesday are, 2.58, -0.25, -2.33, 0.17, and -0.17, respectively; while the computed main effects for Business, News, and Sports are 0.77, 0.57, and -1.33, respectively.

The computed interaction effects are

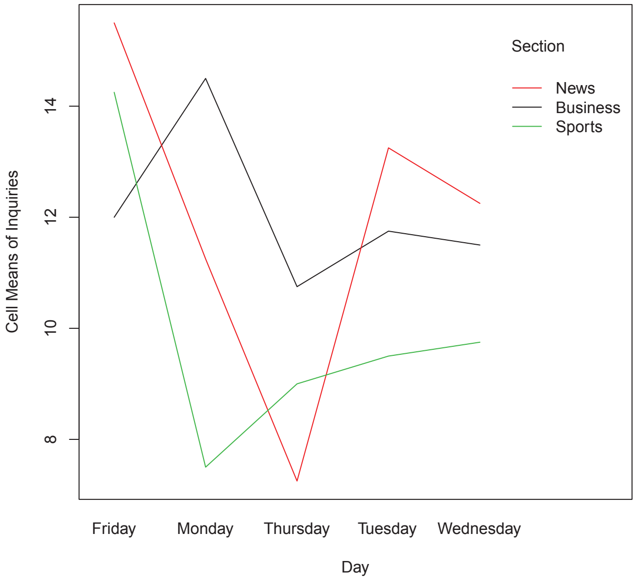
	Section		
Day	Business	News	Sports
Friday	-2.6833333	1.0166667	1.66666667
Monday	2.6500000	-0.4000000	-2.25000000
Thursday	0.9833333	-2.3166667	1.33333333
Tuesday	-0.5166667	1.1833333	-0.66666667
Wednesday	-0.4333333	0.5166667	-0.08333333

The overall best day to put a newspaper ad is on Friday, and the overall best newspaper section is Business.

(b) The interaction plot with the levels of the factor day being traced:



The interaction plot with the levels of the factor section being traced:



These plots show that there are interaction effect between the factor day and section, and the (Friday, News) combination has the most inquiries.

Chapter 2
Introduction to Probability

2.2 Sample Spaces, Events, and Set Operations

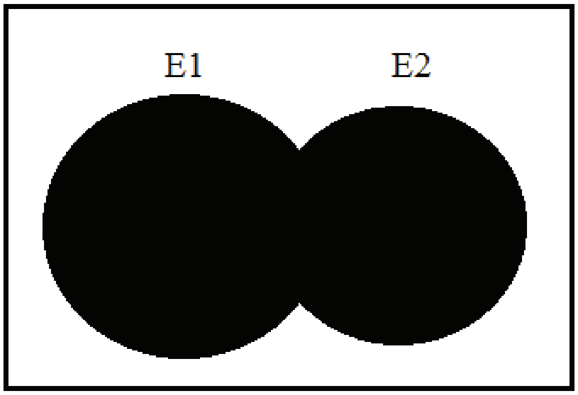
1. (a) The sample space is $\{(1,1),(1,2),\cdots,(1,6),\cdots,\cdots,(6,1),(6,2),\cdots,(6,6)\}$.

   (b) The sample space is $\{2,3,4,\cdots,12\}$.

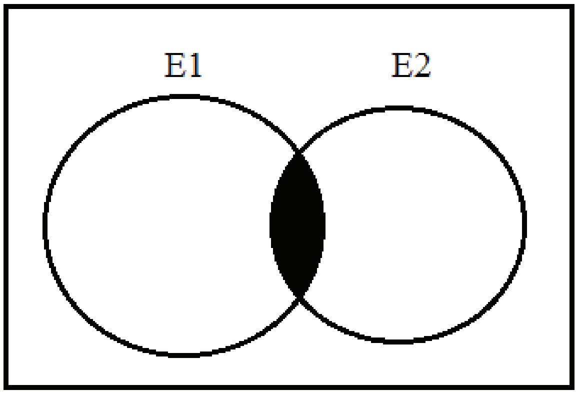
   (c) The sample space is $\{0,1,2,\cdots,6\}$.

   (d) The sample space is $\{1,2,3,\cdots\}$.

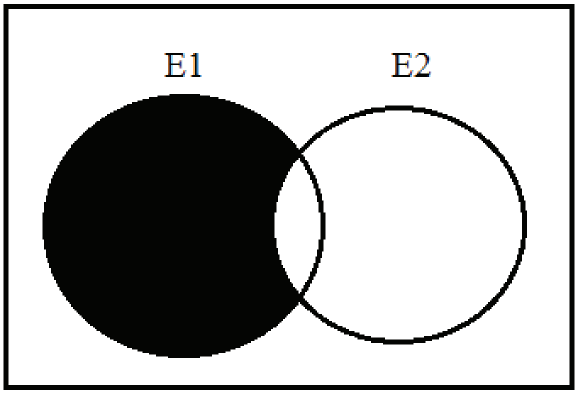
2. (a) The Venn diagram is shown as



   (b) The Venn diagram is shown as



(c)The Venn diagram is shown as



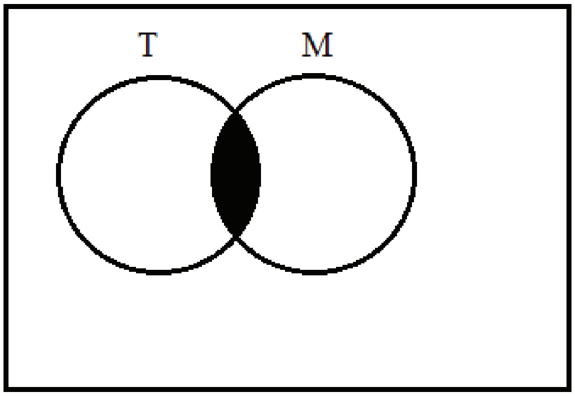
3. (a) The events are represented as

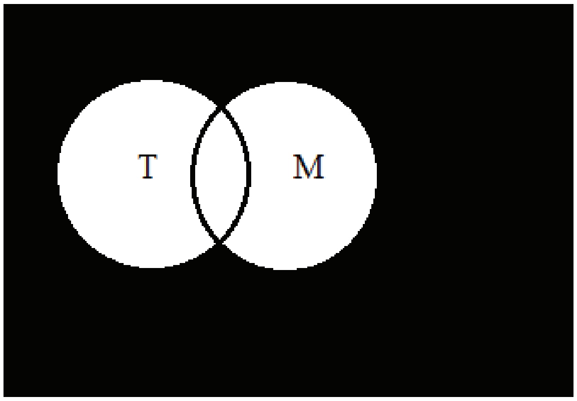
   (i) $T \cap M$

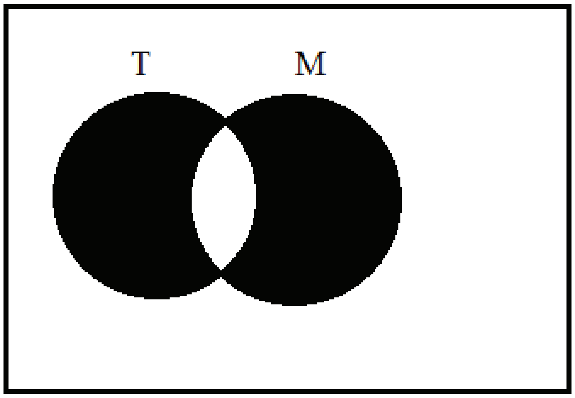
   (ii) $T^c \cap M^c$

   (iii) $(T \cap M^c) \cup (T^c \cap M)$

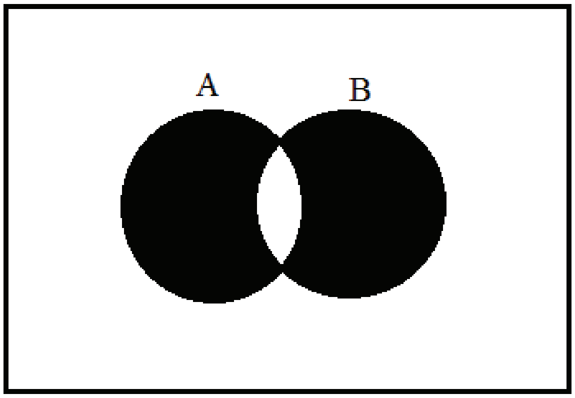
   (b) The Venn diagrams for part (a) are shown as







4. Both of the Venn diagrams should be similar to



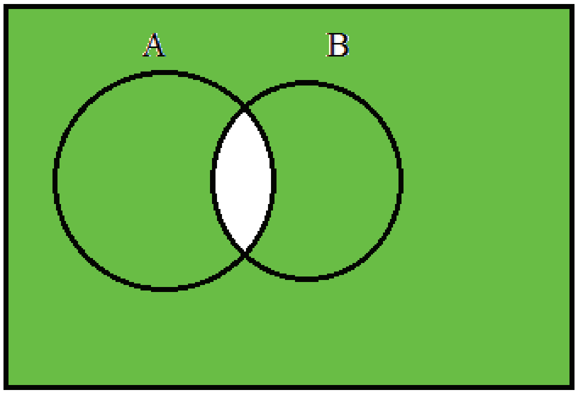
5. (a) $A^c = \{x | x \geq 75\}$, the component will last at least 75 time units.

   (b) $A \cap B = \{x | 53 < x < 75\}$, the component will last more than 53 units but less than 75 time units.

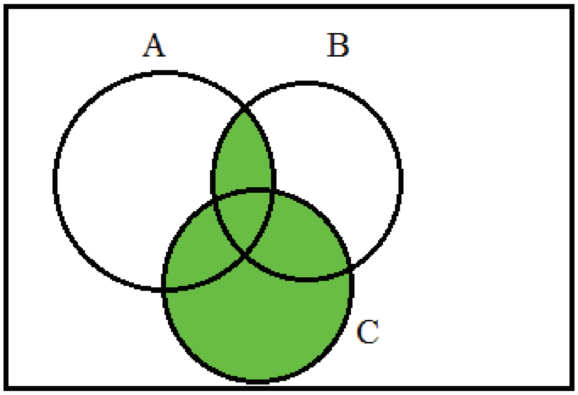
   (c) $A \cup B = S$, the sample space.

   (d) $(A - B) \cup (B - A) = \{x | x \geq 75 \text{ or } x \leq 53\}$, the component will last either at most 53 or at least 75 time units.

6. Both of the Venn diagrams should be similar to



7. Both of the Venn diagrams should be similar to



8. (a) Prove that $(A - B) \cup (B - A) = (A \cup B) - (A \cap B)$:

$$\begin{aligned} x \in (A - B) \cup (B - A) &\Leftrightarrow x \in A - B \text{ or } x \in B - A \\ &\Leftrightarrow x \in A \text{ but } x \notin B \text{ or } x \in B \text{ but } x \notin A \\ &\Leftrightarrow x \in A \text{ or } x \in B \text{ but not in both} \\ &\Leftrightarrow x \in A \cup B \text{ and } x \notin A \cap B \\ &\Leftrightarrow x \in (A \cup B) - (A \cap B). \end{aligned}$$

(b) Prove that $(A \cap B)^c = A^c \cup B^c$:

$$\begin{aligned} x \in (A \cap B)^c &\Leftrightarrow x \notin A \cap B \\ &\Leftrightarrow x \in A - B \text{ or } x \in B - A \text{ or } x \in (A \cup B)^c \\ &\Leftrightarrow [x \in A - B \text{ or } x \in (A \cup B)^c] \text{ or } [x \in B - A \text{ or } x \in (A \cup B)^c] \\ &\Leftrightarrow x \in B^c \text{ or } x \in A^c \Leftrightarrow x \in A^c \cup B^c. \end{aligned}$$

(c) Prove that $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$:

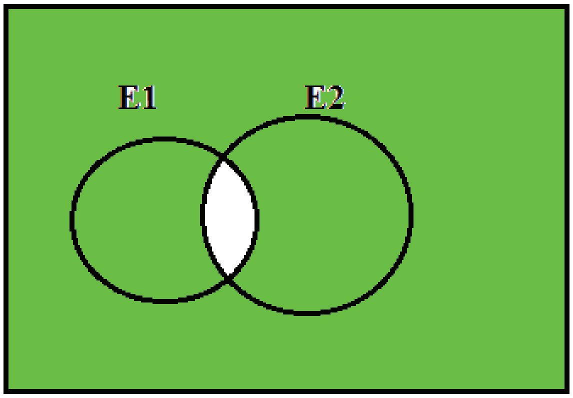
$$\begin{aligned} x \in (A \cap B) \cup C &\Leftrightarrow x \in A \cap B \text{ or } x \in C \\ &\Leftrightarrow x \in A \text{ and } x \in B \text{ or } x \in C \\ &\Leftrightarrow [x \in A \text{ or } x \in C] \text{ and } [x \in B \text{ or } x \in C] \\ &\Leftrightarrow x \in A \cup C \text{ and } x \in B \cup C \\ &\Leftrightarrow x \in (A \cup C) \cap (B \cup C). \end{aligned}$$

9. (a) The sample space is $S = \{(x_1, x_2, x_3, x_4, x_5) | x_i = 5.3, 5.4, 5.5, 5.6, 5.7, i = 1, 2, 3, 4, 5\}$. The size of the sample space is $5^5 = 3125$.

(b) The sample space is the collection of the distinct averages, $(x_1 + x_2 + x_3 + x_4 + x_5)/5$, formed from the elements of S. The R commands *s=c(5.3,5.4,5.5, 5.6,5.7); Sa= expand.grid(x1=s,x2=s,x3=s, x4=s,x5=s);*
length(table(rowSums(Sa))) return 21 for the size of the sample space of the averages.

10. (a) The number of disks in E_1 is $5+16 = 21$, the number of disks in E_2 is $5+9 = 14$, and the number of disks in E_3 is $5 + 16 + 9 = 30$.

    (b) Both of the Venn diagrams should be similar to



    (c) $E_1 \cap E_2$ is the event that "the disk has low hardness and low shock absorption," $E_1 \cup E_2$ is the event that "the disk has low hardness or low shock absorption," $E_1 - E_2$ is the event that "the disk has low hardness but does not have low shock absorption," and $(E_1 - E_2) \cup (E_2 - E_1)$ is the event that "the disk has low hardness or low shock absorption but does not have low hardness and low shock absorption at the same time."

    (d) The number of disks in $E_1 \cap E_2$ is 5, the number of disks in $E_1 \cup E_2$ is 30, the number of disks in $E_1 - E_2$ is 16, and the number of disks in $(E_1 - E_2) \cup (E_2 - E_1)$ is 25.

2.3 Experiments with Equally Likely Outcomes

1. $P(E_1) = 0.5$, $P(E_2) = 0.5$, $P(E_1 \cap E_2) = 0.3$, $P(E_1 \cup E_2) = 0.7$, $P(E_1 - E_2) = 0.2$, $P((E_1 - E_2) \cup (E_2 - E_1)) = 0.4$.

2. (a) If we select two wafers with replacement, then

    (i) The sample space for the experiment that records the doping type is {(n-type, n-type), (n-type, p-type), (p-type, n-type), (p-type, p-type)} and the corresponding probabilities are 0.25, 0.25, 0.25, and 0.25.

    (ii) The sample space for the experiment that records the number of n-type wafers is {0, 1, 2} and the corresponding probabilities are 0.25, 0.50, and 0.25.

    (b) If we select four wafers with replacement, then

    (i) The sample space for the experiment that records the doping type is all of the 4-component vectors, with each element being n-type or p-type. The size of the sample space can be found by the R commands

G=expand.grid(W1=0:1, W2=0:1, W3=0:1, W4=0:1); length(G$W1)

and the result is 16. The probability of each outcome is 1/16.

(ii) The sample space for the experiment that records the number of n-type wafer is {0, 1, 2, 3, 4}. The PMF is given by

x	0	1	2	3	4
$p(x)$	0.0625	0.2500	0.3750	0.2500	0.0625

(iii) The probability of at most one n-type wafer is 0.0625+0.25 = 0.3125.

3. $E_1 = \{6.8, 6.9, 7.0, 7.1\}$ and $E_2 = \{6.9, 7.0, 7.1, 7.2\}$. Thus, $P(E_1) = P(E_2) = 4/5$. $E_1 \cap E_2 = \{6.9, 7.0, 7.1\}$ and $P(E_1 \cap E_2) = 3/5$. $E_1 \cup E_2 = S$ and $P(E_1 \cup E_2) = 1$. $E_1 - E_2 = \{6.8\}$ and $P(E_1 - E_2) = 1/5$. Finally, $(E_1 - E_2) \cup (E_2 - E_1) = \{6.8, 7.2\}$, so $P(E_1 - E_2) \cup (E_2 - E_1)) = 2/5$.

4. (a) If the water PH level is measured over the next two irrigations, then

(i) The sample space is $S = \{(x_1, x_2) : x_1 = 6.8, 6.9, 7.0, 7.1, 7.2, \text{and } x_2 = 6.8, 6.9, 7.0, 7.1, 7.2\}$. The size of the sample space is 25.

(ii) The sample space of the experiment that records the average of the two PH measurements is $S = \{6.8, 6.85, 6.9, 6.95, 7, 7.05, 7.1, 7.15, 7.2\}$ and the PMF is

x	6.8	6.85	6.9	6.95	7	7.05	7.1	7.15	7.2
$p(x)$	0.04	0.08	0.12	0.16	0.20	0.16	0.12	0.08	0.04

(b) The probability mass function of the experiment that records the average of the PH measurements taken over the next five irrigations is

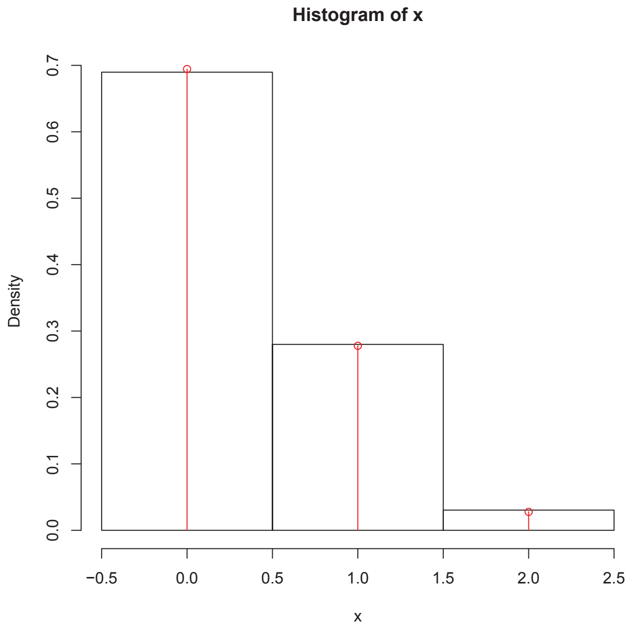
x	6.8	6.82	6.84	6.86	6.88	6.9	6.92
$p(x)$	0.00032	0.00160	0.00480	0.01120	0.02240	0.03872	0.05920
x	6.94	6.96	6.98	7	7.02	7.04	7.06
$p(x)$	0.08160	0.10240	0.11680	0.12192	0.11680	0.10240	0.08160
x	7.08	7.1	7.12	7.14	7.16	7.18	7.2
$p(x)$	0.05920	0.03872	0.02240	0.01120	0.00480	0.00160	0.00032

5. (a) The R command is *sample(0:2, size =10, replace=T, prob=pr)* and the following gives one possible result: 1, 0, 0, 1, 1, 0, 1, 0, 0, 0.

(b) The relative frequency based on 10,000 replications is

0	1	2
0.6897	0.2799	0.0304

(c) The histogram of the relative frequencies and line graph of the probability mass function is given on the next page.



This figure shows that all relative frequencies are good approximations to corresponding probabilities and we have empirical confirmation of the limiting relative frequency interpretation of probability.

6. (a) The number of ways to finish the test is $2^5 = 32$.

   (b) The sample space for the experiment that records the test score is $S = \{0, 1, 2, 3, 4, 5\}$.

   (c) The PMF of X is given by

X	0	1	2	3	4	5
$p(x)$	0.03125	0.15625	0.31250	0.31250	0.15625	0.03125

7. The number of assignments is

$$\binom{4}{1,1,1,1} = 24.$$

8. The probability is

$$\frac{26^2 \times 10^3}{26^3 \times 10^4} = 0.0038.$$

9. (a) The number of possible committees is

$$\binom{12}{4} = 495.$$

(b) The number of committees consisting of 2 biologists, 1 chemist, and 1 physicist is

$$\binom{5}{2}\binom{4}{1}\binom{3}{1} = 120.$$

(c) The probability is $120/495 = 0.2424$.

10. (a) The number of possible selections is

$$\binom{10}{5} = 252.$$

(b) The number of divisions of the 10 players into two teams of 5 is $252/2 = 126$.

(c) The number of handshakes is

$$\binom{12}{2} = 66.$$

11. (a) In order to go from the lower left corner to the upper right corner, we need to totally move 8 steps, with 4 steps to the right and 4 steps upwards. Thus, the total number of paths is

$$\binom{8}{4} = 70.$$

(b) We decompose the move as two stages: stage 1 is from lower left corner to circled point, which needs 5 steps with 3 steps to the right and 2 steps upwards; stage 2 is from the circled point to the upper right corner, which needs 3 steps with 1 step to the right and 2 steps upwards. Thus, the total number of paths passing the circled point is

$$\binom{5}{3}\binom{3}{1} = 30.$$

(c) The probability is $30/70 = 3/7$.

12. (a) In order to keep the system working, the nonfunctioning antennas cannot be next to each other. There are 8 antennas functioning; thus, the 5 nonfunctioning antennas must be in the 9 spaces created by the 8 functioning antennas. The number of arrangements is

$$\binom{9}{5} = 126.$$

(b) The total number of the 5 nonfunctioning antennas is $\binom{13}{5} = 1287$. Thus, the required probability is $126/1287 = 0.0979$.

13. (a) The total number of selections is

$$\binom{15}{5} = 3003.$$

(b) The number of selections containing three defective buses is

$$\binom{4}{3}\binom{11}{2} = 220.$$

(c) The asked probability is $220/3003 = 0.07326$.

(d) The probability all five buses are free of the defect is calculated as

$$\frac{\binom{11}{5}}{\binom{15}{5}} = 0.1538.$$

14. (a) The number of samples of size five is $\binom{30}{5} = 142506$.

(b) The number of samples that include two of the six tagged moose is $\binom{6}{2}\binom{24}{3} = 30360$.

(c)

(i) The probability is

$$\frac{\binom{6}{2}\binom{24}{3}}{\binom{30}{5}} = \frac{30360}{142506} = 0.213.$$

(ii) The probability is

$$\frac{\binom{24}{5}}{\binom{30}{5}} = \frac{30360}{142506} = 0.298.$$

15. (a) The probability is

$$\frac{48}{\binom{52}{5}} = 1.85 \times 10^{-5}.$$

(b) The probability is

$$\frac{\binom{4}{2}\binom{4}{2}44}{\binom{52}{5}} = 0.00061.$$

(c) The probability is

$$\frac{\binom{4}{3}\binom{12}{2}4^2}{\binom{52}{5}} = 0.0016.$$

16. The total number of possible assignments is $\binom{10}{2,2,2,2,2} = 113400$.

17. (a) There are $3^{15} = 14348907$ ways to classify the next 15 shingles in tow three grades.

(b) The number of ways to classify into three high, five medium and seven low grades is

$$\binom{15}{3,5,7} = 360360.$$

(c) The probability is $360360/14348907 = 0.0251$.

18. (a)

$$2^n = (1+1)^n = \sum_{k=0}^{n} \binom{n}{k} 1^k 1^{n-k} = \sum_{k=0}^{n} \binom{n}{k}.$$

(b)

$$\begin{aligned}(a^2+b)^4 &= \binom{4}{0}(a^2)^0 b^{4-0} + \binom{4}{1}(a^2)^1 b^{4-1} + \binom{4}{2}(a^2)^2 b^{4-2} + \binom{4}{3}(a^2)^3 b^{4-3}\\ &\quad + \binom{4}{4}(a^2)^4 b^{4-4}\\ &= b^4 + 4a^2b^3 + 6a^4b^2 + 4a^6b + a^8\end{aligned}$$

19.

$$\begin{aligned}(a_1^2+2a_2+a_3)^3 &= \binom{3}{0,0,3}(a_1^2)^0(2a_2)^0a_3^3 + \binom{3}{0,1,2}(a_1^2)^0(2a_2)^1a_3^2\\ &\quad + \binom{3}{0,2,1}(a_1^2)^0(2a_2)^2a_3^1 + \binom{3}{0,3,0}(a_1^2)^0(2a_2)^3a_3^0\\ &\quad + \binom{3}{1,0,2}(a_1^2)^1(2a_2)^0a_3^2 + \binom{3}{1,1,1}(a_1^2)^1(2a_2)^1a_3^1\\ &\quad + \binom{3}{1,2,0}(a_1^2)^1(2a_2)^2a_3^0 + \binom{3}{2,0,1}(a_1^2)^2(2a_2)^0a_3^1\\ &\quad + \binom{3}{2,1,0}(a_1^2)^2(2a_2)^1a_3^0 + \binom{3}{3,0,0}(a_1^2)^3(2a_2)^0a_3^0\\ &= a_3^3 + 6a_2a_3^2 + 12a_2^2a_3 + 8a_2^3 + 3a_1^2a_3^2 + 12a_1^2a_2a_3\\ &\qquad + 12a_1^2a_2^2 + 3a_1^4a_3 + 6a_1^4a_2 + a_1^6.\end{aligned}$$

2.4 Axioms and Properties of Probabilities

1. $P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0.37 + 0.23 - 0.47 = 0.13$.

2. (a) $P(A_1) = \cdots = P(A_m) = 1/m$.

   (b) If $m = 8$, $P(A_1 \cup A_2 \cup A_3 \cup A_4) = P(A_1) + P(A_2) + P(A_3) + P(A_4) = 4 \times 1/m = 1/2$.

3. (a) The R commands are

```
t = c(50,51,52,53); G=expand.grid(X1=t,X2=t,X3=t); attach(G)
table((X1+X2+X3)/3)/length(X1)
```

The resulting PMF is

x	50	50.33	50.67	51	51.33
$p(x)$	0.015625	0.046875	0.093750	0.156250	0.187500
x	51.67	52	52.33	52.67	53
$p(x)$	0.187500	0.156250	0.093750	0.046875	0.015625

(b) The probability that the average gas mileage is at least 52 MPG is $0.156250 + 0.093750 + 0.046875 + 0.015625 = 0.3125$.

4. (a)

   (i) $E_1 = \{5, 6, 7, 8, 9, 10, 11, 12\}$. $P(E_1) = 4/36+5/36+6/36+5/36+4/36+3/36+2/36+1/36 = 5/6$.

   (ii) $E_2 = \{2, 3, 4, 5, 6, 7, 8\}$. $P(E_2) = 1/36 + 2/36 + 3/36 + 4/36 + 5/36 + 6/36 + 5/36 = 13/18$.

   (iii) $E_3 = E_1 \cup E_2 = \{2, \cdots, 12\}$, $P(E_3) = 1$. $E_4 = E_1 - E_2 = \{9, 10, 11, 12\}$, $P(E_4) = 4/36+3/36+2/36+1/36 = 5/18$. $E_5 = E_1^c \cap E_2^c = \emptyset$, $P(E_5) = 0$.

   (b) $P(E_3) = P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 30/36 + 26/36 - (4/36 + 5/36 + 6/36 + 5/36) = 1$.

   (c) $P(E_5) = P(E_1^c \cap E_2^c) = P((E_1 \cup E_2)^c) = P(E_3^c) = 1 - P(E_3) = 1 - 1 = 0$.

5. (a)

   (i) $E_1 = \{(> 3, V), (< 3, V)\}$, $P(E_1) = 0.25 + 0.3 = 0.55$.

   (ii) $E_2 = \{(< 3, V), (< 3, D), (< 3, F)\}$, $P(E_2) = 0.3 + 0.15 + 0.13 = 0.58$.

   (iii) $E_3 = \{(> 3, D), (< 3, D)\}$, $P(E_3) = 0.1 + 0.15 = 0.25$.

   (iv) $E_4 = \{(> 3, V), (< 3, V), (< 3, D), (< 3, F)\}$, $P(E_4) = 0.25 + 0.3 + 0.15+0.13 = 0.83$. $E_5 = \{(> 3, V), (< 3, V), (< 3, D), (< 3, F), (> 3, D)\}$, $P(E_5) = 0.25 + 0.3 + 0.15 + 0.13 + 0.1 = 0.93$.

   (b) $P(E_4) = P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.55 + 0.58 - 0.3 = 0.83$.

   (c) $P(E_5) = P(E_1 \cup E_2 \cup E_3) = P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3) = 0.55 + 0.58 + 0.25 - 0.3 - 0 - 0.15 + 0 = 0.93$.

6. (a) The probability that, in any given hour, only machine A produces a batch with no defects is

$$P(E_1 \cap E_2^c) = P(E_1) - P(E_1 \cap E_2) = 0.95 - 0.88 = 0.07.$$

   (b) The probability, in that any given hour, only machine B produces a batch with no defects is

$$P(E_2 \cap E_1^c) = P(E_2) - P(E_1 \cap E_2) = 0.92 - 0.88 = 0.04.$$

(c) The probability that exactly one machine produces a batch with no defects is

$$P((E_1 \cap E_2^c) \cup (E_2 \cap E_1^c)) = P(E_1 \cap E_2^c) + P(E_2 \cap E_1^c) = 0.07 + 0.04 = 0.11.$$

(d) The probability that at least one machine produces a batch with no defects is

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2) = 0.95 + 0.92 - 0.88 = 0.99.$$

7. The probability that at least one of the machines will produce a batch with no defectives is

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P(E_1) + P(E_2) + P(E_3) - P(E_1 \cap E_2) - P(E_1 \cap E_3) \\ &\quad - P(E_2 \cap E_3) + P(E_1 \cap E_2 \cap E_3) \\ &= 0.95 + 0.92 + 0.9 - 0.88 - 0.87 - 0.85 + 0.82 = 0.99. \end{aligned}$$

8. (a)

(i) $P(E_1) = 0.10 + 0.04 + 0.02 + 0.08 + 0.30 + 0.06 = 0.6.$

(ii) $P(E_2) = 0.10 + 0.08 + 0.06 + 0.04 + 0.30 + 0.14 = 0.72.$

(iii) $P(E_1 \cap E_2) = 0.1 + 0.04 + 0.08 + 0.3 = 0.52.$

(b) The probability mass function for the experiment that records only the online monthly volume of sales category is given as

Online Sales	0	1	2
Probability	0.16	0.44	0.4

9. Let

$$E_4 = \{\text{at least two of the original four components work}\},$$

and

$$\begin{aligned} E_5 = \{&\text{at least three of the original four components work}\} \\ &\cup \{\text{two of the original four components work} \\ &\quad \text{and the additional component works}\}. \end{aligned}$$

Then $E_4 \not\subset E_5$ because

$$\begin{aligned} B =\{&\text{exactly two of the original four components work} \\ &\text{and the additional component does not work}\}, \end{aligned}$$

which is part of E_4, is not in E_5. Thus, $E_4 \not\subset E_5$ and, hence, it is not necessarily true that $P(E_4) \le P(E_5)$.

10. (a) If two dice are rolled, there are a total of 36 possibilities, among which 6 are tied. Hence, the probability of tie is $6/36 = 1/6$.

(b) By symmetry of the game $P(A \text{ wins}) = P(B \text{ wins})$ and $P(A \text{ wins})+P(B \text{ wins})+P(\text{tie}) = 1$. Using the result of (a), we can solve that $P(A \text{ wins}) = P(B \text{ wins}) = 5/12$.

11. (a) $A > B = \{\text{die A results in 4}\}$, $B > C = \{\text{die C results in 2}\}$,
$C > D = \{\text{die C results in 6, or die C results in 2 and die D results in 1}\}$,
$D > A = \{\text{die D results in 5, or die D results in 1 and die A results in 0}\}$.

(b) $P(A > B) = 4/6$, $P(B > C) = 4/6$, $P(C > D) = 4/6$, $P(D > A) = 4/6$.

12. (a) If the participant sticks with the original choice, the probability of winning the big prize is 1/3.

(b) If the participant chooses to switch his/her choice, the probability of winning the big prize is 2/3. This is because that if the first choice was actually the minor prize, then, after switching, he/she will win the big prize. If the first choice was actually the big prize, after switching he/she will win the minor prize. While the first choice being the minor prize has a probability of 2/3, consequently, switching leads to a probability of 2/3 to win the big prize.

13. To prove that $P(\emptyset) = 0$, let $E_1 = S$ and $E_i = \emptyset$ for $i = 2, 3, \cdots$. Then $E_1, E_2, \cdots$ is a sequence of disjoint events. By Axiom 3, we have

$$P(S) = P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = P(S) + \sum_{i=2}^{\infty} P(\emptyset),$$

which implies that $\sum_{i=2}^{\infty} P(\emptyset) = 0$, and we must have $P(\emptyset) = 0$.

To prove (2) of Proposition 2.4-1, let $E_i = \emptyset$ for $i = n+1, n+2, \cdots$. Then $E_1, E_2, \cdots$ is a sequence of disjoint events and $\bigcup_{i=1}^{\infty} E_i = \bigcup_{i=1}^{n} E_i$. By Axiom 3, we have

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i) = \sum_{i=1}^{n} P(E_i) + \sum_{i=n+1}^{\infty} P(E_i) \\ &= \sum_{i=1}^{n} P(E_i) + \sum_{i=n+1}^{\infty} P(\emptyset) = \sum_{i=1}^{n} P(E_i), \end{aligned}$$

which is what to be proved.

2.5 Conditional Probability

1. The probability can be calculated as

$$P(>3|>2) = \frac{P((>3) \cap (>2))}{P(>2)} = \frac{P(>3)}{P(>2)} = \frac{(1+3)^{-2}}{(1+2)^{2}} = 9/16.$$

2. Let $B = \{\text{system re-evaluation occurs}\}$ and $C = \{\text{a component is individually replaced}\}$. Consider a new experiment with reduced sample space $A = B \cup C$. The desired probability is the probability of B in this new experiment, which is calculated as

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(B)}{P(B) + P(C)} = \frac{0.005}{0.005 + 0.1} = 0.048.$$

3. (a) $P(A) = 0.132 + 0.068 = 0.2$.

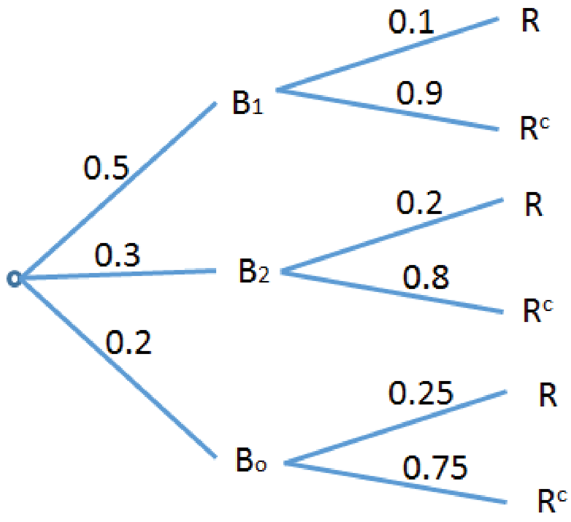
   (b) $P(A \cap B) = 0.132$, thus $P(B|A) = P(A \cap B)/P(A) = 0.132/0.2 = 0.66$.

   (c) $P(X = 1) = 0.2$, $P(X = 2) = 0.3$, $P(X = 3) = 0.5$.

4. We let B_1, B_2, B_O be the event that the TV is brand 1, brand 2, and other brand, respectively. Let R be the event that the TV needs warranty repair.

   (a) $P(B_1 \cap R) = P(B_1)P(R|B_1) = 0.5 \times 0.1 = 0.05$.

   (b) The tree diagram is

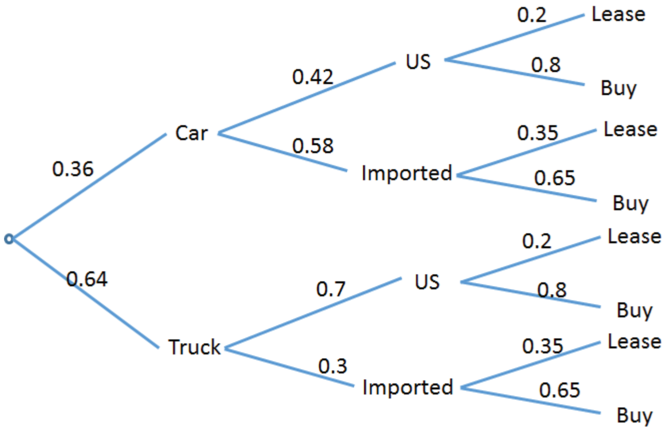


   (c) Using the diagram

$$\begin{aligned} P(R) &= P(B_1)P(R|B_1) + P(B_2)P(R|B_2) + P(B_O)P(R|B_O) \\ &= 0.5 \times 0.1 + 0.3 \times 0.2 + 0.2 \times 0.25 = 0.16. \end{aligned}$$

5. (a) The probability is $0.36 \times 0.58 = 0.2088$.

   (b) The tree diagram is



(c) By using the tree diagram, the probability that the next consumer will lease his/her vehicle is

$$0.36\times 0.42\times 0.2+0.36\times 0.58\times 0.35+0.64\times 0.7\times 0.2+0.64\times 0.3\times 0.35=0.26.$$

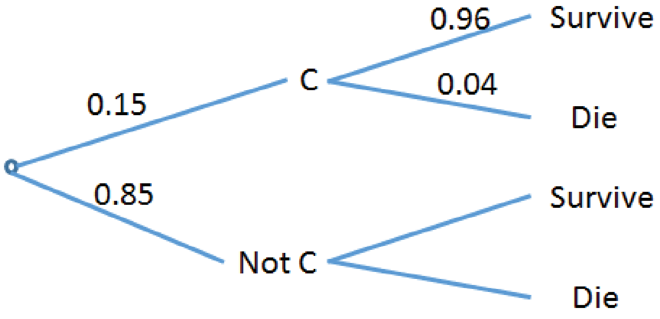
6. (a) $P(\text{no defect}\cap A)=P(\text{no defect}|A)P(A)=0.99\times 0.3=0.297.$

(b)$P(\text{no defect}\cap B)=P(\text{no defect}|B)P(B)=0.97\times 0.3=0.291$, and $P(\text{no defect}\cap C)=P(\text{no defect}|C)P(C)=0.92\times 0.3=0.276.$

(c) $P(\text{no defect})=P(\text{no defect}\cap A)+P(\text{no defect}\cap B)+P(\text{no defect}\cap C)=0.297+0.291+0.276=0.864.$

(d) $P(C|\text{no defect})=P(\text{no defect}\cap C)/P(\text{no defect})=0.276/0.864=0.3194.$

7. (a) The tree diagram is



(b) From the given information, we have

$$P(\text{survive})=0.15\times 0.96+0.85\times P(\text{Survive}|\text{Not C-section})=0.98.$$

Solving this equation gives us $P(\text{Survive}|\text{Not C-section})=0.984$.

8. Let B be the event that the credit card holds monthly balance, then $P(B)=0.7$ and $P(B^c)=0.3$. Let L be the event that the card holder has annual income less than \$20,000, then $P(L|B)=0.3$ and $P(L|B^c)=0.2$.

(a) $P(L) = P(L|B)P(B) + P(L|B^c)P(B^c) = 0.3 \times 0.7 + 0.2 \times 0.3 = 0.27$.

(b) $P(B|L) = P(L|B)P(B)/P(L) = 0.3 \times 0.7/0.27 = 0.778$.

9. Let A be the event that the plant is alive and let W be the roommate waters it. Then, from the given information, $P(W) = 0.85$ and $P(W^c) = 0.15$; $P(A|W) = 0.9$ and $P(A|W^c) = 0.2$.

   (a) $P(A) = P(A|W)P(W) + P(A|W^c)P(W^c) = 0.9 \times 0.85 + 0.2 \times 0.15 = 0.795$.

   (b) $P(W|A) = P(A|W)P(W)/P(A) = 0.9 \times 0.85/0.795 = 0.962$.

10. Let D_1 be the event that the first is defective and D_2 the event that the second is defective.

    (a) $P(\text{no defective}) = P(D_1^c \cap D_2^c) = P(D_2^c|D_1^c)P(D_1^c) = 6/9 \times 7/10 = 0.467$.

    (b) X can be 0, 1, or 2. We already calculated $P(X = 0) = P(\text{no defective}) = 0.467$. $P(X = 2) = P(D_1 \cap D_2) = P(D_2|D_1)P(D_1) = 2/9 \times 3/10 = 0.067$. Thus, $P(X = 1) = 1 - P(X = 0) - P(X = 2) = 0.466$.

    (c) $P(D_1|X = 1) = P(D_1 \cap D_2^c)/P(X = 1) = P(D_2^c|D_1)P(D_1)/P(X = 1) = 7/9 \times 0.3/0.466 = 0.5$.

11. Let L_1, L_2, L_3, L_4 be the event that the radar traps are operated at the 4 locations, then $P(L_1) = 0.4, P(L_2) = 0.3, P(L_3) = 0.2, P(L_4) = 0.3$. Let S be the person speeding to work, then $P(S|L_1) = 0.2, P(S|L_2) = 0.1, P(S|L_3) = 0.5, P(S|L_4) = 0.2$.

    (a) $P(S) = P(S|L_1)P(L_1) + P(S|L_2)P(L_2) + P(S|L_3)P(L_3) + P(S|L_4)P(L_4) = 0.2 \times 0.4 + 0.1 \times 0.3 + 0.5 \times 0.2 + 0.2 \times 0.3 = 0.27$.

    (b) $P(L_2|S) = P(S|L_2)P(L_2)/P(S) = 0.1 \times 0.3/0.27 = 0.11$.

12. Let D be the event that the aircraft will be discovered, and E be the event that it has an emergency locator. From the problem, $P(D) = 0.7$ and $P(D^c) = 0.3$; $P(E|D) = 0.6$ and $P(E|D^c) = 0.1$.

    (a) $P(E \cap D^c) = P(E|D^c)P(D^c) = 0.1 \times 0.3 = 0.03$.

    (b) $P(E) = P(E|D^c)P(D^c) + P(E|D)P(D) = 0.1 \times 0.3 + 0.6 \times 0.7 = 0.45$.

    (c) $P(D^c|E) = P(E \cap D^c)/P(E) = 0.03/0.45 = 0.067$.

13.

$$\begin{aligned}\text{R.H.S.} &= P(E_1)\frac{P(E_1 \cap E_2)}{P(E_1)}\frac{P(E_1 \cap E_2 \cap E_3)}{P(E_1 \cap E_2)} \cdots \frac{P(E_1 \cap E_2 \cap \cdots \cap E_{n-1} \cap E_n)}{P(E_1 \cap E_2 \cap \cdots \cap E_{n-1})} \\ &= P(E_1 \cap E_2 \cap \cdots \cap E_{n-1} \cap E_n) = \text{L.H.S.}\end{aligned}$$

2.6 Independent Events

1. From the given information $P(E_2) = 2/10$ and $P(E_2|E_1) = 2/9$, thus $P(E_2) \neq P(E_2|E_1)$. Consequently, E_1 and E_2 are not independent.

2. We can calculate from the table that $P(X = 1) = 0.132 + 0.068 = 0.2$ and $P(Y = 1) = 0.132 + 0.24 + 0.33 = 0.702$, thus $P(X = 1)P(Y = 1) = 0.2 \times 0.702 = 0.1404 \neq 0.132 = P(X = 1, Y = 1)$. Thus, the events $[X = 1]$ and $[Y = 1]$ are not independent.

3. (a) The probability is $0.9^{10} = 0.349$.

   (b) The probability is $0.1 \times 0.9^9 = 0.0387$.

   (c) The probability is $10 \times 0.1 \times 0.9^9 = 0.387$

4. A total of 8 fuses being inspected means that the first 7 are not defective and the 8th is defective, thus the probability is calculated as $0.99^7 \times 0.01 = 0.0093$.

5. Assuming the cars assembled on each line are independent, also assume that the two lines are independent. We have

   (a) The probability of finding zero nonconformance in the sample from line 1 is $0.8^4 = 0.410$.

   (b) The probability of finding zero nonconformance in the sample from line 1 is $0.9^3 = 0.729$.

   (c) The probability is $0.8^4 \times 0.9^3 = 0.2986$.

6. Yes. By the given information, $P(T|M) = P(T)$, we see that T and M are independent. Thus, T and $F = M^c$ are also independent; that is, $P(T|F) = P(T)$.

7. (a) The completed table is given as

	Football	Basketball	Track	Total
Male	0.3	0.22	0.13	0.65
Female	0	0.28	0.07	0.35
Total	0.3	0.5	0.2	1

   (b) Let B be the event that the student prefers basketball, then $P(F|B) = P(F \cap B)/P(B) = 0.28/0.5 = 0.56$.

   (c) F and B are not independent because $P(F|B) = 0.56 \neq 0.35 = P(F)$.

8. We can write

$$E_1 = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\},$$

$$E_2 = \{(3,1), (3,2), (3,3), (3,4), (3,5), (3,6)\},$$

and

$$E_3 = \{(1,4),(2,4),(3,4),(4,4),(5,4),(6,4)\}.$$

Thus, $E_1 \cap E_2 = E_1 \cap E_3 = E_2 \cap E_3 = \{(3,4)\}$, and $E_1 \cap E_2 \cap E_3 = \{(3,4)\}$. Hence, $P(E_1) = P(E_2) = P(E_3) = 1/6$, and $P(E_1 \cap E_2) = P(E_1 \cap E_3) = P(E_2 \cap E_3) = 1/36$, this shows that E_1, E_2, E_3 are pairwise independent. But $P(E_1 \cap E_2 \cap E_3) \neq P(E_1)P(E_2)P(E_3)$.

9. Since E_1, E_2, E_3 are independent, we have

$$\begin{aligned} P(E_1 \cap (E_2 \cup E_3)) &= P((E_1 \cap E_2) \cup (E_1 \cap E_3)) = P(E_1 \cap E_2) + P(E_1 \cap E_3) \\ &\qquad - P(E_1 \cap E_2 \cap E_3) \\ &= P(E_1)P(E_2) + P(E_1)P(E_3) - P(E_1)P(E_2)P(E_3) \\ &= P(E_1)[P(E_2) + P(E_3) - P(E_2 \cap E_3)] = P(E_1)P(E_2 \cup E_3), \end{aligned}$$

which proves the independence between E_1 and $E_2 \cup E_3$.

10. Let E_1, E_2, E_3, E_4 be the events that components 1, 2, 3, 4 function, respectively, then

$$\begin{aligned} P(\text{system functions}) &= P((E_1 \cap E_2) \cup (E_3 \cap E_4)) = P(E_1 \cap E_2) + P(E_3 \cap E_4) \\ &\qquad - P(E_1 \cap E_2 \cap E_3 \cap E_4) \\ &= P(E_1)P(E_2) + P(E_3)P(E_4) - P(E_1)P(E_2)P(E_3)P(E_4) \\ &= 2 \times 0.9^2 - 0.9^4 = 0.9639. \end{aligned}$$

11. Let A denote the event that the system functions and A_i denote the event that component i functions, $i = 1, 2, 3, 4$. In mathematical notations

$$\begin{aligned} A = (A_1 \cap A_2 \cap A_3 \cap A_4^c) \cup (A_1 \cap A_2 \cap A_3^c \cap A_4) \cup (A_1 \cap A_2^c \cap A_3 \cap A_4) \\ \cup (A_1^c \cap A_2 \cap A_3 \cap A_4) \cup (A_1 \cap A_2 \cap A_3 \cap A_4). \end{aligned}$$

Thus

$$\begin{aligned} P(A) = P(A_1 \cap A_2 \cap A_3 \cap A_4^c) + P(A_1 \cap A_2 \cap A_3^c \cap A_4) + P(A_1 \cap A_2^c \cap A_3 \cap A_4) \\ + P(A_1^c \cap A_2 \cap A_3 \cap A_4) + P(A_1 \cap A_2 \cap A_3 \cap A_4) \\ = 4 \times 0.9^3 \times 0.1 + 0.9^4 = 0.9477. \end{aligned}$$

Chapter 3

Random Variables and Their Distributions

3.2 Describing a Probability Distribution

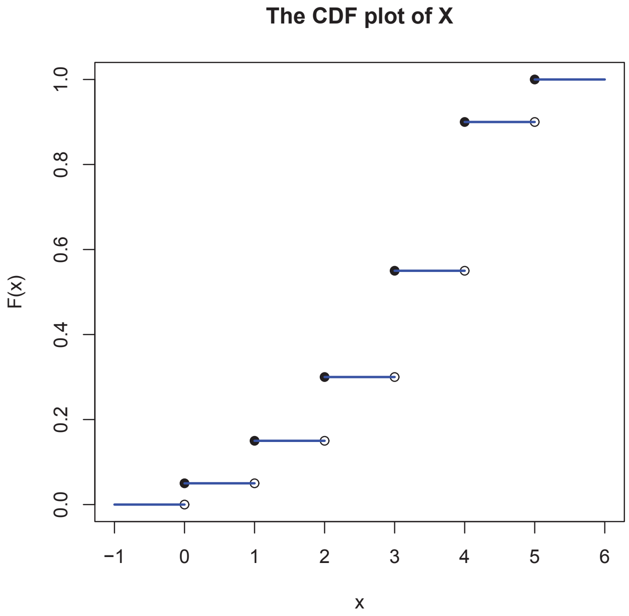
1. (a) $p_1(x)$ is not a valid probability mass function but $p_2(x)$ is.

   (b) To find k, solve the equation $0.2k + 0.3k + 0.4k + 0.2k = 1$. The solution is $k = 1/1.1$.

2. (a) The CDF of X is

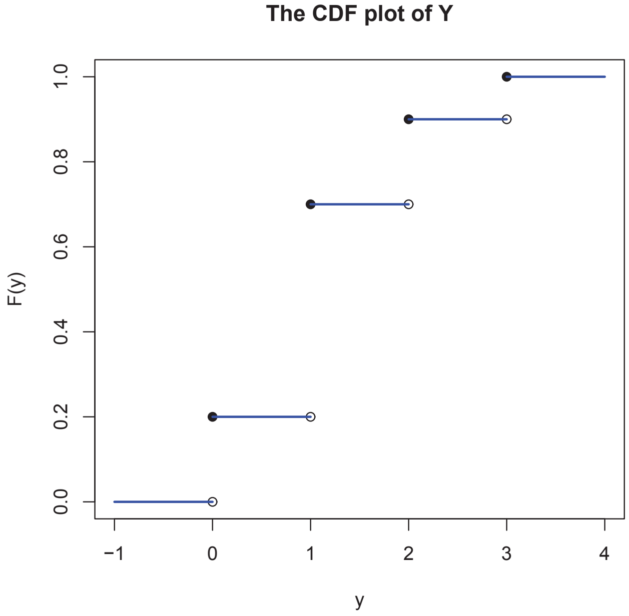
$$F(x) = \begin{cases} 0, & \text{if } x < 0 \\ 0.05, & \text{if } 0 \le x < 1 \\ 0.15, & \text{if } 1 \le x < 2 \\ 0.3, & \text{if } 2 \le x < 3 \\ 0.55, & \text{if } 3 \le x < 4 \\ 0.9, & \text{if } 4 \le x < 5 \\ 1, & \text{if } x \ge 5 \end{cases}$$

   The plot is given on the next page.



(b) $P(1 \leq X \leq 4) = P(X \leq 4) - P(X < 1) = F(4) - F(0) = 0.9 - 0.05 = 0.85.$

3. (a) $P(Y \geq 2) = 1 - P(Y < 2) = 1 - 0.7 = 0.3$. The plot is given below.



(b) The possible values of Y are the jumping points 0, 1, 2, 3, and the probability at the jumping point is the jumping size. Thus $p(0) = 0.2 - 0 = 0.2$, $p(1) = 0.7 - 0.2 = 0.5$, $p(2) = 0.9 - 0.7 = 0.2$, and $p(3) = 1 - 0.9 = 0.1$.

4. X can assume values 0, 1, 2, and 3. We have the PMF of X as

$$p(0) = P(X = 0) = \frac{\binom{7}{3}}{\binom{10}{3}} = 0.292, \qquad p(1) = P(X = 1) = \frac{\binom{3}{1}\binom{7}{2}}{\binom{10}{3}} = 0.525,$$

$$p(2) = P(X = 2) = \frac{\binom{3}{2}\binom{7}{1}}{\binom{10}{3}} = 0.175, \qquad p(3) = P(X = 3) = \frac{\binom{3}{3}}{\binom{10}{3}} = 0.008.$$

The CDF $F(x) = 0$ if $x < 0$; if $0 \le x < 1$, $F(x) = p(0) = 0.292$; if $1 \le x < 2$, $F(x) = p(0)+p(1) = 0.292+0.525 = 0.817$; if $2 \le x < 3$, $F(x) = p(0)+p(1)+p(2) = 0.292 + 0.525 + 0.175 = 0.993$; if $3 \le x$, $F(x) = p(0) + p(1) + p(2) + p(3) = 1$.

5. (a) $f_1(x)$ is not a valid PDF because for some $x \in (0, 2)$, $f_x(x) < 0$. For example, $f_1(1.9) = -0.58$. $f_2(x)$ is a valid PDF because it is easy to verify that $f_2(x) \ge 0$ and $\int_0^2 f_2(x)dx = 1$.

(b)

(i) To find k, we must have $\int_{-\infty}^{\infty} f(x)dx = 1$, that is $\int_8^{10} kxdx = 1$. Solving the equation, we have $k = 1/18$.
It is clear that $F_X(x) = 0$ if $x < 8$ and $F_X(x) = 1$ if $x > 10$. For $x \in [8, 10]$,

$$F_X(x) = \int_8^x f(t)dt = \frac{1}{18}\int_8^x tdt = \frac{1}{36}\, t^2\Big|_8^x = \frac{x^2 - 64}{36}.$$

Using the CDF,

$$P(8.6 \le X \le 9.8) = F_X(9.8)-F_X(8.6) = \frac{9.8^2 - 64}{36} - \frac{8.6^2 - 64}{36} = 0.61333.$$

(ii)

$$\begin{aligned} P(X \le 9.8|X \ge 8.6) &= \frac{P(8.6 \le X \le 9.8)}{P(X \ge 8.6)} = \frac{P(8.6 \le X \le 9.8)}{1 - P(X < 8.6)} \\ &= \frac{P(8.6 \le X \le 9.8)}{1 - F_X(8.6)} = \frac{0.61333}{1 - \frac{8.6^2-64}{36}} = 0.8479. \end{aligned}$$

6. $X \sim U(0, 1)$ and $Y = 3 + 6X$, then clearly the sample space for Y is (3, 9). Thus, the CDF of Y, $F_Y(y) = 0$ if $y < 3$ and $F_Y(y) = 1$ if $y > 9$. For $y \in (3, 9)$, we calculate $F_Y(y)$ as

$$F_Y(y) = P(Y \le y) = P(3 + 6X \le y) = P\left(X \le \frac{y-3}{6}\right) = \frac{y-3}{6} = \frac{y-3}{9-3}.$$

Comparing to the CDF of $U(A, B)$ in Examples 3.2-5, we can see that $F_Y(y)$ is the CDF of $U(3, 9)$, hence $Y \sim U(3, 9)$.

7. The sample space of Y is $(0, \infty)$. If $y < 0$, clearly, the CDF $F_Y(y) = 0$. For $y > 0$,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(-\log X \leq y) = P(X \geq e^{-y}) \\ &= 1 - P(X < e^{-y}) = 1 - F_X(e^{-y}) = 1 - e^{-y}. \end{aligned}$$

The PDF of Y is $f_Y(y) = 0$ if $y < 0$ and $f_Y(y) = e^{-y}$ for $y > 0$.

8. (a) $P(0.5 \leq X \leq 2) = F(1) - F(0.5) = 1/4 - 1/16 = 0.1875$.

(b) Taking derivative to $F(x)$, we find the PDF of X is

$$f(x) = \begin{cases} \frac{x}{2}, & \text{if } 0 < x \leq 2 \\ 0, & \text{otherwise.} \end{cases}$$

(c) Since Y is in seconds, $Y = 60X$. Thus, $F(y) = 0$ for $y \leq 0$ and $F(y) = 1$ for $y > 120$, for $0 < y \leq 120$,

$$F(y) = P(Y \leq y) = P(60X < y) = P\left(x < \frac{y}{60}\right) = \frac{1}{4}\left(\frac{y}{60}\right)^2 = \frac{y^2}{14400}.$$

Taking derivative to $F(y)$, we find the PDF of Y is

$$f(y) = \begin{cases} \frac{y}{7200}, & \text{if } 0 < x \leq 120 \\ 0, & \text{otherwise.} \end{cases}$$

9. (a)

$$P(X > 10) = P(30/D > 10) = P(D < 3) = \frac{\pi 3^2}{\pi 9^3} = 1/9.$$

(b) Since D must between 0 and 9, then the sample space of $X = 30/D$ is $(30/9, \infty)$. We first calculate the CDF of X. Clearly, $F_X(x) = 0$ if $x < 30/9$. For $x \geq 30/9$, we have

$$\begin{aligned} F_X(x) &= P(X \leq x) = 1 - P(X > x) = 1 - P\left(\frac{30}{D} > x\right) \\ &= 1 - P\left(D < \frac{30}{x}\right) = 1 - \frac{\pi(\frac{30}{x})^2}{\pi 9^2} = 1 - \frac{100}{9x^2}. \end{aligned}$$

Differentiating $F_X(x)$, we have $f_X(x) = 0$ for $x < 10/3$, otherwise, $f_X(x) = 200/(9x^3)$.

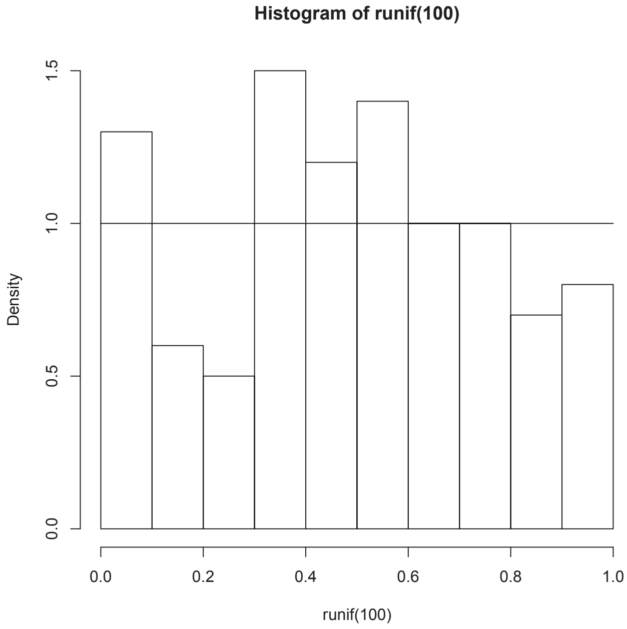
10. (a) The event "no cost" means that $X < 24 \times 3 = 72$. Thus, the probability is

$$\begin{aligned} P(X < 72) &= \int_{48}^{72} f(x)dx = \int_{48}^{72} 0.02e^{-0.02(x-48)}dx = [-e^{-0.02(x-48)}]\Big|_{48}^{72} \\ &= 1 - e^{-0.02 \times 24} = 0.3812. \end{aligned}$$

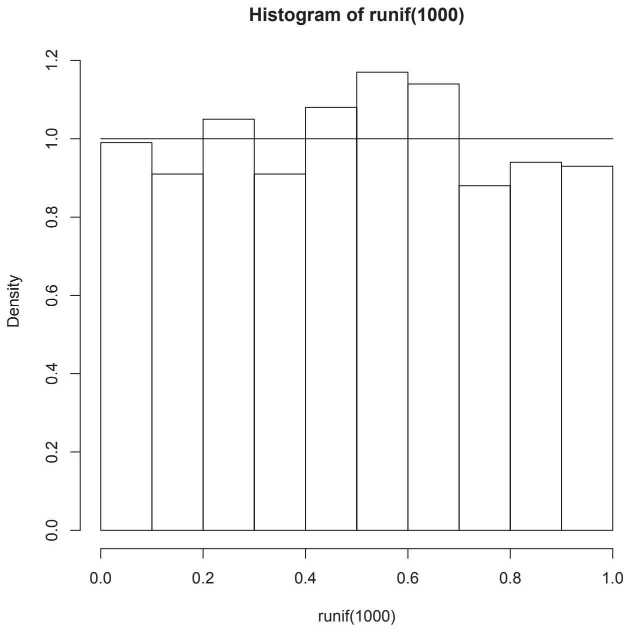
(b) The additional cost is between \$400 and \$800. This means that it takes more than 4 days but less than 7 days for the fixture to arrive. That is, $4 \times 24 < X < 7 \times 24$, or $96 < X < 168$. Thus, the corresponding probability is

$$\begin{aligned} P(96 < X < 168) &= \int_{96}^{168} f(x)dx = \int_{96}^{168} 0.02e^{-0.02(x-48)}dx = [-e^{-0.02(x-48)}]\Big|_{96}^{168} \\ &= e^{-0.02\times 96} - e^{-0.02\times 168} = 0.1119. \end{aligned}$$

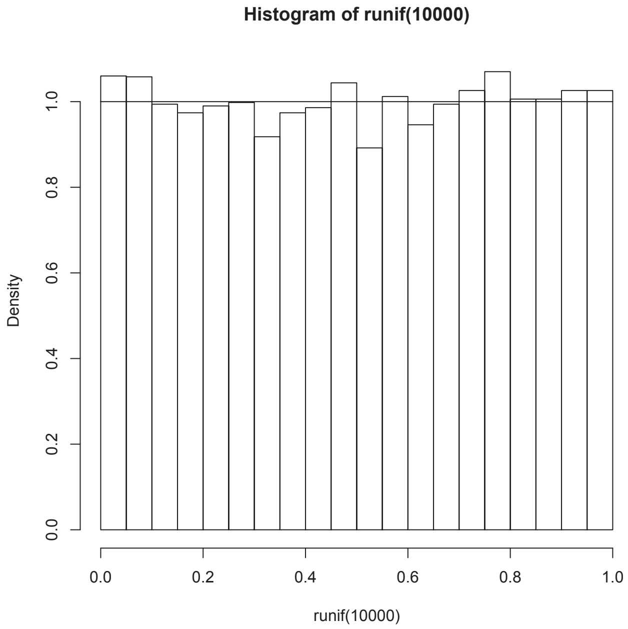
11. When using a sample size of 100, the resulting histogram superimposed with the PDF is given as below.



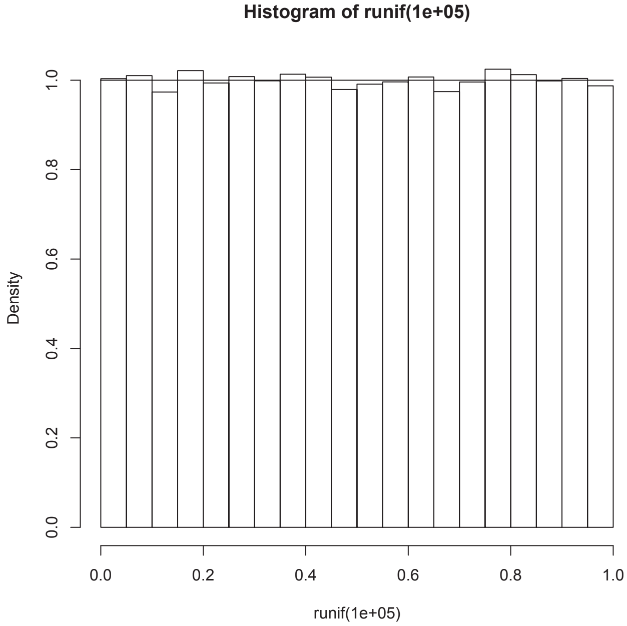
When using a sample size of 1000, the resulting histogram superimposed with the PDF is given on the next page.



When using a sample size of 10000, the resulting histogram superimposed with the PDF is given as below.



When using a sample size of 100000, the resulting histogram superimposed with the PDF is given on the next page.



From these figures, we can see that when the sample size is only 100, the histogram is not reasonably close to the PDF curve. It seems that if the sample size is larger than 1000, the histogram starts to get close to the PDF.

3.3 Parameters of Probability Distributions

1. (a) The random variable X can take 0, 1, 2, 3. The PMF can be calculated as

$$p(0) = P(X = 0) = \frac{\binom{16}{3}}{\binom{20}{3}} = 0.4912, \qquad p(1) = P(X = 1) = \frac{\binom{4}{1}\binom{16}{2}}{\binom{20}{3}} = 0.4211,$$

$$p(2) = P(X = 2) = \frac{\binom{4}{2}\binom{16}{1}}{\binom{20}{3}} = 0.0842, \qquad p(3) = P(X = 3) = \frac{\binom{4}{3}}{\binom{20}{3}} = 0.0035.$$

   (b) $E(X) = \sum_0^3 xp(x) = 0 \times 0.491 + 1 \times 0.4211 + 2 \times 0.0842 + 3 \times 0.0035 = 0.6$, $Var(X) = E(X^2) - E(X)^2 = 0 \times 0.491 + 1 \times 0.4211 + 4 \times 0.0842 + 9 \times 0.0035 - 0.6^2 = 0.429$.

2. (a) $E(X) = 1 \times 0.4 + 2 \times 0.3 + 3 \times 0.1 + 4 \times 0.2 = 2.1$, and $E(1/X) = 1/1 \times 0.4 + 1/2 \times 0.3 + 1/3 \times 0.1 + 1/4 \times 0.2 = 0.63333$.

   (b) We need to compare the expectation of $1000/E(X)$ and $1000/X$: $E(1000/E(X)) = 1000/E(X) = 1000/2.1 = 476.19$, while $E(1000/X) = 1000E(1/X) = 1000 \times 0.63333 = 633.33$. Thus, the player should choose $1000/X$.

3. (a) The sample space of X is $S_X = \{0, 400, 750, 800, 1150, 1500\}$. We can find the PMF as

$$\begin{aligned} p(0) &= P(\text{both do not buy TV}) \\ &= 0.7 \times 0.7 = 0.49, \end{aligned}$$

$$\begin{aligned} p(400) &= P(\text{one buys \$400 TV, the other one does not buy TV}) \\ &= 2 \times 0.7 \times 0.3 \times 0.4 = 0.168, \end{aligned}$$

$$\begin{aligned} p(750) &= P(\text{one buys \$400 TV, the other one does not buy TV}) \\ &= 2 \times 0.7 \times 0.3 \times 0.6 = 0.252, \end{aligned}$$

$$\begin{aligned} p(800) &= P(\text{both buy \$400 TV}) \\ &= 0.3 \times 0.4 \times 0.3 \times 0.4 = 0.0144, \end{aligned}$$

$$\begin{aligned} p(1150) &= P(\text{one buys \$400 TV, the other buys \$750 TV}) \\ &= 2 \times 0.3 \times 0.4 \times 0.3 \times 0.6 = 0.0432, \end{aligned}$$

and

$$\begin{aligned} p(1500) &= P(\text{both buy \$750 TV}) \\ &= 0.3 \times 0.6 \times 0.3 \times 0.6 = 0.0324, \end{aligned}$$

(b) $E(X) = \sum xp(x) = 0 \times 0.49 + 400 \times 0.168 + 750 \times 0.252 + 800 \times 0.0144 + 1150 \times 0.0432 + 1500 \times 0.0324 = 366$.

$Var(X) = 0^2 \times 0.49 + 400^2 \times 0.168 + 750^2 \times 0.252 + 800^2 \times 0.0144 + 1150^2 \times 0.0432 + 1500^2 \times 0.0324 - 366^2 = 173,922$.

4. (a) $E(X) = 0 \times 0.05 + 1 \times 0.1 + 2 \times 0.15 + 3 \times 0.25 + 4 \times 0.35 + 5 \times 0.1 = 3.05$, and $E(X^2) = 0^2 \times 0.05 + 1^2 \times 0.1 + 2^2 \times 0.15 + 3^2 \times 0.25 + 4^2 \times 0.35 + 5^2 \times 0.1 = 11.05$, thus, $\text{Var}(X) = E(X^2) - E(X)^2 = 1.7475$.

(b) Let Y be the bonus, then $Y = 15000X$, thus $E(Y) = E(15000X) = 15000 \times 3.05 = 45750$, and $\text{Var}(Y) = \text{Var}(15000X) = 15000^2\text{Var}(X) = 15000^2 \times 1.7475 = 393187500$.

5. Use the commands *g=function(x)* $\{$*0.01*x*x*exp(-0.1*x)*$\}$*; integrate(g,lower=0, upper=Inf)* to find $E(X)$, and the result is 20. Similarly, we find $E(X^2) = 600$, thus $\sigma_X^2 = 600 - 20^2 = 200$.

6. (a) Since $\tilde{T} = T + 5$ and $T > 0$. Thus, the sample space of $\tilde{T}$ is $(5, \infty)$. The CDF of $\tilde{T}$ is $F_{\tilde{T}}(\tilde{t}) = 0$ if $\tilde{t} < 5$, for $\tilde{t} > 5$

$$F_{\tilde{T}}(\tilde{t}) = P(\tilde{T} \le \tilde{t}) = P(T + 5 \le \tilde{t}) = P(T \le \tilde{t} - 5) = \int_0^{\tilde{t}-5} f_T(t)dt$$
$$= \int_0^{\tilde{t}-5} 0.1e^{-0.1t}dt.$$

Differentiating the CDF of $\tilde{T}$, we get the PDF of $\tilde{T}$ as

$$f_{\tilde{T}}(\tilde{t}) = \begin{cases} 0.1e^{-0.1(\tilde{t}-5)}, & \text{if } \tilde{t} \ge 5 \\ 0, & \text{otherwise.} \end{cases}$$

(b) The expected cost is

$$E(\tilde{Y}) = E(h(\tilde{T})) = \int_{-\infty}^{\infty} h(\tilde{T}) f_{\tilde{T}}(\tilde{t})d\tilde{t}$$
$$= \int_5^{15} 5(15 - \tilde{t})0.1e^{-0.1(\tilde{t}-5)}d\tilde{t} + \int_{15}^{\infty} 10(\tilde{t} - 15)0.1e^{-0.1(\tilde{t}-5)}d\tilde{t} = 55.1819.$$

The two integrals are calculated by the commands

```
g=function(x){5*(15-x)*0.1*exp(-0.1*x+0.5)}
integrate(g, lower=5, upper=15)
```

and

```
g=function(x) {10*(x-15)*0.1*exp(-0.1*x+0.5)}
integrate(g, lower=15, upper=Inf)
```

the company's plan to delay the work on the project does reduce the expected cost.

7. (a) To find the median, solve the equation $F(\tilde{\mu}) = 0.5$, which is $\tilde{\mu}^2/4 = 0.5$, and results in $\tilde{\mu} = \sqrt{2}$. The 25th percentile is the solution to $F(x_{0.25}) = 0.25$, which is $x_{0.25}^2/4 = 0.25$, and results in $x_{0.25} = 1$; The 75th percentile is the solution to $F(x_{0.75}) = 0.75$, which is $x_{0.75}^2/4 = 0.75$, and results in $x_{0.75} = \sqrt{3}$. Thus, the IQR is $IQR = x_{0.75} - x_{0.725} = \sqrt{3} - 1 = 0.732$.

(b) We can get the PDF $f(x) = x/2$ for $0 \le x \le 2$, and otherwise, $f(x) = 0$. So

$$E(X) = \int_0^2 xf(x)dx = \int_0^2 \frac{x^2}{2}dx = \left.\frac{x^3}{6}\right|_0^2 = 4/3 = 1.333,$$

and

$$\sigma_X^2 = \int_0^2 x^2 f(x)dx - E(X)^2 = \int_0^2 \frac{x^3}{2}dx - E(X)^2 = \left.\frac{x^4}{8}\right|_0^2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}.$$

Thus, $\sigma_X = \sqrt{2}/3 = 0.4714$.

8. (a) Since $Y = 60X$ and the sample space for X is $(0, 3)$, the sample space for Y is (0, 180). Thus, the CDF of Y is $F_Y(y) = 0$ if $y < 0$, $F_Y(y) = 1$ if $y \geq 180$, for $y \in (0, 180)$, we have

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(60X \leq y) = P(X \leq y/60) = F_X(y/60) \\ &= \frac{\log(1 + y/60)}{\log 4}. \end{aligned}$$

Thus, by differentiating $F_Y(y)$, we have the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{(60+y)\log 4}, & \text{if } 0 < y < 180 \\ 0, & \text{otherwise.} \end{cases}$$

(b) $V = h(y)$, and $h(y) = 0$ for $0 \leq y \leq 120$, and $h(y) = 200 + 6(y - 120)$ for $y > 120$. Then we can calculate

$$\begin{aligned} E(V) = E(h(Y)) &= \int_{-\infty}^{\infty} h(y) f_Y(y) dy \\ &= \int_{120}^{180} (200 + 6(y - 120)) \frac{1}{(60 + y)\log 4} dy = 77.0686. \end{aligned}$$

The integral can be calculated using the following R commands

```
g=function(y)(200+6*(y-120))/((60+y)*log(4))
integrate(g, lower=120, upper=180)
```

$$\begin{aligned} E(V^2) = E(h(Y)^2) &= \int_{-\infty}^{\infty} h(y)^2 f_Y(y) dy \\ &= \int_{120}^{180} (200 + 6(y - 120))^2 \frac{1}{(60 + y)\log 4} dy = 30859.97. \end{aligned}$$

The integral can be calculated using the following R commands

```
g=function(y)(200+6*(y-120))**2/((60+y)*log(4))
integrate(g, lower=120, upper=180)
```

As a result, $\sigma_V^2 = E(V^2) - E(V)^2 = 30859.97 - 77.0686^2 = 24920.4$.

(c) Let D be the fine expressed in dollars, then $D = Y/100$. Thus, $E(D) = E(Y)/100 = 0.7707$, and $\sigma_D^2 = \sigma_V^2/100^2 = 2.492$.

9. (a)

$$E(P) = \int_0^1 p f_P(p) dp = \int_0^1 \theta p^\theta dp = \theta \frac{1}{\theta + 1} p^{\theta+1}\Big|_0^1 = \frac{\theta}{\theta + 1},$$

and

$$E(P^2) = \int_0^1 p^2 f_P(p)dp = \int_0^1 \theta p^{\theta+1}dp = \theta\frac{1}{\theta+2}\, p^{\theta+2}\big|_0^1 = \frac{\theta}{\theta+2},$$

hence

$$\sigma_P^2 = E(P^2) - E(P)^2 = \frac{\theta}{\theta+2} - \left(\frac{\theta}{\theta+1}\right)^2 = \frac{\theta}{(\theta+2)(\theta+1)^2.}$$

(b) Clearly, if $p \le 0$, $F_P(p) = 0$, and if $p \ge 1$, $F_P(p) = 1$. If $0 < p < 1$,

$$F_P(p) = \int_0^p f_P(t)dt = \int_0^p \theta t^{\theta-1}dt = t^\theta\big|_0^p = p^\theta.$$

(c) Denote the 25th percentile and 75th percentile as $p_{0.25}$ and $p_{0.75}$, respectively. Then $F_P(p_{0.25}) = 0.25$ and $F_P(p_{0.75}) = 0.75$, or, $p_{0.25}^\theta = 0.25$ and $p_{0.75}^\theta = 0.75$. Thus, $p_{0.25} = 0.25^{1/\theta}$ and $p_{0.75} = 0.75^{1/\theta}$. So $IQR = p_{0.75} - p_{0.25} = 0.75^{1/\theta} - 0.25^{1/\theta}$.

3.4 Models for Discrete Random Variables

1. (a) X is Binomial R.V.

   (b) The sample space is $S_X = \{0, 1, \cdots, 5\}$ and the PMF is $p(x) = \binom{5}{x}0.3^x 0.7^{5-x}$ for $x = 0, 1, 2, 3, 4, 5$.

   (c) $E(X) = 5 \times 0.3 = 1.5$ and $\text{Var}(X) = 5 \times 0.3 \times 0.7 = 1.05$.

   (d)

   (i) The probability that there are more than 2 fails, $P(X > 2)$, can be calculated by the R command *1-pbinom(2, 5, 0.3)*, which gives the result 0.163.

   (ii) Let Y be the cost from failed grafts, then $Y = 9X$. Thus, $E(Y) = E(9X) = 9E(X) = 13.5$, and $\text{Var}(Y) = \text{Var}(9X) = 81\text{Var}(X) = 85.05$.

2. (a) X is Binomial R.V., $X \sim \text{Bin}(15, 0.3)$.

   (b) $E(X) = 15 \times 0.3 = 4.5$ and $\text{Var}(X) = 15 \times 0.3 \times 0.7 = 3.15$.

   (c) Use *dbinom(6, 15, 0.3)* for $P(X = 6)$, which gives 0.147236. Use *1-pbinom(5, 15, 0.3)* for $P(X \ge 6)$, which gives 0.2783786.

3. (a) X is Binomial R.V.

   (b) The parameters are $n = 20$ and $p = 0.01$. $P(\text{refunded}) = P(X \ge 2)$ and can be calculated using R command *1-pbinom(1, 20, 0.01)*, which gives 0.01685934.

4. (a) X is Binomial R.V. with $n = 20$ and $p = 0.5$. Thus, $E(X) = 10 \times 0.5 = 5$ and $\text{Var}(X) = 10 \times 0.5 \times 0.5 = 2.5$.

   (b) $P(X = 5)$ can be calculated using R command *dbinom(5, 10, 0.5)*, which gives 0.2461.

   (c) This is $P(X \leq 5) = F(5)$, which can be calculated by the command *pbinom(5, 10, 0.5)* and the result is 0.6230.

   (d) Y is the total number of incorrectly answered questions.

   (e) $P(2 \leq Y \leq 5) = P(2 \leq 10 - X \leq 5) = P(5 \leq X \leq 8) = P(X \leq 8) - P(X \leq 4) = F(8) - F(4)$. This can be calculated by the command *pbinom(8, 10, 0.5)-pbinom(4, 10, 0.5)* and the result is 0.6123.

5. (a) X is Binomial R.V. with $n = 10$ and $p = 0.9$.

   (b) $E(X) = 10 \times 0.9 = 9$ and $\text{Var}(X) = 10 \times 0.9 \times 0.1 = 0.9$.

   (c) $P(X \geq 7) = 1 - P(X \leq 6)$, the R command is *1-pbinom(6, 10, 0.9)*, and the result is 0.9872.

   (d) Let Y be the catering cost, then $Y = 100+10X$, thus $E(Y) = 100+10E(X) = 190$, and $\text{Var}(Y) = 100\text{Var}(X) = 90$.

6. (a) Let X be the number of guilty votes when the defendant is guilty, then X has a Binomial distribution with $n = 9$ and $p = 0.9$. In order to convict a guilty defendant, we must have $X > 4$. Thus, the probability of convicting is $P(X > 4) = 1 - P(X \leq 4)$, which could be calculated by *1- pbinom(4,9,0.9)* resulting in 0.9991. Similarly, the probability of convicting an innocent defendant is *1 - pbinom(4,10,0.1) = 0.00089*. Thus, the proportion of all defendants convicted is $0.4 \times 0.00089 + 0.6 \times 0.9991 = 0.5998$.

   (b) Let G and I be the defendant is guilty and innocent, respectively and let VG and VI be the events that the defendant is voted as guilty and innocent, respectively. Then, $P(G) = 0.6$, $P(I) = 0.4$ and $P(VG|G) = 0.9991$ from part (a). Let Y be the number of guilty votes when the defendant is innocent. Then Y has a Binomial distribution with $n = 9$ and $p = 0.1$. Thus, $P(VI|I) = P(Y \leq 4)$, which could be calculated by *pbinom(4,9,0.1)* resulting in 0.9991. Thus,

$$\begin{aligned} P(\text{Correct}) &= P((VG \cap G) \cup (VI \cap I)) = P(VG|G)P(G) + P(VI|I)P(I) \\ &= 0.9991 \times 0.6 + 0.9991 \times 0.4 = 0.9991. \end{aligned}$$

7. (a) The R. V. X follows Negative Binomial distribution with parameter $r = 1$ and $p = 0.3$.

   (b) The sample space is $S = \{1, 2, \cdots\}$, and $p(x) = P(X = x) = p(1 - p)^{x-1}$.

   (c) $E(X) = 1/p = 3.33$, and $\text{Var}(X) = (1 - p)/p^2 = 7.778$.

8. (a) The R. V. X follows Negative Binomial distribution with parameter $r = 5$ and $p = 0.05$.

   (b) The sample space is $S = \{5, 6, \cdots\}$, and $p(x) = P(X = x) = \binom{x-1}{4} p^5 (1-p)^{x-5}$.

   (c) The R command is *1-pnbinom(30, 5, 0.05)* and it gives 0.971.

9. (a) Let X be the number of games needed for team A to win twice, then X has the negative binomial distribution with $r = 3$ and $p = 0.6$. Team A will win the series if $X = 3$ or $X = 4$ or $X = 5$. Thus, the probability can be calculated using the command *sum(dnbinom(0:2, 3, 0.6))*, which gives 0.6826.

   (b) The probability for a better team to win a best-of-five series is larger. With more games played, the better team will win more games.

10. (a) The R. V. X follows Negative Binomial distribution with parameter $r = 1$ and $p = 0.01$.

    (b) The sample space is $S = \{1, 2, \cdots\}$ and $p(x) = P(X = x) = p(1-p)^{x-1}$.

    (c) $E(X) = 1/p = 100$.

11. (a) The R. V. Y follows Negative Binomial distribution with parameter $r = 5$ and $p = 0.01$.

    (b) $E(Y) = r/p = 500$ and $\text{Var}(X) = r(1-p)/p^2 = 49500$.

12. (a) The R. V. X follows Hypergeometric distribution with parameter $n = 5$, $M_1 = 6$ and $M_2 = 9$. $N = M_1 + M_2 = 15$.

    (b) The sample space is $S_X = \{0, 1, 2, \cdots, 5\}$ and the PMF is

$$p(x) = P(X = x) = \frac{\binom{6}{x}\binom{9}{5-x}}{\binom{15}{5}}.$$

    (c) The R command is *sum(dhyper(2:4, 6, 9, 5))* and it gives 0.7043.

    (d) $E(X) = nM_1/N = 5 \times 6/15 = 2$, and

$$\text{Var}(X) = n\frac{M_1}{N}\left(1 - \frac{M_1}{N}\right)\frac{N-n}{N-1} = 5\frac{6}{15}\left(1 - \frac{6}{15}\right)\frac{15-5}{15-1} = 0.8571.$$

13. (a) The R. V. X follows Hypergeometric distribution with parameter $n = 5$, $M_1 = 3$ and $M_2 = 17$. $N = M_1 + M_2 = 20$.

    (b) The sample space is $S_X = \{0, 1, 2, 3\}$ and the PMF is

$$p(x) = P(X = x) = \frac{\binom{3}{x}\binom{17}{5-x}}{\binom{20}{5}}.$$

    (c) The R command is *dhyper(1, 3, 17, 5)* and it gives 0.4605.

(d) $E(X) = nM_1/N = 5 \times 3/20 = 0.75$, and

$$\text{Var}(X) = n\frac{M_1}{N}\left(1 - \frac{M_1}{N}\right)\frac{N-n}{N-1} = 5\frac{3}{20}\left(1 - \frac{3}{20}\right)\frac{20-5}{20-1} = 0.5033.$$

14. (a) The R. V. X follows Hypergeometric distribution with parameter $n = 20$, $M_1 = 200$ and $M_2 = 800$. $N = M_1 + M_2 = 1000$.

    (b) The R command is *phyper(4, 200, 800, 20)* and it gives 0.6301.

    (c) We can use the Binomial distribution with parameter $n = 20$ and $p = M_1/N = 0.2$ to approximate the distribution of X.

    (d) The R command is *pbinom(4, 20, 0.2)* and it gives 0.6296. The result is very close to that in part (b).

15. (a) The R. V. X follows Hypergeometric distribution with parameter $n = 50$, $M_1 = 300$ and $M_2 = 9700$. $N = M_1 + M_2 = 10000$.

    (b) The R command is *phyper(3, 300, 9700, 50)* and it gives 0.9377.

    (c) We can use the Binomial distribution with parameter $n = 50$ and $p = M_1/N = 0.03$ to approximate the distribution of X.

    (d) The R command is *pbinom(3, 50, 0.03)* and it gives 0.9372. The result is very close to that in part (b).

16. (a) Poisson.

    (b) $E(Y) = \text{Var}(Y) = \lambda = 1800 \times 0.6 = 1080$.

    (c) The R command is *1-ppois(1100, 1080)* and it gives 0.2654.

17. Let X be the number of loads during the next quarter, then X has a Poisson distribution with $\lambda = 0.5$. We are looking for $P(X > 2)$, the R command *1-ppois(2, 0.5)* gives us 0.0144.

18. (a) The R. V. X follows Poisson distribution with parameter $\lambda = 1.6 \times 3 = 4.8$.

    (b) $E(X) = \text{Var}(X) = \lambda = 4.8$.

    (c) The R command *ppois(9, 4.8)-ppois(4, 4.8)* gives us 0.4986.

    (d) $Y = 5000X$, thus $E(Y) = 5000E(X) = 24000$ and $\text{Var}(Y) = 5000^2\text{Var}(X) = 1.2 \times 10^8$.

19. (a) $\text{Var}(X_1) = 2.6$ and $\text{Var}(X_2) = 3.8$.

    (b) Let T_1 and T_2 be the event that the article is handled by typesetter 1 and 2, respectively. Then

$$P(\text{No error}|T_1) = e^{-\lambda_1}\frac{\lambda_1^0}{0!} = e^{-2.6} \text{ and } P(\text{No error}|T_2) = e^{-\lambda_2}\frac{\lambda_2^0}{0!} = e^{-3.8}.$$

Thus

$$\begin{aligned} P(\text{No error}) &= P(\text{No error}|T_1)P(T_1) + P(\text{No error}|T_2)P(T_2) \\ &= e^{-2.6} \times 0.6 + e^{-3.8} \times 0.4 = 0.0535. \end{aligned}$$

(c) By Bayes' Rule

$$P(T_2|\text{No error}) = \frac{P(\text{No error}|T_2)P(T_2)}{P(\text{No error})} = \frac{e^{-3.8} \times 0.4}{0.0535} = 0.1672.$$

20. (a) The R. V. X follows Binomial distribution with parameter $n = 1500$ and $p = 0.002$.

    (b) The R command is *sum(dbinom(4:8, 1500, 0.002))* and it gives 0.3490.

    (c) The distribution of X can be approximated by Poisson with parameter $\lambda = np = 3$.

    (d) The R command is *sum(dpois(4:8, 3))* and it gives 0.3490.

    (e) The exact probability of no faulty is calculated by *dbinom(0, 1500, 0.002)*, which gives 0.0496, and the approximate probability of no faulty is calculated by *dpois(0, 3)*, which results in 0.0498.

21. (a) The random variable Y has a hypergeometric(300, 9700, 200) distribution, which can be approximated by a binomial(200, 0.03) distribution, which can be approximated by a Poisson(6) distribution.

    (b) The R command for the exact probability is *phyper(10, 300, 9700, 200)* which gives 0.9615; the R command for binomial approximation is *pbinom(10, 200, 0.03)* and the result is 0.9599; the R command for Poisson approximation is *ppois(10, 6)* and the result is 0.9574. The two approximations are quite good.

22. (a) Each fish has a small probability of being caught. It makes sense to assume that each fish behaves independently, thus we have a large number of Bernoulli trials with a small probability of success. As a consequence of Proposition 3.4-1, the number of fish caught by an angler is modeled as a Poisson random variable.

    (b) The probability of each disabled vehicle being abandoned on I95 is small and it makes sense to assume that each vehicle owner behaves independently. Thus we have a large number of Bernoulli trials with a small probability of success. As a consequence of Proposition 3.4-1, the number of disabled vehicle being abandoned on I95 in one year is modeled as a Poisson random variable.

    (c) Each person has a small probability of dialing a wrong telephone number and it makes sense to assume that each person behaves independently. Thus, we have a large number of Bernoulli trials with a small probability of success. As a consequence of Proposition 3.4-1, the number of wrongly dialed number in a city in one hour is modeled as a Poisson random variable.

(d) Each person has a small probability of living 100 years and it makes sense to assume that each person behaves independently. Thus, we have a large number of Bernoulli trials with a small probability of success. As a consequence of Proposition 3.4-1, the number of people who reach age 100 in a city is modeled as a Poisson random variable.

23. (a) Both of the two events mean that there is one event happened in $[0, t]$ and there is no event happened in $(t, 1]$.

(b) From Proposition 3.4-2, $X(0.6)$ has a Poisson distribution with $\lambda_1 = 2 \times 0.6 = 1.2$ and $X(1) - X(0.6)$ has a Poisson distribution with $\lambda_2 = 2 \times 0.4 = 0.8$. Furthermore, $X(0.6)$ and $X(1) - X(0.6)$ are independent, thus

$$\begin{aligned} P([X(t) = 1] \cap [X(1) - X(t) = 0]) &= P([X(t) = 1])\, P([X(1) - X(t) = 0]) \\ &= e^{-1.2}\frac{1.2^1}{1!}e^{-0.8} = 0.1624. \end{aligned}$$

(c)

(i) Both $T \le t$ and $X(t) = 1$ say that the event happened before or at time t.

(ii) The proof uses Proposition 3.4-2 and the results in (i) and Part (a):

$$\begin{aligned} P(T \le t | X(1) = 1) &= \frac{P([T \le t] \cap [X(1) = 1])}{P(X(1) = 1)} = \frac{P([X(t) = 1] \cap [X(1) = 1])}{P(X(1) = 1)} \\ &= \frac{P([X(t) = 1] \cap [X(1) - X(t) = 0])}{P(X(1) = 1)} \\ &= \frac{P([X(t) = 1])P([X(1) - X(t) = 0])}{P(X(1) = 1)} \\ &= \frac{e^{-\lambda t}\frac{(\lambda t)^1}{1!}e^{-\lambda(1-t)}\frac{(\lambda(1-t))^0}{0!}}{e^{-\lambda\times 1}\frac{(\lambda\times 1)^1}{1!}} = t. \end{aligned}$$

3.5 Models for Continuous Random Variables

1. Let T be the life time. From the problem, we know that T has a exponential distribution with parameter $\lambda = 1/6$.

(a) $P(T > 4) = 1 - P(T \le 4) = 1 - F(4) = 1 - (1 - e^{-\lambda 4}) = e^{-\lambda 4} = 0.513$.

(b) $Var(T) = 1/\lambda^2 = 36$. To find the 95th percentile, we solve the equation $F(t_{0.95}) = 0.95$, or, $1 - e^{-\lambda t_{0.95}} = 0.95$. $t_{0.95} = 17.9744$.

(c) Let T_r be the remaining life time. By the memoryless property of exponential distribution, T_r still has exponential distribution with $\lambda = 1/6$. Thus,

(i) $P(T_r > 5) = e^{-\lambda 5} = 0.4346$.

(ii) $E(T_r) = 1/\lambda = 6$.

2. Let T_1 be the time, in month, of the first arrival, then $T_1 \sim \text{Exp}(1)$.

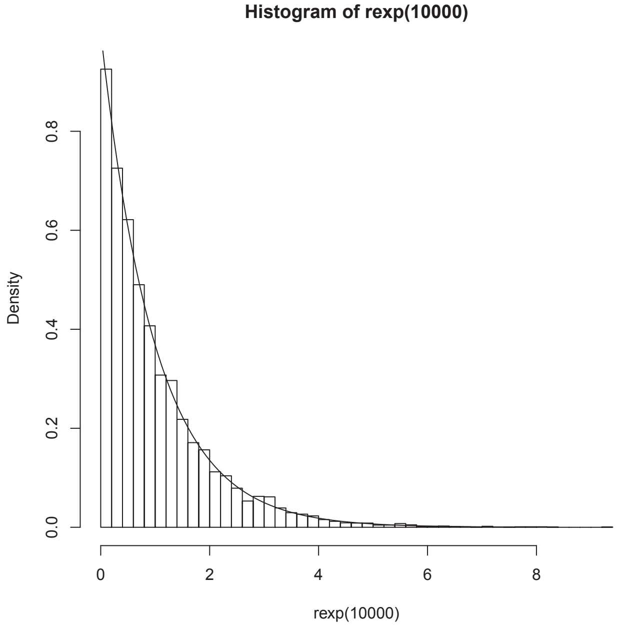
   (a) We are looking for $P(7/30 \leq T_1 \leq 14/30)$, by the CDF of exponential distribution, we have

   $$P\left(\frac{7}{30} \leq T_1 \leq \frac{14}{30}\right) = \exp\left(-1 \times \frac{7}{30}\right) - \exp\left(-1 \times \frac{14}{30}\right) = 0.1648.$$

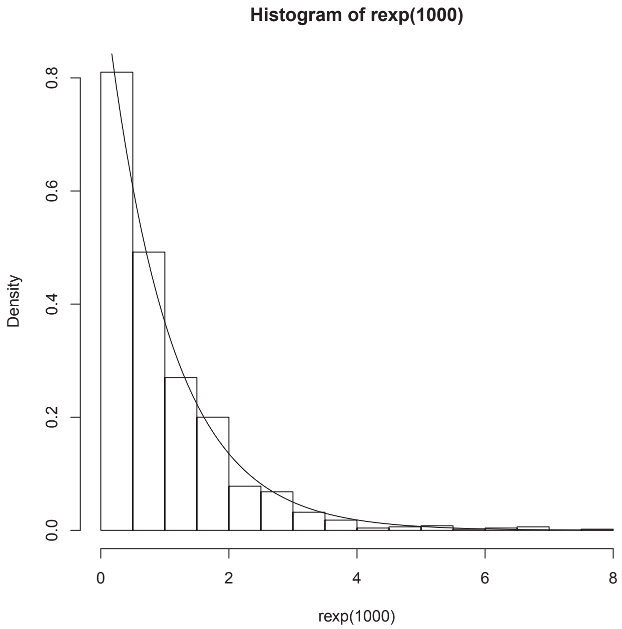
   (b) Let T_2 be the remaining waiting time. By the memoryless property of exponential distribution, $T_2 \sim \text{Exp}(1)$. Thus, $E(T_2) = 1$, and $Var(T_2) = 1$.

3. From (3.5.3), $P(X > s+t|X > s) = P(X > t)$, we must have $1 - P(X > s+t|X > s) = 1 - P(X > t)$, or, $P(X \leq s+t|X > s) = P(X \leq t) = F(t)$. Using the expression for $F(T)$ given in (3.5.1), we have $P(X \leq s+t|X \geq s) = 1 - e^{-\lambda t}$.

4. When using a sample size of 10,000, the resulting histogram superimposed with the PDF is given as below. This shows that the histogram provides very close approximation to the PDF.



When using a sample size of 1000, the resulting histogram superimposed with the PDF is given on the next page. The histogram is still reasonably close to the PDF.



5. (a) We could calculate this value by R command *qnorm(0.25, 43, 4.5)*, which gives 39.9648.

   (b) We could calculate this value by R command *qnorm(0.9, 43, 4.5)*, which gives 48.76698.

   (c) According to the requirement, $43+c$ must be the 99.5 percentile, which can be calculated using the command *qnorm(0.995, 43, 4.5)*, resulting in 54.59123. Thus, $c = 11.59123$.

   (d) The probability of a randomly selected A36 steel having strength less than 43 is 0.5. Let Y be the number of A36 steels having strength less than 43 among 15 randomly selected steels, then we are calculating $P(Y \leq 3)$. This can be calculated using the R command *sum(dbinom(0:3, 15, 0.5))* and the result is 0.01757812.

6. (a) We could calculate this value by R command *1-pnorm(8.64, 8, 0.5)*, which gives 0.100.

   (b) The probability is $0.1^3 = 0.001$.

7. (a) We could calculate this value by R command *pnorm(9.8, 9, 0.4)-pnorm(8.6, 9, 0.4)*, which gives 0.8185946.

   (b) Let Y be the number of acceptable resistors, then Y has the binomial distribution with $n = 4$ and probability of success as calculated in part (a). By the binomial probability formula, $P(Y = 2) = 0.1323$. We could also get this result by using R command.

8. (a) We could calculate this value by R command *1-pnorm(600, 500, 80)*, which gives 0.1056498.

   (b) We could calculate this value by R command *qnorm(0.99, 500, 80)*, which gives 686.1078.

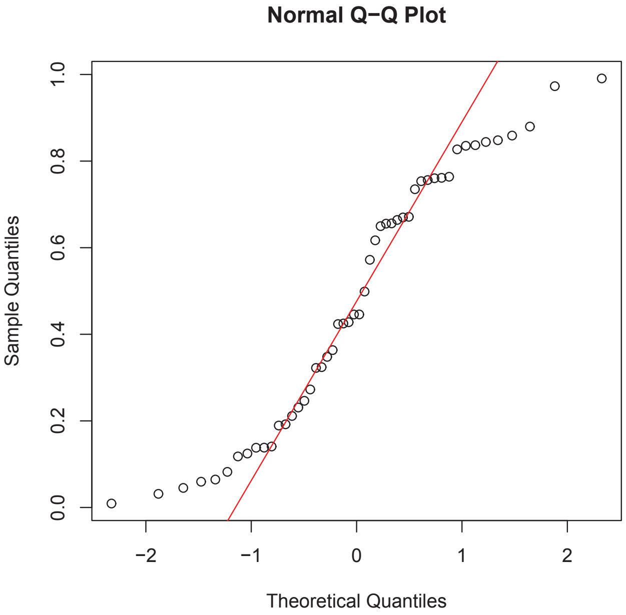
9. (a) We could calculate this value by R command *qnorm(0.1492, 10, 0.03)*, which gives 9.968804.

   (b) We could calculate this value by R command *pnorm(10.06, 10, 0.03)*, which gives 0.9772499.

10. (a) We could calculate this value by R command *pnorm(7.9, 10, 2)*, which gives 0.1468591.

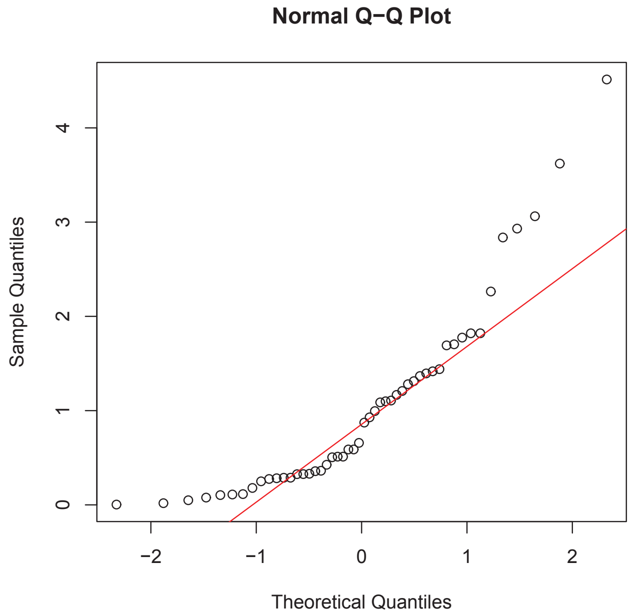
    (b) We could calculate this value by R command *qnorm(0.3, 10, 2)*, which gives 8.951199.

11. (a) The resulting normal Q-Q plot is shown below.



The plotted points are not close to the straight line, thus it indicates that the data does not come from a normal distribution.

(b) The resulting normal Q-Q plot is given on the next page.



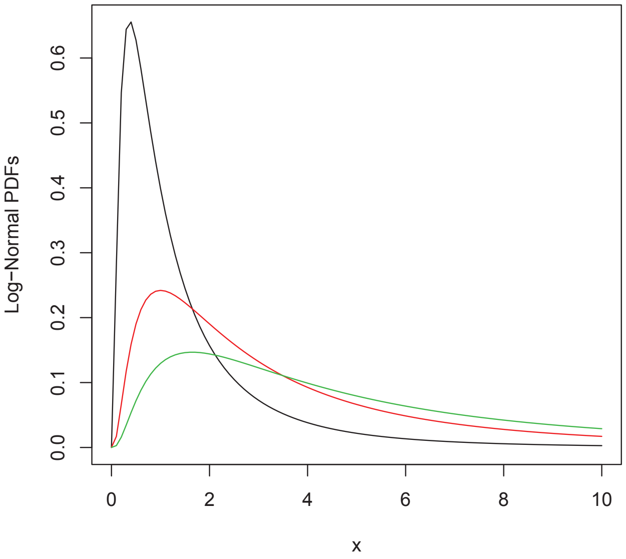
The plotted points are not close to the straight line, thus it indicates that the data does not come from a normal distribution.

12. (a) If $T \sim \text{log-norm}(\mu_{\ln}, \sigma_{\ln})$, the CDF is $F_T(t) = 0$ for $t < 0$. For $t > 0$

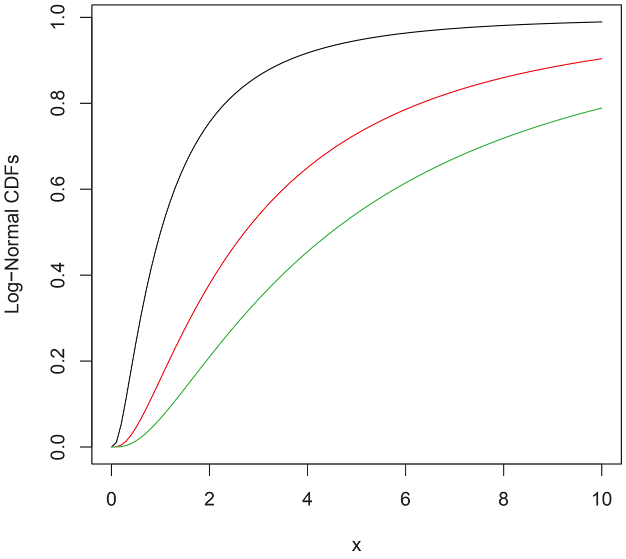
$$F_T(t) = P(T < t) = P(\log T < \log t) = \Phi\left(\frac{\log t - \mu_{\ln}}{\sigma_{\ln}}\right),$$

since $\log T \sim N(\mu_{\ln}, \sigma_{\ln})$. This is to be proved.

(b) The three PDFs are given as



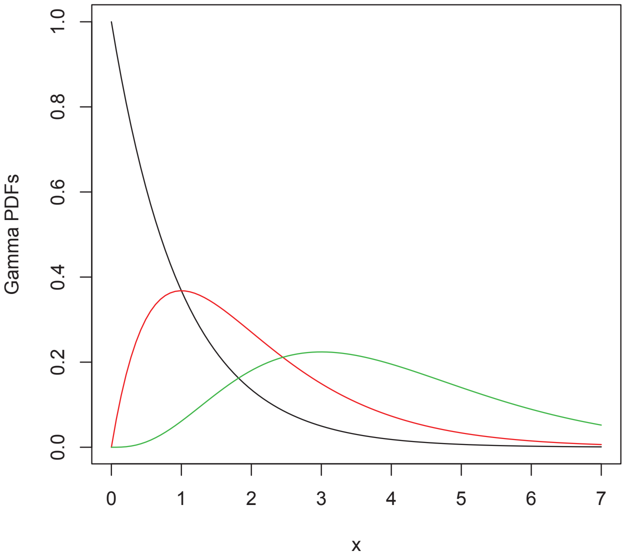
The three CDFs are given as



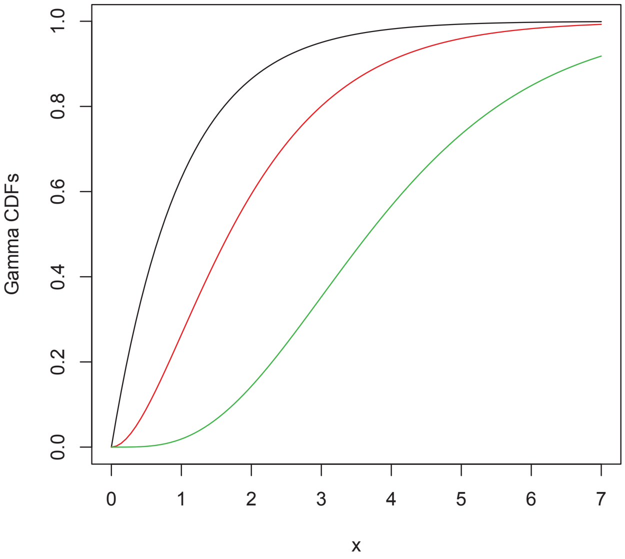
(c) For log-norm(0,1), we have $\mu = \sqrt{e} = 1.6487$ and $\sigma^2 = e(e-1) = 4.6708$. For log-norm(5,1), we have $\mu = e^5\sqrt{e} = 244.692$ and $\sigma^2 = e^{11}(e-1) = 102880.6$. For log-norm(5,2), we have $\mu = e^7 = 1096.633$ and $\sigma^2 = e^{14}(e^4-1) = 64457365$.

(d) The R commands *log(qlnorm(0.95))* and *qnorm(0.95)* indeed return the same value. The reason is that if T is a log-norm(0,1) random variable, then $\log T$ is a standard normal random variable.

13. (a) The three PDFs are given as



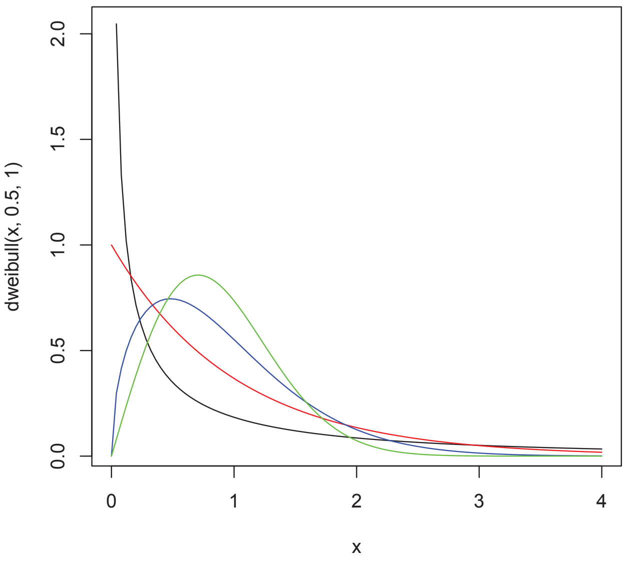
The three CDFs are given as



(b) For gamma(2,1), $\mu = 2$ and $\sigma^2 = 2$. For gamma(2,2), $\mu = 4$ and $\sigma^2 = 8$. For gamma(3,1), $\mu = 3$ and $\sigma^2 = 3$. For gamma(3,2), $\mu = 6$ and $\sigma^2 = 12$;

(c) The 95th percentile for gamma(2,1), gamma(2,2), gamma(3,1), and gamma(3,2) are respectively 4.743865, 2.371932, 6.295794, and 3.147897.

14. (a) The four PDFs are given as



(b) Using the commands, we have $\mu_T = 1200$ and $\sigma_T^2 = 361440000$.

(c) Using the CDF,

$$\begin{aligned} P(20 \le T < 30) &= F_T(30) - F_T(20) \\ &= [1 - \exp(-(30/10)^{0.2})] - [1 - \exp(-(20/10)^{0.2})] \\ &= 0.02931867. \end{aligned}$$

The R command gives the same answer.

(d) To find the 95th percentile, the equation is

$$F_T(t_{0.05}) = 0.95, \qquad 1 - \exp(-(t_{0.05}/10)^{0.2}) = 0.95,$$

the solution is $t_{0.05} = 2412.765$. The R command gives the same answer.

Chapter 4

Jointly Distributed Random Variables

4.2 Describing Joint Probability Distributions

1. (a) $P(X > 1, Y > 2) = P(X = 2, Y = 3) + P(X = 3, Y = 3) = 0.11 + 0.09 = 0.2$, $P(X > 1 \text{ or } Y > 2) = P(X = 2, Y = 1) + P(X = 3, Y = 1) + P(X = 2, Y = 2) + P(X = 3, Y = 2) + P(X = 2, Y = 3) + P(X = 3, Y = 3) + P(X = 1, Y = 3) = 0.79$, $P(X > 2, Y > 2) = P(X = 3, Y = 3) = 0.09$.

   (b) The marginal PMF of X is $p_X(1) = 0.09 + 0.12 + 0.13 = 0.34$, $p_X(2) = 0.12 + 0.11 + 0.11 = 0.34$, $p_X(3) = 0.13 + 0.10 + 0.09 = 0.32$.

   The marginal PMF of Y is $p_Y(1) = 0.09 + 0.12 + 0.13 = 0.34$, $p_Y(2) = 0.12 + 0.11 + 0.10 = 0.33$, $p_Y(3) = 0.13 + 0.11 + 0.09 = 0.33$.

2. (a) The marginal PMF of X is $p_X(0.0) = 0.388 + 0.009 + 0.003 = 0.4$, $p_X(1.0) = 0.485 + 0.010 + 0.005 = 0.5$, $p_X(2.0) = 0.090 + 0.006 + 0.004 = 0.1$.

   The marginal PMF of Y is $p_Y(0) = 0.388 + 0.485 + 0.090 = 0.963$, $p_Y(1) = 0.009 + 0.010 + 0.006 = 0.025$, $p_Y(2) = 0.003 + 0.005 + 0.004 = 0.012$.

   (b) (i) The probability that a randomly selected rat has one tumor is $P(Y = 1) = p_Y(1) = 0.025$.

   (ii) The probability that a randomly selected rat has at least one tumor is $P(Y \geq 1) = p_Y(1) + p_Y(2) = 0.037$.

   (c) (i) This is the conditional probability that $P(Y = 0|X = 1.0) = P(X = 1.0, Y = 0)/P(X = 1.0) = 0.485/0.5 = 0.97$.

   (ii) This is the conditional probability that $P(Y > 0|X = 1.0) = 1 - P(Y = 0|X = 1.0) = 0.03$.

3. (a) $P(X \leq 10, Y \leq 2) = P(X = 8, Y = 1.5) + P(X = 8, Y = 2) + P(X = 10, Y = 1.5) + P(X = 10, Y = 2) = 0.3 + 0.12 + 0.15 + 0.135 = 0.705$, $P(X \leq 10, Y = 2) = P(X = 8, Y = 2) + P(X = 10, Y = 2) = 0.12 + 0.135 = 0.255$.

   (b) The marginal PMF of X is $p_X(8) = 0.3 + 0.12 + 0.0 = 0.42$, $p_X(10) = 0.15 + 0.135 + 0.025 = 0.31$, $p_X(12) = 0.03 + 0.15 + 0.09 = 0.27$.

The marginal PMF of Y is $p_Y(1.5) = 0.3 + 0.15 + 0.03 = 0.48$, $p_Y(2) = 0.12 + 0.135 + 0.15 = 0.405$, $p_Y(2.5) = 0 + 0.025 + 0.09 = 0.115$.

(c) We want to find $P(X \leq 10|Y = 2)$, using Bayes' rule and parts (a) and (b),

$$P(X \leq 10|Y = 2) = \frac{P(X \leq 10, Y = 2)}{P(Y = 2)} = \frac{0.255}{0.405} = 0.6296.$$

4. (a) The table for the CDF is

			y	
	$F(x,y)$	1	2	3
	1	0.09	0.21	0.34
x	2	0.21	0.44	0.68
	3	0.34	0.67	1

(b) In this problem $F_X(x) = F(x, \infty) = F(x, 3)$, so $F_X(1) = 0.34$, $F_X(2) = 0.68$, $F_X(3) = 1$. By the same reason, we have $F_Y(1) = 0.34$, $F_Y(2) = 0.67$, $F_Y(3) = 1$.

(c) $P(X = 2, Y = 2) = F(2, 2) - F(2, 1) - F(1, 2) + F(1, 1) = 0.44 - 0.21 - 0.21 + 0.09 = 0.11$, as shown in the table of Exercise 1.

5. $P_{X_1}(0)$ is the sum of values in the first 3 by 3 cells and we have $P_{X_1}(0) = 0.3$. Similarly, $P_{X_1}(1) = 0.3$ and $P_{X_1}(2) = 0.4$.

$P_{X_2}(0)$ is the sum of values in the first row, and we have $P_{X_2}(0) = 0.27$. Similarly $P_{X_2}(1) = 0.38$, and $P_{X_2}(2) = 0.35$.

$P_{X_3}(1)$ is the sum of values in the three columns which is "$X_3 = 1$", and we have $P_{X_3}(1) = 0.29$. Similarly $P_{X_3}(2) = 0.34$, and $P_{X_3}(3) = 0.37$.

6. (a) The sample space is $\{(x_1, x_2, x_3, x_4)|x_i \text{ are integers}, 0 \leq x_1 \leq 3, 0 \leq x_2 \leq 2, 0 \leq x_3 \leq 1, x_1 + x_2 + x_3 + x_4 = 4\}$.

(b) The joint PMF of X_1, X_2, and X_3 is

$$\begin{aligned} p(x_1, x_2, x_3) &= P(X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = 4 - x_1 - x_2 - x_3) \\ &= \frac{\binom{3}{x_1}\binom{2}{x_2}\binom{1}{x_3}\binom{9}{4-x_1-x_2-x_3}}{\binom{15}{4}} \quad \text{for } x_1 + x_2 + x_3 \leq 4, \end{aligned}$$

and $p(x_1, x_2, x_3) = 0$ if $x_1 + x_2 + x_3 > 4$.

7. (a) To find k, we use $\iint f(x, y)dxdy = 1$, thus there is

$$\begin{aligned} 1 &= \iint f(x, y)dxdy = \int_0^2 \int_x^3 kxy^2 dydx = \int_0^2 kx\frac{1}{3}(3^3 - x^3)dx \\ &= \frac{9}{2}kx^2 \,|_0^2 - \frac{k}{15}x^5 \,|_0^2 = \frac{238}{15}k, \end{aligned}$$

hence, $k = 15/238$.

(b) The joint CDF of X and Y is

$$\begin{aligned} F(x,y) &= \int_0^x \int_u^y kuv^2 dv du = \int_0^x \frac{k}{3}uv^3\Big|_u^y du = \int_0^x \frac{ku}{3}(y^3 - u^3)du \\ &= \left(\frac{ku^2y^3}{6} - \frac{ku^5}{15}\right)\Bigg|_0^x = kx^2y^3/6 - kx^5/15. \end{aligned}$$

8. (a) Let region $R = \{(x,y)|0 \le x \le y\} \cap \{(x,y)|x+y \le 3\}$, then

$$\begin{aligned} P(X+Y \le 3) &= \iint_R f(x,y)dxdy = \int_0^{1.5}\int_x^{3-x} 2e^{-x-y}dydx = \int_0^{1.5} 2e^{-x}(e^{-x} - e^{x-3})dx \\ &= \int_0^{1.5} 2e^{-2x}dx - 2e^{-3} \times 1.5 = 1 - e^{-3} - 3e^{-3} = 1 - 4e^{-3}. \end{aligned}$$

(b) The marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_x^{\infty} 2e^{-x-y}dy = 2e^{-x}[-e^{-y}]\,|_x^{\infty} = 2e^{-2x} \quad \text{for} \quad x \ge 0,$$

and $f_X(x) = 0$ for $x < 0$.
The marginal PDF of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^y 2e^{-x-y}dx = 2e^{-y}[-e^{-x}]\,|_0^y = 2e^{-y}[1-e^{-y}] \quad \text{for} \quad y \ge 0,$$

and $f_Y(y) = 0$ for $y < 0$.

4.3 Conditional Distributions

1. (a) The marginal PMF of X is $p_X(0) = 0.06 + 0.04 + 0.2 = 0.3$, $p_X(1) = 0.08 + 0.3 + 0.06 = 0.44$, $p_X(2) = 0.1 + 0.14 + 0.02 = 0.26$.

The marginal PMF of Y is $p_Y(0) = 0.06+0.08+0.1 = 0.24$, $p_Y(1) = 0.04+0.3+0.14 = 0.48$, $p_Y(2) = 0.2 + 0.06 + 0.02 = 0.28$. X and Y are not independent, for example $P(X = 0, Y = 0) = 0.06 \ne p_X(0)p_Y(0)$.

(b) The conditional PMF $p_{Y|X=0}(y) = P(X = 0, Y = y)/p_X(0)$. That is, the first row of the table divides $p_X(0)$, thus $p_{Y|X=0}(0) = 0.06/0.3 = 0.2$, $p_{Y|X=0}(1) = 0.133$, and $p_{Y|X=0}(2) = 0.667$.

By the same reason, we have

$p_{Y|X=1}(0) = 0.1818$, $p_{Y|X=1}(1) = 0.6818$, and $p_{Y|X=1}(2) = 0.1364$.

$p_{Y|X=2}(0) = 0.3846$, $p_{Y|X=2}(1) = 0.5385$, and $p_{Y|X=2}(2) = 0.0769$.

The conditional PMF of Y depends on the value of X, thus X and Y are not independent.

(c) $E(Y|X=1) = 0p_{Y|X=1}(0) + 1p_{Y|X=1}(1) + 2p_{Y|X=1}(2) = 0.9546$, and $E(Y^2|X=1) = 0^2p_{Y|X=1}(0) + 1^2p_{Y|X=1}(1) + 2^2p_{Y|X=1}(2) = 1.2274$, Thus, $\mathrm{Var}(Y|X=1) = E(Y^2|X=1) - E(Y|X=1)^2 = 1.2274 - 0.9546^2 = 0.3161$.

2. (a) The regression function of Y on X is $E(Y|X=0) = 0p_{Y|X=0}(0) + 1p_{Y|X=0}(1) + 2p_{Y|X=0}(2) = 1.4667$, $E(Y|X=1) = 0.9546$ as calculated in 1 (c), $E(Y|X=2) = 0p_{Y|X=2}(0) + 1p_{Y|X=2}(1) + 2p_{Y|X=2}(2) = 0.6923$.

   (b) By the law of total expectation, $E(Y) = E(Y|X=0)p_X(0) + E(Y|X=1)p_X(1) + E(Y|X=2)p_X(2=) = 1.4667 \times 0.3 + 0.9546 \times 0.44 + 0.6923 \times 0.26 = 1.04$.

3. (a) The regression function is

$$E(Y|X=8) = \sum_y yp_{Y|X=8}(y) = \sum_y yp_{X,Y}(8,y)/p_X(8)$$
$$= 1.5 \times 0.3/0.42 + 2 \times 0.12/0.42 + 2.5 \times 0/0.42 = 1.643,$$

$$E(Y|X=10) = \sum_y yp_{Y|X=10}(y) = \sum_y yp_{X,Y}(10,y)/p_X(10)$$
$$= 1.5 \times 0.15/0.31 + 2 \times 0.135/0.31 + 2.5 \times 0.025/0.31 = 1.798,$$

and

$$E(Y|X=12) = \sum_y yp_{Y|X=12}(y) = \sum_y yp_{X,Y}(12,y)/p_X(12)$$
$$= 1.5 \times 0.03/0.27 + 2 \times 0.15/0.27 + 2.5 \times 0.09/0.27 = 2.111.$$

   (b) By the Law of Total Expectation

$$\begin{aligned} E(Y) &= \sum_x E(Y|X=x)p_X(x) \\ &= E(Y|X=8)p_X(8) + E(Y|X=10)p_X(10) + E(Y|X=12)p_X(12) \\ &= 1.643 \times 0.42 + 1.798 \times 0.31 + 2.111 \times 0.27 = 1.817. \end{aligned}$$

   (c) From part (a), we see that the regression function of Y on X depends on the value of X. Thus from part (1) of Proposition 4.3-3, the amount of tip left is not independent of the price of the meal.

4. (a) The conditional PMF of Y given $X = 1$ can be calculated as

$$p_{Y|X=1}(0) = \frac{p_{X,Y}(1,0)}{p_X(1)} = \frac{0.485}{0.5} = 0.97,$$

$$p_{Y|X=1}(1) = \frac{p_{X,Y}(1,1)}{p_X(1)} = \frac{0.01}{0.5} = 0.02,$$

and

$$p_{Y|X=1}(2) = \frac{p_{X,Y}(1,2)}{p_X(1)} = \frac{0.005}{0.5} = 0.01.$$

(b) The regression function is

$$\begin{aligned} E(Y|X=0) &= \sum_y y p_{Y|X=0}(y) = \sum_y y p_{X,Y}(0,y)/p_X(0) \\ &= 0 \times 0.388/0.4 + 1 \times 0.009/0.4 + 2 \times 0.003/0.4 = 0.0375, \end{aligned}$$

$$\begin{aligned} E(Y|X=1) &= \sum_y y p_{Y|X=1}(y) \\ &= 0 \times 0.97 + 1 \times 0.02 + 2 \times 0.01 = 0.04, \end{aligned}$$

and

$$\begin{aligned} E(Y|X=2) &= \sum_y y p_{Y|X=2}(y) = \sum_y y p_{X,Y}(2,y)/p_X(2) \\ &= 0 \times 0.09/0.1 + 1 \times 0.006/0.1 + 2 \times 0.004/0.1 = 0.14. \end{aligned}$$

(c) By the Law of Total Expectation

$$\begin{aligned} E(Y) &= \sum_x E(Y|X=x)p_X(x) \\ &= E(Y|X=0)p_X(0) + E(Y|X=1)p_X(1) + E(Y|X=2)p_X(2) \\ &= 0.0375 \times 0.4 + 0.04 \times 0.5 + 0.14 \times 0.1 = 0.049. \end{aligned}$$

5. (a) The conditional PMF of Y depends on the value of X, thus X and Y are not independent.

(b) The table for the joint PMF with the marginal PMFs is

		y		$p_X(x)$
	$P(x,y)$	0	1	
	0	0.3726	0.1674	0.54
x	1	0.1445	0.0255	0.17
	2	0.2436	0.0464	0.29
	$p_Y(y)$	0.7607	0.2393	

X and Y are not independent because $P(X=x, Y=y) = p_X(x)p_Y(y)$ does not hold for all the (x,y) combinations.

6. (a) The regression function of Y on X is $\mu_{Y|X}(x) = E(Y|X=x) = 0p_{Y|X=x}(0) + 1p_{Y|X=x}(1) = p_{Y|X=x}(1)$, thus, we have $E(Y|X=0) = 0.31$, $E(Y|X=1) = 0.15$, $E(Y|X=2) = 0.16$.

(b) By the law of total expectation, $E(Y) = E(Y|X = 0)p_X(0) + E(Y|X = 1)p_X(1)+E(Y|X = 2)p_X(2 =) = 0.31\times0.54+0.15\times0.17+0.16\times0.29 = 0.2393$.

7. (a) $E(Y|X = 1) = 1p_{Y|X=1}(1)+2p_{Y|X=1}(2) = 1\times0.66+2\times0.34 = 1.34$, $E(Y^2|X = 1) = 1^2p_{Y|X=1}(1) + 2^2p_{Y|X=1}(2) = 1 \times 0.66 + 2 \times 0.34 = 2.02$, thus $\text{Var}(Y|X = 1) = E(Y^2|X = 1) - E(Y|X = 1)^2 = 2.02 - 1.34^2 = 0.2244$.

(b) The table for the joint PMF is

		y	
	$P(x, y)$	1	2
	1	0.132	0.068
x	2	0.24	0.06
	3	0.33	0.17

(c) This problem is asking for $P(Y = 1) = p_Y(1) = 0.132 + 0.24 + 0.33 = 0.702$.

(d) This is asking for $P(X = 1|Y = 1) = P(X = 1, Y = 1)/P(Y = 1) = 0.132/0.702 = 0.188$.

8. (a) The regression function of Y on X is $\mu_{Y|X}(x) = E(Y|X = x) = 1p_{Y|X=x}(1) + 2p_{Y|X=x}(2)$. Thus, we have $E(Y|X = 0) = 1.34$, $E(Y|X = 1) = 1.2$, $E(Y|X = 2) = 1.34$.

(b) By the law of total expectation, $E(Y) = E(Y|X = 0)p_X(0) + E(Y|X = 1)p_X(1) + E(Y|X = 2)p_X(2 =) = 1.34 \times 2 + 1.2 \times 0.3 + 1.34 \times 0.5 = 1.298$.

9. (a) We know the marginal PMF of X is $p_X(4) = 0.3$, $p_X(5) = 0.5$, and $p_X(6) = 0.2$. Thus,

$$p_{X,Y}(x, 1) = P(Y = 1|X = x)p_X(x) = \frac{(-0.8 + 0.04x)^4}{1 + (-0.8 + 0.04x)^4}p_X(x),$$

and

$$p_{X,Y}(x, 0) = P(Y = 0|X = x)p_X(x) = \frac{1}{1 + (-0.8 + 0.04x)^4}p_X(x).$$

Using these formulas, we could get the following table for the joint PMF:

		y	
	$P(x, y)$	0	1
	4	0.2569	0.0431
x	5	0.4426	0.0574
	6	0.1821	0.0179

X and Y are not independent because the conditional distribution of Y depends on X.

(b) We have the marginal PMF of Y as $p_Y(0) = 0.2569 + 0.4426 + 0.1821 = 0.8816$ and $p_Y(1) = 0.1184$. Thus,

$$\begin{aligned} E(X|Y=1) &= \sum_{x=4}^{6} x p_{X|Y=1}(x) = \sum_{x=4}^{6} x p_{X,Y}(x,1)/p_Y(1) \\ &= 4 \times 0.0431/0.1184 + 5 \times 0.0574/0.1184 + 6 \times 0.0179/0.1184 \\ &= 4.7872, \end{aligned}$$

and

$$\begin{aligned} E(X|Y=0) &= \sum_{x=4}^{6} x p_{X|Y=0}(x) = \sum_{x=4}^{6} x p_{X,Y}(x,0)/p_Y(0) \\ &= 4 \times 0.2569/0.8816 + 5 \times 0.4426/0.8816 + 6 \times 0.1821/0.8816 \\ &= 4.9152. \end{aligned}$$

10. (a) By the definition of Binomial distribution, it is clear that given $X = x$, $Y \sim \text{Bin}(x, 0.6)$. The joint PMF can be calculated using the formula

$$p_{X,Y}(x,y) = \binom{x}{y} 0.6^y 0.4^{x-y} p_X(x), \text{ with } 0 \le y \le x \le 4.$$

According to this formula and the given marginal distribution of X, we have

$$p_{X,Y}(0,0) = 0.1,\ p_{X,Y}(1,0) = 0.08,\ p_{X,Y}(1,1) = 0.12,$$

$$p_{X,Y}(2,0) = 0.048,\ p_{X,Y}(2,1) = 0.144,\ p_{X,Y}(2,2) = 0.108$$

$$p_{X,Y}(3,0) = 0.016,\ p_{X,Y}(3,1) = 0.072,\ p_{X,Y}(3,2) = 0.108,$$

$$p_{X,Y}(3,3) = 0.054,\ p_{X,Y}(4,0) = 0.00384,\ p_{X,Y}(4,1) = 0.02304,$$

and

$$p_{X,Y}(4,2) = 0.05184,\ p_{X,Y}(4,3) = 0.05184,\ p_{X,Y}(4,4) = 0.01944.$$

(b) Given $X = x$, $Y \sim \text{Bin}(x, 0.6)$, thus the regression function $E(Y|X = x) = 0.6x$.

(c) By the law of total expectation, $E(Y) = \sum_x E(Y|X = x)p_X(x) = 0.6 \sum_x x p_X(x) = 0.6 \times [0 \times 0.1 + 1 \times 0.2 + 2 \times 0.3 + 3 \times 0.25 + 4 \times 0.15] = 1.29$

11. (a) The marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 (x+y)dy = x + \frac{1}{2}, \quad \text{for} \quad 0 < x < 1.$$

Thus, the conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{x+y}{x+1/2}, \quad \text{for} \quad 0 < x < 1, 0 < y < 1.$$

Hence,

$$P(0.3 < Y < 0.5|X = x) = \int_{0.3}^{0.5} f_{Y|X=x}(y)dy = \int_{0.3}^{0.5} \frac{x+y}{x+1/2}dy$$
$$= \frac{0.2x + 0.08}{x + 1/2}.$$

(b) By (4.3.16),

$$P(0.3 < Y < 0.5) = \int_0^1 P(0.3 < Y < 0.5|X = x)f_X(x)dx$$
$$= \int_0^1 \frac{0.2x + 0.08}{x+1/2}(x + 1/2)dx = 0.18.$$

12. (a) Independent

(b) Not independent

(c) Not independent

13. Since T_1 and T_2 are the first two inter-arrival times of a Poisson process $X(s)$, $s \geq 0$, with rate α, according to Proposition 3.5-1, both T_1 and T_2 have an exponential distribution with PDF $f(t) = \alpha e^{-\alpha t}$, $t > 0$. To show T_1 and T_2 are independent, let us consider the event $T_2 > t$ given $T_1 = s$. This means that the first arrival occurs at s while the second arrival occurs after time t, thus there is no event in $(s, s+t]$. Therefore

$$P(T_2 > t|T_1 = s) = P(\text{No events in } (s, s+t]|T_1 = s).$$

By the third postulate in definition 3.4-1 of a Poisson process, the events " no event in $(s, s+t]$" and "$T_1 = s$" are independent, thus

$$P(\text{No events in } (s, s+t]|T_1 = s) = P(\text{No events in } (s, s+t]) = P(X(s+t)-X(s) = 0).$$

According to part (2) of Proposition 3.4-2, $X(s+t)-X(s)$ has Poisson distribution with parameter $\alpha(s + t - t) = \alpha t$, then

$$P(X(s + t) - X(s) = 0) = e^{-\alpha t}.$$

Combining these arguments, we have

$$P(T_2 > t|T_1 = s) = e^{-\alpha t}.$$

Hence, the conditional PDF of T_2 given $T_1 = s$ is

$$f_{T_2|T_1=s}(t) = \frac{d}{dt}P(T_2 < t|T_1 = s) = -\frac{d}{dt}P(T_2 > t|T_1 = s) = \alpha e^{-\alpha t} = f_{T_2}(t),$$

and this shows that T_1 and T_2 are independent.

14. (a) T_2 is the distance from the first pothole to the second one and, according to Proposition 3.5-1, T_2 is exponential(0.16). Since the first pothole found is 8 miles from the start, the second pothole will be found between 14 and 19 miles from the star if and only if $14-8 \leq T_2 \leq 19-8$. Thus, the desired probability is $P(6 \leq T_2 \leq 11) = 0.2108$ by the command *pexp(11,0.16) - pexp(6,0.16)*.

    (b) T_1 and T_2 are both exponential (0.16) random variables and the result from Exercise 13 shows that T_1 and T_2 are independent. Thus, $E(T_2|T_1 = x) = E(T_2) = 1/0.16 = 6.25$. Hence, the regression function of Y on X is

$$E(Y|X = x) = E(T_1 + T_2|T_1 = x) = x + E(T_2|T_1 = x) = x + 6.25.$$

15. The conditional PDF of X given $Y = y$ is

$$f_{X|Y=y}(x) = \frac{1}{y}e^{-x/y} \quad \text{for } x > 0,$$

and $f_{X|Y=y}(x) = 0$ otherwise. Thus, $f_{X|Y=y}(x)$ depends on y. According to Proposition 4.3-2 (4), X and Y are not independent.

16. (a) The regression function is

$$E(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)dy = \int_0^{\infty} yxe^{-xy}dy = \frac{1}{x}\int_0^{\infty} ze^{-z}dz = \frac{1}{x}\Gamma(2) = \frac{1}{x},$$

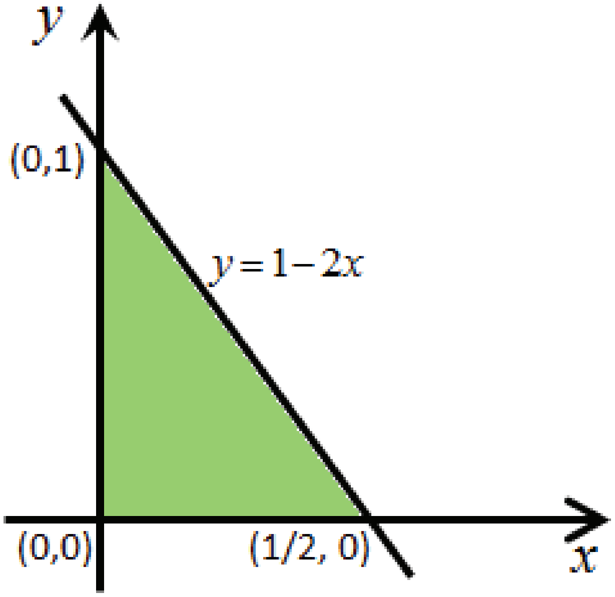
    where we used the variable transformation $z = xy$ and the definition of Gamma function.

    Clearly, $E(Y|X = 5.1) = 1/5.1 = 0.1961$.

    (b) By the law of total expectation,

$$E(Y) = \int_{-\infty}^{\infty} E(Y|X = x)f_X(x)dx = \int_5^6 \frac{1}{x}\frac{1}{\log 6 - \log 5}\frac{1}{x}dx = \frac{1/5 - 1/6}{\log 6 - \log 5}.$$

17. (a) The support of the joint PDF is given on the following page.



(b) The support of the joint PDF is not a rectangle, thus X and Y are not independent.

(c) We have

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^{1-2x} 24x dy = 24x(1-2x), \quad \text{for} \quad 0 \le x \le 0.5,$$

and $f_X(x) = 0$ otherwise.

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^{(1-y)/2} 24x dx = 12x^2 \mid_0^{(1-y)/2} = 3(1-y)^2,$$

for $0 \le y \le 1$, and $f_Y(y) = 0$ otherwise. Thus,

$$E(X) = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^{0.5} 24x^2(1-2x)dx = \int_0^1 3t^2(1-t)dt$$
$$= 3\frac{2!1!}{4!} = \frac{1}{4},$$

and

$$E(Y) = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 3y(1-y)^2 dy = 3\frac{2!1!}{4!} = \frac{1}{4}.$$

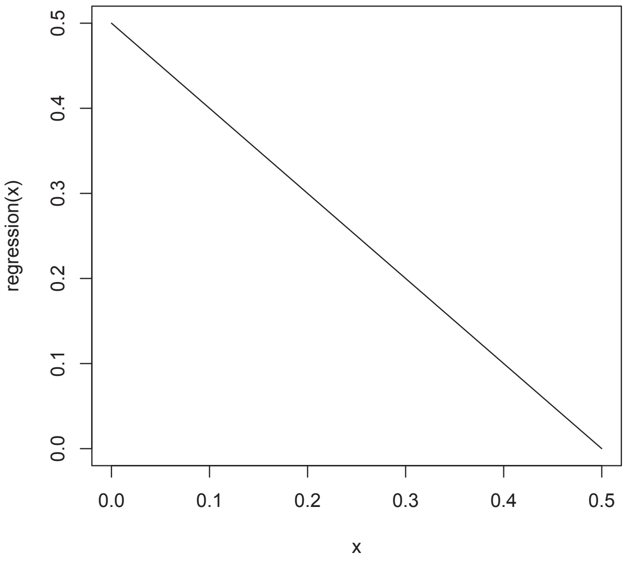
18. (a) The conditional PDF of Y given $X = x$ is

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{24x}{24x(1-2x)} = \frac{1}{1-2x},$$

for $0 \le y \le 1-2x$, and $f_{Y|X=x}(y) = 0$ otherwise. The regression function is

$$E(Y|X=x) = \int_{-\infty}^{\infty} y f_{Y|X=x}(y)dy = \int_0^{1-2x} y\frac{1}{1-2x}dy = \frac{1-2x}{2} = \frac{1}{2} - x.$$

The plot for the regression function is given as follows:



By the expression, $E(Y|X=0.3) = 0.5 - 0.3 = 0.2$.

(b) By the law of total expectation,

$$\begin{aligned} E(Y) &= \int_{-\infty}^{\infty} E(Y|X=x)f_X(x)dx = \int_0^{0.5} \frac{1-2x}{2} 24x(1-2x)dx \\ &= \int_0^{0.5} 12x(1-2x)^2dx = \int_0^1 3t(1-t)^2dt = 3\frac{2!1!}{4!} = \frac{1}{4}. \end{aligned}$$

4.4 Mean Value of Functions of Random Variables

1. The price the person pays is $P = \min\{X, Y\}$. So

$$\begin{aligned} E(P) = {} & \min(150,150) \times 0.25 + \min(150,135) \times 0.05 + \min(150,120) \times 0.05 \\ & + \min(135,150) \times 0.05 + \min(135,135) \times 0.2 + \min(135,120) \times 0.1 \\ & + \min(120,150) \times 0.05 + \min(120,135) \times 0.1 + \min(120,120) \times 0.15 = 132, \end{aligned}$$

and

$$\begin{aligned}E(P^2) &= \min(150,150)^2 \times 0.25 + \min(150,135)^2 \times 0.05 + \min(150,120)^2 \times 0.05 \\ &\quad + \min(135,150)^2 \times 0.05 + \min(135,135)^2 \times 0.2 + \min(135,120)^2 \times 0.1 \\ &\quad + \min(120,150)^2 \times 0.05 + \min(120,135)^2 \times 0.1 + \min(120,120)^2 \times 0.15 \\ &= 17572.5.\end{aligned}$$

Therefore $\text{Var}(P) = E(P^2) - E(P)^2 = 17572.5 - 132^2 = 148.5$.

2. (a) Let X and Y be the times components A and B fail, respectively, so the system fails at time $T = \max\{X, Y\}$. Then, the CDF of T is $F_T(t) = 0$ if $t \notin [0,1]$; if $t \in [0,1]$,

$$\begin{aligned}F_T(t) &= P(T \le t) = P(\max\{X,Y\} \le T) = P(X \le T, Y \le T) \\ &= P(X \le t)P(Y \le t) = t^2,\end{aligned}$$

where we used the independence of X and Y, and since X and Y are uniform(0,1) random variables, $P(X \le t) = P(Y \le t) = t$.

Thus, the PDF of T $f_T(t) = 2t$ for $t \in [0,1]$ and $f_T(t) = 0$ otherwise.

(b) $E(T) = \int t f_T(t)dt = \int_0^1 2t^2 dt = 2/3$ and

$$E(T^2) = \int t^2 f_T(t)dt = \int_0^1 2t^3 dt = 1/2,$$

thus

$$\text{Var}(T) = E(T^2) - E(T)^2 = 1/2 - (2/3)^2 = 1/18.$$

3. The volume of the cylinder is $h(X,Y) = \pi Y^2 X$. In Example 4.3-17 it was found that $E[h(X,Y)] = (13/16)\pi$. We also have

$$\begin{aligned}E[h^2(X,Y)] &= \iint h^2(x,y)f(x,y)dxdy = \int_0^3 \int_{0.5}^{0.75} \frac{3x}{8y^2}\pi^2 y^4 x^2 dydx \\ &= \frac{3\pi^2}{8}\int_0^3 x^3 dx \int_{0.5}^{0.75} y^2 dy = \frac{3\pi^2}{8}\left[\frac{1}{4}x^4 \,|_0^3\right]\left[\frac{1}{3}y^3 \,|_{0.5}^{0.75}\right] \\ &= \frac{1539}{2048}\pi^2.\end{aligned}$$

Thus,

$$\text{Var}[h(X,Y)] = E[h^2(X,Y)] - E[h(X,Y)]^2 = \frac{1539}{2048}\pi^2 - \frac{13^2}{16^2}\pi^2 = 0.901.$$

4. (a) The total waiting time is $T = X_1 + X_2 + \cdots + X_5 + Y_1 + Y_2 + Y_3$.

(b) The expected value is

$$\begin{aligned}E(T) &= E(X_1) + \cdots + E(X_5) + E(Y_1) + E(Y_2) + E(Y_3)\\ &= 3 \times 5 + 6 \times 3 = 33\end{aligned}$$

and the variance is

$$\begin{aligned}\mathrm{Var}(T) &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_5) + \mathrm{Var}(Y_1) + \mathrm{Var}(Y_2) + \mathrm{Var}(Y_3)\\ &= 2 \times 5 + 4 \times 3 = 22.\end{aligned}$$

In order to make the calculation valid, we have to assume that the waiting times are independent.

5. (a) Let X be the height of a randomly selected segment, then X is a uniform(35.5, 36.5) random variable. Thus $E(X) = (35.5 + 36.5)/2 = 36$, and $\mathrm{Var}(X) = (35.5 - 36.5)^2/12 = 1/12$.

   (b) Let H_1 be the height of tower 1, then $H_1 = X_1 + \cdots + X_{30}$. Thus, $E(H_1) = E(X_1) + \cdots + E(X_{30}) = 36 \times 30 = 1080$, and $\mathrm{Var}(H_1) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_{30}) = 30/12 = 2.5$.

   (c) Let $Y_1, \cdots, Y_{30}$ be the heights of the segments used in tower 2, and let H_2 be the height of tower 2, then $H_2 = Y_1 + \cdots + Y_{30}$. As in part (b), we can find $E(H_2) = 1080$ and $\mathrm{Var}(H_2) = 2.5$. Let D be the difference of the heights of the two towers, then $D = H_1 - H_2$. It makes sense to assume that the concrete segments are independent, thus H_1 and H_2 are independent. Then, $E(D) = E(H_1) - E(H_2) = 0$ and $\mathrm{Var}(D) = \mathrm{Var}(H_1 - H_2) = \mathrm{Var}(H_1) + \mathrm{Var}(H_2) = 5$.

6. The number of injuries in a month is $X_1 + X_2 + \cdots + X_N$ and $\mu = E(X_i) = 1.5$. Thus the expected number to injuries in a month is $E(X_1 + X_2 + \cdots + X_N) = E(N)\mu = 7 \times 1.5 = 10.5$.

7. The total tips is $T = X_1 + \cdots + X_{N_1} + Y_1 + \cdots + Y_{N_2}$, and we know that $\mu_1 = E(X_i) = 20$, $\mu_2 = E(Y_j) = 10$, N_1 is Poisson(4), N_2 is Poisson(6). Thus, the expected value of the total amount of tips is

$$\begin{aligned}E(T) &= E(X_1 + \cdots + X_{N_1}) + E(Y_1 + \cdots + Y_{N_2}) = E(N_1)\mu_1 + E(N_2)\mu_2\\ &= 4 \times 20 + 6 \times 10 = 140.\end{aligned}$$

8. The marginal PDF of X was derived in Example 4.3-9 as $f_X(x) = 12x(1-x)^2$ for $0 \le x \le 1$ and zero otherwise. Then,

$$E(X) = \int x f_X(x)dx = \int_0^1 12x^2(1-x)^2dx = \frac{2}{5}.$$

By the symmetry of the joint PDF, the marginal PDF of Y is the same as that of X. It follows that $E(Y) = E(X) = 2/5$. We calculate

$$E(XY) = \iint xyf(x,y)dxdy = \int_0^1 dx \int_0^{1-x} 24x^2y^2 dxdy = 8\int_0^1 x^2(1-x)^3 dx$$
$$= \frac{2}{15}.$$

Hence,

$$\text{Cov}(X,Y) = E(XY) - E(Y)E(X) = \frac{2}{15} - \frac{2}{5}\frac{2}{5} = -\frac{2}{75}.$$

9. $\text{Cov}(X,Y) = \text{Cov}(X, 9.3 + 1.5X + \epsilon) = 1.5\text{Cov}(X,X) + \text{Cov}(X,\epsilon) = 1.5\sigma_X^2 = 1.5 \times 9 = 13.5$, $\text{Cov}(\epsilon, Y) = \text{Cov}(\epsilon, 9.3 + 1.5X + \epsilon) = 1.5\text{Cov}(\epsilon, X) + \text{Cov}(\epsilon,\epsilon) = \sigma_\epsilon^2 = 16$.

10. For a randomly selected customer, the total cost of the meal is $T = X + Y$. Thus the expected value is

$$E(T) = E(X+Y) = \sum_x\sum_y (x+y)p(x,y) = (8+1.5)\times 0.3 + (8+2.0)\times 0.12$$
$$+ (8+2.5)\times 0 + (10+1.5)\times 0.15 + (10+2.0)\times 0.135 + (10+2.5)\times 0.025$$
$$+ (12+1.5)\times 0.03 + (12+2.0)\times 0.15 + (12+2.5)\times 0.09$$
$$= 11.5175,$$

and

$$E(T^2) = E[(X+Y)^2] = \sum_x\sum_y (x+y)^2 p(x,y) = (8+1.5)^2\times 0.3 + (8+2.0)^2\times 0.12$$
$$+ (8+2.5)^2\times 0 + (10+1.5)^2\times 0.15 + (10+2.0)^2\times 0.135$$
$$+ (10+2.5)^2\times 0.025 + (12+1.5)^2\times 0.03 + (12+2.0)^2\times 0.15$$
$$+ (12+2.5)^2\times 0.09 = 136.0487.$$

Therefore, $\text{Var}(T) = E(T^2) - E(T)^2 = 136.0487 - 11.5175^2 = 3.3959$.

11. Similar to the previous exercise,

$$E(8X+10Y) = \sum_x\sum_y (8x+10y)p(x,y) = (8\times 0 + 10\times 0)\times 0.06$$
$$+ (8\times 0 + 10\times 1)\times 0.04 + (8\times 0 + 10\times 2)\times 0.2$$
$$+ (8\times 1 + 10\times 0)\times 0.08 + (8\times 1 + 10\times 1)\times 0.3$$
$$+ (8\times 1 + 10\times 2)\times 0.06 + (8\times 2 + 10\times 0)\times 0.1$$
$$+ (8\times 2 + 10\times 1)\times 0.14 + (8\times 2 + 10\times 2)\times 0.02$$
$$= 18.08$$

and

$$E[(8X+10Y)^2] = \sum_x \sum_y (8x+10y)^2 p(x,y) = (8\times 0+10\times 0)^2 \times 0.06$$
$$+ (8\times 0+10\times 1)^2 \times 0.04 + (8\times 0+10\times 2)^2 \times 0.2$$
$$+ (8\times 1+10\times 0)^2 \times 0.08 + (8\times 1+10\times 1)^2 \times 0.3$$
$$+ (8\times 1+10\times 2)^2 \times 0.06 + (8\times 2+10\times 0)^2 \times 0.1$$
$$+ (8\times 2+10\times 1)^2 \times 0.14 + (8\times 2+10\times 2)^2 \times 0.02$$
$$= 379.52.$$

Thus,

$$\text{Var}(8X+10Y) = E[(8X+10Y)^2] - E(8X+10Y)^2 = 379.52 - 18.08^2 = 52.6336.$$

12. The joint PMF is

		y	
	$P(x,y)$	1	2
	1	0.132	0.068
x	2	0.24	0.06
	3	0.33	0.17

Thus,

$$E(C) = E(2\sqrt{X}+3Y^2) = \sum_{x=1}^{3}\sum_{y=1}^{2}(2\sqrt{x}+3y^2)p(x,y) = (2\times\sqrt{1}+3\times 1^2)\times 0.132$$
$$+ (2\times\sqrt{1}+3\times 2^2)\times 0.068 + (2\times\sqrt{2}+3\times 1^2)\times 0.24$$
$$+ (2\times\sqrt{2}+3\times 2^2)\times 0.06 + (2\times\sqrt{3}+3\times 1^2)\times 0.33$$
$$+ (2\times\sqrt{3}+3\times 2^2)\times 0.17 = 8.662579,$$

and

$$E(C^2) = E[(2\sqrt{X}+3Y^2)^2] = \sum_{x=1}^{3}\sum_{y=1}^{2}(2\sqrt{x}+3y^2)^2 p(x,y) = (2\times\sqrt{1}+3\times 1^2)^2\times 0.132$$
$$+ (2\times\sqrt{1}+3\times 2^2)^2\times 0.068 + (2\times\sqrt{2}+3\times 1^2)^2\times 0.24$$
$$+ (2\times\sqrt{2}+3\times 2^2)^2\times 0.06 + (2\times\sqrt{3}+3\times 1^2)^2\times 0.33$$
$$+ (2\times\sqrt{3}+3\times 2^2)^2\times 0.17 = 92.41633.$$

Hence, $\text{Var}(C) = E(C^2) - E(C)^2 = 92.41633 - 8.662579^2 = 17.37606.$

13. (a) Since X, Y, and Z are independent uniform(0,1) random variables, we have $\text{Var}(X) = \text{Var}(Y) = \text{Var}(Z) = 1/12$. Thus,

$$\text{Var}(X_1) = \text{Var}(X + Z) = \text{Var}(X) + \text{Var}(Z) = 1/6,$$

$$\text{Var}(Y_1) = \text{Var}(Y + 2Z) = \text{Var}(Y) + 4\text{Var}(Z) = 5/12,$$

and

$$\begin{aligned}\text{Cov}(X_1, Y_1) &= \text{Cov}(X + Z, Y + 2Z) = \text{Cov}(X, Y + 2Z) + \text{Cov}(Z, Y + 2Z)\\ &= 2\text{Cov}(Z, Z) = 2\text{Var}(Z) = 1/6.\end{aligned}$$

Hence,

$$\text{Var}(X_1{+}Y_1) = \text{Var}(X_1){+}\text{Var}(Y_1){+}2\text{Cov}(X_1, Y_1) = 1/6{+}5/12{+}2{\times}1/6 = 11/12,$$

and

$$\text{Var}(X_1{-}Y_1) = \text{Var}(X_1){+}\text{Var}(Y_1){-}2\text{Cov}(X_1, Y_1) = 1/6{+}5/12{-}2{\times}1/6 = 1/4.$$

(b) Use the following commands:

```
set.seed=111; x=runif(10000); y=runif(10000); z=runif(10000);
x1 = x+z; y1 = y+2*z; var(x1 + y1); var(x1 - y1)
```

The commands give the sample variances of a sample of 10,000 $X_1 + Y_1$ values and a sample of 10,000 $X_1 - Y_1$ values. They are 0.923218 and 0.2471397, respectively. These values are close to the calculation in part (a).

14. First we find

$$\begin{aligned}\text{Cov}(\bar{X}, \bar{Y}) &= \text{Cov}\left(\frac{X_1 + X_2 + X_3}{3}, \frac{Y_1 + Y_2 + Y_3}{3}\right) = \frac{1}{9}\sum_{i=1}^{3}\sum_{j=1}^{3}\text{Cov}(X_i, Y_j)\\ &= \frac{1}{9}[\text{Cov}(X_1, Y_1) + \text{Cov}(X_2, Y_2) + \text{Cov}(X_3, Y_3)] = \frac{5}{3},\end{aligned}$$

$\text{Var}(\bar{X}) = \sigma_X^2/3 = 3$ and $\text{Var}(\bar{Y}) = \sigma_Y^2/3 = 4/3$. Thus,

$$\text{Var}(\bar{X} + \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) + 2\text{Cov}(\bar{X}, \bar{Y}) = 3 + \frac{4}{3} + 2\frac{5}{3} = \frac{23}{3}.$$

15. The hypergeometric random variable X with parameters n, M_1, and M_2 can be thought of as a sum of n Bernoulli random variables $X_1, X_2, \cdots, X_n$, each with probability of success $p = M_1/(M_1 + M_2)$, i.e. $X = X_1 + X_2 + \cdots + X_n$. Thus,

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = \frac{nM_1}{M_1 + M_2}.$$

16. From the model, $X_1, X_2, \cdots, X_r$ are independent and all of them have geometric distribution with success probability p. Thus, $E(X_i) = 1/p$ and $\mathrm{Var}(X_i) = (1-p)/p^2$. Since $X = X_1 + \cdots + X_r$, we have

$$E(X) = E(X_1) + \cdots + E(X_r) = \frac{r}{p},$$

and

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_r) = \frac{r(1-p)}{p^2}.$$

4.5 Quantifying Dependence

1. We have the joint PMF and the marginal PMF as

			y		
	$P(x,y)$	0	1	2	$p_X(x)$
	0	0.06	0.04	0.2	0.3
x	1	0.08	0.3	0.06	0.44
	2	0.1	0.14	0.02	0.26
	$p_Y(y)$	0.24	0.48	0.28	

Thus,

$$E(X) = 0 \times 0.3 + 1 \times 0.44 + 2 \times 0.26 = 0.96,$$
$$E(X^2) = 0^2 \times 0.3 + 1^2 \times 0.44 + 2^2 \times 0.26 = 1.48,$$
$$E(Y) = 0 \times 0.24 + 1 \times 0.48 + 2 \times 0.28 = 1.04,$$
$$E(Y^2) = 0^2 \times 0.24 + 1^2 \times 0.48 + 2^2 \times 0.28 = 1.6,$$
$$\sigma_X^2 = E(X^2) - E(X)^2 = 1.48 - 0.96^2 = 0.5584,$$
$$\sigma_Y^2 = E(Y^2) - E(Y)^2 = 1.6 - 1.04^2 = 0.5184,$$
$$E(XY) = 1 \times 0.3 + 1 \times 2 \times 0.06 + 2 \times 1 \times 0.14 + 2 \times 2 \times 0.02 = 0.78,$$

and

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0.78 - 0.96 \times 1.04 = -0.2184.$$

Hence, the linear correlation coefficient of X and Y is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{-0.2184}{\sqrt{0.5584}\sqrt{0.5184}} = -0.4059.$$

2. (a) With more drug administered, we would expect the laboratory rats to develop more tumors, thus X and Y are expected to be positively correlated.

   We have the joint PMF and the marginal PMF as

			y		
	$P(x,y)$	0	1	2	$p_X(x)$
	0	0.388	0.009	0.003	0.4
x	1	0.485	0.01	0.005	0.5
	2	0.09	0.006	0.004	0.1
	$p_Y(y)$	0.963	0.025	0.012	

Thus,

$$E(X) = 0 \times 0.4 + 1 \times 0.5 + 2 \times 0.1 = 0.7,$$

$$E(X^2) = 0^2 \times 0.4 + 1^2 \times 0.5 + 2^2 \times 0.1 = 0.9,$$

$$E(Y) = 0 \times 0.963 + 1 \times 0.025 + 2 \times 0.012 = 0.049,$$

$$E(Y^2) = 0^2 \times 0.963 + 1^2 \times 0.025 + 2^2 \times 0.012 = 0.073,$$

$$\sigma_X^2 = E(X^2) - E(X)^2 = 0.9 - 0.7^2 = 0.41,$$

$$\sigma_Y^2 = E(Y^2) - E(Y)^2 = 0.073 - 0.049^2 = 0.070599,$$

$$E(XY) = 1 \times 0.01 + 1 \times 2 \times 0.005 + 2 \times 1 \times 0.006 + 2 \times 2 \times 0.004 = 0.048,$$

and

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0.048 - 0.7 \times 0.049 = 0.0137.$$

The positive covariance shows that X and Y are positively correlated.

(b) The linear correlation coefficient of X and Y is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{0.0137}{\sqrt{0.41}\sqrt{0.070599}} = 0.0805.$$

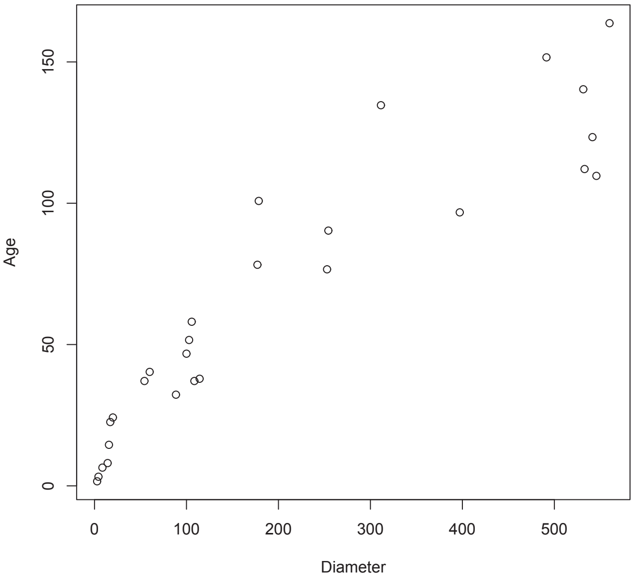
3. (a) Using the R commands

```
x = c(12.8, 12.9, 12.9, 13.6, 14.5, 14.6, 15.1, 17.5, 19.5, 20.8)
y = c(5.5, 6.2, 6.3, 7.0, 7.8, 8.3, 7.1, 10.0, 10.8, 11.0)
var(x); var(y); cov(x,y); cor(x,y),
```

we get $S_X^2 = 8.268$, $S_Y^2 = 3.907$, $S_{X,Y} = 5.46$, and $r_{X,Y} = 0.9607$.

(b) If the distances had been given in inches, S_X^2, S_Y^2, and $S_{X,Y}$ would be changed by a factor of 12^2, but $r_{X,Y}$ would be the same.

4. (a) We would expect the diameter and age to be positively correlated because, intuitively, if a tree is older, it generally has bigger diameter. The scatterplot of the data is given on the next page and it confirms that the diameter and age are positively correlated.



(b) Using the command *cov(x,y); cor(x,y)*, we get the sample covariance and linear correlation of diameter and age to be 9308.47 and 0.9262236, respectively. On the basis of the scatterplot, we can conclude that linear correlation correctly captures the strength of the diameter-age dependence.

5. The commands give the correlation matrix as

	Head.L	Head.W	Neck.G	Chest.G	Weight
Head.L	1.0000000	0.7677513	0.8932822	0.8584959	0.8374185
Head.W	0.7677513	1.0000000	0.8138328	0.8109276	0.8012839
Neck.G	0.8932822	0.8138328	1.0000000	0.9575036	0.9672750
Chest.G	0.8584959	0.8109276	0.9575036	1.0000000	0.9599134
Weight	0.8374185	0.8012839	0.9672750	0.9599134	1.0000000

It is observed that the variables Neck.G and Chest.G have the largest correlations with the variable Weight. Thus, we would say that the variables Neck.G and Chest.G are the two best single predictors of the variable Weight.

6. (a) The marginal distribution of X is Bernoulli(0.3).

(b) If $X = 1$, that is, the first selection is defective, then there are 2 defective and 7 non-defective products left. Thus, $Y|X = 1 \sim$ Bernoulli(2/9). By same reason, $Y|X = 0 \sim$ Bernoulli(3/9).

(c) $p_{X,Y}(1,1) = P(Y = 1|X = 1)P(X = 1) = 2/9{\times}0.3$, $p_{X,Y}(1,0) = P(Y = 0|X = 1)P(X = 1) = 7/9 \times 0.3$, $p_{X,Y}(0,1) = P(Y = 1|X = 0)P(X = 0) = 3/9 \times 0.7$, and $p_{X,Y}(0,0) = P(Y = 0|X = 0)P(X = 0) = 6/9 \times 0.7$.

(d) $p_Y(1) = p_{X,Y}(1,1) + p_{X,Y}(0,1) = 2/9 \times 0.3 + 3/9 \times 0.7 = 0.3$, thus the marginal distribution of Y is Bernoulli(0.3) , which is the same as X.

(e) From the joint distribution of X and Y, we have

$$E(XY) = \sum_{x=0}^{1} \sum_{y=0}^{2} xyp_{X,Y}(x,y) = p_{X,Y}(1,1) = 2/9 \times 0.3.$$

Thus,

$$\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{2}{9} \times 0.3 - 0.3 \times 0.3 = -0.02333,$$

and the linear correlation coefficient is

$$r_{X,Y}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{\frac{2}{9} \times 0.3 - 0.3 \times 0.3}{\sqrt{0.3 \times (1-0.3) \times 0.3 \times (1-0.3)}} = -\frac{1}{9}.$$

7. (a) From the marginal distributions, we calculate

$$E(X) = \int x f_X(x) dx = \int_0^{0.5} 24x^2(1-2x)dx = 3\int_0^1 t^2(1-t)dt = 3\frac{2!1!}{4!} = \frac{1}{4},$$

$$E(X^2) = \int x^2 f_X(x) dx = \int_0^{0.5} 24x^3(1-2x)dx = \frac{3}{2}\int_0^1 t^3(1-t)dt = \frac{3}{2}\frac{3!1!}{5!} = \frac{3}{40},$$

$$E(Y) = \int y f_Y(y) dy = \int_0^1 3y(1-y)^2 dx = 3\frac{1!2!}{4!} = \frac{1}{4},$$

and

$$E(Y^2) = \int y^2 f_Y(y) dy = \int_0^1 3y^2(1-y)^2 dx = 3\frac{2!2!}{5!} = \frac{1}{10}.$$

Thus, $\sigma_X^2 = E(X^2) - E(X)^2 = 3/40 - 1/16 = 1/80$, and $\sigma_Y^2 = E(Y^2) - E(Y)^2 = 1/10 - 1/16 = 3/80$. Further

$$\begin{aligned} E(XY) &= \iint xyf(x,y)dxdy = \int_0^{0.5}\int_0^{1-2x} 24x^2 y dy dx = \int_0^{0.5} 12x^2(1-2x)^2 dx \\ &= \frac{3}{2}\int_0^1 t^2(1-t)^2 dt = \frac{3}{2}\frac{2!2!}{5!} = \frac{1}{20}, \end{aligned}$$

therefore, $\sigma_{X,Y} = E(XY) - E(X)E(Y) = 1/20 - 1/4 \times 1/4 = -1/80$.

(b) The linear correlation coefficient is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{-1/80}{\sqrt{1/80}\sqrt{3/80}} = -\frac{\sqrt{3}}{3}.$$

(c) Given $X = x$, we have the conditional PDF of Y is $f_{Y|X=x}(y) = 0$ if $y \notin [0, 1-2x]$, otherwise

$$f_{Y|X=x}(y) = \frac{f(x,y)}{f_X(x)} = \frac{24x}{24x(1-2x)} = \frac{1}{1-2x}.$$

Therefore, given $X = x$, Y is uniformly distributed on $[0, 1-2x]$, hence the regression function of Y and X is

$$E(Y|X=x) = \frac{1}{2(1-2x)}.$$

The dependence between X and Y is not linear, thus it is not appropriate to use $\rho_{X,Y}$.

8. It is clear that $f(x)$ is an even function on $[-1,1]$, thus $xf(x)$ and $x^3f(x)$ are odd functions on $[-1,1]$. Then $E(X) = \int_{-1}^{1} xf(x)dx = 0$, by the same reason $E(X^3) = 0$. Hence,

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(Y) = 0.$$

Consequently,

$$\rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_x \sigma_Y} = 0.$$

9. (a) It is seen that $f(x)$ is an even function on $[-1,1]$, thus $xf(x)$ and $x^3f(x)$ are odd functions on $[-1,1]$. Then $E(X) = \int_{-1}^{1} xf(x)dx = 0$, by the same reason $E(X^3) = 0$. Hence,

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y) = E(X^3) - E(X)E(Y) = 0.$$

(b) When the value of X is given as x, the value of Y is known as x^2. Thus $E(Y|X=x) = x^2$, without any calculation.

(c) The dependence between X and Y is not linear, thus it is not appropriate to use $\rho_{X,Y}$.

4.6 Models for Joint Distributions

1. (a) From the model, we know that given $P = p$, the distribution of Y is $\text{Bin}(n,p)$. Thus the joint PMF of P and Y is

$$p_{P,Y}(p,y) = P(Y=y|P=p)P(P=p) = \binom{n}{y} p^y (1-p)^{n-y} P(P=p).$$

In detail, for $y = 0, 1, \cdots, n$,

$$p_{P,Y}(0.6, y) = 0.2 \binom{n}{y} 0.6^y 0.4^{n-y},$$

$$p_{P,Y}(0.8, y) = 0.5\binom{n}{y}0.8^y 0.2^{n-y},$$

and

$$p_{P,Y}(0.9, y) = 0.3\binom{n}{y}0.9^y 0.1^{n-y}.$$

(b) We have the general formula $p_Y(y) = p_{P,Y}(0.6, y) + p_{P,Y}(0.8, y) + p_{P,Y}(0.9, y)$. Thus, when $n = 3$,

$$p_Y(0) = 0.2\binom{3}{0}0.6^0 0.4^3 + 0.5\binom{3}{0}0.8^0 0.2^3 + 0.3\binom{3}{0}0.9^0 0.1^3 = 0.0171,$$

$$p_Y(1) = 0.2\binom{3}{1}0.6^1 0.4^2 + 0.5\binom{3}{1}0.8^1 0.2^2 + 0.3\binom{3}{1}0.9^1 0.1^2 = 0.1137,$$

$$p_Y(2) = 0.2\binom{3}{2}0.6^2 0.4^1 + 0.5\binom{3}{2}0.8^2 0.2^1 + 0.3\binom{3}{2}0.9^2 0.1^1 = 0.3513,$$

$$p_Y(3) = 0.2\binom{3}{3}0.6^3 0.4^0 + 0.5\binom{3}{3}0.8^3 0.2^0 + 0.3\binom{3}{3}0.9^3 0.1^0 = 0.5179.$$

2. (a) From the model, we know that given $P = p$, the distribution of Y is $\text{Bin}(n, p)$. Thus the joint density of P and Y is the conditional PMF of Y given $P = p$ times the marginal PDF of P, that is,

$$f_{P,Y}(p, y) = P(Y = y|P = p)f_P(p) = \binom{n}{y}p^y(1-p)^{n-y}.$$

(b) The marginal PMF of Y is given by

$$p_Y(y) = \int_0^1 f_{P,Y}(p, y)dp = \int_0^1 \binom{n}{y}p^y(1-p)^{n-y}dp = \frac{1}{n+1}.$$

3. We have $Y_1 \sim N(9.3 + 1.5 \times 20, 16) = N(39.3, 16)$ and $Y_2 \sim N(9.3 + 1.5 \times 25, 16) = N(46.8, 16)$.

(a) The 95th percentile of Y_1 can be found by *qnorm(0.95, 39.3, 4)*, which gives 45.879.

(b) Since Y_1 and Y_2 are independent, $Y_2 - Y_1 \sim N(46.8-39.3, 16+16) = N(7.5, 32)$. Thus $P(Y_2 > Y_1) = P(Y_2 - Y_1 > 0)$, which can be found by the command *1 - pnorm(0, 7.5, sqrt(32))*, and it gives 0.9076.

4. (a) From the regression model given, we have $E(Y|X = x) = 9.3 + 1.5x$, thus the marginal mean of Y is $E(Y) = 9.3 + 1.5\mu_X = 9.3 + 1.5 \times 24 = 45.3$.

(b) From the model, we have $\beta_1 = 1.5$. Thus, by (4.6.8), the covariance of X and Y is $\sigma_{X,Y} = \beta_1 \sigma_X^2 = 1.5 \times 9 = 13.5$. The correlation coefficient is

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{13.5}{3 \times \sqrt{36.25}} = 0.7474.$$

5. (a) Since the marginal distribution of X is normal, the conditional distribution of Y given $X = x$ is also normal, the joint PDF of X and Y is bivariate normal distribution with parameters $\mu_X = 24$, $\mu_Y = 45.3$, $\sigma_X^2 = 9$, $\sigma_Y^2 = 36.25$, and $\sigma_{X,Y} = 13.5$.

   (b) Use the command *pmnorm(c(25, 45), c(24, 45.3), matrix(c(9, 13.5, 13.5, 36.25), 2))*, and it gives the probability as 0.42612.

6. By the second mean plus error expression in (4.6.6), $Y = \beta_0 + \beta_1(X - \mu_X) + \epsilon$. Because any (intrinsic) error variable has mean value zero, that is, $E(\epsilon) = 0$, we have

$$E(Y) = E(\beta_0 + \beta_1(X - \mu_X) + \epsilon) = \beta_0 + \beta_1(E(X) - \mu_X) + E(\epsilon) = \beta_0 + \beta_1(\mu_X - \mu_X) = \beta_0.$$

7. In the exponential regression, the conditional distribution of Y given $X = x$ is given as

$$f_{Y|X=x} = \lambda(x) e^{-\lambda(x) y}.$$

Hence, the regression function of Y on X is

$$E(Y|X = x) = \frac{1}{\lambda(x)}.$$

   (a) Given $X = x$,

$$E(Y|X = x) = \frac{1}{\lambda(x)} = \frac{1}{\exp(\alpha + \beta x)} = \frac{1}{\exp(4.2 + 3.1x)}.$$

   X has a uniform distribution on (2, 6), thus $f(x) = 1/4$ on (2, 6) and zero otherwise. By the Law of Total Expectation,

$$\begin{aligned} E(Y) &= \int E(Y|X = x) f(x) dx = \int_2^6 \frac{1}{4} \frac{1}{\exp(4.2 + 3.1x)} dx \\ &= \frac{1}{4} e^{-4.2} \frac{-1}{3.1}\, e^{-3.1x}\Big|_2^6 = 2.454222 \times 10^{-6}. \end{aligned}$$

   (b) By the analysis above, the joint PDF of X and Y is $f_{X,Y}(x, y) = 0.25\lambda(x) e^{-\lambda(x) y}$, with $2 \le x \le 6$ and $y > 0$ and $f_{X,Y}(x, y) = 0$ otherwise.

8. (a) The marginal distribution of N_4 is Binomial with parameters $n = 16$ and $p = 0.02$. Thus, the probability that exactly one of the 16 planned chemical reactions will not be performed due to unusable raw materials is $P(N_4 = 1)$, which can be calculated by the command *dbinom(1, 16, 0.02)*, and it gives 0.2363421.

(b) The probability that 10 chemical reactions will be performed with recent materials, 4 with moderately aged materials and 2 with aged materials, is $P(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0)$, and it is calculated as

$$(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0) = \binom{16}{10,4,2,0} 0.6^{10} 0.3^4 0.08^2 0.02^0 = 0.03765241.$$

(c) The R command is *dmultinom(c(10, 4, 2, 0), prob=c(0.6,0.3,0.08,0.02))* and it gives exactly the same answer as in part (b).

(d) The covariance can be calculated as

$$\begin{aligned}\mathrm{Cov}(N_1 + N_2, N_3) &= \mathrm{Cov}(N_1, N_3) + \mathrm{Cov}(N_2, N_3) = -np_1p_3 - np_2p_3 \\ &= -16 \times 0.6 \times 0.08 - 16 \times 0.3 \times 0.08 = -1.152.\end{aligned}$$

In general, because we have fixed number of total items, if we have more items being aged, we would expect that to have fewer recent and moderately aged items. Thus, it is reasonable for the covariance to be negative.

(e) The random variable $N_1 + N_2 + N_3$ has Binomial distribution with $n = 16$ and $p = p_1 + p_2 + p_3 = 0.98$. Thus, $\mathrm{Var}(N_1 + N_2 + N_3) = np(1 - p) = 16 \times 0.98 \times 0.02 = 0.3136$.

9. (a) The marginal distribution of N_3 is Binomial with parameter $n = 15$ and $p = p_3 = 0.54$. Thus, the probability that exactly 10 children use a child seat is $P(N_3 = 10)$, which can be calculated by *dbinom(10, 15, 0.54)*, and it gives 0.1304.

(b) The probability that exactly 10 children use a child seat and five use a seat belt is $P(N_1 = 0, N_2 = 5, N_3 = 10)$, and the R command is *dmultinom(c(0,5,10), prob=c(0.17,0.29,0.54))*, which gives 0.01298622.

(c) The random variable $N_2 + N_3$ has binomial distribution with $n = 15$ and $p = p_2 + p_3 = 0.83$, thus $\mathrm{Var}(N_2 + N_3) = np(1-p) = 15 \times 0.83 \times 0.17 = 2.1165$, and

$$\begin{aligned}\mathrm{Cov}(N_1, N_2 + N_3) &= \mathrm{Cov}(N_1, N_2) + \mathrm{Cov}(N_1, N_3) = -np_1p_2 - np_1p_3 \\ &= -15 \times 0.17 \times 0.29 - 15 \times 0.17 \times 0.54 = -2.1165.\end{aligned}$$

Alternatively,

$$\begin{aligned}\mathrm{Cov}(N_1, N_2 + N_3) &= \mathrm{Cov}(n - (N_2 + N_3), N_2 + N_3) = -\mathrm{Cov}((N_2 + N_3), N_2 + N_3) \\ &= -\mathrm{Var}(N_2 + N_3) = -2.1165.\end{aligned}$$

10. (a) From (4.6.7), we have $\sigma_Y^2 - \sigma_\epsilon^2 = \beta_1^2 \sigma_X^2$. Combining this with (4.6.8), there is

$$\rho^2 = \frac{\beta_1^2 \sigma_X^2}{\sigma_Y^2} = \frac{\sigma_Y^2 - \sigma_\epsilon^2}{\sigma_Y^2} = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2},$$

that is, $1 - \rho^2 = \sigma_\epsilon^2 / \sigma_Y^2$.

(b) By Proposition 4.6-1, $\beta_0 = \mu_Y$, and from part (a), $\sigma_\epsilon = \sqrt{1-\rho^2}\sigma_Y$. Then (4.6.3) is

$$\begin{aligned}
f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma_\epsilon\sigma_X}\exp\left\{-\frac{(y-\beta_0-\beta_1(x-\mu_X))^2}{2\sigma_\epsilon^2}-\frac{(x-\mu_X)^2}{2\sigma_X^2}\right\}\\
&= \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_Y\sigma_X}\exp\left\{-\frac{(y-\mu_Y-\beta_1\tilde{x})^2}{2(1-\rho^2)\sigma_Y^2}-\frac{\tilde{x}^2}{2\sigma_X^2}\right\}\\
&\quad \text{(Plug in } \beta_0=\mu_Y \text{ and } \sigma_\epsilon=\sqrt{1-\rho^2}\sigma_Y\text{)}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{(\tilde{y}-\beta_1\tilde{x})^2}{2(1-\rho^2)\sigma_Y^2}-\frac{\tilde{x}^2}{2\sigma_X^2}\right\}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{\tilde{y}^2-2\beta_1\tilde{x}\tilde{y}+\beta_1^2\tilde{x}^2}{2(1-\rho^2)\sigma_Y^2}-\frac{\tilde{x}^2}{2\sigma_X^2}\right\}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{\tilde{y}^2}{2(1-\rho^2)\sigma_Y^2}+2\beta_1\frac{\tilde{x}\tilde{y}}{2(1-\rho^2)\sigma_Y^2}\right.\\
&\qquad\left.-\frac{\beta_1^2\sigma_X^2\tilde{x}^2}{2(1-\rho^2)\sigma_Y^2\sigma_X^2}-\frac{\tilde{x}^2(1-\rho^2)\sigma_Y^2}{2\sigma_X^2(1-\rho^2)\sigma_Y^2}\right\}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{\tilde{y}^2}{2(1-\rho^2)\sigma_Y^2}+\rho\frac{\sigma_Y}{\sigma_X}\frac{\tilde{x}\tilde{y}}{(1-\rho^2)\sigma_Y^2}\right.\\
&\qquad\left.-\frac{\rho^2\sigma_Y^2\tilde{x}^2}{2(1-\rho^2)\sigma_Y^2\sigma_X^2}-\frac{\tilde{x}^2(1-\rho^2)\sigma_Y^2}{2\sigma_X^2(1-\rho^2)\sigma_Y^2}\right\}\\
&\quad \text{(Plug in the relation } \beta_1\sigma_X=\rho\sigma_Y \text{ from Proposition 4.6-3)}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{-\frac{\tilde{y}^2}{2(1-\rho^2)\sigma_Y^2}+\frac{\rho\tilde{x}\tilde{y}}{(1-\rho^2)\sigma_X\sigma_Y}\right.\\
&\qquad\left.-\frac{\sigma_Y^2\tilde{x}^2}{2(1-\rho^2)\sigma_Y^2\sigma_X^2}\right\}\\
&= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\exp\left\{\frac{-1}{1-\rho^2}\left[\frac{\tilde{x}^2}{2\sigma_X^2}-\frac{\rho\tilde{x}\tilde{y}}{\sigma_X\sigma_Y}+\frac{\tilde{y}^2}{2\sigma_Y^2}\right]\right\},
\end{aligned}$$

which is (4.6.13), as to be proved.

(c) In (4.6.14),

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_Y^2 \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix},$$

since $\rho = \sigma_{X,Y}/(\sigma_X\sigma_Y)$. It is easy to calculate that the determinant of Σ is

$$|\Sigma| = \sigma_X^2\sigma_Y^2 - (\rho\sigma_X\sigma_Y)^2 = (1-\rho^2)\sigma_X^2\sigma_Y^2,$$

and the inverse of Σ is

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix} = \frac{1}{(1-\rho^2)\sigma_X^2\sigma_Y^2}\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}.$$

Thus,

$$\begin{aligned}
\frac{1}{2}(x-\mu_X, y-\mu_Y)\Sigma^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix} &= \frac{1}{2(1-\rho^2)\sigma_X^2\sigma_Y^2}(\tilde{x},\tilde{y})\begin{pmatrix} \sigma_Y^2 & -\rho\sigma_X\sigma_Y \\ -\rho\sigma_X\sigma_Y & \sigma_X^2 \end{pmatrix}\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \\
&= \frac{(\sigma_Y^2\tilde{x}-\rho\sigma_X\sigma_Y\tilde{y}, \sigma_X^2\tilde{y}-\rho\sigma_X\sigma_Y\tilde{x})}{2(1-\rho^2)\sigma_X^2\sigma_Y^2}\begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \\
&= \frac{\sigma_Y^2\tilde{x}^2-2\rho\sigma_X\sigma_Y\tilde{y}\tilde{x}+\sigma_X^2\tilde{y}^2}{2(1-\rho^2)\sigma_X^2\sigma_Y^2} \\
&= \frac{1}{1-\rho^2}\left[\frac{\tilde{x}^2}{2\sigma_X^2}-\frac{\rho\tilde{x}\tilde{y}}{\sigma_X\sigma_Y}+\frac{\tilde{y}^2}{2\sigma_Y^2}\right].
\end{aligned}$$

Hence,

$$\begin{aligned}
\frac{1}{2\pi\sqrt{|\Sigma|}}\exp\left\{-\frac{1}{2}(x-\mu_X, y-\mu_Y)\Sigma^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\right\} &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
&\times\exp\left\{\frac{-1}{1-\rho^2}\left[\frac{\tilde{x}^2}{2\sigma_X^2}-\frac{\rho\tilde{x}\tilde{y}}{\sigma_X\sigma_Y}+\frac{\tilde{y}^2}{2\sigma_Y^2}\right]\right\}.
\end{aligned}$$

Chapter 5
Some Approximation Results

5.2 The LLN and the Consistency of Averages

1. (a) The proof is

$$P(|X-\mu| > a\sigma) \leq \frac{\text{Var}(X)}{(a\sigma)^2} = \frac{\sigma^2}{(a\sigma)^2} = \frac{1}{a^2}.$$

(b) Using the inequality in part (a), we get the upper bounds of 1, 0.25, and 0.11, respectively. To calculate the exact probability, since $X \sim N(\mu, \sigma^2)$, we have

$$\begin{aligned} P(|X-\mu| > a\sigma) &= P\left(\frac{|X-\mu|}{\sigma} > a\right) = P(|Z| > a) = P(Z < -a) + P(Z > a) \\ &= \Phi(-a) + 1 - \Phi(a) = 2\Phi(-a), \end{aligned}$$

where $Z \sim N(0,1)$ and $\Phi(\cdot)$ is the CDF of $N(0,1)$. Using R commands, we can calculate the exact value when $a = 1$, 2, and 3, as 0.3173105, 0.04550026, and 0.002699796. The upper bounds are much worse.

2. (a) By the LLN, $\bar{X}$ should be approximately equal to the expected life span of a randomly selected component, which is

$$E(X) = \frac{1}{\lambda} = 1/0.013 = 76.92,$$

since the mean of a random variable having exponential distribution with parameter λ is $1/\lambda$.

(b) The problem asks the probability that

$$P(|\bar{X}-\mu| < 15.38) = 1 - P(|\bar{X}-\mu| \geq 15.38).$$

By Chebyshev's inequality

$$P(|\bar{X}-\mu| \geq 15.38) \leq \frac{\text{Var}(\bar{X})}{15.38^2} = \frac{\frac{\text{Var}(X)}{n}}{15.38^2} = \frac{\frac{1}{100}\frac{1}{\lambda^2}}{15.38^2} = 0.25.$$

Thus, $P(|\bar{X}-\mu| < 15.38) = 1 - P(|\bar{X}-\mu| \geq 15.38) \geq 1 - 0.25 = 0.75$. We can say that the probability that $\bar{X}$ will be within 15.38 units from the population mean is at least 0.75.

3. (a) Let X be a Poisson random variable with mean 1, that is $\lambda = 1$, then $\mathrm{Var}(X) = 1$. Then $E(\bar{X}) = \lambda = 1$, and $\mathrm{Var}(\bar{X}) = \lambda/n = 1/10 = 0.1$. Thus, the calculation is

$$\begin{aligned} P(0.5 \leq \bar{X} \leq 1.5) &= P(-0.5 \leq \bar{X} - 1 \leq 0.5) = P(|\bar{X} - 1| \leq 0.5) \\ &= 1 - P(|\bar{X} - 1| > 0.5) \geq 1 - \frac{\mathrm{Var}(\bar{X})}{0.5^2} = 0.6. \end{aligned}$$

(b) Using the fact that $Y = \sum_{i=1}^{10} X_i$ is a Poisson random variable with mean 10, we have

$$P(0.5 \leq \bar{X} \leq 1.5) = P(5 \leq \sum_{i=1}^{10} X_i \leq 15) = P(Y \leq 15) - P(Y \leq 4) = 0.922,$$

which is calculated by the R command *ppois(15, 10)-ppois(4, 10)*.

5.3 Convolutions

1. (a) Given $X = k$, we have $Z = X + Y = k + Y$, thus the sample space of Z is $\{k, k+1, \cdots, k+n_2\}$. Hence, for z in the sample space

$$\begin{aligned} P(Z = z|X = k) &= P(Y = z - k|X = k) = P(Y = z - k) \\ &= \binom{n_2}{z-k} p^{z-k}(1-p)^{n_2-(z-k)}. \end{aligned}$$

(b) Since $Z = X + Y$, the sample space of Z is $\{1, 2, \cdots, n_1 + n_2\}$. Hence, for z in the sample space, by the total probability formula

$$\begin{aligned} P(Z = z) &= \sum_{k=0}^{n_1} P(Z = z|X = k)P(X = K) \\ &= \sum_{k=0}^{n_1} \binom{n_2}{z-k} p^{z-k}(1-p)^{n_2-(z-k)} \binom{n_1}{k} p^k(1-p)^{n_1-k} \\ &= p^z(1-p)^{n_1+n_2-z} \sum_{k=0}^{n_1} \binom{n_2}{z-k}\binom{n_1}{k} \\ &= \binom{n_1+n_2}{z} p^z(1-p)^{n_1+n_2-z}, \end{aligned}$$

which shows that $Z \sim \mathrm{Bin}(n_1 + n_2, p)$.

2. Let $Y = X_1 + X_2$, since X_1 and X_2 are both exponential random variables, the sample space of Y is $(0, \infty)$, thus $f_Y(y) = 0$ for $y < 0$. For $y > 0$, we have

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1 = \int_0^y f_{X_2}(y - x_1) f_{X_1}(x_1) dx_1 \\ &= \int_0^y \lambda \exp(-\lambda(y - x_1)) \lambda \exp(-\lambda x_1) dx_1 = \lambda^2 y \exp(-\lambda y). \end{aligned}$$

3. (a) By Proposition 5.3-1, the distribution of $X_1 + X_2 + X_3$ is $N(\mu_{X_1} + \mu_{X_2} + \mu_{X_3}, \sigma^2_{X_1} + \sigma^2_{X_2} + \sigma^2_{X_3}) = N(180, 36)$. Thus, $P(X_1 + X_2 + X_3 > 185)$ can be found by *1 - pnorm(185, 180, 6)*, which gives 0.2023.

   (b) By Corollary 5.3-1, $\bar{X} \sim N(\mu_1, \sigma_1^2/3) = N(60, 4)$, and $\bar{X} \sim N(\mu_2, \sigma_2^2/3) = N(65, 5)$. Since X_i and Y_j are independent for all i and j, $\bar{X}$ and $\bar{Y}$ are independent. Apply Proposition 5.3-1 again, we have $\bar{Y} - \bar{X} \sim N(5, 9)$. Thus, $P(\bar{Y} - \bar{X} > 8)$ can be found by *1 - pnorm(8, 5, 3)*, which gives 0.1587.

4. (a) The total duration is $X_1 + X_2 + X_3$ which follows $N(\mu_{X_1} + \mu_{X_2} + \mu_{X_3}, \sigma^2_{X_1} + \sigma^2_{X_2} + \sigma^2_{X_3}) = N(21, 9)$, by Proposition 5.3-1. Thus, the 95th percentile of the total duration can be found by the command *qnorm(0.95, 21, 3)*, which gives 25.93 hours.

   (b) This problem is asking for $P(X_1 + X_2 + X_3 < 25)$, which can be found by the command *pnorm(25, 21, 3)*, which gives 0.9088.

   (c) Let p be the probability that the flashlight will last more than 25 hours, then using part (b), $p = 1 - 0.9088 = 0.0912$. Let Y be the number of trips that the batteries will last more than 25 hours, then $Y \sim \text{Bin}(5, p)$. The problem is asking for $P(Y = 3)$, which can be found by the command *dbinom(3, 5, 0.0912)*, which gives 0.0063.

5. In probabilistic notation, we want to determine the sample size n so that $P(|\bar{X} - \mu| < 0.005) = 0.95$. According to Corollary 5.3-1,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

   Thus, we can write the desired probability as

$$P(|\bar{X} - \mu| < 0.005) = P\left(\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| < \frac{0.005}{\sigma/\sqrt{n}}\right) = P\left(|Z| < \frac{0.005}{\sigma/\sqrt{n}}\right) = 0.95,$$

   where $Z \sim N(0, 1)$. Hence, $\frac{0.005}{\sigma/\sqrt{n}} = z_{0.025} = 1.96$, solving for n gives us

$$n = \left(\frac{1.96\sigma}{0.005}\right)^2 = 138.30.$$

   Finally, we choose to use $n = 139$.

5.4 The Central Limit Theorem

1. (a) In all 10 repetitions, the smallest and largest order statistics are outliers. The following lists the results of 10 runs:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2006.0000	-1.2910	-0.0498	-8.2870	0.8515	95.2400
-2635.0000	-0.7402	0.1411	-5.1150	1.0730	83.3200
-322.3000	-0.6833	0.1635	-0.9070	1.2130	127.6000
-509.9000	-0.8157	0.0690	-0.5815	1.1880	267.1000
-990.6000	-0.9281	0.0867	-1.5660	1.0510	111.4000
-889.2000	-0.7637	0.1762	-0.5776	0.9609	301.9000
-88.95000	-1.09000	-0.01795	-0.68860	0.96320	29.37000
-128.40000	-0.97310	0.08502	0.52190	1.14600	189.80000
-1191.0000	-1.0610	0.0103	-3.1000	1.0710	651.3000
-384.5000	-0.9337	0.0127	-0.3989	1.0870	130.1000

(b) The commands are repeated 10 times and the results are as follows:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2913.0000	-1.0700	0.0694	-5.3770	0.9376	440.3000
-301.4000	-1.1450	-0.0652	0.4325	1.0480	659.3000
-233.50000	-1.11500	-0.09916	-0.47300	0.89380	118.60000
-468.2000	-0.9738	0.0045	2.6000	0.8916	1374.0000
-648.3000	-0.9662	0.0563	-1.9120	0.9866	64.4600
-1444.0000	-1.1050	-0.0946	-3.2830	1.0250	102.3000
-67.04000	-0.95250	-0.02498	0.61270	0.97650	190.80000
-187.3000	-1.0690	-0.0209	5.8330	0.9322	2874.0000
-217.2000	-0.7372	0.1418	0.3848	1.2980	125.8000
-221.4000	-0.8928	0.0006	1.5640	1.0280	335.0000
-11040.000	-0.830	0.033	-21.480	0.986	235.200

As in part (a), the distribution of the averages seems to have several outliers.

2. Since $X_1, \cdots, X_{30}$ are independent Poisson random variables having mean 1. By CLT, $X_1 + \cdots + X_{30} \sim N(30, 30)$, thus $P(X_1 + \cdots + X_{30} \leq 35)$ can be found by *pnorm(35, 30, sqrt(30))* if no continuity correction is used and this gives 0.8193; if using continuity correction, the command is *pnorm(35.5, 30, sqrt(30))*, which gives 0.8423.

(b) By the property of Poisson random variable, $X_1 + \cdots + X_{30}$ has Poisson distribution with mean 30. Thus, $P(X_1 + \cdots + X_{30} \leq 35)$ can be found by *ppois(35,30)* which gives 0.8426 as the exact value. Clearly, the approximation with continuity correction gives a more accurate value.

3. Since mean and variance of a uniform(0, 10) distribution are 5 and 100/12 respectively, according to CLT, the total waiting time S of 120 times has the distribution

$$S \sim N\left(5 \times 120, 120 \times \frac{100}{12}\right) = N(600, 1000).$$

Thus, the 95th percentile of the total waiting time can be found by the R command *qnorm(0.95, 600, sqrt(1000))*, which gives 652.0148.

4. (a) The gamma distribution with parameters α and β has mean $\alpha\beta$ and variance $\alpha\beta^2$. Thus, by the CLT,

$$\bar{X_1} \sim N\left(\alpha_1\beta_1, \frac{\alpha_1\beta_1^2}{n_1}\right) = N(2 \times 2, 2 \times 2^2/36) = N(4, 0.2222),$$

and

$$\bar{X_2} \sim N\left(\alpha_2\beta_2, \frac{\alpha_2\beta_2^2}{n_2}\right) = N(1 \times 3, 1 \times 3^2/42) = N(3, 0.2143).$$

The two types of materials are independent, thus $\bar{X_1}$ and $\bar{X_2}$ are independent. By the property of normal distribution

$$\bar{X_1} - \bar{X_2} \sim N(4 - 3, 0.2222 + 0.2143) = N(1, 0.4365).$$

(b) $P(\bar{X_1} > \bar{X_2}) = P(\bar{X_1} - \bar{X_2} > 0) = 0.9349$, by the command *1-pnorm(0, 1, sqrt(0.4365))*.

5. Let S_1 and S_2 be the total height of tower 1 and 2, respectively, and let X be the height of a randomly selected segment. Then $X \sim \text{uniform}(35.5, 36.5)$, hence $\mu = 36$ and $\sigma^2 = (36.5 - 35.5)^2/12 = 1/12$.

Clearly, $S_1 = X_1 + X_2 + \cdots + X_{30}$, where $X_1, X_2, \cdots, X_{30}$ are the heights of the randomly selected 30 segments. By CLT, $S_1 \sim N(30\mu, 30\sigma^2)$. By the same reason, $S_2 \sim N(30\mu, 30\sigma^2)$. Since the segments for tower 1 and 2 are independent, we have $S_1 - S_2 \sim N(0, 2 \times 30\sigma^2) = N(0, 5)$.

The roadway can be laid when $|S_1 - S_2| < 4$. This probability is

$$P(|S_1 - S_2| < 4) = P(S_1 - S_2 < 4) - P(S_1 - S_2 < -4) = 0.9264,$$

which is calculated by the command *pnorm(4, 0, sqrt(5)) - pnorm(-4, 0, sqrt(5))*. This probability is approximated.

6. The mean and variance of the tip from a random customer are 1.8175 and 0.1154, respectively. Let S be the total tips from the 70 customers then, by CLT, $S \sim N(n\mu, n\sigma^2) = N(70 \times 1.8175, 70 \times 0.1154)$. Thus, the probability for her tips to exceed \$120 is $P(S > 120) = 0.9945$ by the command *1 - pnorm(120, 70*1.8175, sqrt(70*0.1154))*.

7. Let R be the round-off error, then R has a uniform distribution on $(-0.5, 0.5)$. Thus, the mean value and variance of R are $\mu = 0$ and $\sigma^2 = (0.5 - (-0.5))^2/12 = 1/12$. Let $R_1, \cdots, R_{50}$ be the round-off errors of the 50 numbers. By CLT, the average round-off error $\bar{R}$ has a normal distribution $\bar{R} \sim N(\mu, \sigma^2/n) = N(0, 1/600)$. The event that the resulting average differs from the exact average of the 50 numbers by more than 0.1 happens if $|\bar{R}| > 0.1$, thus the corresponding probability is

$$P(|\bar{R}| > 0.1) = P(\bar{R} > 0.1) + P(\bar{R} < -0.1) = 2P(\bar{R} < -0.1) = 0.0143,$$

which is calculated by the command *2*pnorm(-0.1, 0, sqrt(1/600))*.

8. let $T = X_1 + \cdots + X_n$ be the combined duration of n components, we want $P(T > 3000) = 0.95$. By the CLT, $T \sim N(n\mu, n\sigma^2) = N(100n, 900n)$. Thus,

$$\frac{T - 100n}{30\sqrt{n}} \sim N(0, 1).$$

Then,

$$P(T > 3000) = P\left(\frac{T - 100n}{30\sqrt{n}} > \frac{3000 - 100n}{30\sqrt{n}}\right) = P\left(Z > \frac{3000 - 100n}{30\sqrt{n}}\right) = 0.95.$$

Thus, $\frac{3000-100n}{30\sqrt{n}}$ is the 5th percentile of $N(0, 1)$, which is -1.645, i.e.

$$\frac{3000 - 100n}{30\sqrt{n}} = -1.645,$$

or

$$10n - 3 \times 1.645\sqrt{n} - 300 = 0.$$

Using the command *polyroot(c(-300, -3*1.645, 10))* gives us $\sqrt{n} = 5.73$, thus $n = 5.73^2 = 32.8$. Finally, we should take $n = 33$.

9. (a) Let Y be the coating thickness, then $Y = X_1 + X_2 + \cdots + X_{36}$, where $X_1, \cdots, X_{36}$ are the thickness of the layers, and they are independent, have the same distribution with mean $\mu = 0.5$ and variance $\sigma^2 = 0.04$. Since $n = 36 > 30$, we have approximately

$$Y \sim N(n\mu, n\sigma^2) = N(36 \times 0.5, 36 \times 0.04) = N(18, 1.44).$$

   (b) The proportion is $P(Y < 16)$, which can be calculated using the R command *pnorm(16, 18, 1.2)* and the result is 0.0478.

10. (a) Let $\bar{X}$ be the average diameters of 100 rods, according to CLT, $\bar{X} \sim N(\mu, \sigma^2/n) = N(0.503, 0.03^2/100)$. The probability that the batch passes the inspection is $P(0.495 < \bar{X} < 0.505) = 0.7437$, which is given by the command *pnorm(0.505, 0.503, 0.003) - pnorm(0.495, 0.503, 0.003)*.

(b)

(i) X has a binomial distribution with $n = 40$ and $p = 0.7437$. The exact value for $P(X \leq 30)$ can be calculated by the command *pbinom(30, 40, 0.7437)* and the result is 0.5963.

(ii) By the DeMoivre-Laplace Theorem, X approximately follows $N(np, np(1-p) = N(40 \times 0.7437, 40 \times 0.7437 \times (1 - 0.7437))$. Without the continuity correction, the command is *pnorm(30, 40*0.7437, sqrt(40*0.7437*(1-0.7437)))*, which gives 0.5364; with the continuity correction, the command is *pnorm(30.5, 40*0.7437, sqrt(40*0.7437*(1-0.7437)))*, which gives 0.6073. The method with continuity correction gives more accurate result.

11. (a) X has a binomial distribution with $n = 500$ and $p = 0.6$. The exact value for $P(270 \leq X \leq 320) = P(X \leq 320) - P(X \leq 269)$ can be calculated by the command *pbinom(320, 500, 0.6)-pbinom(269, 500, 0.6)* and the result is 0.9671.

(b) By the DeMoivre-Laplace Theorem, X approximately follows $N(np, np(1-p) = N(500 \times 0.6, 500 \times 0.6 \times (1 - 0.6)) = N(300, 120)$. Without the continuity correction, the command is *pnorm(320, 300, sqrt(120))-pnorm(270, 300, sqrt(120))*, which gives 0.9630; with the continuity correction, the command is *pnorm(320.5, 300, sqrt(120))-pnorm(269.5, 300, sqrt(120))*, which gives 0.9667. The method with continuity correction gives a more accurate result.

12. Let X be the thickness of a randomly selected tire, then $X \sim (10, 4)$, and let p be the rejection probability for a randomly selected tire, then $p = P(X < 7.9)$, which can be calculated by *pnorm(7.9, 10, 2)*. Let Y be the number of rejections among the 100 tires, then $Y \sim \text{Bin}(100, p)$, and we want to calculate $P(Y \leq 10)$. By the DeMoivre-Laplace Theorem, Y approximately follows $N(100p, 100p(1-p))$. To calculate the probability, we apply continuity correction and the R command is *pnorm(10.5, 100*pnorm(7.9, 10, 2), sqrt(100*pnorm(7.9, 10, 2)*(1-pnorm(7.9, 10, 2))))*, which gives 0.1185 as the probability.

13. (a) Let X_1 and X_2 be the number of defective items found in line A and B, respectively. Then, $X_1 \sim \text{Bin}(200, 0.1)$ and $X_2 \sim \text{Bin}(1000, 0.01)$. By the DeMoivre-Laplace Theorem, X_1 approximately follows $N(200 \times 0.1, 200 \times 0.1 \times 0.9) = N(20, 18)$, and X_2 approximately follows $N(1000 \times 0.01, 1000 \times 0.01 \times 0.99) = N(10, 9.9)$. By the independence of the two lines, we have $X_1 + X_2$ approximately follows $N(20 + 10, 18 + 9.9) = N(30, 27.9)$. The problem is to calculate $P(X_1 + X_2) \leq 35$. Using continuity correction, the R command is *pnorm(35.5, 30, sqrt(27.9))*, which gives 0.8511.

(b) The commands are listed as

```
S = expand.grid(X=0:200,Y=0:1000)
```

```
P = expand.grid(px=dbinom(0:200, 200, .1), py=dbinom(0:1000, 1000, .01))
  P$pxy = P$px*P$py; attach(P); attach(S); sum(pxy[which(X+Y<=35)])
```

The exact probability is given as 0.8509.

Chapter 6
Fitting Models to Data

6.2 Some Estimation Concepts

1. Using the commands *OZ = read.table("OzoneData.txt", header=T)* to read data and using *mean(OZ$OzoneData)* to get the sample mean, we get 286.3571. The command *sd(OZ$OzoneData)/sqrt(14)* gets an estimated standard error of 17.07244.

2. The difference of the average maximum penetration between the two types is estimated as $0.49 - 0.36 = 0.13$ and the estimated standard error of $\bar{X}_1 - \bar{X}_2$ is calculated as

$$S_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{0.19^2}{48} + \frac{0.16^2}{42}} = 0.0369.$$

3. The proof is straightforward:

$$\begin{aligned} E(\hat{\sigma^2}) = E\left[\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\right] &= \frac{(n_1-1)E(S_1^2) + (n_2-1)E(S_2^2)}{n_1+n_2-2} \\ = \frac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2}{n_1+n_2-2} &= \sigma^2. \end{aligned}$$

4. (a) The parameter of interest is the proportion of all credit card customers who had incurred an interest charge in the previous year due to an unpaid balance. The empirical estimator is the proportion in a sample of credit card customers who had incurred an interest charge in the previous year due to an unpaid balance. Using the provided information, we can get the estimate as $\hat{p} = 136/200 = 0.68$.

   (b) Yes, it is unbiased.

   (c) The estimated standard error is

$$S_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.68 \times (1-0.68)}{200}} = 0.033.$$

5. The standard error is $S_{\hat{\theta}} = S_{2\bar{X}} = 2S_{\bar{X}} = 2\theta/\sqrt{12n}$. $\hat{\theta}$ is unbiased because $E(\hat{\theta}) = E(2\bar{X}) = 2E(\bar{X}) = 2 \times \theta/2 = \theta$.

6. (a) $E(\hat{p_1} - \hat{p_2}) = E(\hat{p_1}) - E(\hat{p_2}) = E(X)/m - E(Y)/n = mp_1/m - np_2/n = p_1 - p_2$, thus $\hat{p_1} - \hat{p_2}$ is unbiased estimator of $p_1 - p_2$.

   (b) The standard error of $\hat{p_1} - \hat{p_2}$ is

$$\sigma_{\hat{p_1}-\hat{p_2}} = \sqrt{\sigma^2_{\hat{p_1}} + \sigma^2_{\hat{p_2}}} = \sqrt{\frac{p_1(1-p_1)}{m} + \frac{p_2(1-p_2)}{n}}.$$

   The estimated standard error is

$$S_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{m} + \frac{\hat{p_2}(1-\hat{p_2})}{n}}.$$

   (c) From the data, we have $\hat{p_1} = X/m = 70/100 = 0.7$ and $\hat{p_2} = Y/n = 160/200 = 0.8$, thus the estimator for $p_1 - p_2$ is $\hat{p_1} - \hat{p_2} = -0.1$ and the estimated standard error is

$$S_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{0.7 \times (1-0.7)}{100} + \frac{0.8 \times (1-0.8)}{200}} = 0.054.$$

7. (a) The model-free estimation is 0.5.

   (b) Using the commands *x=c(2.08, 2.10, 1.81, 1.98, 1.91, 2.06); 1-pnorm(2.05, mean(x),sd(x))*, we get a model based estimation of 0.2979.

8. (a) The average of the 10,000 variances was computed as 0.083 (it might be different at a different time and with a different computer), which is very close to the population variance. On the other hand, the average of the 10,000 sample standard deviations was computed as 0.235, which is not as close to the population version. Thus, we conclude that S^2 is unbiased but S is biased.

   (b) In part (a), the bias of S is $0.235 - 0.2887 = -0.0537$. When using the sample size $n = 5$, the average of the 10,000 sample standard deviations was computed as 0.278, with bias $0.278 - 0.2887 = -0.0107$. Thus, we conclude that the bias of S decreases as the sample size increases.

9. (a) Using R commands we can get $P(12 < X \leq 16) = 0.2956$ and the 15th, 25th, 55th, and 95th percentiles are 6.85, 8.30, 11.50, and 17.58, respectively.

   (b) The estimated values for $P(12 < X \leq 16)$ is 0.38 and the estimated 15th, 25th, 55th, and 95th percentiles are 6.37, 7.95, 13.09, and 18.89.

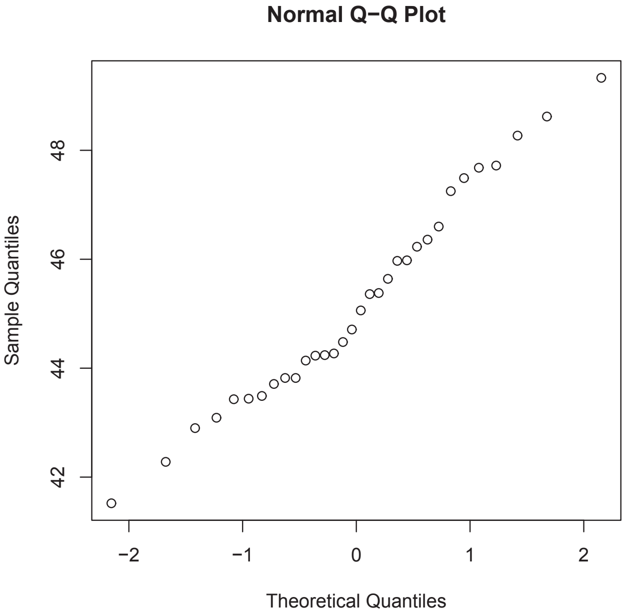
   (c) We use the following commands

```
m = mean(x); s = sd(x);
pnorm(16, m, s) - pnorm(12, m, s)
qnorm(c(0.15, 0.25, 0.55, 0.95), m, s)
```

The model-based estimation for $P(12 < X \leq 16)$ is 0.289 and the model-based estimation for 15th, 25th, 55th, and 95th percentiles are 6.67, 8.40, 12.22, and 19.46.

By comparing the results in (b) and (c) to those in (a), we can see that, in general, the model-based estimators are closer to the population values.

10. (a) The normal Q-Q plot is given below.



The figure suggests that the normal model for the data is appropriate.

(b) The model based estimation for $P(44 < X \leq 46)$ is 0.39 and the model based estimation for median and 75th percentile are 45.20 and 46.52, respectively.

(c) The model-free estimation for $P(12 < X \leq 16)$ is 0.375 and the model-free estimation for median and 75th percentile are 44.885 and 46.420 , respectively.

(d) Since the Q-Q plot suggests that normal assumption is appropriate, we would prefer the model-based estimation.

6.3 Methods for Fitting Models to Data

1. For the exponential(λ) distribution, $\mu = 1/\lambda$. Letting $\bar{X} = 1/\lambda$, we can solve for the method of moment estimator for λ as $\hat{\lambda} = 1/\bar{X}$. It is not unbiased estimator because $E(1/\bar{X}) \neq 1/E(\bar{X})$.

2. (a) The commands to fit Weibull(α,β) distribution are

```
t=read.table(“RobotReactTime.txt”, header=T); t1=t$Time[t$Robot==1];
fn=function(a)
{(mu/gamma(1+1/a))**2*(gamma(1+2/a)-gamma(1+1/a)**2)-var}
library(nleqslv); mu=mean(t1); var=var(t1);
nleqslv(13, fn); mu/gamma(1+1/32.39172)
```

The fitted model parameters are $\hat{\alpha} = 32.39$, and $\hat{\beta} = 31.05$.

(b) To fit the exponential(λ) distribution, using the results in Example 6.3-5, we have $\hat{\lambda} = 1/\bar{X} = 0.0328$.

(c) The model-based estimate of 80th population percentile is 31.51 under model (a) (using command *qweibull(0.8,32.39,31.05)*) and it is 49.07 under model (b) (using command *qexp(0.8, 0.0328)*. As for the probability $P(28.15 \leq X \leq 29.75)$, the estimates under the two models are 0.1805 and 0.0203, respectively.

(d) Using the commands *quantile(t1,0.8); sum(t1>=28.15&t1<=29.75)/length(t1)*, we get the empirical estimate for the 80th population percentile and the probability $P(28.15 \leq X \leq 29.75)$ as 31.522 and 0.2727, respectively.

3. For gamma(α, β) distribution, we have $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Thus, $\beta = \sigma^2/\mu$, and $\alpha = \mu^2/\sigma^2$. We get an estimator of $\hat{\alpha} = \bar{X}^2/S^2$ and $\hat{\beta} = S^2/\bar{X}$. For the given problem, $\hat{\alpha} = 113.5^2/1205.55 = 10.686$ and $\hat{\beta} = 1205.55/113.5 = 10.622$.

4. (a) Since $\mu = \theta\sqrt{\pi/2}$, there is $\theta = \mu\sqrt{2/\pi}$. Thus, the method of moments estimator for θ is $\hat{\theta} = \bar{X}\sqrt{2/\pi}$. It is unbiased because

$$E(\hat{\theta}) = E(\bar{X})\sqrt{\frac{2}{\pi}} = E(X)\sqrt{\frac{2}{\pi}} = \theta\sqrt{\frac{\pi}{2}}\sqrt{\frac{2}{\pi}} = \theta.$$

(b) A model based estimator of the population variance is

$$\hat{\sigma}^2 = \hat{\theta}^2\frac{4-\pi}{2} = \bar{X}^2\frac{2}{\pi}\frac{4-\pi}{2} = \bar{X}^2\frac{4-\pi}{\pi}.$$

$\hat{\sigma}^2$ is not an unbiased estimator of σ^2 because

$$\begin{aligned}
E(\hat{\sigma}^2) &= E(\bar{X}^2)\frac{4-\pi}{\pi} = (Var(\bar{X}) + E(\bar{X})^2)\frac{4-\pi}{\pi} = \left(\frac{\sigma^2}{n} + \mu^2\right)\frac{4-\pi}{\pi} \\
&= \left(\frac{\sigma^2}{n} + \frac{\pi}{2}\theta^2\right)\frac{4-\pi}{\pi} = \left(\frac{\sigma^2}{n} + \frac{\pi}{2}\frac{2}{4-\pi}\sigma^2\right)\frac{4-\pi}{\pi} \\
&= \sigma^2\left(1 + \frac{4-\pi}{n\pi}\right) \neq \sigma^2.
\end{aligned}$$

5. (a) Since $X \sim \text{Bin}(n,p)$, $E(X) = np$. Thus, we can estimate p by $\hat{p} = X/n$. It is unbiased because $E(\hat{p}) = E(X)/n = np/n = p$.

(b) By part (a), $\hat{p} = 24/37 = 0.6486$.

(c) The system lasts more than 350 hours if and only if both of the two components can last more than 350 hours. By the independence of the two components, this probability is p^2 and can be estimated as $\hat{p}^2$. Given the information in (b), $\hat{p}^2 = (24/37)^2 = 0.4207$.

(d) $\hat{p}^2$ is not unbiased estimator for p^2 because $E(\hat{p}^2) = Var(\hat{p}) + E(\hat{p})^2 = \sqrt{p(1-p)/n} + p^2 \neq p^2$.

6. (a) Since the PMF of Poisson(λ) is $e^{-\lambda}\lambda^x/x!$ The likelihood function is

$$\text{lik}(\lambda) = e^{-\lambda}\frac{\lambda^{x_1}}{x_1!}e^{-\lambda}\frac{\lambda^{x_2}}{x_2!}\cdots e^{-\lambda}\frac{\lambda^{x_n}}{x_n!} = e^{-n\lambda}\frac{\lambda^{x_1+x_2+\cdots+x_n}}{\prod_{i=1}^n x_i!}$$

and the log-likelihood function is

$$\mathcal{L}(\lambda) = -n\lambda + \left(\sum_{i=1}^n x_i\right)\log\lambda - \sum_{i=1}^n \log x_i!.$$

Setting the first derivative of the log-likelihood function to zero yields the equation

$$-n + \left(\sum_{i=1}^n x_i\right)\frac{1}{\lambda} = 0.$$

Solving this equation with respect to λ yields the MLE $\hat{\lambda} = \bar{X}$ of λ.

(b) The MLE estimate of λ is $\bar{X} = 2.24$.

(c) The model-based population variance is $\hat{\sigma}^2 = \hat{\lambda} = \bar{X} = 2.24$, and the sample variance is 1.533. Assuming the Poisson model correctly describes the population distribution, we would prefer the model-based estimate.

7. (a) There are $X + 5$ helmets and the last one has flaw, among the rest $X + 4$ helmets, there are 4 with flaw and X flawless, thus, we have the probability

$$P(X = x) = \binom{x+4}{4}p^5(1-p)^x.$$

Therefore, the log-likelihood function is

$$\mathcal{L}(p) = \log\binom{X+4}{4} + 5\log p + X\log(1-p).$$

Setting the first derivative of the log-likelihood function to zero yields the equation

$$\frac{5}{p} - \frac{X}{1-p} = 0.$$

Solving this equation yields the MLE $\hat{p} = 5/(5+X)$.

(b) The distribution of X is easily identified as Negative binomial with $r = 5$ and parameter p (compare to formula (3.4.15)). Thus, $E(X) = r/p = 5/p$. In method of moment estimation, set $X = 5/p$, and we can solve for the estimator $\hat{p} = 5/X$.

(c) If $X = 47$, the MLE (a) gives $\hat{p} = 5/(5 + 47) = 0.096$ and the method of moment formula in (b) gives $\hat{p} = 5/47 = 0.106$.

8. (a) For uniform(0, θ) distribution, $E(X) = \theta/2$, thus the method of moment estimator is $\hat{\theta} = 2\bar{X}$. For the commands *set.seed(3333); x=runif(20, 0, 10); mean(x)*, we have $\bar{X} = 5.359$, thus $\hat{\theta} = 10.718$. The model-based estimator of the population variance σ^2 is $\hat{\sigma}^2 = \hat{\theta}^2/12$, thus, for this dataset, the estimate is $10.718^2/12 = 9.573$.

   (b) The sample variance is 6.781. Compared to the true value of the population variance, $10^2/12 = 8.333$, the model-based estimate overestimates 1.24, while the model-free estimate underestimates 1.552. Thus, the model-based estimate provides a better approximation.

9. (a) To get the moments estimator for θ, solve the equation $\hat{P} = E(P)$, that is $\hat{P} = \theta/(1 + \theta)$, and we have the estimator

$$\hat{\theta} = \frac{\hat{P}}{1 - \hat{P}}.$$

   (b) For the given data, the estimate of θ is $\hat{\theta} = 0.202$.

10. (a) The regression coefficients are

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2} = \frac{11 \times 400.5225 - 263.53 \times 36.66}{11 \times 9677.4709 - 263.53^2} = -0.1420,$$

and

$$\hat{\alpha}_1 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{36.66}{11} - (-0.1420)\frac{263.53}{11} = 6.735.$$

Thus, the regression line is $\hat{y} = \hat{\alpha}_1 + \hat{\beta}_1 x = 6.735 - 0.142x$.

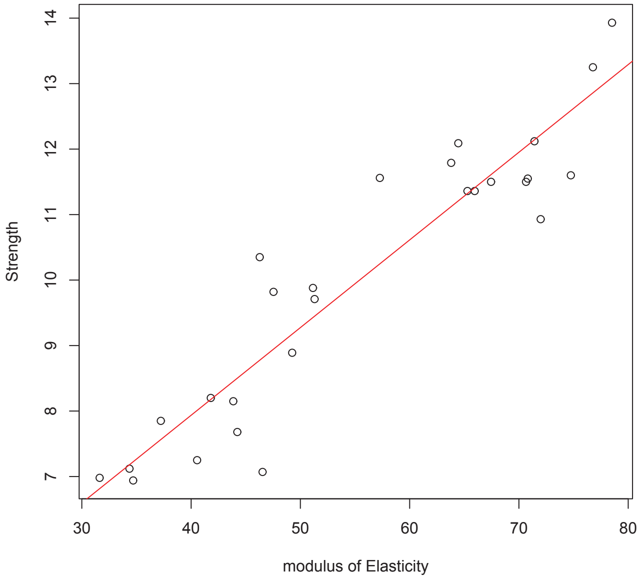
   (b) Since the observed concentrations are in the range of 2.50 to 55.00 and the concentrations 4.5 and 34.7 are in this range, but 62.8 is not in the range, we can conclude that it is appropriate to use the regression line to 4.5 and 34.7. The estimated expected corrosion rate at 4.5 is $6.735 - 0.142 \times 4.5 = 6.096$, and at 34.7 is $6.735 - 0.142 \times 34.7 = 1.808$.

11. (a) Using the commands *x = c(498,526,559,614); y=c(16, 25, 34, 39); lm(y∼x)*, we find the estimated regression line is $\hat{y} = -78.7381 + 0.1952x$. The expected number of manatee deaths in a year with 550,000 powerboat registrations is estimated as $-78.7381 + 0.1952 \times 550 = 28.62$.

(b) The R command for (6.3.11) is *sum(y**2)+78.7381*sum(y) - 0.1952*sum(x*y)* and it gives 4.82 as the error sum of squares. The intrinsic error variance is $SSE/(n-2) = 24.82/(4-2) = 12.41$.

(c) The command *lm(y∼x)$fitted* gives the fitted values as 18.49371, 23.96056, 30.40364, and 41.14209. The command *lm(y∼x)$resid* gives the residuals as -2.493712, 1.039438, 3.596365, and -2.142091. The command *sum((lm(y∼x)$resid)**2)* gives the sum of squared residuals and is the same as in part (b).

12. (a) The following shows the scatterplot of the data with the fitted regression line drawn through it.



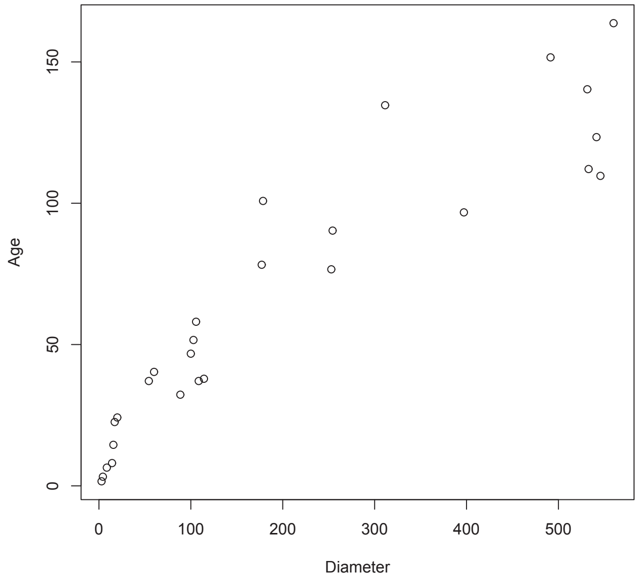
From this graph, the linearity of the regression function and homoscedasticity appear to hold.

(b) The LSE regression coefficients are $\hat{\alpha}_1 = 2.5801$ and $\hat{\beta}_1 = 0.1339$. We can estimate the expected strength at modulus of elasticity $X = 60$ as $2.5801 + 0.1339 \times 60 = 10.6141$.

(c) Using the commands *out=lm(y∼x); sum(out$resid**2); sum(out$resid**2)/out$df.resid*, we get the error sum of squares and the estimator of the intrinsic error variance are 15.16757 and 0.6067028, respectively.

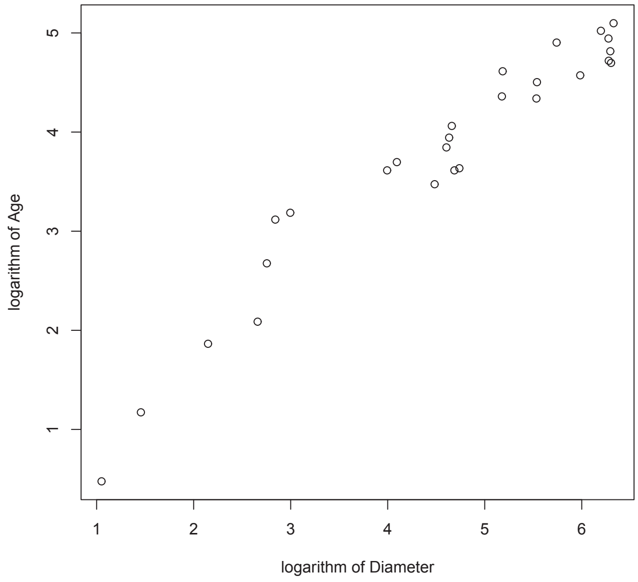
13. (a) The LSE regression coefficients are $\hat{\alpha}_1 = 19.9691$ and $\hat{\beta}_1 = 0.2255$. We can estimate the expected age at diameter x as $\hat{y} = 19.9691 + 0.2255 \times x$.

(b) The scatterplot of the data is shown on the next page.



This figure suggests that the age of the tree increases with the diameter of the tree at approximately linear fashion. Thus, the assumption of linearity of the regression function seems to be, at least approximately, satisfied. On the other hand, the variability in age of trees seems to increase with the diameter of tree. Thus, the homoscedasticity assumption appears to be violated for this data set.

(c) The following shows the scatterplot of the transformed data.



After the log-transformation, the assumptions of the simple linear regression model seem to be valid.

6.4 Comparing Estimators: The MSE Criterion

1. (a) $\text{Bias}(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta = 2E(\bar{X}) - \theta = 2E(X) - \theta = 2 \times \theta/2 - \theta = 0$. The bias for $\hat{\theta}_2$ is $\text{Bias}(\hat{\theta}_2) = E(\hat{\theta}_2) - \theta = n\theta/(n+1) - \theta = -\theta/(n+1)$. Thus, $\hat{\theta}_1$ is unbiased while $\hat{\theta}_2$ is biased.

   (b) For $\hat{\theta}_1$, we have

   $$\text{MSE}(\hat{\theta}_1) = \text{Var}(\hat{\theta}_1) + \text{Bias}(\hat{\theta}_1)^2 = \text{Var}(2\bar{X}) = 4\text{Var}(\bar{X}) = 4\frac{\sigma^2}{n} = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

   For $\hat{\theta}_2$,

   $$\begin{aligned}\text{MSE}(\hat{\theta}_2) = \text{Var}(\hat{\theta}_2) + \text{Bias}(\hat{\theta}_2)^2 &= \frac{n}{(n+1)^2(n+2)}\theta^2 + \left(-\frac{\theta}{n+1}\right)^2 \\ &= \frac{2\theta^2}{(n+1)(n+2)}.\end{aligned}$$

   (c) When $n = 5$ and true value of θ is 10, we have $\text{MSE}(\hat{\theta}_1) = 10^2/(3 \times 5) = 6.67$, while $\text{MSE}(\hat{\theta}_2) = 2 \times 10^2/[((5+1)(5+2)] = 4.76$. According to the MSE selection criterion, $\hat{\theta}_2$ is preferable.

2. From the distributions of $X_1, \cdots, X_{10}$ and $Y_1, \cdots, Y_{10}$, we have $E(\bar{X}) = E(\bar{Y}) = \mu$, $\text{Var}(\bar{X}) = \sigma^2/10$, and $\text{Var}(\bar{Y}) = 4\sigma^2/10$. $\bar{X}$ and $\bar{Y}$ are also independent. Thus,

   (a) For any $0 \le \alpha \le 1$, $E(\hat{\mu}) = E(\alpha\bar{X} + (1-\alpha)\bar{Y}) = \alpha E(\bar{X}) + (1-\alpha)E(\bar{Y}) = \alpha\mu + (1-\alpha)\mu = \mu$. Thus, $\hat{\mu}$ is unbiased for μ.

   (b) Since $\hat{\mu}$ is unbiased,

   $$\begin{aligned}\text{MSE}(\hat{\mu}) = \text{Var}(\hat{\mu}) &= \text{Var}(\alpha\bar{X} + (1-\alpha)\bar{Y}) = \alpha^2\text{Var}(\bar{X}) + (1-\alpha)^2\text{Var}(\bar{Y}) \\ &= \alpha^2\frac{\sigma^2}{10} + (1-\alpha)^2\frac{4\sigma^2}{10} = (5\alpha^2 - 8\alpha + 4)\frac{\sigma^2}{10}\end{aligned}$$

   (c) The estimator $0.5\bar{X} + 0.5\bar{Y}$ corresponds to $\hat{\mu}$ with $\alpha = 0.5$. The MSE is

   $$\text{MSE}(0.5\bar{X} + 0.5\bar{Y}) = (5 \times 0.5^2 - 8 \times 0.5 + 4)\frac{\sigma^2}{10} = 1.25\frac{\sigma^2}{10}.$$

   Since $\text{MSE}(\bar{X}) = \text{Var}(\bar{X}) = \sigma^2/10 < \text{MSE}(0.5\bar{X} + 0.5\bar{Y})$, $\bar{X}$ is a preferable estimator.

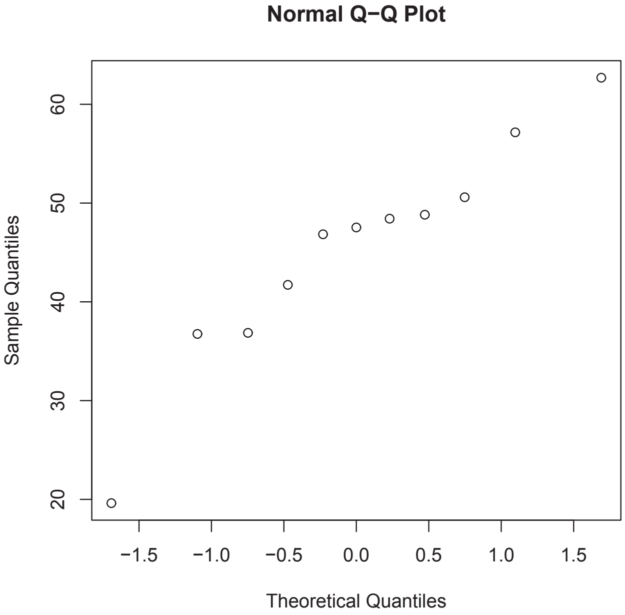
Chapter 7

Confidence and Prediction Intervals

7.3 Type of Confidence Intervals

1. (a) The 95% CI for the mean μ is $\bar{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$, or $45.18 \pm t_{11-1,0.05/2} 11.48/\sqrt{11}$, which is calculated as (37.47, 52.89). In order to make the CI valid, we need to assume that the data are distributed approximately normal.

   (b) The normal Q-Q plot is shown below.



   The Q-Q plot shows that the normal assumption is approximately valid.

2. (a) Using the R commands *x=c(649, 832, 418, 530, 384, 899, 755); confint(lm(x∼1),level=0.9)*, we find the 90% CI for the true mean histamine content for all worker bees of this age as (489.9886, 786.2971). In order to make the CI valid, we need to assume that the data are distributed approximately normal.

(b) False

3. (a) The 80% CI for the mean breaking strength μ is $\bar{X} \pm t_{n-1,\alpha/2}S/\sqrt{n}$, or $210 \pm t_{50-1,0.1/2}18/\sqrt{50}$, which is calculated as (206.69, 213.31). Since the sample size $n = 50$ is large enough, the normality assumption is not necessary.

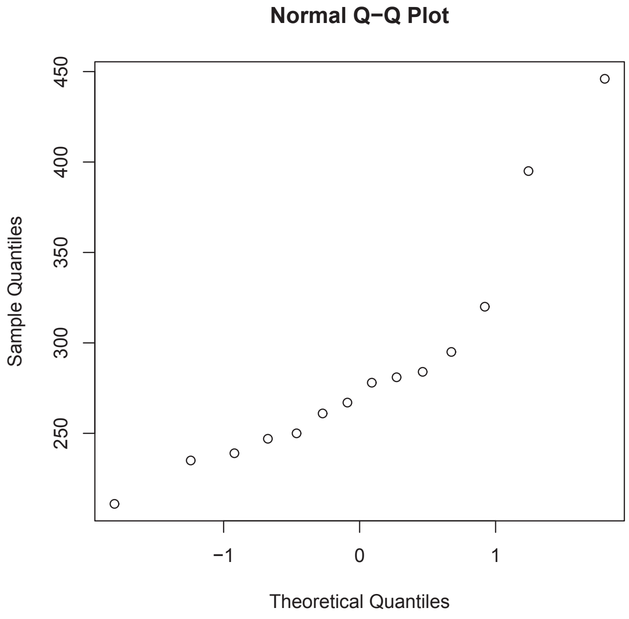
   (b) Yes

   (c) No

4. (a) In Exercise 1, after sorting, the data is 19.62, 36.75, 36.86, 41.72, 46.84, 47.53, 48.42, 48.82, 50.59, 57.16, 62.69. Thus, the CI (36.86, 50.59) is $X_{(3)}, X_{(9)}$. By the command *2*(1-pbinom(11-3, 11, 0.5))*, we find $\alpha = 0.06542969$, thus the confidence level is $(1 - \alpha)100\% = 93.46\%$. The CI (36.75, 57.16) is $X_{(2)}, X_{(10)}$ and the confidence level is 98.83%.

   (b) In Exercise 2, after sorting, the data is 384, 418, 530, 649, 755, 832, 899. Thus, the CI (418, 832) is $X_{(2)}, X_{(6)}$. By the command *2*(1-pbinom(7-2, 7, 0.5))*, we find $\alpha = 0.125$, thus the confidence level is $(1 - \alpha)100\% = 87.5\%$.

5. (a) The normal Q-Q plot is shown below.



   The Q-Q plot shows that the normal assumption is not quite valid.

   (b) The 90% CI for the mean ozone level is given by the command *confint(lm(x∼1),level=0.9)* and the result is (256.123, 316.5913).

   (c) The 90% CI for the median ozone level is given by the command *library(BSDA); SIGN.test(x, alternative="two.side", conf.level=0.9)* and the result is (248.0466, 291.1626).

(d) The length for the 90% CI for the mean is 60.468, while the length for the 90% CI for the median is 48.116. Clearly, the 90% CI for the median is shorter. Since the normal assumption of the data seems not appropriate, we would prefer the 90% CI for the median.

6. (a) The 95% CI for the mean solar intensity is given by the command *confint(lm(x∼1),level=0.95)* and the result is (706.4223, 721.9277). Since the sample size $n = 40$ is large enough, the normality assumption is not necessary.

   (b) The 90% CI for the median solar intensity is given by the command *library(BSDA); SIGN.test(x, alternative="two.side", conf.level=0.95)* and the result is (704.8189, 728.0892). No assumption is needed.

   (c) The interpretation of the confidence level is wrong for both the CI in (a) and (b).

7. (a) For Poisson(λ) distribution, $\mu = \lambda$. Thus, the 95% CI for λ is the same as the 95% CI for μ. We use the following command:

   x=c(rep(0,4), rep(1,12),rep(2,11),rep(3,14),rep(4,9));
   confint(lm(x∼1),level=0.95)

   to find the CI as (1.888116, 2.591884).

   (b) For Poisson(λ) distribution, $\sigma^2 = \lambda$. Thus, the 95% CI for σ is $(\sqrt{1.888116}, \sqrt{2.591884})$, or (1.374087, 1.609933).

8. (a) The 95% CI for the mean eruption duration is given by the command *confint(lm(ed∼1),level=0.95)* and the result is (3.351534, 3.624032).

   (b) The 90% CI for the median eruption duration is given by the command *library(BSDA); SIGN.test(x, alternative="two.side", conf.level=0.95)* and the result is (3.833, 4.1115).

   (c) The sample proportion, $\hat{p}$, can be found with the R command *phat=sum(ed>4.42)/length(ed)*, which gives 0.2683824. Then using the following commands

   alpha=0.05;
   *phat-qnorm(1-alpha/2)*sqrt(phat*(1-phat)/length(ed));*
   *phat+qnorm(1-alpha/2)*sqrt(phat*(1-phat)/length(ed)),*

   gives the 95% CI for the probability that an eruption duration will last more than 4.42 min as (0.2157221, 0.3210426).

9. (a) To find the 95% confidence interval for the proportion, p, of customers who qualify, we use the following commands:

```
n=500; phat = 40/n; alpha=0.05;
phat-qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n);
phat+qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n),
```

The obtained CI is (0.05622054, 0.1037795).

(b) In order to make the CI valid, there should be at least 8 customers who qualify and at least 8 customers who do not qualify in the sample, which is satisfied by the data.

10. (a) To find the 95% confidence interval for the proportion, p, of young adult US citizens who drink beer, wine, or hard liquor on a weekly basis, we use the following commands:

```
n=1516; phat = 985/n; alpha=0.05;
phat-qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n);
phat+qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n),
```

The obtained CI is (0.6257221, 0.6737502).

(b) False

11. (a) To find the 95% confidence interval for the proportion, p, that a randomly selected component lasts more than 350 hours, we use the following commands:

```
n=37; phat = 24/n; alpha=0.05;
phat-qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n);
phat+qnorm(1-alpha/2)*sqrt(phat*(1-phat)/n),
```

The obtained CI is (0.4948251, 0.8024722).

(b) Assuming that the life spans of the two components in the system are independent, the probability that the system lasts more than 350 hours is p^2. Thus, the 95% CI for the probability that the system lasts more than 350 hours is $(0.4948251^2, 0.8024722^2)$.

12. By (7.3.12), the $(1-\alpha)100\%$ CI for $\mu_{Y|X}(x)$ is

$$\hat{\mu}_{Y|X}(x) \pm t_{n-2,\alpha/2} S_{\hat{\mu}_{Y|X}(x)}$$

with

$$S_{\hat{\mu}_{Y|X}(x)} = S_\epsilon \sqrt{\frac{1}{n} + \frac{n(x-\bar{X})^2}{n\sum X_i^2 - (\sum X_i)^2}}.$$

Since $\mu_{Y|X}(0) = \alpha_1$ and $\hat{\mu}_{Y|X}(0) = \hat{\alpha}_1$, we can let $x = 0$ in the above formula to find the $(1-\alpha)100\%$ CI for α_1 as

$$\hat{\alpha}_1 \pm t_{n-2,\alpha/2} S_\epsilon \sqrt{\frac{1}{n} + \frac{n\bar{X}^2}{n\sum X_i^2 - (\sum X_i)^2}}.$$

13. (a) From the given information, we have

$$\hat{\beta}_1 = \frac{n\sum X_iY_i - (\sum X_i)(\sum Y_i)}{n\sum X_i^2 - (\sum X_i)^2} = \frac{10 \times 2418968 - 3728 \times 5421}{10 \times 1816016 - 3728^2} = 0.9338$$

and

$$\hat{\alpha}_1 = \bar{Y} - \hat{\beta}_1\bar{X} = \frac{5421}{10} - 0.9338 \times \frac{3728}{10} = 193.9643.$$

The LSE for σ_ϵ^2 is

$$\begin{aligned} S_\epsilon^2 &= \frac{1}{n-2}\left[\sum Y_i^2 - \hat{\alpha}_1 \sum Y_i - \hat{\beta}_1 \sum X_iY_i\right] \\ &= \frac{1}{8}[3343359 - 193.9643 \times 5421 - 0.9338 \times 2418968] = 4130.776. \end{aligned}$$

(b) From the data, we have

$$\begin{aligned} S_{\hat{\beta}_1} &= S_\epsilon\sqrt{\frac{n}{n\sum X_i^2 - (\sum X_i)^2}} \\ &= \sqrt{10 \times 4130.776/(10 \times 1816016 - 3728^2)} = 0.09844647, \end{aligned}$$

thus the 95% CI for the true slope of the regression line is

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}S_{\hat{\beta}_1} = 0.9338 \pm t_{8,0.025} \times 0.09844647,$$

which is calculated as (0.7068, 1.1608). In order to make the CI valid, we need the assumption that the error terms are normally distributed.

(c) When the surface conductivity is 500, the expected sediment conductivity is

$$\hat{\mu}_{Y|X}(500) = \hat{\alpha}_1 + \hat{\beta}_1 \times 500 = 193.9643 + 0.9338 \times 500 = 660.8643,$$

and the estimated standard error is

$$\begin{aligned} S_{\hat{\mu}_{Y|X}(x)} &= S_\epsilon\sqrt{\frac{1}{n} + \frac{n(x-\bar{X})^2}{n\sum X_i^2 - (\sum X_i)^2}} \\ &= \sqrt{4130.776}\sqrt{\frac{1}{10} + \frac{10 \times (500 - 372.8)^2}{10 \times 1816016 - 3728^2}} = 23.87232. \end{aligned}$$

Thus, the 95% CI for the the expected sediment conductivity at the surface conductivity 500 is

$$\hat{\mu}_{Y|X}(500) \pm t_{n-2,\alpha/2}S_{\hat{\mu}_{Y|X}(500)} = 660.8643 \pm t_{8,0.025} \times 23.87232,$$

which is calculated as (605.8146, 715.914).

When the surface conductivity is 900, the expected sediment conductivity is

$$\hat{\mu}_{Y|X}(900) = \hat{\alpha}_1 + \hat{\beta}_1 \times 900 = 193.9643 + 0.9338 \times 900 = 1034.384,$$

and the estimated standard error is

$$S_{\hat{\mu}_{Y|X}(x)} = S_\epsilon\sqrt{\frac{1}{n}+\frac{n(x-\bar{X})^2}{n\sum X_i^2-(\sum X_i)^2}}$$
$$= \sqrt{4130.776}\sqrt{\frac{1}{10}+\frac{10\times(900-372.8)^2}{10\times 1816016-3728^2}} = 55.73858.$$

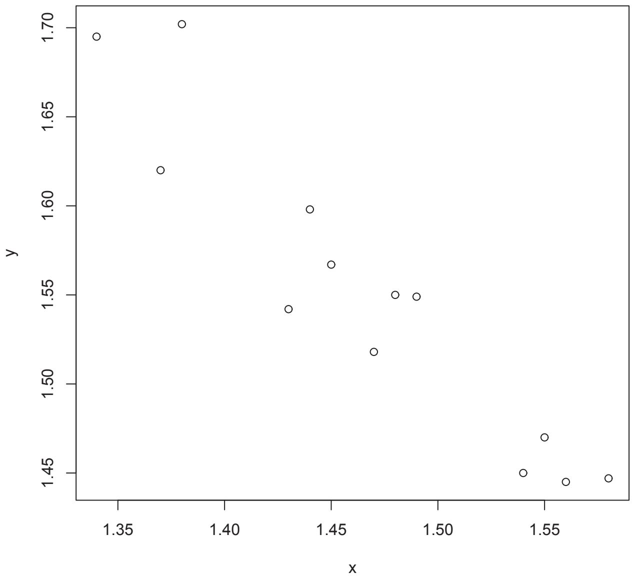
Thus, the 95% CI for the the expected sediment conductivity at the surface conductivity 500 is

$$\hat{\mu}_{Y|X}(900) \pm t_{n-2,\alpha/2}S_{\hat{\mu}_{Y|X}(900)} = 1034.384 \pm t_{8,0.025}\times 55.73858,$$

which is calculated as (905.8506, 1162.917).

The CI at X = 900 is not appropriate because 900 is not in the range of X-values in the data set.

14. (a) The scatterplot is given below.



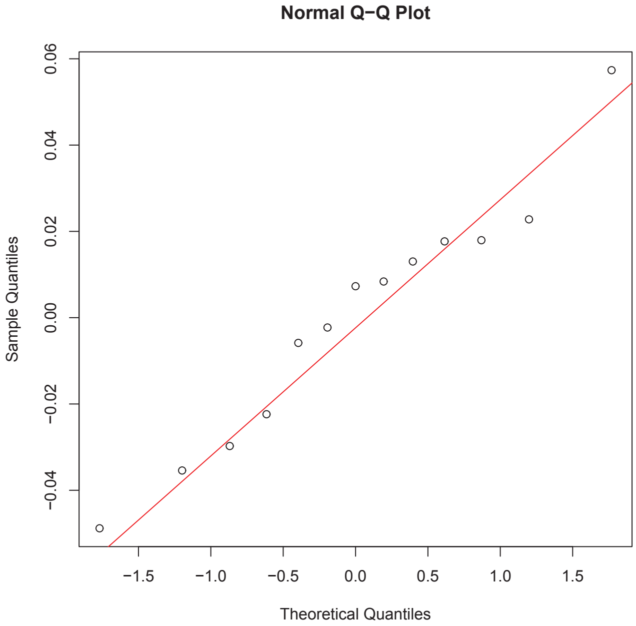
The scatterplot of the data shows that the linearity and homoscedasticity assumptions of the simple linear regression model seem to be valid.

(b) Using the R commands *out=lm(y∼x); out$coef*, we have the LSE estimates $\hat{\alpha}_1 = 3.130417$ and $\hat{\beta}_1 = -1.076647$. Using the command *sqrt(sum(out$resid**2)/out$df.resid)*, we find the LSE for σ_ϵ as 0.0298461.

(c) Using the R command *confint(out, level =0.9)*, we find the 90% CI for β_1 as (-1.279482, -0.873812). Since the mean difference of the strengths at water/cement ratios 1.55 and 1.35 is $\mu_{Y|X(1.55)} - \mu_{Y|X(1.35)} = 0.2\beta_1$, we calculate the 90% CI for the mean difference of the strengths at water/cement ratios 1.55 and 1.35 as (-0.2558964, -0.1747624).

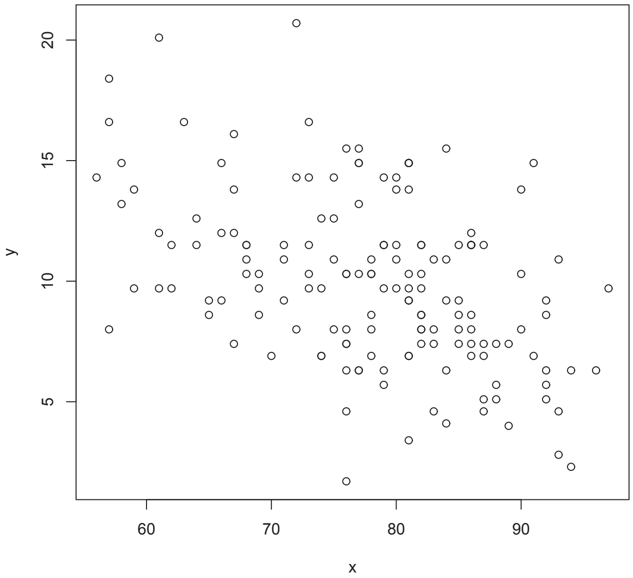
(d) Using the R commands *t=data.frame(x=c(1.35,1.45,1.55)); predict(out, t, interval="confidence", level=0.9)*, we find the 90% CIs for the mean strength at water/cement ratios 1.35, 1.45, and 1.55 as (1.648821, 1.705066), (1.553986, 1.584572), and (1.439260, 1.483969), respectively.

(e) Using the R commands *qqnorm(out$resid); qqline(out$resid, col="red")*, we get the following normal Q-Q plot of the residuals:



The plot does not suggest serious departure from the normality assumption and thus, the above CIs can be reasonably trusted.

15. (a) The scatterplot is given on the following page.



The scatterplot of the data shows that the linearity and homoscedasticity assumptions of the simple linear regression model seem to be valid.

(b) Using the R commands *out=lm(y∼x); confint(out, level =0.95)*, we find the 95% CI for β_1 as (-0.2236649, -0.117264). Using the R commands *t=data.frame(x=80); predict(out, t, interval="confidence", level=0.95)*, we find the 95% CI for the expected wind speed on an 80F day as (9.082136, 10.11093).

16. Using the command *2*(1-pbinom(30-10, 30, 0.5))*, we find $\alpha = 0.04277395$, thus the confidence level is $(1-\alpha)100\% = 95.72\%$.

17. Using the command *n=16; a=4; 1-2*(1-pbinom(n-a,n,0.5))*, we have the result 97.87%; changing a to 5, we have the result 92.32%. These are consistent with Example 7.3-7.

18. From (7.3.19), we find the $(1-\alpha)100\%$ CI for σ is

$$\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}S < \sigma < \sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}}S.$$

Thus, we can write the R commands as

n= 15; S = 0.64; a = 0.05;

*L = sqrt((n-1)/qchisq(1-a/2, n-1))*S; U=sqrt((n-1)/qchisq(a/2, n-1))*S;*

We find the 95% CI for σ as (0.468561, 1.009343). In order to make the CI valid, we need to assume that the population distribution is normal.

19. From (7.3.19), we find the $(1-\alpha)100\%$ CI for σ is

$$\sqrt{\frac{n-1}{\chi^2_{n-1,\alpha/2}}}S < \sigma < \sqrt{\frac{n-1}{\chi^2_{n-1,1-\alpha/2}}}S.$$

Thus, we can write the R commands as

```
n= 35; S = 0.117; a = 0.05;
L = sqrt((n-1)/qchisq(1-a/2, n-1))*S; U=sqrt((n-1)/qchisq(a/2, n-1))*S;
```

We find the 95% CI for σ as (0.09463803, 0.1532936). The traditional value of 0.1 lies within the CI.

20. Using the R commands

```
rt=read.table("RobotReactTime.txt", header=T); t2=rt$Time[rt$Robot==2];
n= length(t2); S2 = var(t2); a = 0.05;
L = (n-1)*S2/qchisq(1-a/2, n-1); U=(n-1)*S2/qchisq(a/2, n-1);
```

We find the 95% CI for the population variance of reaction times of Robot 2 as (0.4916012, 1.696162).

7.4 The Issue of Precision

1. We use the R command *library(BSDA); nsize(b=0.2, sigma=1.2, conf.level=0.98, type="mu")*; the desired sample size is 195.

2. We use the R command *library(BSDA); nsize(b=4/2, sigma=18, conf.level=0.9, type="mu")*; the desired sample size is 220.

3. (a) We use the R command *library(BSDA); nsize(b=0.1/2, p=9/40, conf.level=0.9, type="pi")*; the desired sample size is 189.

   (b) If no prior information is given, we use the R command *library(BSDA); nsize(b=0.1/2, p=0.5, conf.level=0.9, type="pi")*; the desired sample size is 271.

4. (a) We use the R command *library(BSDA); nsize(b=0.03, p=75/193, conf.level=0.95, type="pi")*; the desired sample size is 1015.

   (b) If no prior information is given, we use the R command *library(BSDA); nsize(b=0.03, p=0.5, conf.level=0.95, type="pi")*; the desired sample size is 1068.

7.5 Prediction Intervals

1. The problem is to find a prediction interval for the next chocolate chip cookie. Using (7.5.4), the prediction interval is

$$\bar{Y} \pm t_{n-1,\alpha/2} S \sqrt{1 + \frac{1}{n}}.$$

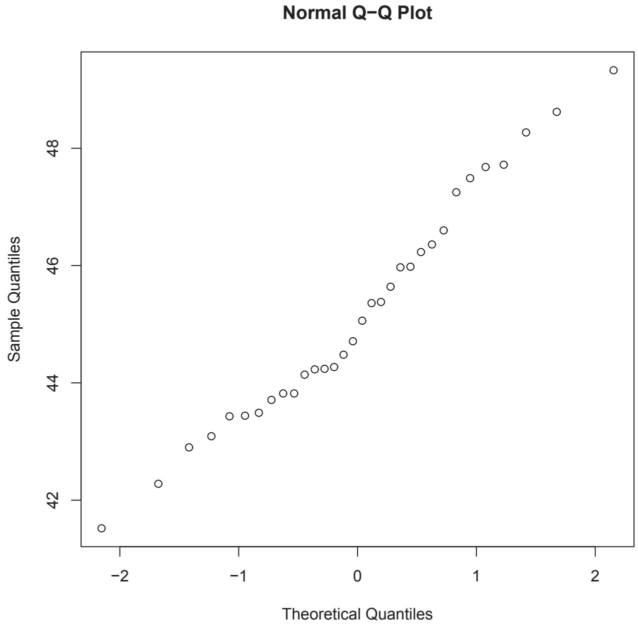
Using the R commands *Ybar = 3.1; S = 0.3; n=16; a=0.1; L=Ybar - qt(1-a/2, n-1)*S*sqrt(1+1/n); U=Ybar + qt(1-a/2, n-1)*S*sqrt(1+1/n)*, we find the 90% prediction interval as (2.557899, 3.642101). In order to make the CI valid, we need to assume that the population distribution is normal.

2. (a) Using the R commands *Xbar = 30.79; S = 6.53; n=8; a=0.1; L=Xbar - qt(1-a/2, n-1)*S*sqrt(1+1/n); U=Xbar + qt(1-a/2, n-1)*S*sqrt(1+1/n)*, we find the 90% prediction interval as (17.66794, 43.91206). In order to make the CI valid, we need to assume that the population distribution is normal.

   (b) Using the R commands *L=Xbar - qt(1-a/2, n-1)*S; U=Xbar + qt(1-a/2, n-1)*S*, we find the 90% confidence interval for the mean heat flux as (18.4184, 43.1616). The prediction interval has a length of 26.24412, and the confidence interval has a length of 24.7432. It is clear that the prediction interval is longer.

3. (a) Using the R commands *predict(lm(y∼1), data.frame(1), interval="predict", level=0.95)*, we find the 95% prediction interval for the compressive strength of the next concrete specimen as (41.17, 49.24).

   (b) The normal Q-Q plot for the data is shown as follows:



The plot shows the normality assumption holds.

4. (a) The expected separation distance between the next cyclist, whose distance from the roadway center line is 15 feet, and a passing car can be estimated as

$$\hat{\mu}_{Y|X=15} = -2.1825 + 0.6603 \times 15 = 7.722.$$

Using formula (7.5.6), we calculate the prediction interval for Y when $X = 15$ as

$$\hat{\mu}_{Y|X=15} \pm t_{10-2,0.1/2}\sqrt{0.3389}\sqrt{1 + \frac{1}{10} + \frac{10 \times (15 - 15.42)^2}{10 \times 2452.18 - 154.2^2}} = (6.59, 8.86).$$

(b) A distance of 12 feet is not in the range of X values in the data set, so the desired PI would not be reliable.

5. (a) We use the command

predict(lm(y∼1), data.frame(1), interval="predict", level=0.95)

to find (-36.24556, 441.9256) as the prediction interval for the weight of the next bear that will be captured during the same time period. Since the weight cannot be negative, we use (0, 441.9256) as the prediction interval instead.

(i) No assumption is need for the validity of the prediction because $\bar{Y}$ is the best estimate for μ under MSE criterion.

(ii) To make the PI valid, we need to assume that the weights of bears are normally distributed.

(b) We use the command *predict(lm(y∼x), data.frame(x=40), interval="predict", level=0.95)* to find (165.9967, 301.5883) as the prediction interval for the weight of the next bear that will be captured during the same time period if its chest girth measures 40 cm.

(i) To make the prediction valid, we need to assume that the weight and chest girth have a linear relation.

(ii) The validity of the PI requires that all the assumptions of the normal simple linear regression model, that is, the additional assumptions of homoscedasticity and normality of the intrinsic error variables, be satisfied.

(c) The prediction interval in part (a) is much longer than that in part (b).

Chapter 8
Testing of Hypotheses

8.2 Setting Up a Test Procedure

1. (a) Let μ be the mean soil heat flux, then the null and alternative hypotheses are

$$H_0 : \mu \leq 31 \quad \text{vs.} \quad H_a : \mu > 31.$$

   (b) If the null hypothesis is rejected, we should use the coal dust cover.

2. (a) The null and alternative hypotheses are

$$H_0 : p \leq 0.25 \quad \text{vs.} \quad H_a : p > 0.25.$$

   (b) If the null hypothesis is rejected, we should adopt the modified bumper design.

3. (a) If the CEO wants to adopt it unless there is evidence that it has a lower protection index, then $H_a : \mu < \mu_0$.

   (b) If the CEO does not want to adopt it unless there is evidence that it has a higher protection index, then $H_a : \mu > \mu_0$.

   (c) If the null hypothesis is rejected for part (a), the CEO should not adopt the new grille guard and for part (b), the CEO should adopt the new grille guard.

4. (a) If the manufacturer does not want to buy the new machine unless there is evidence it is more productive than the old one, then $H_a : \mu > \mu_0$.

   (b) If the manufacturer wants to buy the new machine unless there is evidence it is less productive than the old one, then $H_a : \mu < \mu_0$.

   (c) If the null hypothesis is rejected for part (a), the manufacturer should buy the new machine and for part (b) the manufacturer should not buy the new machine.

5. (a) Let p be the proportion of all customers who qualify for membership, then the hypotheses are

$$H_0 : p \geq 0.05 \quad \text{vs.} \quad H_a : p < 0.05.$$

(b) If the null hypothesis is rejected, the airline should not proceed with the establishment of the traveler's club.

6. (a) The statement is true.

(b) We determine C from the requirement that the probability of incorrectly rejecting H_0 is no more than 0.05 or, in mathematical notation,

$$P(\bar{X} \geq C) \leq 0.05 \text{ if } H_0 \text{ is true.}$$

Over the range of μ values specified by H_0 (i.e., $\mu \leq 28,000$), the probability $P(\bar{X} \geq C)$ is largest when $\mu = 28,000$. Thus, the requirement will be satisfied if C is chosen so that when $\mu = 28,000$, $P(\bar{X} \geq C) = 0.05$. This is achieved by choosing C to be the 95th percentile of the distribution of $\bar{X}$ when $\mu = 28,000$. Recall that σ is assumed to be known, this yields $C = 28,000 + z_{0.05}\sigma/\sqrt{n}$.

(c) Let

$$Z_{H_0} = \frac{\bar{X} - 28,000}{\sigma/\sqrt{n}}.$$

Then the standardized version of the rejection region is $Z_{H_0} \geq z_{0.05}$.

7. (a) The rejection region is of the form $\hat{\mu}_{Y|X}(x) \geq C$ for some constant C.

(b) We determine C from the requirement that the probability of incorrectly rejecting H_0 is no more than 0.05 or, in mathematical notation,

$$P(\hat{\mu}_{Y|X}(x) \geq C) \leq 0.05 \text{ if } H_0 \text{ is true.}$$

Thus, the requirement will be satisfied if C is chosen so that under H_0 (i.e. $\mu_{Y|X}(x) = \mu_{Y|X}(x)_{x_0}$, $P(\hat{\mu}_{Y|X}(x) \geq C) = 0.05$. This is achieved by choosing C to be the 95th percentile of the distribution of $\hat{\mu}_{Y|X}(x)$ when $\mu_{Y|X}(x) = \mu_{Y|X}(x)_{x_0}$. Recall that the distribution of $\hat{\mu}_{Y|X}(x)$

$$\frac{\hat{\mu}_{Y|X}(x) - \mu_{Y|X}(x)_{x_0}}{S_{\hat{\mu}_{Y|X}(x)}} \sim t_{n-2},$$

this yields the selection of $C = \mu_{Y|X}(x)_{x_0} + t_{n-2,0.05}S_{\hat{\mu}_{Y|X}(x)}$.

8. (a) Let p be the proportion of all detonators that will ignite, then the null and alternative hypotheses are

$$H_0 : p \geq 0.9 \quad \text{vs.} \quad H_a : p < 0.9.$$

(b) The standardized test statistic is

$$Z_{H_0} = \frac{\hat{p} - 0.9}{\sqrt{0.9 \times 0.1/n}}.$$

(c) The statement is false.

9. (a) Since the $(1-\alpha)100\%$ CI for β_1 is $\hat{\beta}_1 \pm t_{n-2,\alpha/2} S_{\hat{\beta}_1}$, $H_0 : \beta_1 = \beta_{1,0}$ is rejected if

$$\beta_{1,0} < \hat{\beta}_1 - t_{n-2,\alpha/2} S_{\hat{\beta}_1} \quad \text{or} \quad \beta_{1,0} > \hat{\beta}_1 + t_{n-2,\alpha/2} S_{\hat{\beta}_1}.$$

These inequalities could be rewritten as

$$\frac{\hat{\beta}_1 - \beta_{1,0}}{S_{\hat{\beta}_1}} > t_{n-2,\alpha/2} \quad \text{or} \quad \frac{\hat{\beta}_1 - \beta_{1,0}}{S_{\hat{\beta}_1}} < -t_{n-2,\alpha/2}.$$

Let

$$T_{H_0} = \frac{\hat{\beta}_1 - \beta_{1,0}}{S_{\hat{\beta}_1}},$$

the CI based RR could be written as $|T_{H_0}| > t_{n-2,\alpha/2}$.

(b) Since the $(1-\alpha)100\%$ CI for $\mu_{Y|X}(x)$ is $\hat{\mu}_{Y|X}(x) \pm t_{n-2,\alpha/2} S_{\hat{\mu}_{Y|X}(x)}$,
$H_0 : \mu_{Y|X}(x) = \mu_{Y|X}(x)_0$ is rejected if

$$\mu_{Y|X}(x)_0 < \hat{\mu}_{Y|X}(x) - t_{n-2,\alpha/2} S_{\hat{\mu}_{Y|X}(x)} \quad \text{or} \quad \mu_{Y|X}(x)_0 > \hat{\mu}_{Y|X}(x) - t_{n-2,\alpha/2} S_{\hat{\mu}_{Y|X}(x)}.$$

These inequalities could be rewritten as

$$\frac{\hat{\mu}_{Y|X}(x) - \mu_{Y|X}(x)_0}{S_{\hat{\mu}_{Y|X}(x)}} > t_{n-2,\alpha/2} \quad \text{or} \quad \frac{\hat{\mu}_{Y|X}(x) - \mu_{Y|X}(x)_0}{S_{\hat{\mu}_{Y|X}(x)}} < -t_{n-2,\alpha/2}.$$

Let

$$T_{H_0} = \frac{\hat{\mu}_{Y|X}(x) - \mu_{Y|X}(x)_0}{S_{\hat{\mu}_{Y|X}(x)}}.$$

Then the CI based RR could be written as $|T_{H_0}| > t_{n-2,\alpha/2}$.

10. (a) The value of the test statistic is

$$Z_{H_0} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{28640 - 28000}{900/\sqrt{25}} = 3.556.$$

Since the RR is $Z_{H_0} \geq z_\alpha$, the smallest level at which H_0 is rejected is found by solving $3.556 = z_\alpha$ for α. Let $\Phi(\cdot)$ be the cumulative distribution function of $N(0,1)$. The solution to this equation, which is also the p-value, is

$$p - \text{value} = 1 - \Phi(3.556) = 0.00019.$$

(b) Since p-value< 0.05, the null hypothesis should be rejected at a 0.05 level of significance.

11. (a) The standardized test statistic is

$$Z_{H_0} = \frac{8/50 - 0.25}{\sqrt{0.25 \times 0.75/50}} = -1.469694.$$

To sketch the figure, draw a $N(0,1)$ PDF and shade the right of -1.469694, which represents the p-value.

(b) The p-value is calculated as 0.9292. Since p-value> 0.05, the null hypothesis should not be rejected at a 0.05 level of significance.

8.3 Types of Tests

1. (a) The value of test statistic is

$$T_{H_0} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{9.8 - 9.5}{1.095/\sqrt{50}} = 1.94.$$

The RR is $T_{H_0} > t_{n-1,\alpha} = t_{49,0.05} = 1.68$. Since $1.94 > 1.68$, we should reject H_0.

(b) Since the sample size $n = 50 > 30$, no additional assumptions are needed.

2. (a) Let μ be the average permissible exposure, then the null and alternative hypotheses are

$$H_0 : \mu \leq 1 \quad \text{vs.} \quad H_a : \mu > 1.$$

(b) The value of test statistic is

$$T_{H_0} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{2.1 - 1}{4.1/\sqrt{36}} = 1.61.$$

The RR is $T_{H_0} > t_{n-1,\alpha} = t_{35,0.05} = 1.69$. Since $1.69 > 1.61$, we should not reject H_0. Since the sample size $n = 36 > 30$, no additional assumptions are needed.

(c) By using Table A4 the p-value should be between 0.05 and 0.1. Using the R command *1-pt(1.61,35)*, the exact p-value is 0.058.

3. (a) Let μ be the (population) mean penetration, then the null and alternative hypotheses are

$$H_0 : \mu \leq 50 \quad \text{vs.} \quad H_a : \mu > 50.$$

(b) The value of test statistic is

$$T_{H_0} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{52.7 - 50}{4.8/\sqrt{16}} = 2.25.$$

The RR is $T_{H_0} > t_{n-1,\alpha} = t_{15,0.1} = 1.34$. Since $2.25 > 1.34$, we should reject H_0. In order to make the test valid, we need the assumption that the population is normal.

(c) By using Table A4 p-value should be between 0.01 and 0.025. Using R command *1-pt(2.25, 15)*, the exact p-value is 0.02.

4. (a) The value of test statistic is

$$T_{H_0} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{30.79 - 29}{6.53/\sqrt{8}} = 0.775.$$

The RR is $T_{H_0} > t_{n-1,\alpha} = t_{7,0.05} = 1.89$. Since $1.89 > 0.775$, we should not reject H_0. Using R command *1-pt(0.775, 7)*, the exact p-value is 0.2318.

(b) In order to make the test valid, we need the assumption that the population is normal.

5. (a) The standardized test statistic is

$$Z_{H_0} = \frac{40/500 - 0.05}{\sqrt{0.05 \times 0.95/500}} = 3.0779.$$

The RR is $Z_{H_0} < -z_\alpha = -z_{0.01} = -2.33$. Since $3.0779 > -2.33$, we should not reject H_0, and thus the traveler's club should be established.

(b) The p-value is calculated as 0.999 by using R command *pnorm(3.0779).*

6. (a) Let p be the proportion of all customers in states east of the Mississippi who prefer the bisque color, then the the null and alternative hypotheses are

$$H_0 : p \leq 0.3 \quad \text{vs.} \quad H_a : p > 0.3.$$

(b) The standardized test statistic is

$$Z_{H_0} = \frac{185/500 - 0.3}{\sqrt{0.3 \times 0.7/500}} = 3.416.$$

The RR is $Z_{H_0} > z_\alpha = z_{0.05} = 1.645$. Since $3.416 > 1.645$, we should reject H_0 at level 0.05.

(c) The p-value is calculated as 0.0003 by using R command *1-pnorm(3.416).* Since p-value is less than the significant level $\alpha = 0.01$, we should reject H_0.

7. (a) Let p be the proportion of all consumers who would be willing to try this new product, then the the null and alternative hypotheses are

$$H_0 : p \leq 0.2 \quad \text{vs.} \quad H_a : p > 0.2.$$

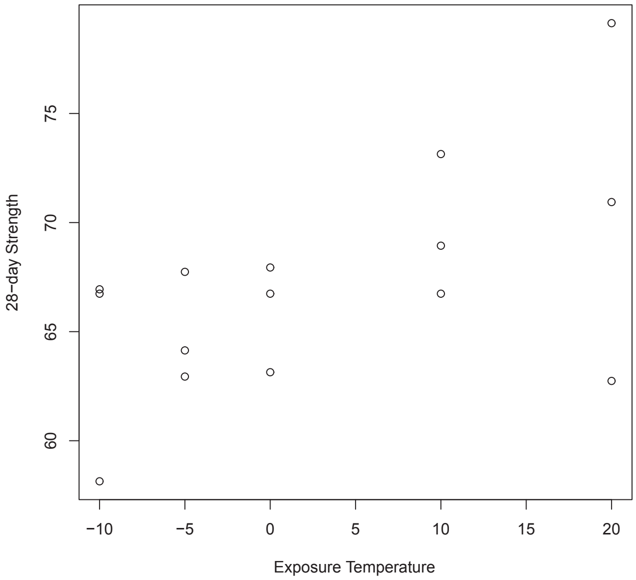
(b) The standardized test statistic is

$$Z_{H_0} = \frac{9/42 - 0.2}{\sqrt{0.2 \times 0.8/42}} = 0.23.$$

The RR is $Z_{H_0} > z_\alpha = z_{0.01} = 2.33$. Since $0.23 < 2.33$, we should not reject H_0. The p-value is calculated as 0.41 by using R command *1-pnorm(0.23).* Thus, there is not enough evidence that the marketing would be profitable.

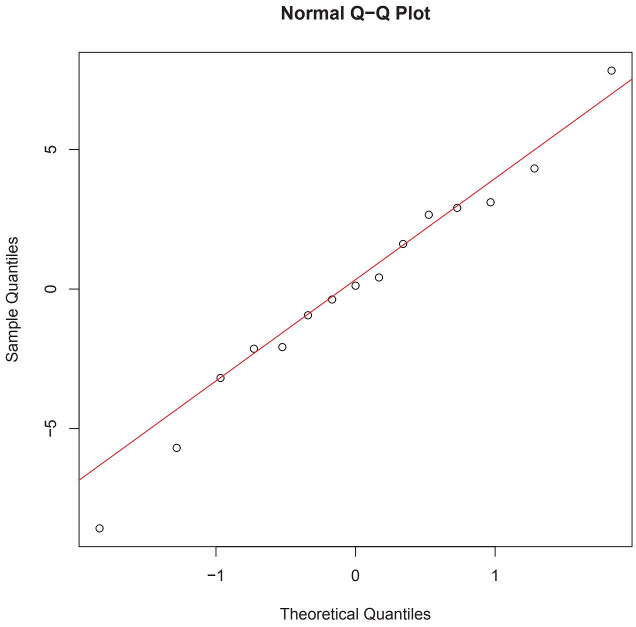
8. (a) The estimated regression line is $\hat{y} = 66.3251 + 0.2494x$.

   (b) The scatter plot is shown as



The scatterplot of the data shows that the linearity and homoscedasticity assumptions of the simple linear regression model seem to be valid.

The normal Q-Q plot of the residuals is shown as



The Q-Q plot shows that the normal assumption seems to be valid.

(c) From the command *anova(out)*, SSE $= 233.88$ and SSR $= 108.25$, and SST $=$ SSE $+$ SSR $= 233.88 + 108.25 = 342.13$. The percent of the total variability is explained by the regression model $R^2 =$ SSR/SST $= 108.25/342.13 = 0.3164$.

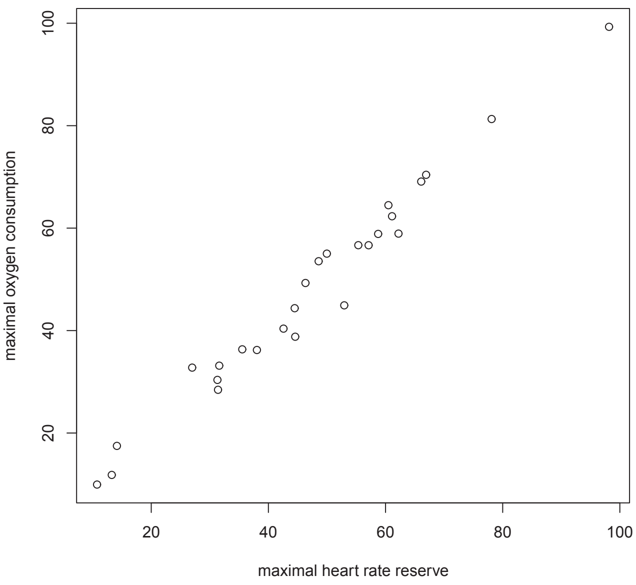
(d) The the value of the F statistic is 6.0169. The command *qf(0.95,1,13)* returns the $F_{1,13,0.05} = 4.667$. Since $F = 6.0169 > 4.667$, the null hypothesis is rejected at level 0.05.

(e) The null and alternative hypotheses are

$$H_0 : \beta_1 \leq 0 \quad \text{vs.} \quad H_a : \beta_1 > 0.$$

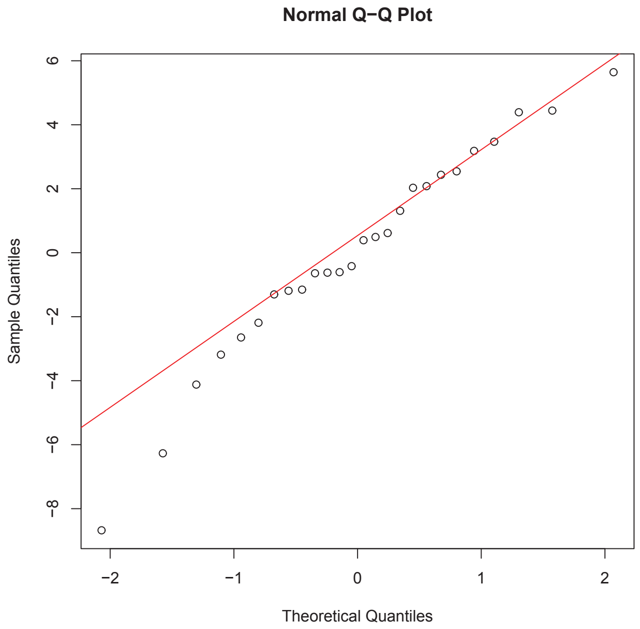
From the command *summary(out)*, the T value for the test statistic is 2.453. Thus, the p-value is returned by *1-pt(2.453, 13)*, which is 0.0145. Since the p-value is less than the significant level $\alpha = 0.05$, we would reject H_0 and conclude that the decrease in temperature would weaken the concrete.

9. (a) The scatter plot is shown as



The scatterplot of the data shows that the linearity and homoscedasticity assumptions of the simple linear regression model seem to be valid.

The normal Q-Q plot of the residuals is shown as



The Q-Q plot shows that the normal assumption seems to be valid, although there are several outliers.

(b) The completed ANOVA table is

	DF	Sum Sq	Mean Sq	F value	Pr
x	1	10630.6	10630.6	925.44	<2.2e-16
Residuals	24	275.7	11.5		

The proportion of the total variability in oxygen consumption is explained by the regression model is $R^2 = 0.9747$.

(c) The fitted regression line is $\hat{y} = -0.4021 + 1.0200x$. When the percentage of maximal heart rate reserve increases by 10 points, the estimated average oxygen consumption increases by 10.2.

(d)

(i) The null and alternative hypotheses are

$$H_0 : \beta_1 \leq 1 \quad \text{vs.} \quad H_a : \beta_1 > 1.$$

(ii) From the command *summary(out)*, we have $\hat{\beta}_1 = 1.02$ and $S_{\hat{\beta}_1} = 0.03353$. Thus, the value of the test statistic is

$$T_{H_0} = \frac{1.02 - 1}{0.03353} = 0.5965.$$

Thus, the p-value is returned by *1-pt(0.5965, 24)*, which is 0.278. Since the p-value is greater than the significant level $\alpha = 0.05$, we would not reject H_0.

10. (a) The sample size $n = 48 + 2 = 50$.

(b) The estimate of the standard deviation of the intrinsic error is $\sqrt{\text{MSE}} = \sqrt{11354/48} = 15.38$.

(c) For the test $H_0 : \alpha_1 = 0$ vs. $H_0 : \alpha_1 \neq 0$, the value of the T test statistic is -2.601 and the corresponding p-value is 0.0123. For the test $H_0 : \beta_1 = 0$ vs. $H_0 : \beta_1 \neq 0$, the value of the T test statistic is 9.464 and the corresponding p-value is 1.490×10^{-12}.

(d) The completed ANOVA table is

	DF	Sum Sq	Mean Sq	F value	Pr
x	1	21186	21186	89.56562	1.490e-12
Residuals	48	11354	236.5417		

(e) The proportion of the total variability of the stopping distance is explained by the regression model is $R^2 = 0.6511$.

11. (a) Let $\tilde{\mu}$ be the median income, then the null and alternative hypotheses are

$$H_0 : \tilde{\mu} = 300 \quad \text{vs} \quad H_a : \tilde{\mu} \neq 300.$$

The converted hypotheses are

$$H_0 : p = 0.5 \quad \text{vs} \quad H_a : p \neq 0.5.$$

Copy the data into R object x and use the command *sum(x>300)/length(x)*, giving us $\hat{p} = 0.3$. The test statistic

$$Z_{H_0} = \frac{\hat{p} - 0.5}{0.5/\sqrt{n}} = \frac{0.3 - 0.5}{0.5/\sqrt{20}} = -1.789.$$

The p-value is returned by *2*pnorm(-1.789)*, which is 0.0736. Since the p-value is greater than the significant level $\alpha = 0.05$, we would not reject H_0 and conclude that the data does not present strong enough evidence to conclude that the claim is false.

(b) Let $\tilde{\mu}$ be the median income, then the null and alternative hypotheses are

$$H_0 : \tilde{\mu} = 300 \quad \text{vs} \quad H_a : \tilde{\mu} < 300.$$

The converted hypotheses are

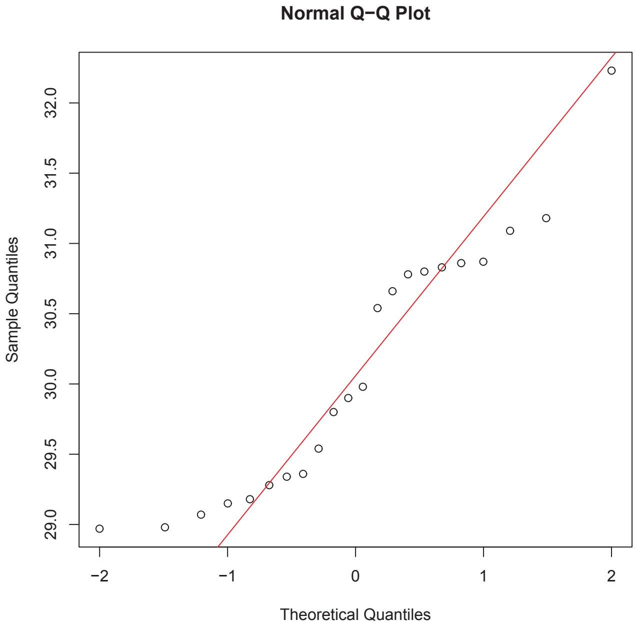
$$H_0 : p = 0.5 \quad \text{vs} \quad H_a : p < 0.5.$$

Copy the data into R object x and use the command *sum(x>300)/length(x)*, giving us $\hat{p} = 0.3$. The test statistic

$$Z_{H_0} = \frac{\hat{p} - 0.5}{0.5/\sqrt{n}} = \frac{0.3 - 0.5}{0.5/\sqrt{20}} = -1.789.$$

The p-value is returned by *pnorm(-1.789)*, which is 0.0368. Since the p-value is less than the significant level $\alpha = 0.05$, we would reject H_0 and conclude that the data presents strong enough evidence to conclude that the median increase is less than 300.

12. (a) The normal Q-Q plot of the data is shown as



The Q-Q plot shows that the data do not come from a normal distribution. We also notice that the dataset size is 22. Thus, it is not appropriate to test $H_0 : \mu = 28$ vs. $H_a : \mu > 28$. Instead, we should test the median, $H_0 : \tilde{\mu} = 28$ vs. $H_a : \tilde{\mu} > 28$.

(b) We first calculate $\hat{p}$ by the command *sum(r2>28)/length(r2)*, which is 1. The test statistic

$$Z_{H_0} = \frac{\hat{p} - 0.5}{0.5/\sqrt{n}} = \frac{1 - 0.5}{0.5/\sqrt{22}} = 4.69.$$

The p-value is returned by *1-pnorm(4.69)*, which is 1.37×10^{-6}. Since the p-value is less than the significant level $\alpha = 0.05$, we would reject H_0.

13. (a) The original hypotheses are

$$H_0 : x_{0.75} = 250 \quad \text{vs} \quad H_a : x_{0.75} < 250,$$

since according to the problem, we have $(1 - \pi)100 = 25$, thus $\pi = 0.75$. The original hypotheses are transformed to

$$H_0 : p = 0.75 \quad \text{vs} \quad H_a : p < 0.75,$$

where p is the probability that an observation is larger than 250. In this problem, we calculate $\hat{p}$ by the command *sum(x>250)/length(x)*, which returns 0.6875. Thus, the test statistic is

$$Z_{H_0} = \frac{\hat{p}-\pi}{\sqrt{\pi(1-\pi)/n}} = \frac{0.6875-0.75}{\sqrt{0.75\times 0.25/16}} = -0.577.$$

The p-value is calculated by *pnorm(-0.577)*, which returns 0.282. Since the p-value is greater than the significant level $\alpha = 0.05$, we cannot reject H_0; that is, there is not enough evidence to show the 25th percentile is smaller than 250.

(b) The original hypotheses are

$$H_0 : x_{0.25} = 104 \quad \text{vs} \quad H_a : x_{0.25} > 104,$$

since, according to the problem, we have $(1-\pi)100 = 75$, thus $\pi = 0.25$. The original hypotheses are transformed to

$$H_0 : p = 0.25 \quad \text{vs} \quad H_a : p > 0.25,$$

where p is the probability that an observation is larger than 104. In this problem, we calculate $\hat{p}$ by the command *sum(x>104)/length(x)*, which returns 0.3333. Thus, the test statistic is

$$Z_{H_0} = \frac{\hat{p}-\pi}{\sqrt{\pi(1-\pi)/n}} = \frac{0.3333-0.25}{\sqrt{0.25\times 0.75/36}} = 1.154.$$

The p-value is calculated by *1-pnorm(1.154)*, which returns 0.1243. Since the p-value is greater than the significant level $\alpha = 0.05$, we cannot reject H_0; that is, there is not enough evidence to show the 75th percentile of the systolic blood pressure is greater than 104.

14. (a) Let σ be the standard deviation of tread life span of the new tires, then the null and alternative hypotheses are

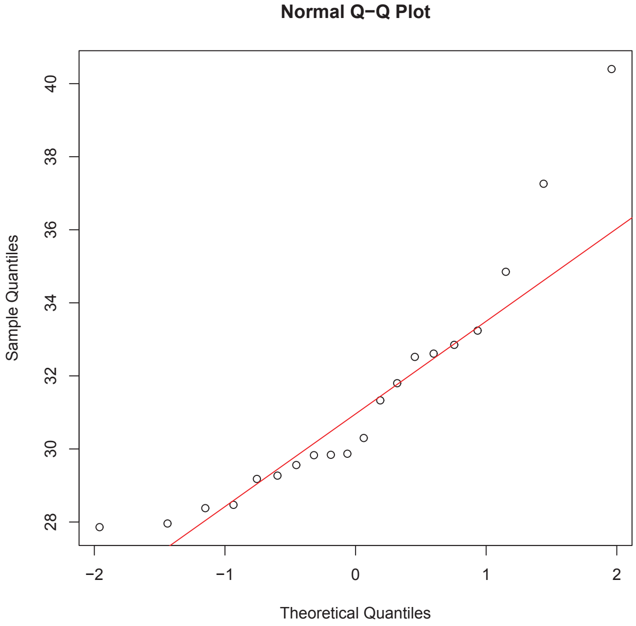
$$H_0 : \sigma \leq 2.5 \quad \text{vs} \quad H_a : \sigma > 2.5.$$

(b) The sample variance of the tread life spans of a sample of 20 new tires is calculated as 10.47185. To test the hypotheses in part (a), we calculate the value of the test statistic as

$$\chi^2_{H_0} = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(20-1)\times 10.47185}{2.5^2} = 31.83.$$

Since H_a is $\sigma > 2.5$, the RR is $\chi^2_{H_0} > \chi^2_{n-1,\alpha} = 32.85$. Thus, the calculated value is not in the rejection region and the null hypothesis should not be rejected. The p-value is calculated by the R command *1-pchisq(31.83, 19)*, which gives 0.033. We conclude that the new design should be adopted.

(c) The normal Q-Q plot of the data is shown as



The Q-Q plot shows that the data do not come from a normal distribution. Thus, the above conclusion might not be reliable.

15. In this problem, $n = 36$, $S = 0.25$, and $\sigma_0 = 0.2$. Thus the value of the test statistic is

$$\chi^2_{H_0} = \frac{(n-1)S^2}{\sigma_0^2} = \frac{(36-1) \times 0.25^2}{0.2^2} = 54.69.$$

Since H_a is $\sigma \neq 0.2$, the RR is $\chi^2_{H_0} > \chi^2_{n-1,\alpha/2} = 53.20$ or $\chi^2_{H_0} < \chi^2_{n-1,1-\alpha/2} = 20.57$. Thus, the calculated value is in the rejection region and the null hypothesis should be rejected. The p-value is calculated by the R command *2*min(pchisq(54.69,35),1-pchisq(54.69,35))*, which gives 0.036.

8.4 Precision in Hypothesis Testing

1. (a) When a null hypothesis is rejected, there is risk of committing Type I error.

   (b) When a null hypothesis is not rejected, there is risk of committing Type II error

2. (a) True

   (b) False

   (c) False

3. (a) To calculate Type I error, we have

$$\begin{aligned} P(\text{Type I Error}) &= P(H_0 \text{ is rejected when it is true}) \\ &= P(X \geq 8|p = 0.25, n = 20) = \sum_{k=8}^{20} \binom{20}{k} 0.25^k 0.75^{20-k}, \end{aligned}$$

because under this situation, the random variable X has a binomial distribution with $n = 20$ and $p = 0.25$. This probability can be calculated using R command *1-pbinom(7,20,0.25)*, which gives us 0.1018 as the probability of Type I error.

(b) We first calculate the probability of Type II error as

$$\begin{aligned} P(\text{Type II Error when } p = 0.3) &= P(H_0 \text{ is not rejected when } p = 0.3) \\ &= P(X < 8|p = 0.3, n = 20) = \sum_{k=0}^{7} \binom{20}{k} 0.3^k 0.7^{20-k}, \end{aligned}$$

because under this situation, the random variable X has a binomial distribution with $n = 20$ and $p = 0.3$. This probability can be calculated using R command *pbinom(7,20,0.3)*, which gives us 0.7723 as the probability of Type II error. Finally, the power is 1-0.7723 = 0.2277.

(c) When $n = 50$ and rejection region is $X \geq 17$, the probability of Type I error can be found by the command *1-pbinom(16,50,0.25)*, which gives us 0.0983. The power at $p = 0.3$ can be found by the command *1-pbinom(16,20,0.3)*, which gives us 0.316. We found that as the sample size increases, we have a smaller probability of Type I error and more power.

4. (a) The R command *1-pwr.t.test(36, (2-1)/4.1, 0.05, power=NULL, "one.sample", "greater")$power* returns 0.583 as the probability of Type II error when the true concentration is 2 ppm.

(b) The R command *pwr.t.test(n=NULL, (2-1)/4.1, 0.05, 0.99, "one.sample", alternative="greater")* returns a sample size of 266.46, which is rounded up to 267.

5. In this problem, the hypotheses tells us that $\mu_0 = 8.5$. We require that the probability of delivering a batch of acidity 8.65 should not exceed 0.05, thus, $\mu_a = 8.65$ and The type II error is 0.05, therefore the power is 0.95. We also know that the standard deviation from a preliminary study is 0.4, therefore, $S_{pr} = 0.4$. Combining this information, we use the commands *library(pwr); pwr.t.test(n=NULL, (8.65-8.5)/0.4, 0.05, 0.95, "one.sample", alternative="greater")*, which returns a sample size of 78.33 and is rounded up to 79.

6. (a) In this test, the testing statistic is

$$Z_{H_0} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

and because of $H_a : p > 0.2$, the rejection region is $Z_{H_0} > z_\alpha$. Thus, the probability of Type II error at $p_a = 0.25$ is

$$\begin{aligned}\beta(0.25) &= P(\text{Type II error}|p = 0.25) = P\left(\left.\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} < z_\alpha \right| p = 0.25\right)\\ &= P\left(\left.\hat{p} < p_0 + z_\alpha\sqrt{p_0(1-p_0)/n}\right| p = 0.25\right)\\ &= \Phi\left(\frac{p_0 + z_\alpha\sqrt{p_0(1-p_0)/n} - p}{\sqrt{p(1-p)/n}}\right).\end{aligned}$$

For the calculation, we use the command
*pnorm((0.2+qnorm(1-0.01)*sqrt(0.2*0.8/42)-0.25)/sqrt(0.25*0.75/42))*
and it gives 0.9193 as the probability of Type II error.

(b) To achieve power of 0.3 at $p_a = 0.25$ while keeping the level of significance at 0.01, we should use the commands *library(pwr); h=2*asin(sqrt(0.25))-2*asin(sqrt(0.2)); pwr.p.test(h, n=NULL, 0.01, 0.3, alternative="greater").* The code returns a sample size of 225.85 and is rounded up to 226.

7. (a) In this test, the testing statistic is

$$Z_{H_0} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

and because of $H_a : p < 0.05$, the rejection region is $Z_{H_0} < -z_\alpha$. Thus, the probability of Type II error at $p_a = 0.04$ is

$$\begin{aligned}\beta(0.04) &= P(\text{Type II error}|p = 0.04) = P\left(\left.\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > -z_\alpha \right| p = 0.04\right)\\ &= P\left(\left.\hat{p} > p_0 - z_\alpha\sqrt{p_0(1-p_0)/n}\right| p = 0.04\right)\\ &= 1 - \Phi\left(\frac{p_0 - z_\alpha\sqrt{p_0(1-p_0)/n} - p}{\sqrt{p(1-p)/n}}\right).\end{aligned}$$

For the calculation, we use the command
*1-pnorm((0.05-qnorm(1-0.01)*sqrt(0.05*0.95/500)-0.04)/sqrt(0.04*0.96/500))*
and it gives 0.926 as the probability of Type II error.

(b) To achieve power of 0.5 at $p_a = 0.04$ while keeping the level of significance at 0.01, we should use the following commands *library(pwr); h=2*asin(sqrt(0.04))-2*asin(sqrt(0.05)); pwr.p.test(h,n=NULL, 0.01, 0.5, alternative="less").* The code returns a sample size of 2318.77 and is rounded up to 2319.

Chapter 9

Comparing Two Populations

9.2 Two-Sample Tests and CIs for Means

1. Let μ_1 be the mean fatigue crack growth in aluminum with thickness of 3 mm and μ_2 be the mean fatigue crack growth in aluminum with thickness of 15 mm.

   (a) The null and alternative hypotheses are

   $$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2.$$

   In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

   $$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{15533^2}{3954^2} = 15.43 > 2.$$

   Thus, the assumption $\sigma_1^2 = \sigma_2^2$ does not hold and, consequently, (9.2.14) cannot be used.

   (b) We should use the statistic given in (9.2.15), with

   $$T_{H_0}^{SS} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{160592 - 159778}{\sqrt{\frac{3954^2}{36} + \frac{15533^2}{42}}} = 0.3275.$$

   The degrees of freedom are

   $$\nu = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \right] = \left[\frac{\left(\frac{3954^2}{36} + \frac{15533^2}{42} \right)^2}{\frac{(3954^2/36)^2}{36-1} + \frac{(15533^2/42)^2}{42-1}} \right] = [47.12] = 47.$$

   From R command *qt(1-0.025,47)* we get $t_{0.025,47} = 2.012$, thus we should not reject H_0 at 5% level. The p-value is calculated by the command *2*(1-pt(0.3275,47))*, which gives us 0.745. Since both the two samples have size greater than 30, there is no additional assumption for the validity of the test procedure.

(c) The 95% CI for the difference in the two means is given by (9.2.10):

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 160592 - 159778 \pm t_{47,0.05/2}\sqrt{\frac{3954^2}{36} + \frac{15533^2}{42}},$$

which gives the CI (-4187, 5815). Since 0 is included in the CI, H_0 could not be rejected.

2. Let μ_1 be the average penetration for material A and μ_2 be the average penetration for material B.

(a) The null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 \le 0.1 \quad \text{vs} \quad H_a : \mu_1 - \mu_2 > 0.1.$$

In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

$$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{0.19^2}{0.16^2} = 1.41 < 2.$$

Thus, the assumption $\sigma_1^2 = \sigma_2^2$ holds and, consequently, (9.2.14) can be used.

(b) We should use the statistic given in (9.2.14). We first calculate the pooled estimate of the variance as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(42 - 1) \times 0.19^2 + (42 - 1) \times 0.16^2}{42 + 42 - 2} = 0.03085$$

and the test statistic is

$$T_{H_0}^{EV} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.49 - 0.36 - 0.1}{\sqrt{0.03085 \times \left(\frac{1}{42} + \frac{1}{42}\right)}} = 0.7827.$$

The degree of freedom is $\nu = n_1 + n_2 - 2 = 82$. From R command *qt(1-0.05, 82)*, we get $t_{0.05,82} = 1.6636$, thus we should not reject H_0 at 5% level. The p-value is calculated by the command *1-pt(0.7827, 82)*, which gives us 0.218. Since both the samples have size greater than 30, there is no additional assumption for the validity of the test procedure.

(c) The 95% CI for the difference in the two means is given by (9.2.10):

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.49 - 0.36 \pm t_{82,0.05/2}\sqrt{0.03085 \times \left(\frac{1}{42} + \frac{1}{42}\right)},$$

which gives the CI (-0.0462, 0.1062).

3. Let μ_1 be the average delivery time for the standard route and μ_2 be that of the new route.

   (a) The null and alternative hypotheses are

   $$H_0 : \mu_1 \le \mu_2 \quad \text{vs} \quad H_a : \mu_1 > \mu_2.$$

   In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

   $$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{20.38^2}{15.62^2} = 1.70 < 2.$$

   Thus, the assumption $\sigma_1^2 = \sigma_2^2$ holds and, consequently, (9.2.14) can be used.

   (b) We should use the statistic given in (9.2.14). We first calculate the pooled estimate of the variance as

   $$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(48-1)\times 20.38^2 + (34-1)\times 15.62^2}{48+34-2} = 344.6584$$

   and the test statistic is

   $$T_{H_0}^{EV} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{432.7 - 403.5}{\sqrt{344.6584 \times \left(\frac{1}{48}+\frac{1}{34}\right)}} = 7.01684.$$

   The degrees of freedom are $\nu = n_1 + n_2 - 2 = 80$. From R command *qt(1-0.05, 80)*, we get $t_{0.05,80} = 1.664$. Since $T_{H_0}^{EV} > t_{0.05,80}$, we should reject H_0 at 5% level. The p-value is calculated by the command *1-pt(7.01684, 80)*, which gives us 3.292957×10^{-10}. Since both the samples have size greater than 30, there is no additional assumption for the validity of the test procedure.

   (c) The 99% CI for the difference in the two means is given by (9.2.10):

   $$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 432.7 - 403.5 \pm t_{80,0.01/2}\sqrt{344.6584 \times \left(\frac{1}{48}+\frac{1}{34}\right)},$$

   which gives the CI (18.22, 40.18).

   (d) The command *t.test(duration~route, data=dd, var.equal=T, alternative="greater")* gives the test statistic value $t = 7.0161$ and p-value $= 3.304 \times 10^{-10}$. The command *t.test(duration~route, data=dd, var.equal=T, conf.level=0.99)* gives the 99 percent confidence interval (18.21940, 40.18477). These are essentially the same as in part (a) and (c), considering the round-off errors.

4. Let μ_1 be the mean strength of new concrete at -8C and μ_2 be the mean strength of new concrete at 15C.

(a) The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2.$$

In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

$$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{4.92^2}{3.14^2} = 2.455 < 5.$$

Thus, the assumption $\sigma_1^2 = \sigma_2^2$ holds and, consequently, (9.2.14) can be used.

(b) We should use the statistic given in (9.2.14). We first calculate the pooled estimate of the variance as

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(9 - 1) \times 3.14^2 + (9 - 1) \times 4.92^2}{9 + 9 - 2} = 17.033$$

and the test statistic is

$$T_{H_0}^{EV} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{62.01 - 67.38}{\sqrt{17.033 \times \left(\frac{1}{9} + \frac{1}{9}\right)}} = -2.76.$$

The degrees of freedom are $\nu = n_1 + n_2 - 2 = 16$. From R command *qt(1-0.05, 16)*, we get $t_{0.05,16} = 1.746$. Since $|T_{H_0}^{EV}| > t_{0.05,16}$, we should reject H_0 at 10% level. The p-value is calculated by the command *2*pt(-2.76, 16)*, which gives us 0.01394434. In order to make the test procedure valid, we need the assumption that the samples are from a normal distribution.

(c) The 90% CI for the difference in the two means is given by (9.2.10):

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 62.01 - 67.38 \pm t_{16,0.1/2}\sqrt{17.033 \times \left(\frac{1}{9} + \frac{1}{9}\right)},$$

which gives the CI (-8.76668, -1.97332).

(d) The command *t.test(cs$Temp1, cs$Temp2, var.equal=T, conf.level=0.9)* gives the test statistic value $t = -2.7628$, p-value $= 0.01386$, and the 90 percent confidence interval (-8.772460, -1.978651). These are essentially the same as in part (a) and (c), considering the round-off errors.

5. Let μ_1 be the mean ultimate tensile strength (UTS) of holed specimens of 7075-T6 wrought aluminum and μ_2 be that of notched specimens.

(a) The null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 \leq 126 \quad \text{vs} \quad H_a : \mu_1 - \mu_2 > 126.$$

In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

$$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{52.12}{25.83} = 2.02 < 3.$$

Thus, the assumption $\sigma_1^2 = \sigma_2^2$ does hold and, consequently, (9.2.14) can be used.

(b) We should use the statistic given in (9.2.14). We first calculate the pooled estimate of the variance as

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(15-1)\times 52.12 + (15-1)\times 25.83}{15+15-2} = 38.975$$

and the test statistic is

$$T_{H_0}^{EV} = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{557.47 - 421.40 - 126}{\sqrt{38.975\times\left(\frac{1}{15}+\frac{1}{15}\right)}} = 4.417.$$

The degrees of freedom are $\nu = n_1 + n_2 - 2 = 28$. From R command *qt(1-0.05, 28)*, we get $t_{0.05,28} = 1.70$. Since $|T_{H_0}^{EV}| > t_{0.05,28}$, we should reject H_0 at 5% level. The p-value is calculated by the command *1-pt(4.417, 28)*, which gives us 6.81×10^{-5}. In order to make the test procedure valid, we need the assumption that the samples are from a normal distribution.

(c) The 95% CI for the difference in the two means is given by (9.2.10)

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 557.47 - 421.40 \pm t_{28,0.05/2}\sqrt{38.975\times\left(\frac{1}{15}+\frac{1}{15}\right)},$$

which gives the CI (132.19, 139.95).

(d) The command *t.test(utsUTS_Holed, utsUTS_Notched, mu=126, var.equal=T, alternative="greater")* implements that test statistic in (9.2.14) and it gives the test statistic value $t = 4.4159$ and p-value $= 6.83 \times 10^{-5}$. The command *t.test(utsUTS_Holed, utsUTS_Notched, mu=126, alternative="greater")* implements that test statistic in (9.2.15) and it gives the test statistic value $t = 4.4159$ and p-value $= 8.377 \times 10^{-5}$.

6. Let μ_1 be the mean mean full weight by machine 1 and μ_2 be that by machine 2.

(a) The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_a : \mu_1 \neq \mu_2.$$

In order to use (9.2.14) for the testing problem, we need to make sure that the assumption $\sigma_1^2 = \sigma_2^2$ holds. According to (9.2.9), there is

$$\frac{\max(S_1^2, S_2^2)}{\min(S_1^2, S_2^2)} = \frac{29.30}{26.24} = 1.12 < 3.$$

Thus, the assumption $\sigma_1^2 = \sigma_2^2$ does hold and, consequently, (9.2.14) can be used.

(b) To use the statistic given in (9.2.14), we first calculate the pooled estimate of the variance as

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(12-1)\times 29.30 + (12-1)\times 26.24}{12+12-2} = 27.77$$

and the test statistic is

$$T_{H_0}^{EV} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{966.75 - 962.33}{\sqrt{27.77\times\left(\frac{1}{12}+\frac{1}{12}\right)}} = 2.0545.$$

The degrees of freedom are $\nu = n_1+n_2-2 = 22$. From R command *qt(1-0.05/2, 22)*, we get $t_{0.025,22} = 2.074$. Since $|T_{H_0}^{EV}| < t_{0.025,22}$, we should not reject H_0 at 5% level. The p-value is calculated by the command *2*(1-pt(2.0545,22))*, which gives us 0.052.

To use the statistic given in (9.2.15), we calculate

$$T_{H_0}^{SS} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} = \frac{966.75 - 962.33}{\sqrt{\frac{29.30}{12}+\frac{26.24}{12}}} = 2.0545.$$

The degrees of freedom are

$$\nu = \left[\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}\right] = \left[\frac{\left(\frac{29.30}{12}+\frac{26.24}{12}\right)^2}{\frac{(29.30/12)^2}{12-1}+\frac{(26.24/12)^2}{12-1}}\right] = [21.93] = 21.$$

From R command *qt(1-0.05/2, 21)*, we get $t_{0.025,21} = 2.08$. Since $|T_{H_0}^{SS}| < t_{0.05,21}$, we should not reject H_0 at 5% level. The p-value is calculated by the command *2*(1-pt(2.0545,22))*, which gives us 0.053.

In order to make the test procedure valid, we need the assumption that the samples are from a normal distribution.

(c) The 95% CI for the difference in the two means is given by (9.2.10)

$$\bar{X}_1 - \bar{X}_2 \pm t_{\nu,\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} = 966.75 - 962.33 \pm t_{22,0.05/2}\sqrt{27.77\times\left(\frac{1}{12}+\frac{1}{12}\right)},$$

which gives the CI (-0.04164, 8.8816). Since 0 is included in the CI, we would not reject H_0 at 5% level. The result is the same as in part (b).

7. Let p_1 be the proportion of having the number four in the first post-decimal digit reported by firms with analyst coverage and p_2 be that reported by firms with no analyst coverage.

   (a) The null and alternative hypotheses are

   $$H_0 : p_1 = p_2 \quad \text{vs} \quad H_a : p_1 \neq p_2.$$

   We use the procedure given in (9.2.20) to test the hypotheses. We calculate

   $$\hat{p}_1 = \frac{692}{9396}, \quad \hat{p}_2 = \frac{1182}{13985}, \quad \text{and} \quad \hat{p} = \frac{692 + 1182}{9396 + 13985}.$$

   The test statistic is calculated as

   $$Z_{H_0}^P = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -3.001.$$

   The critical value is $z_{0.01/2} = 2.576$. Since $|Z_{H_0}^P| > z_{0.01/2}$, we should reject H_0 at 1% level. The p-value is calculated by *2*pnorm(-3.001)*, which returns 0.0027.

   (b) Using (9.2.11), the 99% CI for $p_1 - p_2$ is given by

   $$\hat{p}_1 - \hat{p}_2 \pm z_{0.01/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

   and the calculation gives (-0.0201, -0.0017).

   (c) We use the R command *prop.test(c(692, 1182), c(9396, 13985), correct=F, conf.level=0.99)* and it returns 0.0027 and (-0.0201, -0.0017) for the p-value and 99% CI, respectively.

8. Let p_1 be the success rate of tears greater than 25 millimeters and p_2 be that of the tears less than 25 millimeters.

   (a) The null and alternative hypotheses are

   $$H_0 : p_1 = p_2 \quad \text{vs} \quad H_a : p_1 \neq p_2.$$

   We use the procedure given in (9.2.20) to test the hypotheses. We calculate

   $$\hat{p}_1 = \frac{10}{18}, \quad \hat{p}_2 = \frac{22}{30}, \quad \text{and} \quad \hat{p} = \frac{10 + 22}{18 + 30} = \frac{32}{48}.$$

   The test statistic is calculated as

   $$Z_{H_0}^P = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -1.265.$$

The critical value is $z_{0.1/2} = 1.645$. Since $|Z_{H_0}^P| < z_{0.01/2}$, we should not reject H_0 at 1% level. The p-value is calculated by *2*pnorm(-1.265)*, which returns 0.2059.

(b) Using (9.2.11), the 90% CI for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{0.1/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and the calculation gives (-0.4118, 0.0562).

(c) We use the R command *prop.test(c(10, 22), c(18, 30), correct=F, conf.level=0.9)* and it returns 0.2059 and (-0.4118, 0.0562) for the p-value and 90% CI, respectively.

9. Let p_1 be the proportion of correctly identifying the signal from the right and p_2 be the proportion of correctly identifying the signal from the left.

(a) The null and alternative hypotheses are

$$H_0 : p_1 = p_2 \quad \text{vs} \quad H_a : p_1 \neq p_2.$$

We use the procedure given in (9.2.20) to test the hypotheses. We calculate

$$\hat{p}_1 = \frac{85}{100} = 0.85, \quad \hat{p}_2 = \frac{87}{100} = 0.87, \quad \text{and} \quad \hat{p} = \frac{85+87}{100+100} = 0.86.$$

The test statistic is calculated as

$$Z_{H_0}^P = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.85 - 0.87}{\sqrt{0.86 \times (1-0.86) \times \left(\frac{1}{100} + \frac{1}{100}\right)}} = 0.4076.$$

The critical value is $z_{0.01/2} = 2.576$. Since $|Z_{H_0}^P| < z_{0.01/2}$, we should not reject H_0 at 1% level. The p-value is calculated by *2*(1-pnorm(0.4076))*, which returns 0.6836.

(b) Using (9.2.11), the 99% CI for $p_1 - p_2$ is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{0.01/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and the calculation gives (-0.1463, 0.1063).

(c) We use the R command *prop.test(c(85, 87), c(100, 100), correct=F, conf.level=0.99)* and it returns 0.6836 and (-0.1463, 0.1063) for the p-value and 99% CI, respectively.

10. Let p_1 be the proportion of type A car sustained no visible damage in 10-mph crash test and p_2 be the proportion of type B car sustained no visible damage in 10-mph crash test.

    (a) The null and alternative hypotheses are

    $$H_0 : p_1 \le p_2 \quad \text{vs} \quad H_a : p_1 < p_2.$$

    We use the procedure given in (9.2.20) to test the hypotheses. We calculate

    $$\hat{p}_1 = \frac{19}{85}, \quad \hat{p}_2 = \frac{22}{85}, \quad \text{and} \quad \hat{p} = \frac{19+22}{85+85} = \frac{41}{170}.$$

    The test statistic is calculated as

    $$Z_{H_0}^P = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = -0.5378.$$

    The critical value is $z_{0.05/2} = 1.96$. Since $Z_{H_0}^P > -z_{0.05/2}$, we should not reject H_0 at 5% level. The p-value is calculated by *pnorm(-0.5378)*, which returns 0.2954.

    (b) Using (9.2.11), the 95% CI for $p_1 - p_2$ is given by

    $$\hat{p}_1 - \hat{p}_2 \pm z_{0.05/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}},$$

    and the calculation gives (-0.1431, 0.0726).

    (c) We use the R command *prop.test(c(19, 22), c(85, 85), alternative = "less", correct=F)* and it returns 0.2953 for the p-value. We use the R command *prop.test(c(19, 22), c(85, 85), correct=F, conf.level=0.9)* and it returns (-0.1431, 0.0726) for the 90% CI.

9.3 The Rank-Sum Test Procedure

1. Let $\tilde{\mu}_S$ be the median rainfall of the seeded clouds and $\tilde{\mu}_C$ be that of the control clouds. The null and alternative hypotheses are

   $$H_0 : \tilde{\mu}_S \le \tilde{\mu}_C \quad \text{vs} \quad \tilde{\mu}_S > \tilde{\mu}_C.$$

   The commands

   CSD = read.table("CloudSeedingData.txt", header = T);

   wilcox.test(CSD$Seeded, CSD$Control, alternative = "greater", conf.int=T) give a p-value of 0.007. So H_0 should be rejected at level 0.05. The command *wilcox.test(CSD$Seeded, CSD$Control, conf.int=T)* gives the 95% CI for the median of the difference in rainfall between a seeded and an unseeded cloud as (14.10, 237.60).

2. (a) Let $\tilde{\mu}_E$ be the median pollutant concentration on the east side of the lake and $\tilde{\mu}_W$ be that of the west side. The null and alternative hypotheses are

$$H_0 : \tilde{\mu}_E = \tilde{\mu}_W \quad \text{vs} \quad \tilde{\mu}_E \neq \tilde{\mu}_W.$$

The test procedure (9.3.5) is not recommended for this data set because the data set sizes for east side and west side are both less than 8.

(b) The commands

```
E = c(1.88, 2.60, 1.38, 4.41, 1.87, 2.89);
W= c(1.70, 3.84, 1.13, 4.97, 0.86, 1.93, 3.36);
wilcox.test(E, W, conf.int=T)
```

give the p-value 0.945 and (-1.98, 1.74) as the 95% CI for the median of the difference between a measurement from the eastern part of the lake and one from the western part. Thus, we should not reject H_0 at 0.1 level and conclude that there is not enough evidence to show that the pollutant concentration on the two sides of the lake is significantly different.

3. (a) Let $\tilde{\mu}_S$ be the median total strain amplitude of spheroidal graphite (SG) cast iron and $\tilde{\mu}_C$ be the median total strain amplitude of compacted graphite (CG) cast iron. The null and alternative hypotheses are

$$H_0 : \tilde{\mu}_S = \tilde{\mu}_C \quad \text{vs} \quad \tilde{\mu}_S \neq \tilde{\mu}_C.$$

To conduct the test procedure in (9.3.5), we use the commands

```
S = c(105, 77, 52, 27, 22, 17, 12, 14, 65); n1 = length(S);
C= c(90, 50, 30, 20, 14, 10, 60, 24, 76); n2 = length(C); N=n1+n2;
x=c(S, C); r=rank(x); w1=sum(r[1:n1]); w2=sum(r[n1+1:n2]);
s2r=sum((r-(N+1)/2)**2)/(N-1);
z=(w1/n1-w2/n2)/sqrt(s2r*(1/n1+1/n2)); z
```

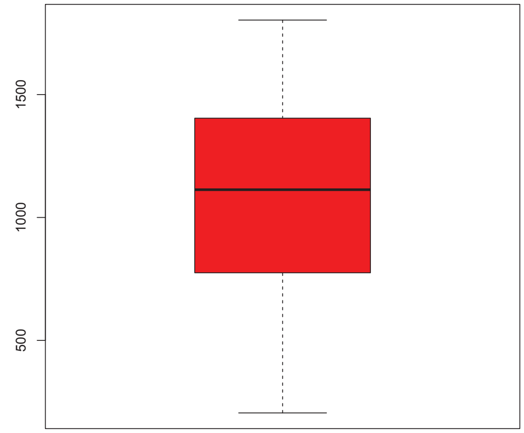
We get the z-value as $Z_{H_0} = 0.088$. The p-value is given by *2*(1-pnorm(0.088))* and it returns 0.93. Thus, we should not reject H_0 at level of significance 0.05 and conclude that there is not enough evidence to show that the total amplitude strain properties of the different types of cast iron significantly different.

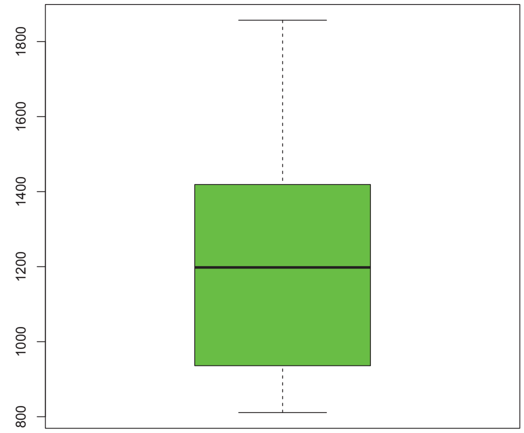
(b) The commands *wilcox.test(S, C, conf.int=T)* give the exact p-value 0.9648 and (-33.00, 38.00) as the 95% CI for the median of the difference between a measurement from SG and one from CG cast iron.

4. (a) We use commands

```
FL=read.table(“FemurLoads.txt”, header=T);
boxplot(FL$X2800lbs, col=“red”);
boxplot(FL$X3200lds, col=“green”)
```

to plot the boxplots in red and green color, respectively. They are shown below.





There is no outlier in the boxplots and the plots look symmetric. Therefore, it seems that the normality assumption is tenable.

(b) Let $\tilde{\mu}_1$ be the median Femur loads for type 1 vehicles and $\tilde{\mu}_1$ be that for type 2 vehicles. The null and alternative hypotheses are

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 \quad \text{vs} \quad \tilde{\mu}_1 \neq \tilde{\mu}_2.$$

The commands *wilcox.test(FL$X2800lbs, FL$X3200lds, conf.int=T)* give the exact p-value 0.4483 and (-469, 270) as the 95% CI for the median of the difference between a femur load measurement from a type 1 vehicle and one from a type 2 vehicle. Thus, we should not reject H_0 at the level 0.1 and conclude that there is not enough evidence to show that the Femur loads for the two types of cars significantly different.

5. For the t-test, the command *dd=read.table("DriveDurat.txt", header=T);*

   t.test(duration∼route, data=dd, var.equal=T, alternative="greater") gives p-value = 3.304×10^{-10}. For MWW rank-sum test, we use the command

   wilcox.test(duration∼route, alternative="greater", data=dd) and it gives the p-value 1.19×10^{-8}. The command *t.test(duration∼route, data=dd, var.equal=T, conf.level=0.9)* gives the 90 percent confidence interval (22.27571, 36.12846) corresponding to the t-test and the command *wilcox.test(duration∼route, conf.int=T, conf.level=0.9, data=dd)* gives the 90 percent confidence interval (22.80, 37.70) corresponding to the MWW rank-sum test.

9.4 Comparing Two Variances

1. Let σ_S^2 be the variance of total strain amplitude of spheroidal graphite (SG) cast iron and σ_C^2 be that of compacted graphite (CG) cast iron. The null and alternative hypotheses are

   $$H_0 : \sigma_S^2 = \sigma_C^2 \quad \text{vs} \quad \sigma_S^2 \neq \sigma_C^2.$$

   To conduct the Levene's test, we use the commands

   S = c(105, 77, 52, 27, 22, 17, 12, 14, 65); n1 = length(S);

   C= c(90, 50, 30, 20, 14, 10, 60, 24, 76); n2 = length(C);

   library(lawstat); x = c(S, C); ind = c(rep(1, n1), rep(2, n2)); levene.test(x, ind).

   The code gives us the p-value 0.7887, thus we cannot reject H_0 at the significant level 0.05.

2. Let σ_1^2 be the variance of fatigue crack growth in aluminum with thickness of 3 mm and σ_2^2 be that of aluminum with thickness of 15 mm. The null and alternative hypotheses are

   $$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad \sigma_1^2 \neq \sigma_2^2.$$

   To conduct the F-test, we calculate

   $$F_{H_0} = \frac{S_1^2}{S_2^2} = \frac{3954^2}{15533^2} = 0.0648 \quad \text{and} \quad \frac{1}{F_{H_0}} = 15.43.$$

The degrees of freedom for F_{H_0} are $\nu_1 = n_1 - 1 = 35$ and $\nu_2 = n_2 - 1 = 41$. Thus, $F_{35,41,0.025} = 1.895$ by the command *qf(0.975, 35, 41)* and $F_{41,35,0.025} = 1.93$ by the command *qf(0.975, 41, 35)*. Since $1/F_{H_0} > F_{41,35,0.025}$, we should reject H_0 at 5% level. To calculate the p-value, we use *2*min(1-pf(0.0648, 35, 41), 1-pf(15.43, 41, 35))* and it gives 4.93×10^{-13}. In order to make the test valid, we need to assume that the data are from normal populations.

3. Let σ_1^2 be the variance of delivery time for the standard route and σ_2^2 be that of the new route. The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{vs} \quad \sigma_1^2 \neq \sigma_2^2.$$

To conduct the F-test, we use the commands

dd=read.table("DriveDurat.txt", header=T); var.test(duration~route, data=dd).

The code gives us the p-value 0.1108, thus we cannot reject H_0 at the significance level 0.05. In order to make the test valid, we need to assume that the data are from normal populations.

9.5 Paired Data

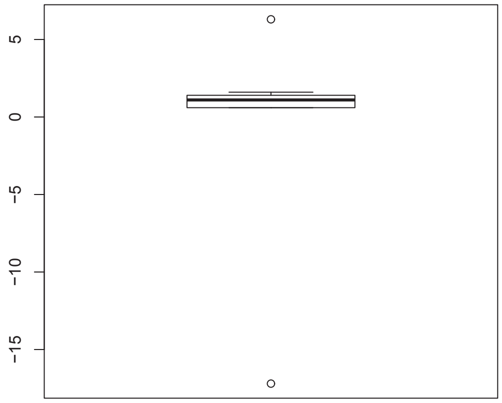
1. Let μ_A and μ_B be the average time required to parallel park type A and type B cars, respectively. The null and alternative hypotheses are

$$H_0 : \mu_A - \mu_B = 0 \quad \text{vs} \quad \mu_A - \mu_B \neq 0.$$

(a) To perform the paired T-test, we use the following commands:

```
A=c(19.0, 21.8, 16.8, 24.2, 22.0, 34.7, 23.8);
B=c(17.8, 20.2, 16.2, 41.4, 21.4, 28.4, 22.7);
t.test(A, B, paired = T)
```

The commands return 0.78 as the p-value, thus we should not reject H_0 at significant level 0.05. In order to make the test procedure valid, we need the assumption that the differences should be normally distributed. To check this assumption, we plot the boxplot of the differences by the command *boxplot(A-B)*, and it is shown on the next page.



This plot shows that there are two outliers and, thus, the normal assumption is suspect.

(b) To apply the signed-rank test, we use the command *wilcox.test(A, B, paired = T)* and it gives us p-value 0.271. Thus we should not reject H_0 at significance level 0.05.

2. Let μ_1 and μ_2 be the mean of the first test and of the second test, respectively. The null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad \mu_1 - \mu_2 \neq 0.$$

The commands *t1 = c(1.2, 1.3, 1.5, 1.4, 1.7, 1.8, 1.4, 1.3); t2 = c(1.4, 1.7, 1.5, 1.3, 2.0, 2.1, 1.7, 1.6); t.test(t1, t2, paired = T)* can be used to solve this problem.

(a) The returned p-value is 0.0103, which is less than 0.05. Thus, we should reject H_0 and conclude that the specialty steel manufacturer should adopt the second method.

(b) The 95% CI for the mean difference is (-0.3569, -0.0681).

3. Let μ_1 and μ_2 be the average percent of soil passing through the sieve for the two locations. Then, the null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad \mu_1 - \mu_2 \neq 0.$$

(a) We use commands *SDN = read.table("SoilDataNhi.txt", header = T); attach(SDN)* to read data. The command *t.test(Soil1, Soil2, paired = T)* gives us 0.037 as the p-value and (-5.490, -0.185) as the 95% CI for the difference of the population means. Thus, we should reject H_0 at the significant level 0.05.

(b) To use the signed-rank procedure, we use the command *wilcox.test(Soil1, Soil2, paired = T, conf.int = T)* and it gives us 0.037 as the p-value and (-5.450, -0.100) as the 95% CI for the difference of the population means. Comparing to part (a), we can see that the p-values and CIs from the two procedures are similar.

(c) The command for the t-test is *t.test(Soil1, Soil2)* and it gives us 0.215 as the p-value and (-7.364, 1.690) as the 95% CI for the difference of the population means. Thus, we should not reject H_0 at the significance level 0.05.

The command for signed-rank procedure is *wilcox.test(Soil1, Soil2, conf.int=T)* and it gives us 0.340 as the p-value and (-6.900, 2.300) as the 95% CI for the difference of the population means. Thus, we should not reject H_0 at the significance level 0.05.

Clearly, these results are very different from those in parts (a) and (b).

4. We use commands *LifeTime = read.table("McycleTiresLifeT.txt", header = T); attach(LifeTime)* to read data.

(a) Let μ_1 and μ_2 be the mean lifetimes of Brand 1 and Brand 2 tires, respectively. The null and alternative hypotheses are

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{vs} \quad \mu_1 - \mu_2 \neq 0.$$

To perform the paired t-test with the 90% CI, we use the command *t.test(Brand1, Brand2, paired=T, conf.level=0.9)* and it returns 0.0986 as the p-value and (4.2661, 1732.4839) as the 90% CI for the mean difference in lifetime. Since the p-value is greater than 0.05, we should not reject H_0 at 5% level. In order to make the test procedure valid, we need the assumption that the differences should be normally distributed.

(b) To use the signed-rank procedure, we use the command *wilcox.test(Brand1, Brand2, paired = T)* and it gives us 0.1094 as the p-value. Since the p-value is greater than 0.05, we should not reject H_0 at 5% level.

5. In this table, we have $n = 1325 + 3 + 13 + 59 = 1400$, $Y_2 = 3$, and $Y_3 = 13$, $\hat{q}_2 = 3/1400$, and $\hat{q}_3 = 13/1400$. According to (9.5.10) McNemar's test statistic is

$$MN = \frac{Y_2 - Y_3}{\sqrt{Y_2 + Y_3}} = \frac{3 - 13}{\sqrt{3 + 13}} = -2.5$$

and, according to (9.5.9), the paired t-test statistic is

$$T_{H_0} = \frac{\hat{q}_2 - \hat{q}_3}{\sqrt{(\hat{q}_2 + \hat{q}_3 - (\hat{q}_2 - \hat{q}_3)^2)/(n-1)}} = -2.505.$$

Because of the large sample size we use $z_{\alpha/2} = z_{0.025} = 1.96$ as the critical point. Since both $|MN|$ and $|T_{H_0}|$ are greater than 1.96, we reject the null hypothesis and conclude that the two algorithms have the same error rates.

6. In this table, we have $n = 260$, $Y_2 = 62$, and $Y_3 = 95$, $\hat{q}_2 = 62/260$, and $\hat{q}_3 = 95/260$. According to (9.5.10) McNemar's test statistic is

$$MN = \frac{Y_2 - Y_3}{\sqrt{Y_2 + Y_3}} = \frac{62 - 95}{\sqrt{62 + 95}} = -2.634$$

and, according to (9.5.9), the paired t-test statistic is

$$T_{H_0} = \frac{\hat{q}_2 - \hat{q}_3}{\sqrt{(\hat{q}_2 + \hat{q}_3 - (\hat{q}_2 - \hat{q}_3)^2)/(n-1)}} = -2.664.$$

Because of the large sample size we use $z_{\alpha/2} = z_{0.025} = 1.96$ as the critical point. Since both $|MN|$ and $|T_{H_0}|$ are greater than 1.96, we reject the null hypothesis and conclude that there was a change in voter attitude.

Chapter 10

Comparing $k > 2$ Populations

10.2 Types of k-Sample Tests

1. (a) Let μ_1, μ_2, μ_3, and μ_4 be the average tread lives of four types of truck tires, respectively. The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

We use the R command *TL = read.table("TireLife1Way.txt", header= T)* to read the data and use the commands *fit = aov(TL$values∼as.factor(TL$ind)); anova(fit)* to get the p-value as 0.1334, which is greater than the given significant level 0.1. Thus, the null hypothesis should not be rejected.

In order to make the test procedure valid, we need the assumptions that the samples are independent and from normal populations, and the population variances are equal.

(b)

(i) The contrast is $\theta = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$. To compare the two brands, the related null and alternative hypotheses are

$$H_0 : \theta = 0 \quad \text{vs} \quad H_a : \theta \neq 0.$$

(ii) To test the hypotheses, we use the commands

```
attach(TL); sm = by(values, ind, mean); svar = by(values, ind, var);
t = (sm[1]+sm[2]-sm[3]-sm[4])/2; st = sqrt(mean(svar)*(1/4*2/7*2));
t -qt(0.95, 24)*st; t+qt(0.95, 24)*st; TS = t/st; 2*(1-pt(abs(TS),24)).
```

The test statistic is given $T_{H_0} = -2.01$ and the corresponding p-value is 0.056, which is less than the given significant level 0.1. Thus, the null hypothesis H_0 should be rejected. The 90% CI is given as $(-1.31, -0.11)$.

(iii) The outcome of the test for the specialized contrast is NOT in agreement with the outcome of the test for the overall null hypothesis in part (a) because $\theta \neq 0$ means that $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ is not true. The t-test for a specialized contrast is more powerful than the F test for the equality of all means.

2. (a) The contrast is $\theta = (\mu_1 + \mu_3)/2 - \mu_2$. To test the hypotheses

$$H_0 : \theta = 0 \quad \text{vs} \quad H_a : \theta \neq 0,$$

we use the commands

fl = read.table("Flammability.txt", header = T); attach(fl);
sm=by(BurnL, Material, mean); sv=by(BurnL, Material, var);
t = (sm[1]+sm[3])/2-sm[2]; st = sqrt(mean(sv)(1/4*2/6+1*1/6));*
TS = t/st; 2(1-pt(abs(TS),15)).*

The commands give a p-value of 0.104. Thus, we should not reject the null hypothesis at significant level 0.05.

(b) The R commands given in the hint return a p-value of 0.035. Thus, we should reject the null hypothesis that the combined populations of materials 1 and 3 is the same as that of material 2.

3. (a) Let μ_1, μ_2, μ_3, and μ_4 be the average REM (rapid eye movement) sleep time of four concentrations of ethanol, respectively. The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

Using hand calculations, we have

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{4} = \frac{79.28 + 61.54 + 47.92 + 32.76}{4} = 55.375$$

and

$$SSTr = \sum_{i=1}^{4} n_i(\bar{X}_i - \bar{X})^2 = 5 \times (79.28 - 55.375)^2 + 5 \times (61.54 - 55.375)^2 + 5 \times (47.92 - 55.375)^2 + 5 \times (32.76 - 55.375)^2 = 5882.358,$$

thus

$$MSTr = \frac{SSTr}{k-1} = \frac{5882.358}{3} = 1960.786.$$

Finally,

$$F_{H_0} = \frac{MSTr}{MSE} = \frac{1960.786}{92.95} = 21.095.$$

The p-value is calculated by *1-pf(21.095,3, 20-4)*, which returns 8.318×10^{-6}. The null hypothesis should be rejected.

In order to make the test procedure valid, we need the assumptions that the samples are independent and from normal populations, and the population variances are equal.

(b) We use the R command *REM = read.table("SleepRem.txt", header= T); attach(REM)* to read the data, and use the commands *fit = aov(values∼as.factor(ind)); anova(fit)* to get the ANOVA table below.

	DF	SS	MS	F	P
as.factor(ind)IM	3	5881.7	1960.58	21.093	8.322×10^{-6}
Residuals	16	1487.1	92.95		

Clearly, the p-value is 8.322×10^{-6} and the null hypothesis should be rejected.

(c) To test the assumptions of equal variances, we use the command

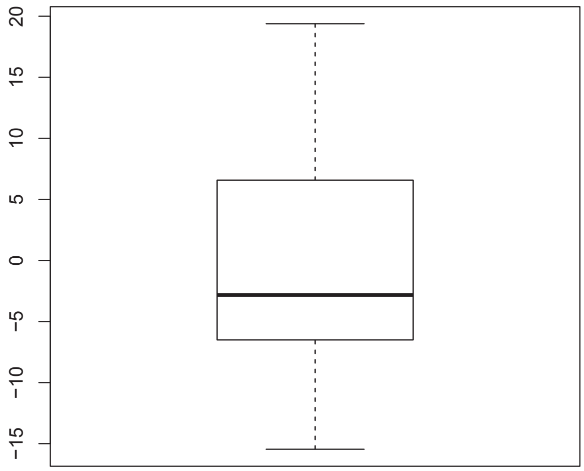
*anova(aov(resid(aov(values∼ind))**2∼ind)),*

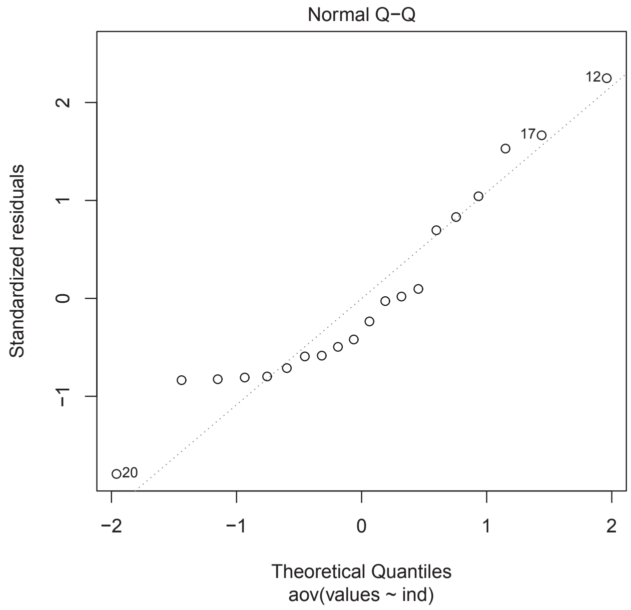
and it returns a p-value of 0.621. This suggests that the assumption of equal variance is approximately valid. To test normality, we use the command

shapiro.test(resid(aov(values∼ind))),

and it returns a p-value of 0.1285. This suggests that the normality assumption approximately holds.

We use the commands *fit = aov(values∼ind); boxplot(resid(fit))* to get the boxplot of the residuals and use the command *plot(fit, which=2)* to get the Q-Q plot for the residuals, shown below.





These plots also suggest that the normality assumption is approximately satisfied, in agreement with the Shapiro-Wilk test p-value.

4. (a) Use the commands *ranks=rank(values); rms=by(ranks, ind, mean)* to get the rank averages as 17.6, 12.6, 7.8, and 4. Thus, the Kruskal-Wallis statistic is calculated as

$$KW = \frac{12}{N(N+1)}\sum_{i=1}^{k} n_i\left(\bar{R}_i - \frac{N+1}{2}\right)^2 = \frac{12}{20\times 21}[(5(17.6-21/2)^2$$
$$+5(12.6-21/2)^2+5(7.8-21/2)^2+5(4-21/2)^2)] = 14.90857.$$

The degrees of freedom are $k-1=3$. Thus, the p-value can be found from the command *1-pchisq(14.90857,3)*, which gives 0.001896473. Since the p-value is less than the given significant level, the null hypothesis should be rejected.

To make the test valid, we need the assumption of continuous population.

(b) The command gives the Kruskal-Wallis statistic as 14.9086 and the p-value as 0.001896, which is less than the given significant level and thus, the null hypothesis should be rejected.

5. (a) Let μ_1, μ_2, and μ_3 be the average oxygen diffusivity at three mole fraction of water levels. The null and alternative hypotheses are

$$H_0: \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_a: H_0 \text{ is not true.}$$

The complete ANOVA table is given on the next page

	DF	Sum Sq	Mean Sq	F-Value
Treatment	2	0.019	0.0095	0.009179113
Residuals	24	24.839	1.034958	

(b) The critical value $F_{k-1,N-k,\alpha}$ can be calculated by *qf(0.95, 2, 24)*, which gives 3.40. Since $F_{H_0} < F_{k-1,N-k,\alpha}$, we should not reject H_0.

(c) The p-value is computed as *1-pf(0.009179113, 2, 24)*, which returns 0.9908664. Thus, the null hypothesis should not be rejected at level 0.05.

6. (a) Let μ_1, μ_2, μ_3, and μ_4 be the average pore size of carbon made at the four different temperatures. The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

Using hand calculations, we have

$$\bar{X} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \bar{X}_4}{4} = \frac{7.43 + 7.24 + 6.66 + 6.24}{4} = 6.8925$$

and

$$SSTr = \sum_{i=1}^{4} n_i(\bar{X}_i - \bar{X})^2 = 5 \times (7.43 - 6.8925)^2 + 5 \times (7.24 - 6.8925)^2 + 5 \times (6.66 - 6.8925)^2 + 5 \times (6.24 - 6.8925)^2 = 4.447375,$$

thus

$$MSTr = \frac{SSTr}{k-1} = \frac{4.447375}{3} = 1.482458,$$

$$MSE = S_p^2 = \frac{(n_1 - 1)S_1^2 + \cdots + n_k S_k^2}{n_1 + \cdots + n_k - k} = \frac{4 \times 0.2245 + 4 \times 0.143 + 4 \times 0.083 + 4 \times 0.068}{16} = 0.129625.$$

Finally,

$$F_{H_0} = \frac{MSTr}{MSE} = \frac{1.482458}{0.129625} = 11.43651.$$

The p-value is calculated by *1-pf(11.43651, 3, 20-4)*, which returns 0.000296. Thus, the null hypothesis should be rejected at significant level 0.05.

In order to make the test procedure valid, we need the assumptions that the samples are independent and from normal populations, and the population variances are equal.

(b) We use the R command *pc = read.table("PorousCarbon.txt", header= T)* to read the data and use the commands *fit = aov(pc$values∼as.factor(pc$temp)); anova(fit)* to get the F-value 11.437 and the p-value as 0.0002958, which is less than the given significant level 0.05. Thus, the null hypothesis should be rejected.

(c) To test the assumptions of equal variances, we use the command

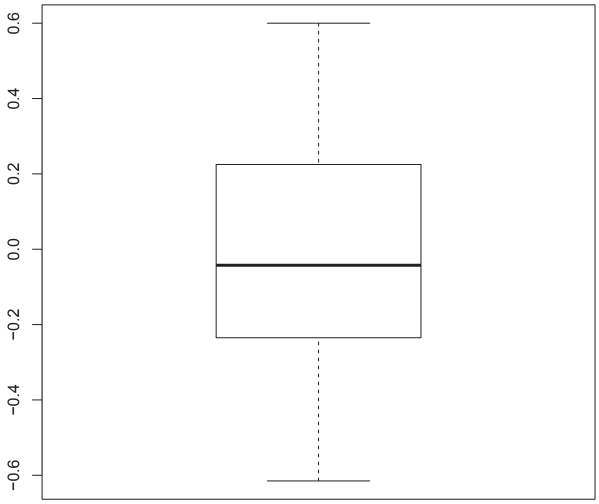
*attach(pc); anova(aov(resid(aov(values∼temp))**2∼temp))*

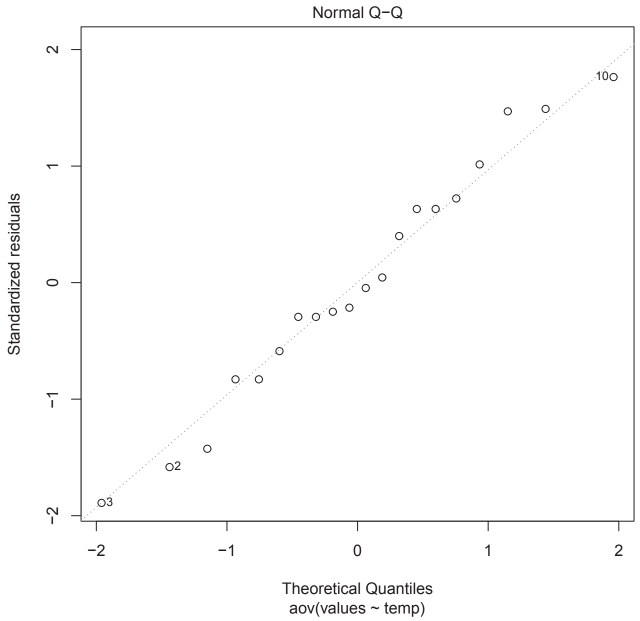
and it returns a p-value of 0.0656. This suggests that the assumption of equal variance is approximately valid. To test normality, we use the command

shapiro.test(resid(aov(values∼temp))),

and it returns a p-value of 0.8255. This suggests that the normality assumption approximately holds.

We use the commands *fit = aov(values∼temp); boxplot(resid(fit))* to get the boxplot of the residuals and use the command *plot(fit, which=2)* to get the Q-Q plot for the residuals, shown below.





These plots also suggest that the normality assumption is approximately satisfied, in agreement with the Shapiro-Wilk test p-value.

7. (a) Using the commands *attach(pc); ranks=rank(values); vranks=var(ranks); rms=by(ranks, temp, mean)*, we get $S^2_{KW} = 34.55263$. The rank averages are 15.9, 14.7, 7.8, and 3.6. Thus, the Kruskal-Wallis statistic is calculated as

$$KW = \frac{1}{S^2_{KW}} \sum_{i=1}^{k} n_i \left(\bar{R}_i - \frac{N+1}{2} \right)^2 = \frac{1}{34.55263}[(5(15.9 - 21/2)^2$$
$$+ 5(14.7 - 21/2)^2 + 5(7.8 - 21/2)^2 + 5(3.6 - 21/2)^2)] = 14.71668.$$

The degrees of freedom are $k - 1 = 3$. Thus, the p-value can be found from the command *1-pchisq(14.71668,3)*, which gives 0.0021. Since the p-value is less than the given significant level, the null hypothesis should be rejected.

To make the test valid, we need the assumption of continuous population.

(b) The command *kruskal.test(values∼temp)* gives the Kruskal-Wallis statistic as 14.7167 and the p-value of 0.002075. Thus, the null hypothesis should be rejected at level 0.05.

8. Let μ_1, μ_2, and μ_3 be average strength using fixed-platen testers for types 1, 2, and 3 of corrugated containers. The null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

From the given information, we calculate

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{3} n_i \bar{X}_i = \frac{1}{36 + 49 + 42}(36 \times 754 + 49 \times 769 + 42 \times 776) = 767.063$$

and

$$SSTr = \sum_{i=1}^{3} n_i(\bar{X}_i - \bar{X})^2 = 36 \times (754 - 767.063)^2$$
$$+ 49 \times (769 - 767.063)^2 + 42 \times (776 - 767.063)^2 = 9681.496.$$

Thus,

$$MSTr = \frac{SSTr}{k-1} = \frac{9681.496}{3-1} = 4840.748.$$

The mean square error is

$$\begin{aligned} MSE &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-k} \\ &= \frac{(36-1)\times 16^2 + (49-1)\times 27^2 + (42-1)\times 38^2}{36+49+42-3} = 831.9032. \end{aligned}$$

Finally, the F statistic is

$$F_{H_0} = \frac{MSTr}{MSE} = \frac{4840.748}{831.9032} = 5.818884$$

and the degrees of freedom are $k-1 = 2$ and $N-k = 36+49+42-3 = 124$. $F_{k-1,N-k,\alpha}$ can be calculated by *qf(1-0.05, 2, 124)*, which is 3.07. Clearly, $F_{H_0} > F_{k-1,N-k,\alpha}$. The p-value is calculated by *1-pf(5.818884, 2, 124)* which returns 0.003841892. Thus, we should reject the null hypothesis at 0.05 level.

9. (a) To test the assumptions of equal variances, we use the command

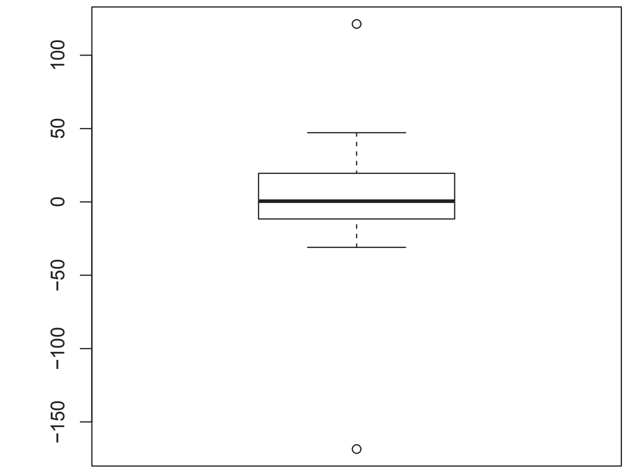
ff= read.table("FlexFatig.txt", header = T); attach(ff);
*anova(aov(resid(aov(values∼ind))**2∼ind))*

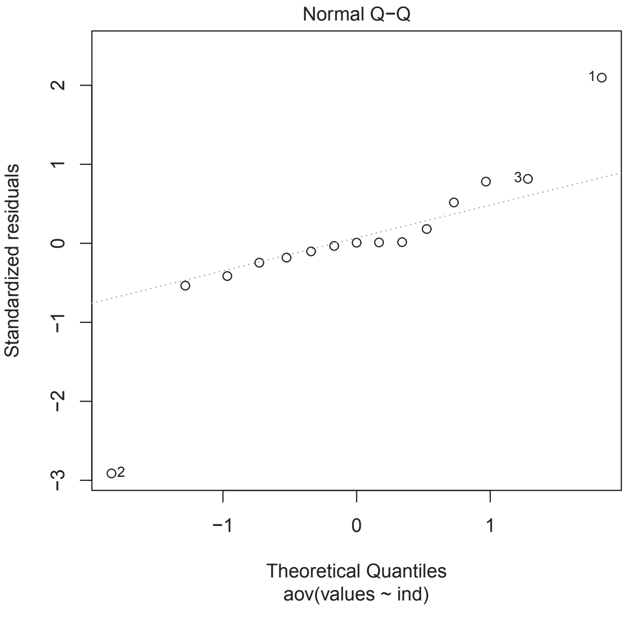
and it returns p-value 0.0407. This suggests that the assumption of equal variance is suspicious. To test normality, we use the command

shapiro.test(resid(aov(values∼ind)))

and it returns a p-value of 0.008378. This suggests that the normality assumption does not hold.

We use the commands *fit = aov(values∼ind); boxplot(resid(fit))* to get the boxplot of the residuals and use the command *plot(fit, which=2)* to get the Q-Q plot for the residuals, shown on the next page.





These plots also suggest that the normality assumption is not satisfied, in agreement with the Shapiro-Wilk test p-value.

(b) Kruskal-Wallis test can be used for this problem because it is applicable with both small and large sample sizes regardless of the normality assumption.

(c) The command *kruskal.test(values∼ind)* gives the Kruskal-Wallis statistic 11.6167 and the p-value as 0.02044. Thus, the null hypothesis should be rejected at level 0.01.

10. Let p_1, p_2, and p_3 be the probability of ignition for all three materials. The null and alternative hypotheses are

$$H_0 : p_1 = p_2 = p_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

To test the hypotheses, we use the R commands *table=matrix(c(37, 74, 28, 57, 21, 79), nrow=2); chisq.test(table)*. The commands return the test statistic as 4.7562 and a p-value of 0.09273, which is greater than the given significant level 0.05, thus the null hypothesis should not be rejected.

11. Let p_1, p_2, $\cdots$, p_5 be the proportion of tractors that require warranty repair work for the five locations. The null and alternative hypotheses are

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

Using the χ^2-test for this problem, we have

$$\hat{p}_1 = 18/50 = 0.36, \hat{p}_2 = 8/50 = 0.16, \hat{p}_3 = 21/50 = 0.42$$

and

$$\hat{p}_4 = 16/50 = 0.32, \hat{p}_5 = 13/50 = 0.26.$$

The overall proportion is

$$\hat{p} = \sum_{i=1}^{5} \frac{n_i}{N}\hat{p}_i = \frac{18+8+21+16+13}{50 \times 5} = 0.304.$$

The test statistic is

$$Q_{H_0} = \sum_{i=1}^{5} \frac{n_i(\hat{p}_i - \hat{p})^2}{\hat{p}(1-\hat{p})} = \frac{50}{0.304(1-0.304)}[(0.36-0.304)^2 + (0.16-0.304)^2$$
$$+ (0.42-0.304)^2 + (0.32-0.304)^2 + (0.26-0.304)^2] = 9.33908.$$

The degrees of freedom are $k-1=4$, thus the p-value can be calculated by the command *1 - pchisq(9.33908, 4)*, which returns 0.05316091. The p-value is greater than the given significant level 0.05 and, therefore, we cannot reject H_0.

The R commands *table=matrix(c(18, 32, 8, 42, 21, 29, 16, 34, 13, 37), nrow=2); chisq.test(table)* return the same value for the test statistic and p-value.

To make the testing procedure valid, we need the assumption that the samples are independent.

12. Let p_i be the probability that inner glass ply breaks (IPBs) for configuration i, $i = 1, 2, 3, 5$. The null and alternative hypotheses are

$$H_0 : p_1 = p_2 = p_3 = p_5 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

To test the hypotheses, we use the R commands *table=matrix(c(91, 14, 128, 20, 46, 41, 62, 31), nrow=2); chisq.test(table)*. The commands return the test statistic as 44.7619 and a p-value of 1.04×10^{-9}, which leads to the null hypothesis being rejected.

10.3 Simultaneous CIs and Multiple Comparisons

1. (a) Let μ_S, μ_C, and μ_G be the mean total strain amplitude properties of the different types of cast iron. The null and alternative hypotheses are

$$H_0 : \mu_S = \mu_C = \mu_G \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

(b) The command gives the p-value as 0.3158 and thus, the null hypothesis should not be rejected at level 0.05.

(c) Since H_0 is not rejected, there is no need to conduct multiple comparisons to determine which pairs of populations differ at experiment-wise level of significance 0.05.

2. (a) There are $m = 3$ contrasts and therefore, we should test each of them at level $\alpha/m = 0.05/3 = 0.0167$.

For $H_{10} : \sigma_1^2 = \sigma_2^2$, we calculate the F-statistic as

$$F_{H_{10}} = \frac{S_1^2}{S_2^2} = \frac{16^2}{27^2} = 0.3512.$$

The degrees of freedom are $n_1 - 1 = 35$ and $n_2 - 1 = 48$. Thus, the p-value is calculated by the command *2*min(1-pf(0.3512, 35, 48), 1-pf(1/0.3512, 48, 35))*, which is 0.00167, and we should reject H_{10}.

For $H_{20} : \sigma_1^2 = \sigma_3^2$, we calculate the F-statistic as

$$F_{H_{20}} = \frac{S_1^2}{S_3^2} = \frac{16^2}{38^2} = 0.1773.$$

The degrees of freedom are $n_1 - 1 = 35$ and $n_3 - 1 = 41$. Thus, the p-value is calculated by the command *2*min(1-pf(0.1773, 35, 41), 1-pf(1/0.1773, 41, 35))*, which is 9.78×10^{-7}. Thus, we should reject H_{20}.

For $H_{30} : \sigma_2^2 = \sigma_3^2$, we calculate the F-statistic as

$$F_{H_{30}} = \frac{S_2^2}{S_3^2} = \frac{27^2}{38^2} = 0.5048.$$

The degrees of freedom are $n_2 - 1 = 48$ and $n_3 - 1 = 41$. Thus, the p-value is calculated by the command *2*min(1-pf(0.5048, 48, 41), 1-pf(1/0.5048, 41, 48))*, which is 0.0234. Thus, we should not reject H_{30}.

Summarizing the results of the testing, we conclude that the homoscedasticity assumption does not hold.

(b) We need to test three hypotheses, $H_{10} : \mu_1 = \mu_2$, $H_{20} : \mu_1 = \mu_3$, and $H_{30} : \mu_2 = \mu_3$, and we should test each of them at level $\alpha/m = 0.05/3 = 0.0167$.

For $H_{10} : \mu_1 = \mu_2$, we calculate

$$T_{H_{10}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{754 - 769}{\sqrt{\frac{16^2}{36} + \frac{27^2}{49}}} = -3.199.$$

The degrees of freedom are

$$\nu = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}} \right] = \left[\frac{\left(\frac{16^2}{36} + \frac{27^2}{49} \right)^2}{\frac{(16^2/36)^2}{36-1} + \frac{(27^2/49)^2}{49-1}} \right] = [79.84] = 79.$$

Thus, the p-value is calculated by *2*pt(-3.199, 79)*, which returns 0.002. Therefore, we should reject H_{10}.

For $H_{20} : \mu_1 = \mu_3$, we calculate

$$T_{H_{20}} = \frac{\bar{X}_1 - \bar{X}_3}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_3^2}{n_3}}} = \frac{754 - 776}{\sqrt{\frac{16^2}{36} + \frac{38^2}{42}}} = -3.415.$$

The degrees of freedom are

$$\nu = \left[\frac{\left(\frac{S_1^2}{n_1} + \frac{S_3^2}{n_3} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_3^2/n_3)^2}{n_3-1}} \right] = \left[\frac{\left(\frac{16^2}{36} + \frac{38^2}{42} \right)^2}{\frac{(16^2/36)^2}{36-1} + \frac{(38^2/42)^2}{42-1}} \right] = [56.86] = 56.$$

Thus, the p-value is calculated by *2*pt(-3.415, 56)*, which returns 0.0012. Therefore, we should reject H_{20}.

For $H_{30} : \mu_2 = \mu_3$, we calculate

$$T_{H_{30}} = \frac{\bar{X}_2 - \bar{X}_3}{\sqrt{\frac{S_2^2}{n_2} + \frac{S_3^2}{n_3}}} = \frac{769 - 776}{\sqrt{\frac{27^2}{49} + \frac{38^2}{42}}} = -0.9974.$$

The degrees of freedom are

$$\nu = \left[\frac{\left(\frac{S_2^2}{n_2} + \frac{S_3^2}{n_3} \right)^2}{\frac{(S_2^2/n_2)^2}{n_2-1} + \frac{(S_3^2/n_3)^2}{n_3-1}} \right] = \left[\frac{\left(\frac{27^2}{49} + \frac{38^2}{42} \right)^2}{\frac{(27^2/49)^2}{49-1} + \frac{(38^2/42)^2}{42-1}} \right] = [72.56] = 72.$$

Thus, the p-value is calculated by *2*pt(-0.9974, 72)*, which returns 0.3219. Therefore, we should not reject H_{30}.

Summarizing the results of the testing, we conclude that at experiment-wise significant level 0.05, μ_2 and μ_3 are not significantly different while the pairs (μ_1, μ_2) and (μ_1, μ_3) are significantly different.

3. (a) Let μ_1, μ_2 and μ_3 be the average spark plug resistance at the three different blow-off pressures, then the null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

From the given information, we calculate

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{3} n_i \bar{X}_i = \frac{1}{150+150+150}(150\times 5.365 + 150\times 5.415 + 150\times 5.883) = 5.554$$

and

$$\begin{aligned} SSTr = \sum_{i=1}^{3} n_i(\bar{X}_i - \bar{X})^2 &= 150 \times (5.365 - 5.554)^2 \\ &+ 150 \times (5.415 - 5.554)^2 + 150 \times (5.883 - 5.554)^2 = 24.492. \end{aligned}$$

Thus,

$$MSTr = \frac{SSTr}{k-1} = \frac{24.492}{3-1} = 12.246.$$

The mean square error is

$$\begin{aligned} MSE &= \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + (n_3-1)S_3^2}{n_1+n_2+n_3-k} \\ &= \frac{(150-1)\times 2.241 + (150-1)\times 1.438 + (150-1)\times 1.065}{150+150+150-3} = 1.581. \end{aligned}$$

Finally, the F statistic is

$$F_{H_0} = \frac{MSTr}{MSE} = \frac{12.246}{1.581} = 7.746$$

and the degrees of freedom are $k-1 = 2$ and $N-k = 150+150+150-3 = 447$. $F_{k-1,N-k,\alpha}$ can be calculated by *qf(1-0.05, 2, 447)*, which is 3.016. Clearly, $F_{H_0} > F_{k-1,N-k,\alpha}$. The p-value is calculated by *1-pf(7.746, 2, 447)* which returns 0.00049. Thus, we should reject the null hypothesis at 0.05 level.

To make the test procedure valid, we need the assumption that the samples are independent and the population variances are equal.

(b) For three difference pressures, we have $m = 3$ mean contrasts: $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$. For each of the three contrasts, we need to test the null hypothesis that the contrast is zero vs the alternative that it is not zero at level of significance $\alpha/m = 0.03/3 = 0.0167$.

For $\mu_1 - \mu_2$:

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(150-1)2.241 + (150-1)1.438}{150+150-2} = 1.8395$$

and

$$T_{H_0} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{5.365 - 5.415}{\sqrt{1.8395\left(\frac{1}{150} + \frac{1}{150}\right)}} = -0.3192646.$$

The p-value is calculated by *2*pt(-0.3192646, 150+150-2)*, which is 0.7497496, which is greater than 0.0167. Thus H_0 should not be rejected and we conclude that μ_1 and μ_2 are not significantly different.

For $\mu_1 - \mu_3$:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_3 - 1)S_3^2}{n_1 + n_3 - 2} = \frac{(150 - 1)2.241 + (150 - 1)1.065}{150 + 150 - 2} = 1.653$$

and

$$T_{H_0} = \frac{\bar{X}_1 - \bar{X}_3}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_3}\right)}} = \frac{5.365 - 5.883}{\sqrt{1.653\left(\frac{1}{150} + \frac{1}{150}\right)}} = -3.489185.$$

The p-value is calculated by *2*pt(-3.489185, 150+150-2)*, which is 0.00056, which is less than 0.0167. Thus H_0 should be rejected and we conclude that μ_1 and μ_3 are significantly different.

For $\mu_2 - \mu_3$:

$$S_p^2 = \frac{(n_2 - 1)S_2^2 + (n_3 - 1)S_3^2}{n_2 + n_3 - 2} = \frac{(150 - 1)1.438 + (150 - 1)1.065}{150 + 150 - 2} = 1.2515$$

and

$$T_{H_0} = \frac{\bar{X}_2 - \bar{X}_3}{\sqrt{S_p^2\left(\frac{1}{n_2} + \frac{1}{n_3}\right)}} = \frac{5.415 - 5.883}{\sqrt{1.2515\left(\frac{1}{150} + \frac{1}{150}\right)}} = -3.622939.$$

The p-value is calculated by *2*pt(-3.622939, 150+150-2)*, which is 0.000342, which is less than 0.0167. Thus H_0 should be rejected and we conclude that μ_2 and μ_3 are significantly different.

4. Let μ_1, μ_2, μ_3, and μ_4 be the average pore size of carbon made at the four different temperatures (300, 400, 500, 600).

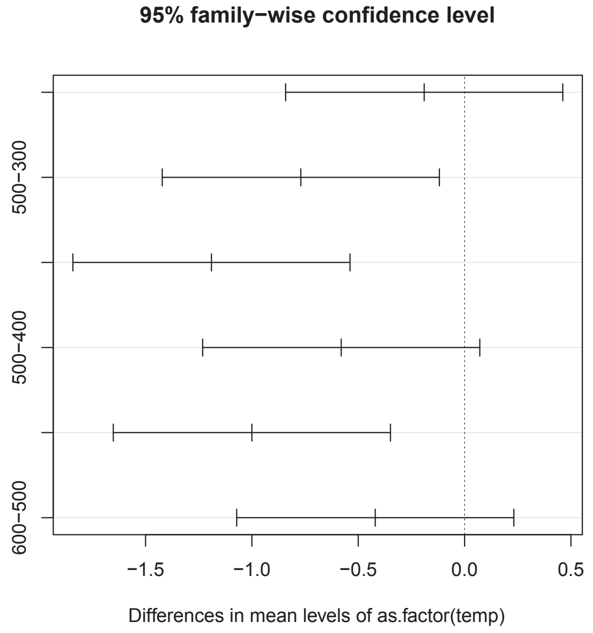
   (a) To compute the Tukey's 95% simultaneous CIs, we use the following commands

   pc = read.table("PorousCarbon.txt", header= T); attach(pc);
   TukeyHSD(aov(values∼as.factor(temp)), conf.level=0.95).

   The results are summarized in the table on the next page.

Contrast	Simultaneous 95% CI	Contains 0?
$\mu_2 - \mu_1$	(-0.841, 0.461)	Yes
$\mu_3 - \mu_1$	(-1.421, -0.119)	No
$\mu_4 - \mu_1$	(-1.841, -0.539)	No
$\mu_3 - \mu_2$	(-1.231, 0.071)	Yes
$\mu_4 - \mu_2$	(-1.651, -0.349)	No
$\mu_4 - \mu_3$	(-1.071, 0.231)	Yes

We see that the pairs (μ_3, μ_1), (μ_4, μ_1), and (μ_4, μ_2) are significant at the experiment-wise level of significant 0.05. The plot is given below.



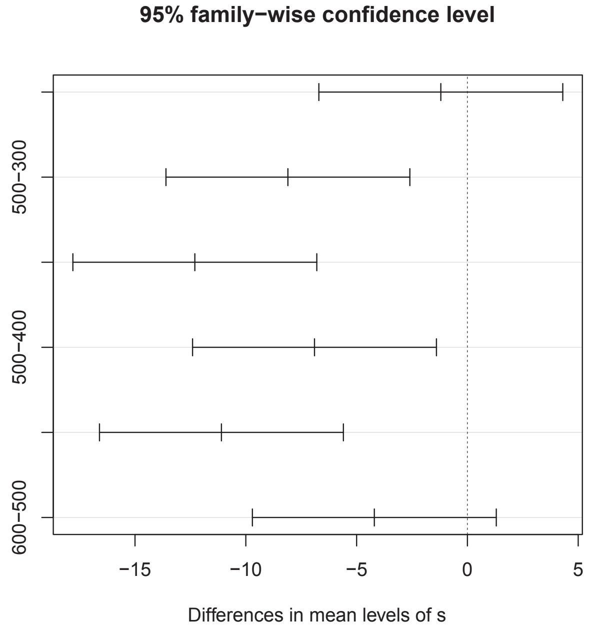
(b) To perform Tukey's multiple comparisons procedure on the ranks, we use the following

r=rank(values); s=as.factor(temp); TukeyHSD(aov(r∼s), conf.level=0.95).

The results are summarized in the table below.

Contrast	Simultaneous 95% CI	Contains 0?
$\tilde{\mu}_2 - \tilde{\mu}_1$	(-6.703, 4.303)	Yes
$\tilde{\mu}_3 - \tilde{\mu}_1$	(-13.603, -2.597)	No
$\tilde{\mu}_4 - \tilde{\mu}_1$	(-17.803, -6.797)	No
$\tilde{\mu}_3 - \tilde{\mu}_2$	(-12.403, -1.397)	No
$\tilde{\mu}_4 - \tilde{\mu}_2$	(-16.603, -5.597)	No
$\tilde{\mu}_4 - \tilde{\mu}_3$	(-9.703, 1.303)	Yes

We see that the pairs ($\tilde{\mu}_3$, $\tilde{\mu}_1$), ($\tilde{\mu}_4$, $\tilde{\mu}_1$), ($\tilde{\mu}_3$, $\tilde{\mu}_2$), and ($\tilde{\mu}_4$, $\tilde{\mu}_2$) are significant at the experiment-wise level of significant 0.05. The plot is given on the next page.



(c) Omitted

5. (a) Let μ_1, μ_2 and μ_3 be the average scores for the three teaching methods, then the null and alternative hypotheses are

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

To perform the ANOVA test, we use the following commands:

GTM= read.table("GradesTeachMeth.txt", header = T); attach(GTM);
anova(aov(score∼as.factor(method))).

The commands give the p-value as 0.01773, which is less than the given significant level 0.05. Thus, the null hypothesis should be rejected.

To make the testing procedure valid, we need the assumptions that the samples are independent and the populations are normal with equal variances.

(b) To construct Tukey's 95% simultaneous CIs for all pairwise contrasts, we use the command *TukeyHSD(aov(score∼as.factor(method)), conf.level=0.95)*. The simultaneous CIs for $\mu_2 - \mu_1$, $\mu_3 - \mu_1$, and $\mu_3 - \mu_2$ are (-3.72, 14.97), (2.28, 20.97), and (-3.35, 15.35), respectively. The result shows that the teaching methods 1 and 3 are significantly different.

(c) To test the assumptions of equal variances, we use the command

*anova(aov(resid(aov(score∼as.factor(method)))**2∼as.factor(method))),*

and it returns p-value 0.8972. This suggests that the assumption of equal variance is plausible. To test normality, we use the command *shapiro.test(resid(aov(score∼as.factor(method))))*,

and it returns a p-value of 0.8471. This suggests that the normality assumption is also plausible.

Thus, the procedures in parts (a) and (b) are valid.

6. (a) To conduct the Kruskal-Wallis test at level 0.05, we use the command

kruskal.test(scores∼as.factor(method)).

The p-value is given as 0.02112. Thus, we should reject the null hypothesis that the distributions are the same. To make the test valid, we need to assume that the populations are continuous.

(b) The three teaching methods yield $m = 3$ possible pairwise median contrasts: $\tilde{\mu}_2 - \tilde{\mu}_1$, $\tilde{\mu}_3 - \tilde{\mu}_1$, and $\tilde{\mu}_3 - \tilde{\mu}_2$. The hypothesis that each of the contrasts is zero vs the two-sided alternative will be tested at an individual level of $\alpha/3 = 0.0167$. We use the following commands: *f1 = score[method==1]; f2 = score[method==2] ; f3 = score[method==3]; wilcox.test(f2, f1); wilcox.test(f3, f1); wilcox.test(f3, f2)*. The results are summarized in the table below.

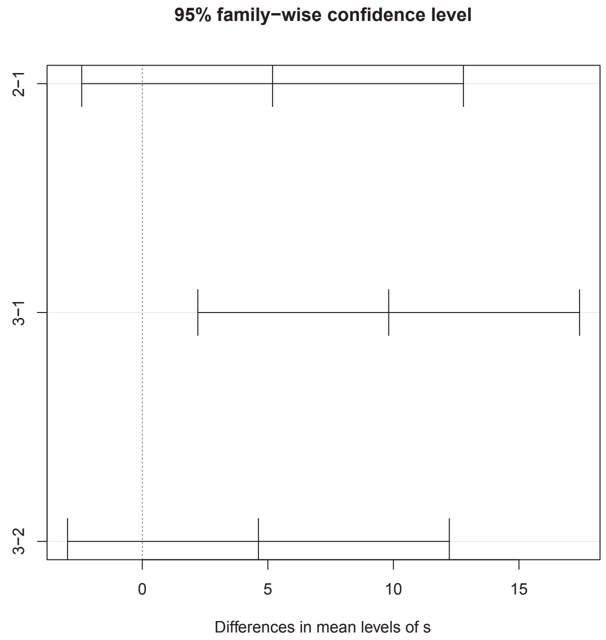
Contrast	p-value for $H_0 : \tilde{\mu}_i - \tilde{\mu}_j = 0$	Less than 0.0167?
$\tilde{\mu}_2 - \tilde{\mu}_1$	0.1049	No
$\tilde{\mu}_3 - \tilde{\mu}_1$	0.01041	Yes
$\tilde{\mu}_3 - \tilde{\mu}_2$	0.1412	No

Thus, methods 3 and 1 are significantly different at experiment-wise level of significance 0.05.

(c) To perform Tukey's multiple comparisons procedure on the ranks, we use the commands

r=rank(score); s=as.factor(method); TukeyHSD(aov(r∼s), conf.level=0.95).

The simultaneous CIs for $\tilde{\mu}_2 - \tilde{\mu}_1$, $\tilde{\mu}_3 - \tilde{\mu}_1$, and $\tilde{\mu}_3 - \tilde{\mu}_2$ are given as (-2.414 12.789), (2.211, 17.414), and (-2.976, 12.226), respectively. Thus, only methods 3 and 1 are significantly different at experiment-wise level of significance 0.05. The plot is given on the next page.



7. We use the following commands:

k=3; alpha=0.05/(k(k-1)/2); o=c(37, 28, 21);n=c(111, 85, 100);*

for(i in 1:(k-1)){ for(j in (i+1):k){

print(prop.test(c(o[i], o[j]), c(n[i], n[j]), conf.level=1-alpha, correct=F)$conf.int)}}

The CIs for $p_1 - p_2$, $p_1 - p_3$, and $p_2 - p_3$ are given as (-0.158, 0.166), (-0.022, 0.268), and (-0.037, 0.276), respectively. The CIs all contain 0 and, thus, all contrasts are not significantly different from zero.

8. We use the following commands:

k=4; alpha=0.05/(k(k-1)/2); o=c(91, 128, 46, 62); n=c(105, 148, 87, 93);*

for(i in 1:(k-1)){ for(j in (i+1):k){

print(prop.test(c(o[i], o[j]), c(n[i], n[j]), conf.level=1-alpha, correct=F)$conf.int)}}

The CIs for $p_1 - p_2$, $p_1 - p_3$, $p_1 - p_5$, $p_2 - p_3$, $p_2 - p_5$, and $p_3 - p_5$ are given as (-0.113, 0.117), (0.172, 0.504), (0.044, 0.356), (0.177, 0.496), (0.049, 0.347), and (-0.329, 0.053). The results show that, p_1 and p_2, p_3 and p_5, are not significantly different at experiment-wise level 0.05. All other contrasts are significantly different from zero.

10.4 Randomized Block Designs

1. (a) The ANOVA F-test procedure in Section 10.2 is not recommended because of the presence of a block effect.

   To incorporate additional information, we use the model

$$X_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}, \qquad i = 1, 2, 3 \quad \text{and} \quad j = 1, \cdots, 9$$

   with $\sum_i \alpha_i = 0$ and $\mathrm{Var}(b_j) = \sigma_b^2$.

   (b) To decide if variation in the mole fraction of water affects the mean diffusivity, we have the null and alternative hypotheses as

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true.}$$

   (c) To test the hypotheses, we use the following commands

   MF = st$ind; temp = as.factor(rep(1:length(CR$MF1),3));
   summary(aov(st$values∼MF+temp)).

   The obtained ANOVA table is:

	DF	SS	MS	F	P
MF	2	0.019	0.0095	128	1.44×10^{-10}
temp	8	24.838	3.1048	41654	$< 2 \times 10^{-16}$
Residuals	16	0.001	0.0001		

   Since the p-value is 1.44×10^{-10}, we should reject H_0.

   (d) In Exercise 5 of Section 10.2, we only considered treatment effect. Using the command *summary(aov(st$values∼MF))*, we have the p-value of 0.991, which is the same as in Exercise 5 of Section 10.2. This analysis is not appropriate because the samples are not independent.

   (e) To construct Tukey's 99% simultaneous CIs, we use the command *TukeyHSD(aov(st$values∼MF+temp), "MF", conf.level=0.99)*, and the results are below.

Comparison	99% Tukey's SCI	Contains 0?
$\mu_2 - \mu_1$	(-0.0027, 0.0249)	Yes
$\mu_3 - \mu_1$	(0.0474, 0.0749)	No
$\mu_3 - \mu_2$	(0.0362, 0.0638)	No

   We can conclude that μ_1 and μ_2 are not significantly different at 99% level, but the other contrasts are significant.

2. (a) An appropriate model for the observation is

$$X_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}, \qquad i = 1, 2, 3 \quad \text{and} \quad j = 1, 2, 3$$

   with $\sum_i \alpha_i = 0$ and $\mathrm{Var}(b_j) = \sigma_b^2$. The random blocks correspond to the technicians.

(b) To test the null hypothesis that the average service time of the three drives is the same, that is,

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 \quad \text{vs} \quad H_a : H_0 \text{ is not true},$$

we use the commands

time=c(44.8, 33.4, 45.2, 47.8, 61.2, 60.8, 73.4, 71.2, 64.6);
Drive = as.factor(c(1, 1, 1, 2, 2, 2, 3, 3, 3)); Tech = as.factor(rep(1:3,3));
anova(aov(time∼Drive+Tech)).

The ANOVA table is given below.

	DF	SS	MS	F	P
Drive	2	1229.66	614.83	10.1183	0.02724
Tech	2	4.92	2.46	0.0404	0.96075
Residuals	4	243.06	60.76		

The p-value is 0.02724.

(c) We use the command *friedman.test(time, Drive, Tech)* for the Friedman's test and we get 6 as the test statistic and 0.04979 as the p-value.

3. (a) The appropriate model is

$$X_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}, \qquad i = 1, 2, 3, 4 \quad \text{and} \quad j = 1, \cdots, 8$$

with $\sum_i \alpha_i = 0$ and $\text{Var}(b_j) = \sigma_b^2$. The parameters α_i specify the treatment (design) effects, and the parameters b_j represent the random block (pilot) effects.

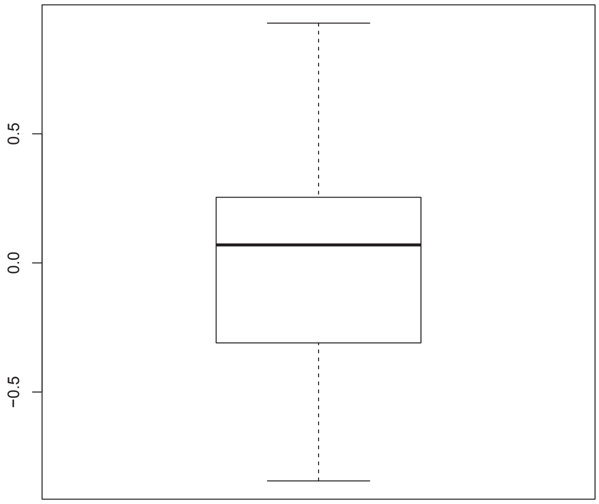
(b) For deciding if the designs differ in terms of the pilot's mean response time, we have the null and alternative hypotheses as

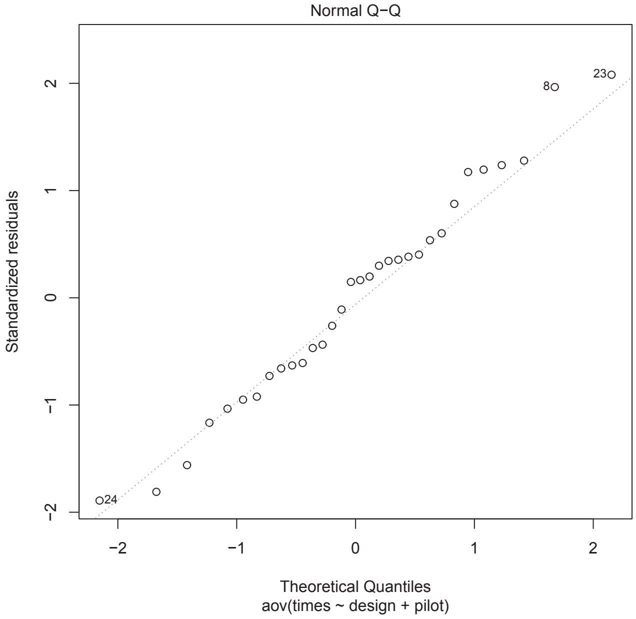
$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 \quad \text{vs} \quad H_a : H_0 \text{ is not true}.$$

(c) Load the data by *PRT = read.table("PilotReacTimes.txt", header = T); attach(PRT)*. Fit the model by *fit = aov(times∼design+pilot)*. For checking the assumptions of homoscedasticity of the intrinsic error variables of the model in (a), we use *anova(aov(resid(fit)**2∼design+pilot))*. The returned p-values are 0.231 and 0.098 for the design and pilot effects on the residual variance, suggesting there is no strong evidence against the homoscedasticity assumption.

To check the normal assumption, we use the command *shapiro.test(resid(fit))*, which returns a p-value of 0.7985, suggesting the normality assumption is reasonable for this data.

We use the commands *boxplot(resid(fit))* to get the boxplot of the residuals, and use the command *plot(fit, which=2)* to get the Q-Q plot for the residuals, shown on the next page.





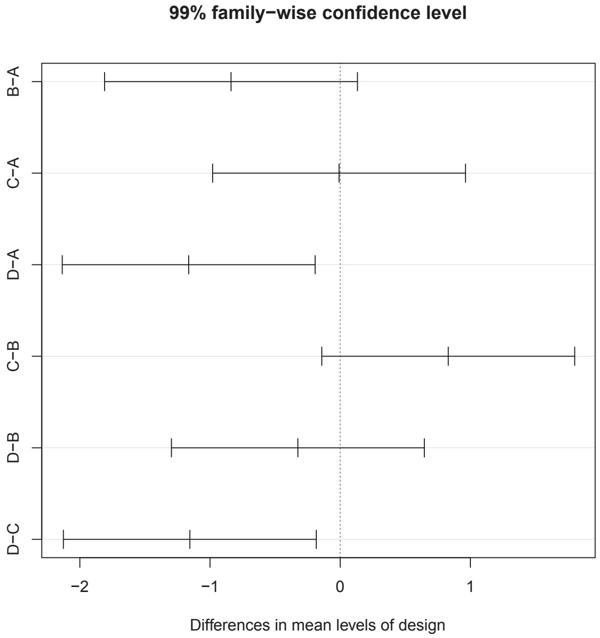
These plots also suggest that the normality assumption is satisfied, in agreement with the p-value.

(d) To test the hypotheses in part (b), we use the command *anova(fit)*, which returns p-values of 0.00044 and 1.49×10^{-6} for the design and pilot effects on the response time. Thus, the null hypothesis in part (b) is rejected.

4. To construct Tukey's 99% simultaneous CIs, we use the command *TukeyHSD(fit, "design", conf.level=0.99)*. The results are listed on the next page.

Comparison	99% Tukey's SCI	Contains 0?
$B - A$	(-1.81, 0.13)	Yes
$C - A$	(-0.98, 0.96)	Yes
$D - A$	(-2.13, -0.19)	No
$C - B$	(-0.14, 1.80,)	Yes
$D - B$	(-1.30, 0.65)	Yes
$D - C$	(-2.13, -0.18)	No

We can conclude that the pairs of designs (D, A) and (D, C) are significantly different at 99% level, but the other pairs of designs are not significantly different. The plot is produced by the command *plot(TukeyHSD(fit, "design", conf.level=0.99))*, and is shown below.



5. (a) "Fabric" is the blocking factor.

   (b) The completed ANOVA table is given below.

	DF	Sum Sq	Mean Sq	F-Value	P
treatment	3	2.4815	0.8271667	19.425	6.72×10^{-5}
block	4	5.4530	1.36325	32.014	2.57×10^{-6}
Residuals	12	0.5110	0.042583		

   Based on the p-value, at level $\alpha = 0.05$, we should reject the null hypothesis that the four chemicals do not differ in terms of the mean strength of the fabric.

6. (a) The command gives 0.0001955 as the p-value for testing the null hypothesis that the four chemicals do not differ in terms of the mean strength of the fabric. The hypothesis be rejected at level $\alpha = 0.01$.

   (b) To conduct Tukey's multiple comparisons on the ranks, we use the command *TukeyHSD(aov(ranks∼fs$chemical+fs$fabric),"fs$chemical", conf.level=0.99)*. The results are listed below.

Comparison	99% Tukey's SCI	Contains 0?
B vs A	(1.22, 11.98)	Yes
C vs A	(-2.88, 7.88)	Yes
D vs A	(3.12, 13.88)	No
C vs B	(-9.48, 1.28)	Yes
D vs B	(-3.48, 7.28)	Yes
D vs C	(0.62, 11.38)	No

   Based on the simultaneous CIs, we find that at experiment-wise level of significance 0.01, the pairs (D, A) and (D, C) are different.

7. (a) We use the commands

```
At = c(19.0, 21.8, 16.8, 24.2, 22.0, 34.7, 23.8);
Bt= c(17.8, 20.2, 16.2, 41.4, 21.4, 28.4, 22.7);
Ct=c(21.3, 22.5, 17.6, 38.1, 25.8, 39.4, 23.9);
t.test(At, Bt, paired=T, conf.level=1-0.05/3)
t.test(At, Ct, paired=T, conf.level=1-0.05/3)
t.test(Bt, Ct, paired=T, conf.level=1-0.05/3).
```

   We get the Bonferroni's 95% SCIs for $\mu_A - \mu_B$, $\mu_A - \mu_C$, and $\mu_B - \mu_C$ are (-10.136, 8.479), (-9.702, 2.188), and (-8.301, 2.444), respectively. Since all of them include 0, none of the differences are significantly different at experiment-wise significance level 0.05.

   (b) To perform Bonferroni's multiple comparisons by the signed-rank test at experiment-wise error rate of 0.05, we use the commands

```
wilcox.test(At- Bt, conf.int = T, conf.level=1-0.05/3)
wilcox.test(At- Ct, conf.int = T, conf.level=1-0.05/3)
wilcox.test(Bt- Ct, conf.int = T, conf.level=1-0.05/3).
```

   We get the Bonferroni's 95% SCIs for $\tilde{\mu}_A-\tilde{\mu}_B$, $\tilde{\mu}_A-\tilde{\mu}_C$, and $\tilde{\mu}_B-\tilde{\mu}_C$ are (-8.30, 3.75), (-13.9, -0.1), and (-11.0, 3.3), respectively. The difference $\tilde{\mu}_A - \tilde{\mu}_C$ is significantly different from zero at experiment-wise error rate of 0.05, but the other differences are not.

Chapter 11
Multifactor Experiments

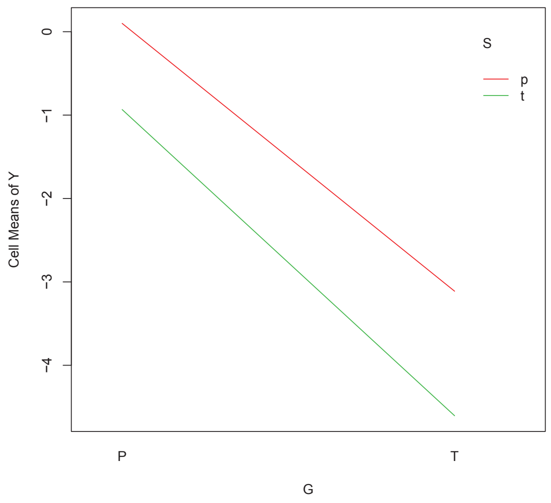
11.2 Two-Factor Designs

1. (a) $F_{H_0}^{GS} = 0.6339$ with p-value of 0.4286; the hypothesis of no interaction effect is not rejected at level of significance 0.05.

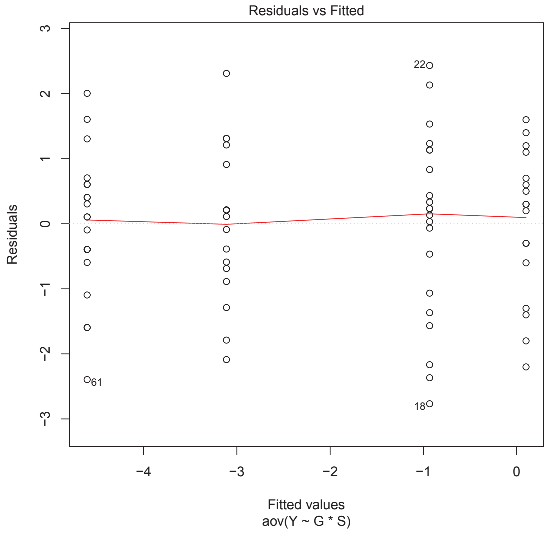
   $F_{H_0}^{G} = 141.78$ with p-value of less than 2.2×10^{-16}; the hypothesis of no main growth hormone effects is rejected at level of significance 0.05.

   $F_{H_0}^{S} = 18.96$ with p-value of 4.46×10^{-5}; the hypothesis of no main sex steroid effects is rejected at level of significance 0.05.

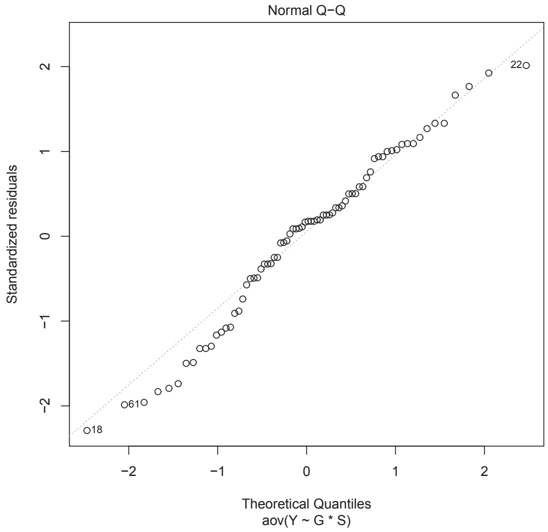
   (b) The interaction plot is given in the following plot. It is observed that the two lines are roughly parallel, which indicates there is no interaction effect, and this is consistent with the formal F-test.



   (c) The residual plot is given in the figure below. This figure shows that the homoscedasticity assumption approximately holds.



The Q-Q plot for the residuals is shown below. The figure shows that the normality assumption holds approximately.



(d) The p-values for testing the hypotheses of no main growth effects, no main sex steroid effects, and no interaction effects on the residual variance are, respectively, 0.346, 0.427, and 0.299; none of these hypotheses are rejected. The p-value for the normality test is 0.199, so the normality assumption appears to be reasonable.

2. (a) $F_{H_0}^{SC} = 1.8192$ with p-value of 0.0910; the hypothesis of no interaction effect is not rejected at level of significance 0.01.

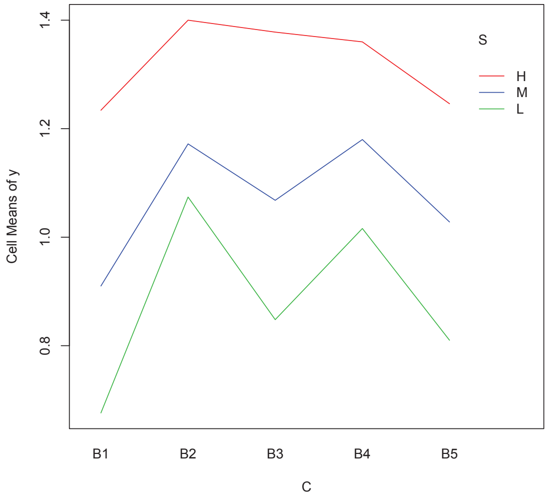
$F_{H_0}^{S} = 146.17$ with p-value of less than 2.2×10^{-16}; the hypothesis of no main signal level effects is rejected at level of significance 0.01.

$F_{H_0}^{C} = 23.13$ with p-value of 1.345×10^{-11}; the hypothesis of no main cell phone type effects is rejected at level of significance 0.01.

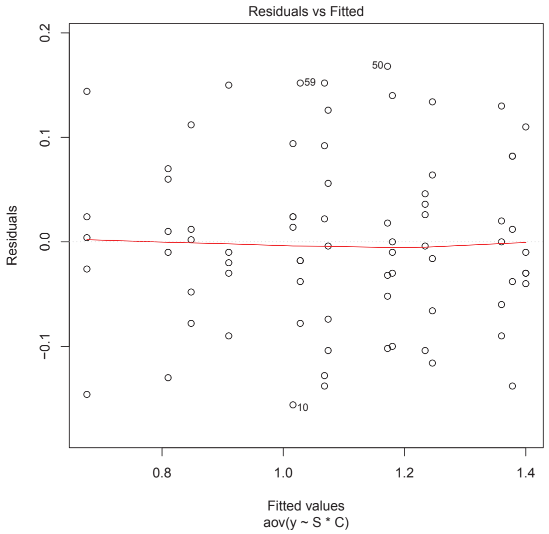
(b) From the result of multiple comparisons, it is seen that the following pairs of cell phone types are significantly different at level 0.01 in terms of their mean effects: B2-B1, B3-B1, B4-B1, B3-B2, B5-B2, and B5-B4. The pairs of signal levels are significantly different at level 0.01 in terms of their mean effects are L-H, M-H, and M-L.

(c) The p-values for testing the hypotheses of no main signal level effects, no main cell phone type effects, and no interaction effects on the residual variance are, respectively, 0.485, 0.954, and 0.801; none of these hypotheses are rejected. The p-value for the normality test is 0.092, so the normality assumption is not strongly contradicted by the data.

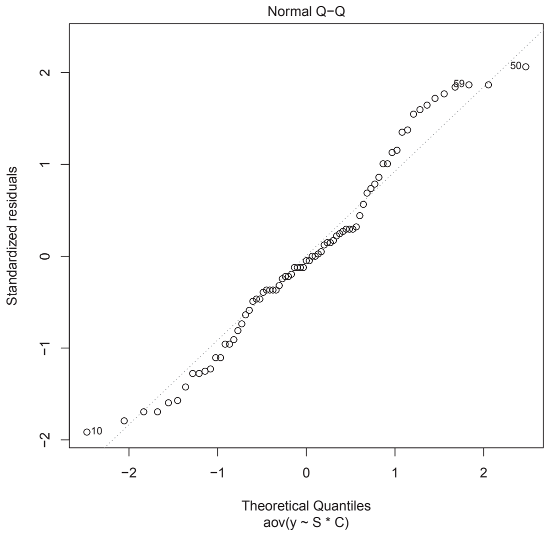
(d) The interaction plot is shown below. It shows that there might be slight interaction effects. This is in agreement with the p-value of 0.091 for the no interaction test.



The residual plot is given below. The figure suggests that the homoscedasticity assumption holds approximately.



The Q-Q plot for the residuals is shown below, and the figure shows that the normality assumption holds approximately.



3. (a) The hypothesis of no interaction between temperature and humidity.

   (b) The hypothesis of no main humidity effect.

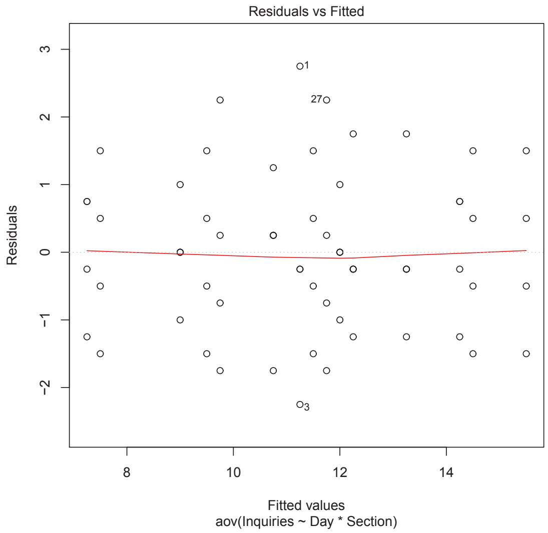
4. (a) This is to test whether or not there is main row effect: $H_0^A : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0$.

   (b) This is to test whether or not there is main column effect: $H_0^A : \beta_1 = \beta_2 = 0$.

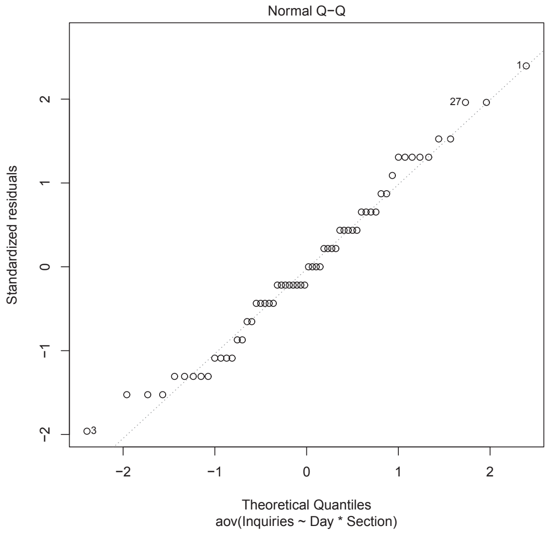
   (c) The F-values for testing H_0^A, H_0^B, and H_0^{AB} are respectively 2.1095, 6.2030, and 3.9496, the corresponding p-values are 0.1392, 0.0241, and 0.0277. Therefore, at level 0.05, we would reject H_0^B and H_0^{AB}.

(d) From part (c), we already concluded that there is no main factor A (humidity level) effects; therefore, it is not necessary to conduct multiple comparisons.

5. (a) The residual plot is given below. The figure shows that the homoscedasticity assumption approximately holds.



The Q-Q plot for the residuals is shown below, and the figure shows that the normality assumption holds approximately.



(b) The p-values for testing the hypotheses of no main day effects, no main section effects, and no interaction effects on the residual variance are, respectively, 0.3634, 0.8096, and 0.6280; none of these hypotheses are rejected. The p-value for the normality test is 0.3147, so the normality assumption appears to be reasonable.

(c) The pairs of days, except for (M, T), (M, W) and (T,W), are significantly different at experiment-wise error rate $\alpha = 0.01$. The pairs of newspaper sections (Sports, Business) and (Sports, News) are also significantly different.

6. (a) The completed ANOVA table is shown below

	DF	SS	MS	F	P
Species	2	146010	73005	5.954	0.016
Size	1	3308	3308	0.270	0.613
Interaction	2	41708	20854	1.701	0.224
Error	12	147138	12261.5		
Total	17	338164			

(b) The statistical model for Y_{ijk} is $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$. The assumption is that the error variables ϵ_{ijk} are independent normal with zero mean and common variance.

(c) To test whether an additive model is appropriate, the null hypothesis is H_0^{AB} : $\gamma_{11} = \cdots = \gamma_{32} = 0$, and the alternative hypothesis is H_1^{AB} : at least one of γ_{ij} is not 0. By the p-value from the table, the null hypothesis is not rejected.

(d) By the p-value from the table, H_0^A is rejected at level 0.05, but H_0^B is not rejected.

7. The completed ANOVA table is shown below

	DF	SS	MS	F	P
Species	2	146010	73005	5.412	0.018
Size	1	3308	3308	0.245	0.628
Error	14	188846	13489		
Total	17	338164			

By the p-value from the table, H_0^A is rejected at level 0.05, but H_0^B is not rejected.

8. The ANOVA table is calculated as

	DF	SS	MS	F	P
A	2	48.6534	24.3267	10.0801	3.34×10^{-4}
B	2	22.0901	11.0451	4.5767	0.0169
A:B	4	5.9626	1.4906	0.6177	0.6528
Residuals	36	86.8800	2.4133		
Total	44	163.5861			

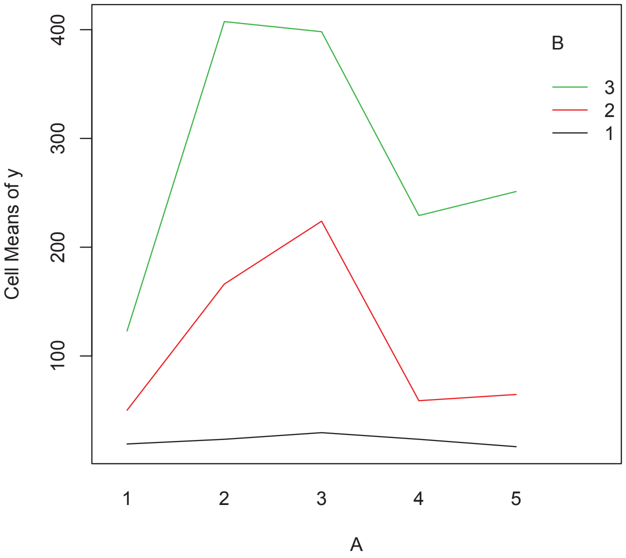
From the result, we can see that H_0^A is rejected at level 0.01, but H_0^B and H_0^{AB} are not rejected.

9. (a) The ANOVA table from the R commands is given as

	DF	SS	MS	F	P
A	4	52066	13016	3.4022	0.0661095
B	2	173333	86667	22.6528	0.0005073
Residuals	8	30607	3826		

From the table, the hypothesis of no main row effect is not rejected, while the hypothesis of no main column effects is rejected.

(b) The interaction plot is shown below, which clearly indicates that there is interactive effect.



The ANOVA table from Tukey's one degree of freedom test is given as

	DF	SS	MS	F	P
A	4	52066	13016	14.491	0.001693
B	2	173333	86667	96.482	8.026×10^{-6}
fitteds	1	24319	24319	27.073	0.001249
Residuals	7	6288	898		

The result suggests the factors interact.

(c) Yes

10. (a) The ANOVA table from the R commands is given as

	DF	SS	MS	F	P
S	2	0.77297	0.38649	63.8469	1.208×10^{-5}
C	4	0.13129	0.03282	5.4224	0.02068
Residuals	8	0.04843	0.00605		

From the table, the hypothesis of no main row effect is rejected, but the hypothesis of no main column effects is not rejected.

(b) All pairs of the signal factor levels are significantly different at level 0.01, and no pair of the cell phone type factor levels is significantly different at level 0.01.

11. (a) The ANOVA table from the R commands is given as

	DF	SS	MS	F	P
A	4	14815305	3703826	31.752	2.347×10^{-11}
B	9	19608061	2178673	18.677	4.612×10^{-11}
Residuals	36	4199367	116649		

From the table, the hypotheses of no main row and no main column effects are both rejected.

(b) All pairs of the Auxin factor levels are significantly different at level 0.01 except (0.5,0.1), (2.5,0.1) and (2.5,0.5).

(c) The ANOVA table from Tukey's one degree of freedom test is given as

	DF	SS	MS	F	P
A	4	14815305	3703826	59.829	3.657×10^{-15}
B	9	19608061	2178673	35.193	6.015×10^{-15}
fitteds	1	2032633	2032633	32.834	1.752×10^{-6}
Residuals	35	2166733	61907		

The result suggests the factors interact.

12. (a) The calculated ANOVA table is

	DF	SS	MS	F	P
A	1	0.125	0.125	0.051	0.8361
B	3	129.375	43.125	17.542	0.0209
Error	3	7.375	2.4583		

(b) At level $\alpha = 0.05$, we would reject H_0^B but retain H_0^A.

11.3 Three-Factor Designs

1. (a) The full model is $X_{ijkl} = \mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}+\epsilon_{ijkl}$. The ANOVA table given by the R commands is

	DF	SS	MS	F	P
MS	1	0.062662	0.062662	55.6612	1.452×10^{-9}
SH	2	0.005973	0.002987	2.6530	0.0807548
MH	1	0.015010	0.015010	13.3331	0.0006433
MS:SH	2	0.024394	0.012197	10.8342	0.0001309
MS:MH	1	0.010693	0.010693	9.4987	0.0034016
SH:MH	2	0.043146	0.021573	19.1629	7.628×10^{-7}
MS:SH:MH	2	0.000008	0.000004	0.0037	0.9962765
Residuals	48	0.054037	0.001126		

It is seen that all the main effect and interaction effects are significant at level 0.05, except for the main effect of SH and the three-factor interaction.

(b) The model without the three factor interaction is $X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + \epsilon_{ijkl}$. The ANOVA table given by the R commands is

	DF	SS	MS	F	P
MS	1	0.062662	0.062662	57.9714	6.606×10^{-10}
SH	2	0.005973	0.002987	2.7631	0.0727448
MH	1	0.015010	0.015010	13.8864	0.0004949
MS:SH	2	0.024394	0.012197	11.2839	9.029×10^{-5}
MS:MH	1	0.010693	0.010693	9.8929	0.0027925
SH:MH	2	0.043146	0.021573	19.9582	4.249×10^{-7}
Residuals	50	0.054046	0.001081		

It is seen that all the main effect and interaction effects are significant at level 0.05, except for the main effect of SH.

(c) When testing the homoscedasticity, the p-values for the main factors, MS, SH, and MH are respectively, 0.010142, 0.108286, and 0.004175. The results suggest the homoscedasticity assumption does not hold. The p-value of the Shapiro-Wilk test for normality is 0.07, but in the presence of heteroscedasticity this is not easily interpretable

(d) After the square root arcsine transformation on the response variable, only the main factor MH effect on the residual variance has a p-value less than 0.05 (0.016). The p-value of the Shapiro-Wilk test is 0.64, suggesting the normality assumption is tenable.

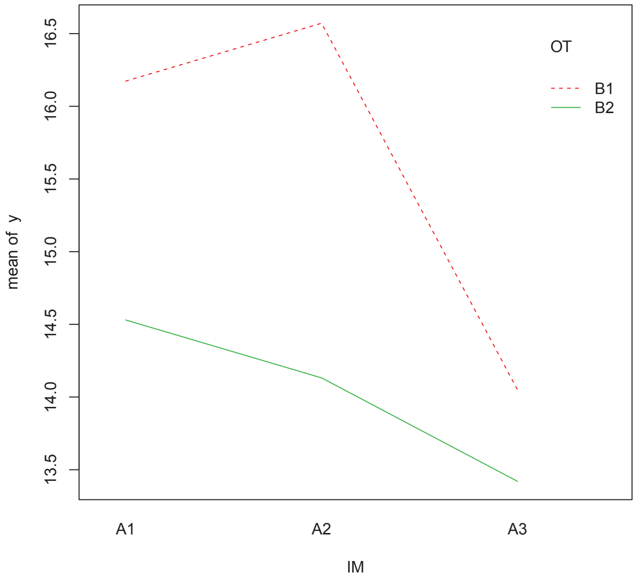
2. (a) The full model is $X_{ijkl} = \mu+\alpha_i+\beta_j+\gamma_k+(\alpha\beta)_{ij}+(\alpha\gamma)_{ik}+(\beta\gamma)_{jk}+(\alpha\beta\gamma)_{ijk}+\epsilon_{ijkl}$. The ANOVA table given by the R commands is

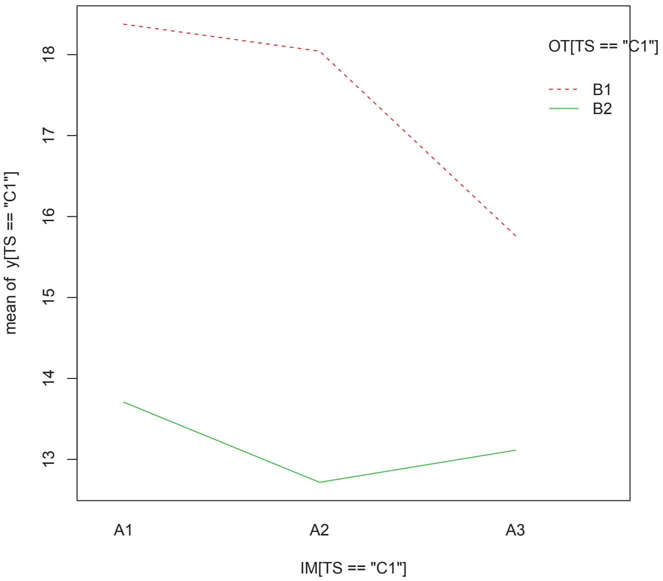
	DF	SS	MS	F	P
IM	2	20.909	10.454	4.4881	0.022087
OT	1	22.216	22.216	9.5370	0.005028
TS	1	8.085	8.085	3.4707	0.074750
IM:OT	2	4.938	2.469	1.0599	0.362173
IM:TS	2	3.574	1.787	0.7672	0.475346
OT:TS	1	62.938	62.938	27.0189	2.519×10^{-5}
IM:OT:TS	2	1.784	0.892	0.3830	0.685938
Residuals	24	55.905	2.329		

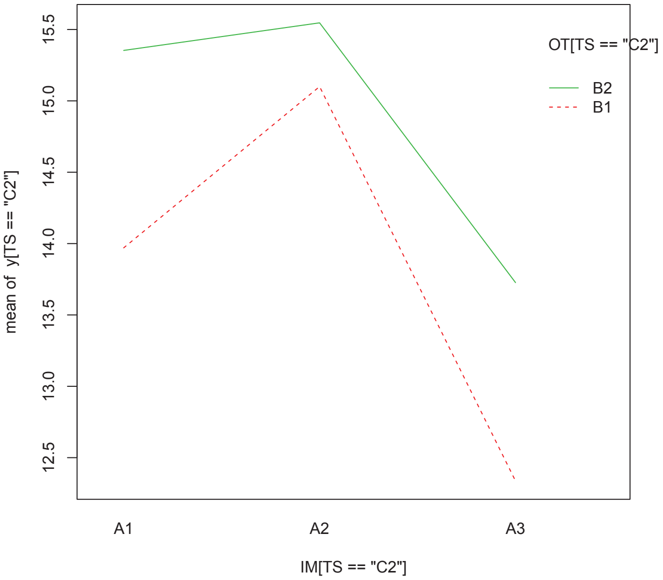
It is seen that the main effect of IM and OT and interaction effects of OT and TS are significant at level 0.05.

(b) When testing the homoscedasticity, only the p-value for the three-factor interaction effect is less than 0.05 (0.01212); thus, we would conclude that the homoscedasticity assumption holds. The p-value of the Shapiro-Wilk test is 0.80, suggesting the normality assumption is tenable.

(c) The three interaction plots are given in the following graphs, and they suggest that there is interaction effects between “insulation type” and “outside temperature”.







(d) The model without the $(\alpha\beta)$, $(\alpha\gamma)$, and $(\alpha\beta\gamma)$ interactions is $X_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\beta\gamma)_{jk} + \epsilon_{ijkl}$. With this model, the resulting ANOVA table is

	DF	SS	MS	F	P
IM	2	20.909	10.454	4.7375	0.016292
OT	1	22.216	22.216	10.0672	0.003473
TS	1	8.085	8.085	3.6636	0.065195
OT:TS	1	62.938	62.938	28.5210	8.906×10^{-6}
Residuals	30	66.202	2.207		

We can see the the main effects of IM, OT are significant at level 0.05 and the interaction effect of OT and TS is significant at level 0.05.

3. (a) $x_{221} = ab = 16$, $x_{112} = c = 12$, and $x_{122} = bc = 18$

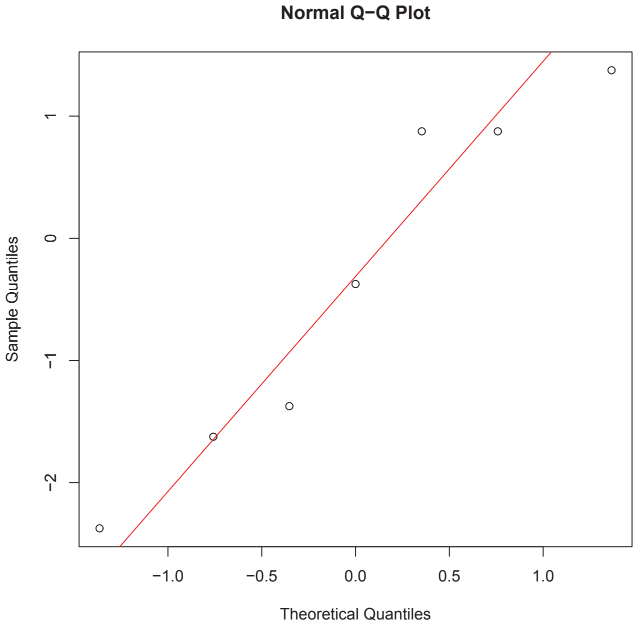
   (b) Same as in Table 11.2

   (c) $\alpha_1 = -1.375$, $\beta_1 = -1.625$, $\gamma_1 = -2.3750$, $(\alpha\beta)_{11} = 0.875$, $(\beta\gamma)_{11} = -0.375$ $(\alpha\gamma)_{11} = 0.875$, and $(\alpha\beta\gamma)_{111} = 1.375$

   (d) $SSA = 16 \times 1.375^2 = 30.25$, $SSB = 16 \times 1.625^2 = 42.25$, $SSC = 16 \times 2.375^2 = 90.25$, $SSAB = 16 \times 0.875^2 = 12.25$, $SSAC = 16 \times 0.375^2 = 2.25$, $SSBC = 16 \times 0.875^2 = 12.25$, and $SSABC = 16 \times 1.375^2 = 30.25$

   (e) $F_{H_0}^{A} = 30.25/((12.25 + 2.25 + 12.25 + 30.25)/4) = 2.12$, $F_{H_0}^{B} = 42.25/((12.25 + 2.25+12.25+30.25)/4) = 2.96$, $F_{H_0}^{C} = 90.25/((12.25+2.25+12.25+30.25)/4) = 6.33$; these test statistics are all less than $F_{1,4,0.05}$, so none of the main effects are significantly different from zero.

   (f) The probability plot of the effects is given below. We notice that the two outliers on the top right corner correspond to the three-factor effect and one of the two-factor effects, which indicate that the assumption of no interaction is not appropriate.



4. (a) Same as in Table 11.2.

   (b) $\alpha_1 = -29.375$, $\beta_1 = 7.75$, $\gamma_1 = -13.25$, $(\alpha\beta)_{11} = 9.125$, $(\beta\gamma)_{11} = -8.875$ $(\alpha\gamma)_{11} = -6.75$, and $(\alpha\beta\gamma)_{111} = -5.875$

   (c) The calculated sum of squares are $SSA = 13806$, $SSB = 961$, $SSC = 2809$, $SSAB = 1332.3$, $SSAC = 1.260.3$, $SSBC = 729$, and $SSABC = 552.25$.

   (d) The calculated F values are $F_{H_0}^{A} = 6.5866$, $F_{H_0}^{B} = 0.4585$, $F_{H_0}^{C} = 1.3401$, $F_{H_0}^{AB} = 0.6356$, $F_{H_0}^{AC} = 0.6012$, $F_{H_0}^{BC} = 0.3478$, and $F_{H_0}^{ABC} = 0.2635$. Since $F_{1,8,0.05} = 5.32$, we conclude that there is main effect of Temperature.

5. (a) True

   (b) True

   (c) $F_{H_0}^{A} = 47.956/((0.276{+}9.456{+}1.156{+}6.126)/4) = 11.27$, $F_{H_0}^{B} = 4.306/((0.276{+}9.456{+}1.156{+}6.126)/4) = 1.01$, $F_{H_0}^{C}12.426/((0.276{+}9.456{+}1.156{+}6.126)/4) = 2.92$. The corresponding p-values are 0.03, 0.37, and 0.16. Thus, the hypothesis of no main factor A effects is rejected and the main effects of factors B and C are not significantly different from zero

6. Since the factor C is fixed at level k, the term γ_k in (11.3.1) is a constant. We collect the constant terms and denote the new constant as $\mu^k = \mu + \gamma_k$, which is (a). Note that this constant depends on the level k. Similarly, we collect all the terms that depend on the main effect of factor A only and denote them as α_i^k, resulting (b) $\alpha_i^k = \alpha_i + (\alpha\gamma)_{ik}$. The other relations could be verified in a similar manner.

11.4 2^r Factorial Experiments

1. Let θ_1 be the effect of block 1 and $\theta_2 = -\theta_1$ be the effect of block 2; then the mean μ_{ijk} of X_{ijk} is

$$\mu_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \theta_1,$$

if (i, j, k) is one of (1, 1, 1), (2, 2, 1), (2, 1, 2), (1, 2, 2), and the same expression with θ_1 replaced by $-\theta_1$ if (i, j, k) is one of the other four sets of indices.

Then $\hat{\alpha}_1$ is the contrast

$$\frac{X_{111} - X_{211} + X_{121} - X_{221} + X_{112} - X_{212} + X_{122} - X_{222}}{8}.$$

In the term $X_{111} - X_{221}$ and $-X_{212} + X_{122}$, θ_1 is cancelled. Similarly, in the other four terms, θ_2 is cancelled.

In the same manner, we can verify that the block effect is not confounded with the other main effects and two-factor interaction effects. However, for the three factor effect, $(\alpha\beta\gamma)_{111}$ is the contrast

$$\frac{X_{111} - X_{211} - X_{121} + X_{221} - X_{112} + X_{212} + X_{122} - X_{222}}{8},$$

where the four positive terms X_{111}, X_{221}, X_{212}, and X_{122} contribute $4\theta_1$, and the four negative terms $-X_{211}$, $-X_{121}$, $-X_{112}$, and $-X_{222}$ contribute $-4\theta_2 = 4\theta_1$. Thus the block effect cannot be cancelled. Hence, the block effect is confounded with the three factor interaction effect.

2. In the following, we assume the block effect for block 1 is θ_1 and for block 2 is $\theta_2 = -\theta_1$.

   (a) With the given command, we will assign $(1, b)$ to block 1 and (a, ab) to block 2. The estimated main effect of factor A is given by

   $$\frac{Y_{11} - Y_{21} + Y_{12} - Y_{22}}{4}.$$

   The positive terms Y_{11} and Y_{12} are in block 1 and contribute $2\theta_1$; the negative terms $-Y_{21}$ and $-Y_{22}$ are in block 2 and contribute $-2\theta_2 = 2\theta_1$. Thus, the block effect cannot be cancelled and it will be confounded with the main effect of factor A.

   (b) With the given command, we will assign $(1, a, bc, abc)$ to block 1 and (b, ab, c, ac) to block 2. The estimated BC interaction effect is

   $$\frac{Y_{111} + Y_{211} - Y_{121} - Y_{221} - Y_{112} - Y_{212} + Y_{122} + Y_{222}}{8}.$$

   The positive terms Y_{111}, Y_{211}, Y_{122}, and Y_{222} are all in block 1 and they contribute $4\theta_1$; the negative terms $-Y_{121}$, $-Y_{221}$, $-Y_{112}$, and $-Y_{212}$ are in block 2 and they contribute $-4\theta_2 = 4\theta_1$. Thus, the block effect cannot be cancelled and it will be confounded with the interaction effect BC.

3. (a) $ABCCDE = ABDE$

   (b) $BCDCDE = BE$

   (c) The R commands for part (a)

   G=rbind(c(1, 1, 1, 0, 0), c(0, 0, 1, 1, 1)); conf.design(G, p=2)

   The R commands for part (b)

   G=rbind(c(0,1, 1, 1, 0), c(0, 0, 1, 1, 1)); conf.design(G, p=2)

4. (a) There are $2^p = 8$ blocks; therefore, there must be $2^p - 1 = 7$ effects to be confounded with the block effects.

   (b) The other four effects to be confounded with the block effects are $ABCBCD = AD$, $ABCCDE = ABDE$, $BCDCDE = BE$, and $ABCBCDCDE = AE$.

   (c) The R commands are

   G=rbind(c(1, 1, 1, 0,0), c(0, 1, 1, 1, 0), c(0,0,1,1,1)); conf.design(G, p=2)

5. (a) The R commands are

```
sr = read.table(”SurfRoughOPtim.txt”, header = T); attach(sr)
b = c(rep(1, 16)); b[c(2,3,5,8,10,11,13,16)] = 2; sr$block = b
anova(aov(y~block+A*B*C, data=sr))
```

The obtained ANOVA table is

	DF	SS	MS	F	P
block	1	0.36	0.36	0.3251	0.584237
A	1	495.06	495.06	447.0090	2.632×10^{-8}
B	1	486.20	486.20	439.0090	2.825×10^{-8}
C	1	90.25	90.25	81.4898	1.812×10^{-5}
A:B	1	13.69	13.69	12.3612	0.007894
A:C	1	0.56	0.56	0.5079	0.496307
B:C	1	1.10	1.10	0.9955	0.347623
Residuals	8	8.86	1.11		

(b) From the ANOVA table, it is clear that the three main effects and the AB interaction effect are significantly different from zero.

6. First use the command `G = rbind(c(1,1,0), c(1,0,1)); conf.design(G,p=2)` to get the block allocation, then apply the commands

```
sr = read.table(”SurfRoughOPtim.txt”, header = T); attach(sr)
b = c(rep(4, 16)); b[c(1,8,9,16)] = 1; b[c(2,7,10,15)] = 2; b[c(3,6,11,14)] = 3;
sr$block = b
anova(aov(y~block+A*B*C, data=sr))
```

The obtained ANOVA table is

	DF	SS	MS	F	P
block	1	1.62	1.62	1.4668	0.260409
A	1	495.06	495.06	447.0090	2.632×10^{-8}
B	1	486.20	486.20	439.0090	2.825×10^{-8}
C	1	90.25	90.25	81.4898	1.812×10^{-5}
A:B	1	13.69	13.69	12.3612	0.007894
A:C	1	0.04	0.04	0.0366	0.853110
A:B:C	1	0.36	0.36	0.3251	0.584237
Residuals	8	8.86	1.11		

From the ANOVA table, it is clear that the three main effects and the AB interaction effect are significantly different from zero.

7. (a) The R commands are

```
G=rbind(c(1, 1, 0,0), c(0,0,1,1)); conf.design(G, p=2)
```

(b) The R commands are

```
AW = read.table(“ArcWeld.txt”, header = T); attach(AW);
w = c(rep(4,16)); w[c(1,4,13,16)] = 1; w[c(2,3,14,15)] = 2;
w[c(5,8,9,12)] = 3; AW$block = w
```

(c) Using the command *anova(aov(y∼block+A*B*C*D, data=AW))* gives the ANOVA table

	DF	SS	MS	F	P
block	1	102	102	0.3103	0.5852
A	1	10513	10513	31.8561	3.660×10^{-5}
B	1	80802	80802	244.8545	4.046×10^{-11}
C	1	50	50	0.1515	0.7022
D	1	91	91	0.2761	0.6065
A:B	1	664	664	2.0128	0.1752
A:C	1	25	25	0.0742	0.7887
B:C	1	61	61	0.1833	0.6742
A:D	1	153	153	0.4640	0.5055
B:D	1	231	231	0.7004	0.4150
A:B:C	1	12	12	0.0379	0.8481
A:B:D	1	6	6	0.0186	0.8933
A:C:D	1	78	78	0.2367	0.6332
B:C:D	1	45	45	0.1367	0.7164
A:B:C:D	1	36	36	0.1095	0.7450
Residuals	16	5280	330		

From the ANOVA table, it is clear that the main effects of the factors A and B are significantly different from zero.

8. First, we consider the contrast

$$\frac{\bar{x}_{111} - \bar{x}_{221} + \bar{x}_{212} - \bar{x}_{122}}{4}.$$

It estimates

$$\begin{aligned}\frac{\mu_{111}-\mu_{221}+\mu_{212}-\mu_{122}}{4} &= \frac{1}{4}[\alpha_1+\beta_1+\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &- \frac{1}{4}[\alpha_2+\beta_2+\gamma_1+(\alpha\beta)_{22}+(\alpha\gamma)_{21}+(\beta\gamma)_{21}+(\alpha\beta\gamma)_{221}]\\ &+ \frac{1}{4}[\alpha_2+\beta_1+\gamma_2+(\alpha\beta)_{21}+(\alpha\gamma)_{22}+(\beta\gamma)_{12}+(\alpha\beta\gamma)_{212}]\\ &- \frac{1}{4}[\alpha_1+\beta_2+\gamma_2+(\alpha\beta)_{12}+(\alpha\gamma)_{12}+(\beta\gamma)_{22}+(\alpha\beta\gamma)_{122}]\\ &= \frac{1}{4}[\alpha_1+\beta_1+\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &- \frac{1}{4}[-\alpha_1-\beta_1+\gamma_1+(\alpha\beta)_{11}-(\alpha\gamma)_{11}-(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &+ \frac{1}{4}[-\alpha_1+\beta_1-\gamma_1-(\alpha\beta)_{11}+(\alpha\gamma)_{11}-(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &- \frac{1}{4}[\alpha_1-\beta_1-\gamma_1-(\alpha\beta)_{11}-(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &= \beta_1+(\alpha\gamma)_{11}.\end{aligned}$$

This shows that β_1 is confounded with $(\alpha\gamma)_{11}$.

Next, consider the contrast

$$\frac{\bar{x}_{111}+\bar{x}_{221}-\bar{x}_{212}-\bar{x}_{122}}{4}.$$

It estimates

$$\begin{aligned}\frac{\mu_{111}+\mu_{221}-\mu_{212}-\mu_{122}}{4} &= \frac{1}{4}[\alpha_1+\beta_1+\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &+ \frac{1}{4}[\alpha_2+\beta_2+\gamma_1+(\alpha\beta)_{22}+(\alpha\gamma)_{21}+(\beta\gamma)_{21}+(\alpha\beta\gamma)_{221}]\\ &- \frac{1}{4}[\alpha_2+\beta_1+\gamma_2+(\alpha\beta)_{21}+(\alpha\gamma)_{22}+(\beta\gamma)_{12}+(\alpha\beta\gamma)_{212}]\\ &- \frac{1}{4}[\alpha_1+\beta_2+\gamma_2+(\alpha\beta)_{12}+(\alpha\gamma)_{12}+(\beta\gamma)_{22}+(\alpha\beta\gamma)_{122}]\\ &= \frac{1}{4}[\alpha_1+\beta_1+\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &+ \frac{1}{4}[-\alpha_1-\beta_1+\gamma_1+(\alpha\beta)_{11}-(\alpha\gamma)_{11}-(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &- \frac{1}{4}[-\alpha_1+\beta_1-\gamma_1-(\alpha\beta)_{11}+(\alpha\gamma)_{11}-(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &- \frac{1}{4}[\alpha_1-\beta_1-\gamma_1-(\alpha\beta)_{11}-(\alpha\gamma)_{11}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{111}]\\ &= \gamma_1+(\alpha\beta)_{11}.\end{aligned}$$

This shows that γ_1 is confounded with $(\alpha\beta)_{11}$.

9. Consider the contrast

$$\frac{-\bar{x}_{211}+\bar{x}_{121}+\bar{x}_{112}-\bar{x}_{222}}{4},$$

which estimates

$$\begin{aligned}\frac{-\mu_{211}+\mu_{121}+\mu_{112}-\mu_{222}}{4} &= -\frac{1}{4}[\alpha_2+\beta_1+\gamma_1+(\alpha\beta)_{21}+(\alpha\gamma)_{21}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{211}]\\ &+\frac{1}{4}[\alpha_1+\beta_2+\gamma_1+(\alpha\beta)_{12}+(\alpha\gamma)_{11}+(\beta\gamma)_{21}+(\alpha\beta\gamma)_{121}]\\ &+\frac{1}{4}[\alpha_1+\beta_1+\gamma_2+(\alpha\beta)_{11}+(\alpha\gamma)_{12}+(\beta\gamma)_{12}+(\alpha\beta\gamma)_{112}]\\ &-\frac{1}{4}[\alpha_2+\beta_2+\gamma_2+(\alpha\beta)_{22}+(\alpha\gamma)_{22}+(\beta\gamma)_{22}+(\alpha\beta\gamma)_{222}]\\ &= -\frac{1}{4}[-\alpha_1+\beta_1+\gamma_1-(\alpha\beta)_{11}-(\alpha\gamma)_{11}+(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &+\frac{1}{4}[\alpha_1-\beta_1+\gamma_1-(\alpha\beta)_{11}+(\alpha\gamma)_{11}-(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &+\frac{1}{4}[\alpha_1+\beta_1-\gamma_1+(\alpha\beta)_{11}-(\alpha\gamma)_{11}-(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &-\frac{1}{4}[-\alpha_1-\beta_1-\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &= \alpha_1-(\beta\gamma)_{11}.\end{aligned}$$

This shows that α_1 is confounded with $(\beta\gamma)_{11}$.

Consider the contrast

$$\frac{\bar{x}_{211}-\bar{x}_{121}+\bar{x}_{112}-\bar{x}_{222}}{4},$$

which estimates

$$\begin{aligned}\frac{\mu_{211}-\mu_{121}+\mu_{112}-\mu_{222}}{4} &= \frac{1}{4}[\alpha_2+\beta_1+\gamma_1+(\alpha\beta)_{21}+(\alpha\gamma)_{21}+(\beta\gamma)_{11}+(\alpha\beta\gamma)_{211}]\\ &-\frac{1}{4}[\alpha_1+\beta_2+\gamma_1+(\alpha\beta)_{12}+(\alpha\gamma)_{11}+(\beta\gamma)_{21}+(\alpha\beta\gamma)_{121}]\\ &+\frac{1}{4}[\alpha_1+\beta_1+\gamma_2+(\alpha\beta)_{11}+(\alpha\gamma)_{12}+(\beta\gamma)_{12}+(\alpha\beta\gamma)_{112}]\\ &-\frac{1}{4}[\alpha_2+\beta_2+\gamma_2+(\alpha\beta)_{22}+(\alpha\gamma)_{22}+(\beta\gamma)_{22}+(\alpha\beta\gamma)_{222}]\\ &= \frac{1}{4}[-\alpha_1+\beta_1+\gamma_1-(\alpha\beta)_{11}-(\alpha\gamma)_{11}+(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &-\frac{1}{4}[\alpha_1-\beta_1+\gamma_1-(\alpha\beta)_{11}+(\alpha\gamma)_{11}-(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &+\frac{1}{4}[\alpha_1+\beta_1-\gamma_1+(\alpha\beta)_{11}-(\alpha\gamma)_{11}-(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &-\frac{1}{4}[-\alpha_1-\beta_1-\gamma_1+(\alpha\beta)_{11}+(\alpha\gamma)_{11}+(\beta\gamma)_{11}-(\alpha\beta\gamma)_{111}]\\ &= \beta_1-(\alpha\gamma)_{11}.\end{aligned}$$

This shows that β_1 is confounded with $(\alpha\gamma)_{11}$.

Consider the contrast

$$\frac{\bar{x}_{211} + \bar{x}_{121} - \bar{x}_{112} - \bar{x}_{222}}{4},$$

which estimates

$$\begin{aligned}
\frac{\mu_{211} + \mu_{121} - \mu_{112} - \mu_{222}}{4} &= \frac{1}{4}[\alpha_2 + \beta_1 + \gamma_1 + (\alpha\beta)_{21} + (\alpha\gamma)_{21} + (\beta\gamma)_{11} + (\alpha\beta\gamma)_{211}] \\
&\frac{1}{4}[\alpha_1 + \beta_2 + \gamma_1 + (\alpha\beta)_{12} + (\alpha\gamma)_{11} + (\beta\gamma)_{21} + (\alpha\beta\gamma)_{121}] \\
&- \frac{1}{4}[\alpha_1 + \beta_1 + \gamma_2 + (\alpha\beta)_{11} + (\alpha\gamma)_{12} + (\beta\gamma)_{12} + (\alpha\beta\gamma)_{112}] \\
&- \frac{1}{4}[\alpha_2 + \beta_2 + \gamma_2 + (\alpha\beta)_{22} + (\alpha\gamma)_{22} + (\beta\gamma)_{22} + (\alpha\beta\gamma)_{222}] \\
&= \frac{1}{4}[-\alpha_1 + \beta_1 + \gamma_1 - (\alpha\beta)_{11} - (\alpha\gamma)_{11} + (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111}] \\
&+ \frac{1}{4}[\alpha_1 - \beta_1 + \gamma_1 - (\alpha\beta)_{11} + (\alpha\gamma)_{11} - (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111}] \\
&- \frac{1}{4}[\alpha_1 + \beta_1 - \gamma_1 + (\alpha\beta)_{11} - (\alpha\gamma)_{11} - (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111}] \\
&- \frac{1}{4}[-\alpha_1 - \beta_1 - \gamma_1 + (\alpha\beta)_{11} + (\alpha\gamma)_{11} + (\beta\gamma)_{11} - (\alpha\beta\gamma)_{111}] \\
&= \gamma_1 - (\alpha\beta)_{11}.
\end{aligned}$$

This shows that γ_1 is confounded with $(\alpha\beta)_{11}$.

In summary, the alias pairs are $[A, BC]$, $[B, AC]$, $[C, AB]$.

10. (a) The design table is given as follows:

A	B	C	D	$E = (ABCD)$
+	+	+	+	+
−	+	+	+	−
+	−	+	+	−
−	−	+	+	+
+	+	−	+	−
−	+	−	+	+
+	−	−	+	+
−	−	−	+	−
+	+	+	−	−
−	+	+	−	+
+	−	+	−	+
−	−	+	−	−
+	+	−	−	+
−	+	−	−	−
+	−	−	−	−
−	−	−	−	+

The set of 15 aliased pairs is

$$[A, BCDE], [B, ACDE], [C, ABDE], [D, ABCE], [E, ABCD],$$

$$[AB, CDE], [AC, BDE], [AD, BCE], [AE, BCD], [BC, ADE],$$

$$[BD, ACE], [BE, ACD], [CD, ABE], [CE, ABD], [DE, ABC].$$

(b) The design table is given as follows:

A	B	C	D	$E(= ABC)$	$F = (BCD)$
$+$	$+$	$+$	$+$	$+$	$+$
$-$	$+$	$+$	$+$	$-$	$+$
$+$	$-$	$+$	$+$	$-$	$-$
$-$	$-$	$+$	$+$	$+$	$-$
$+$	$+$	$-$	$+$	$-$	$-$
$-$	$+$	$-$	$+$	$+$	$-$
$+$	$-$	$-$	$+$	$+$	$+$
$-$	$-$	$-$	$+$	$-$	$+$
$+$	$+$	$+$	$-$	$-$	$-$
$-$	$+$	$+$	$-$	$+$	$-$
$+$	$-$	$+$	$-$	$+$	$+$
$-$	$-$	$+$	$-$	$-$	$+$
$+$	$+$	$-$	$-$	$+$	$+$
$-$	$+$	$-$	$-$	$-$	$+$
$+$	$-$	$-$	$-$	$-$	$-$
$-$	$-$	$-$	$-$	$+$	$-$

The set of 15 aliased groups are

$$[A, BCE, DEF, ABCDF], [B, ACE, CDF, ABDEF], [C, ABE, BDF, ACDEF],$$

$$[D, ABCDE, BCF, AEF], [E, ABC, ADF, BCDEF], [F, BCD, ADE, ABCEF],$$

$$[AB, CE, ACDF, BDEF], [AC, BE, ABDF, CDEF], [AD, BCDE, ABCF, EF],$$

$$[AE, BC, DF, ABCDEF], [BD, ACDE, CF, ABEF], [BE, AC, CDEF, ABDF],$$

$$[CD, ABDE, BF, ACEF], [DE, ABCD, BCEF, AF], [ABD, CDE, ACF, BEF].$$

(c) The design table is omitted. There are $(2^7 - 4)/4 = 31$ groups of four aliased effects.

11. (a) The set of alias pairs are $[A, BCDE]$, $[B, ACDE]$, $[C, ABDE]$, $[D, ABCE]$, $[E, ABCD]$, $[AB, CDE]$, $[AC, BDE]$, $[AD, BCE]$, $[AE, BCD]$, $[BC, ADE]$, $[BD, ACE]$, $[BE, ACD]$, $[CD, ABE]$, $[CE, ABD]$, $[DE, ABC]$.

(b) With the data read into the data frame *df*, the sums of squares of the classes of aliased effects can be obtained by the command *anova(aov(y~A*B*C*D*E, data=df))*. It is not possible to test for the significance of the effects because there are no degrees of freedom for the error sum of squares.

(c) Using the command given in the hint, the obtained ANOVA table is

	DF	SS	MS	F	P
A	1	500.64	500.64	3560.111	0.010669
B	1	0.14	0.14	1.000	0.500000
C	1	489.52	489.52	3481.000	0.010789
D	1	185.64	185.64	1320.111	0.017517
E	1	293.27	293.27	2085.444	0.013938
A:B	1	1233.77	1233.77	8773.444	0.006796
A:D	1	13.14	13.14	93.444	0.065624
A:E	1	26.27	26.27	186.778	0.046499
B:C	1	21.39	21.39	152.111	0.051505
B:D	1	26.27	26.27	186.778	0.046499
B:E	1	43.89	43.89	312.111	0.035997
C:D	1	213.89	213.89	1521.000	0.016320
C:E	1	50.77	50.77	361.000	0.033475
D:E	1	293.27	293.27	2085.444	0.013938
Residuals	1	0.14	0.14		

It is clear that at significant level 0.05, all the effect except B, AD, BC are significant.

Chapter 12
Polynomial and Multiple Regression

12.2 The Multiple Linear Regression Model

1. (a) $\mu_{Y|X_1,X_2}(12, 25) = 3.6 + 2.7 \times 12 + 0.9 \times 25 = 58.5$

   (b) $E(Y) = 3.6 + 2.7E(X_1) + 0.9E(X_2) = 3.6 + 2.7 \times 10 + 0.9 \times 18 = 46.8$

   (c) When X_1 increases by one unit while X_2 remains fixed, the expected change in Y is increasing by 2.7.

   (d) Clearly, $\beta_1 = 2.7$, $\beta_2 = 0.9$, and $\beta_0 = E(Y) = 46.8$.

2. (a) Since $\text{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2)$, we have $E(X_1X_2) = \text{Cov}(X_1, X_2) + E(X_1)E(X_2) = 80 + 10 \times 18 = 260$. Thus, $E(Y) = 3.6 + 2.7E(X_1) + 0.9E(X_2) + 1.5E(X_1X_2) = 3.6 + 2.7 \times 10 + 0.9 \times 18 + 1.5 \times 260 = 436.8$.

   (b) Since β_3 is the coefficient of the term X_1X_2, comparing to the original model, there must be $\beta_3 = 1.5$. Taking expectation in the centered model, there is

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 E(X_1 - \mu_{X_1}) + \beta_2 E(X_2 - \mu_{X_2}) + \beta_3 E[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})] \\ &= \beta_0 + \beta_3 \text{Cov}(X_1, X_2), \end{aligned}$$

   which gives $\beta_0 = E(Y) - \beta_3\text{Cov}(X_1, X_2)$. From the information in part (a), $\beta_0 = E(Y) - \beta_3\text{Cov}(X_1, X_2) = 436.8 - 1.5 \times 80 = 316.8$.

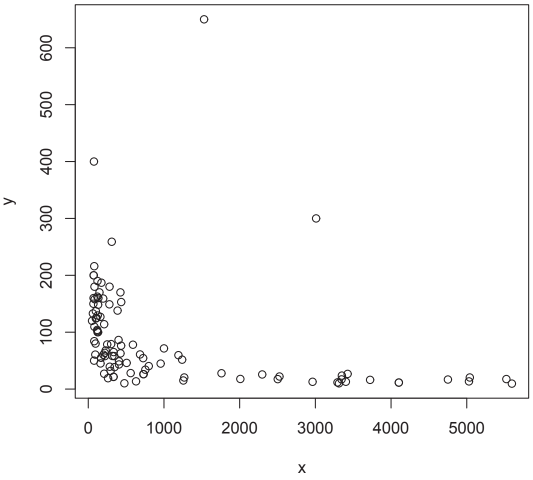
3. (a) The rate of change of the regression function at x is

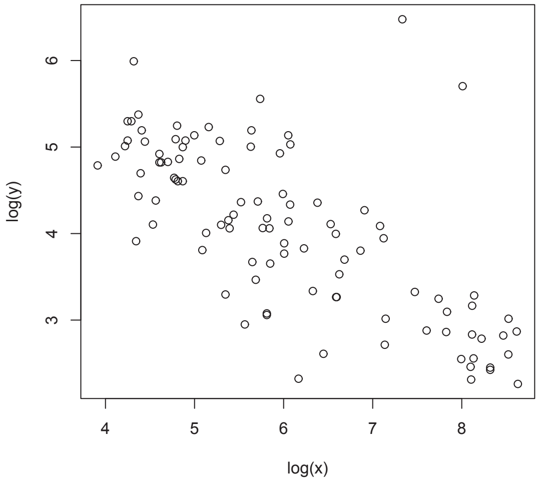
$$\frac{d\mu_{Y|X}(x)}{dx} = -3.2 + 1.4x.$$

   Thus, the rate of change at x=0, 2, and 3 are -3.2, -0.4, and 1, respectively.

   (b) β_2 is the coefficient of the x^2 term, thus $\beta_2 = 0.7$. In the centered model, the coefficient of x is $\beta_1 - 2\beta_2\mu_X$, which must be -3.2. Thus, $\beta_1 = -3.2 + 2\beta_2\mu_X = -3.2 + 2 \times 0.7 \times 2 = -0.4$. $\beta_0 = \mu_{Y|X}(\mu_X) = -8.5 - 3.2 \times 2 + 0.7 \times 2^2 = -12.1$.

4. (a) The scatterplots for (x, y) and $(\log(x), \log(y))$ are given on the next page. The scatterplot of $(\log(x), \log(y))$ suggests a linear relation.





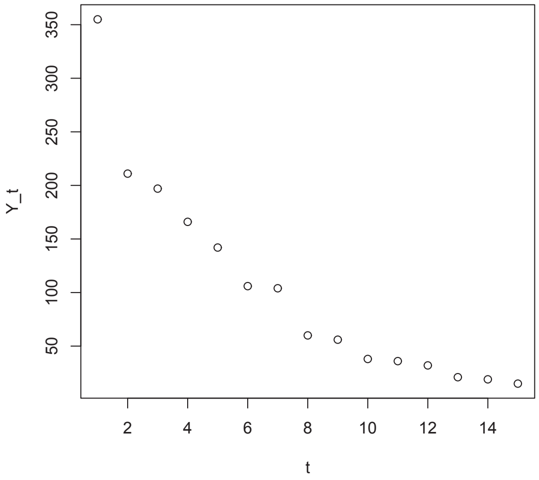
(b) Using the log-transformed data, we can fit the linear regression model as

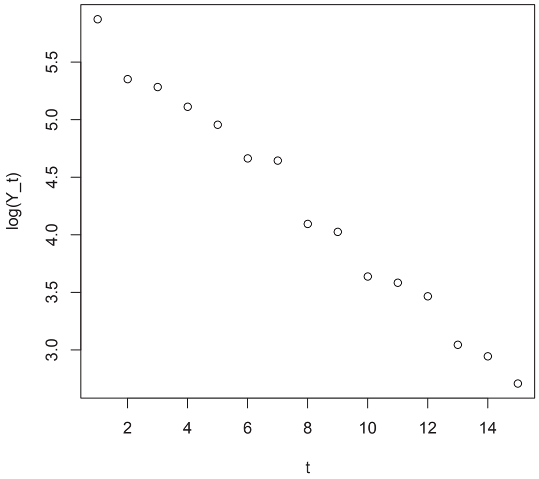
$$\hat{\log} y = 7.1458 - 0.5118 \log(x).$$

(c) When the per capita income is 4000, the infant mortality rate is predicted as

$$\hat{y} = \exp(7.1458 - 0.5118 \log(4000)) = 18.19.$$

5. (a) The scatterplots for (t, y) and $(t, \log(y))$ are given on the next page. The scatterplot of $(t, \log(y))$ suggests a linear relation.





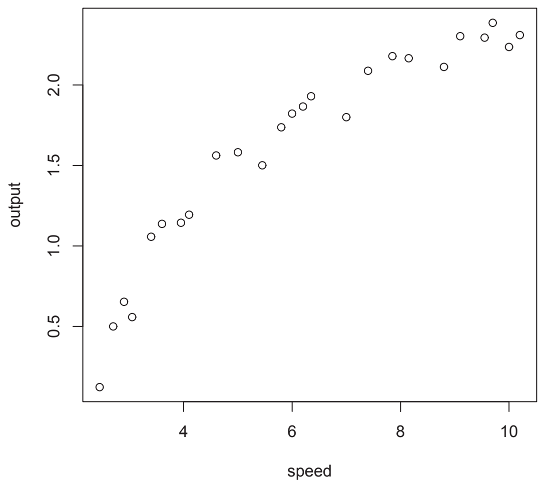
(b) Using the log-transformed data, we can fit the linear regression model as

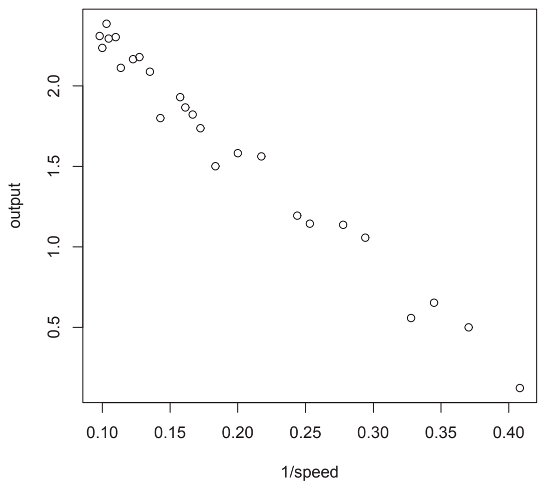
$$\hat{\log} y = 5.9732 - 0.2184t.$$

Thus, to predict the the bacteria count Y at time t, we can use

$$y = \exp(5.9732 - 0.2184t).$$

6. (a) The scatterplots for (x, y) and $(1/x, y)$ are given on the next page. The scatterplot of $(1/x, y)$ suggests a linear relation.





(b) Using the transformed data, we can fit the linear regression model as

$$\hat{y} = 2.979 - 6.935/x.$$

(c) When the wind speed is 8 miles per hour, the current produced is predicted as

$$\hat{y} = 2.979 - 6.935/8 = 2.112125.$$

12.3 Estimating, Testing, and Prediction

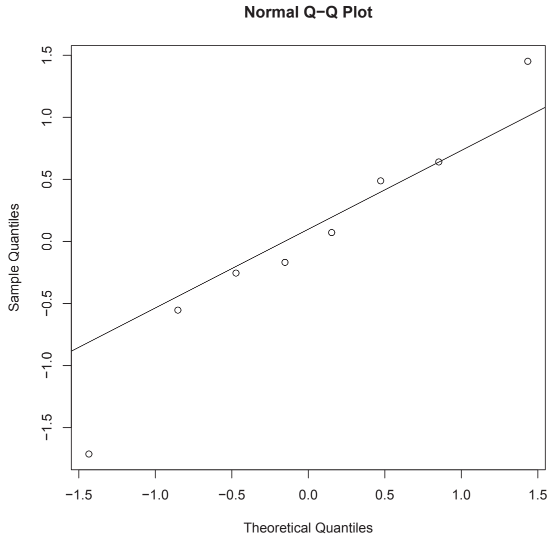
1. (a) Using the R commands *GSP = read.table("GasStatPoll.txt", header = T); attach(GSP); fit =lm(MTBE∼GS+WS+T)*, we can get the fitted model as

$$MTBE = 10.4576978 - 0.0002323GS - 4.7198342WS + 0.0033323T.$$

The R^2 value is reported as 0.8931 and the p-value for the model utility test is 0.02064, thus the model is useful for predicting MTBE concentrations.

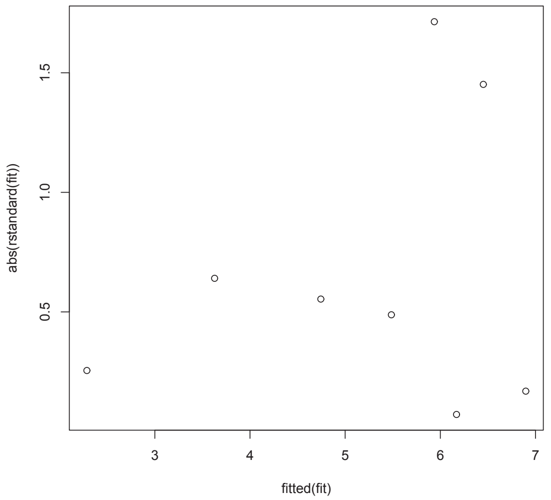
(b) The fitted value and the residual corresponding to the first observation are 5.9367 and -1.0367, respectively.

(c) To test the normal assumption, we use the commands *shapiro.test(rstandard(fit))*, which returns a p-value of 0.9256, and this suggests that the normal assumption is not contradicted by the data. The QQ plot is produced by the commands *qqnorm(rstandard(fit)); qqline(rstandard(fit))* and is given below.



The QQ-plot is consistent with the p-value result. To test the homoscedasticity assumption, we use the commands *r1=lm(abs(rstandard(fit))∼poly(fitted(fit),2)); summary(r1)* and we get a p-value of 0.7506, which suggests that the homoscedasticity assumption is not contradicted by the data. The standardized residual versus the fitted value is given by the command

plot(fitted(fit),abs(rstandard(fit))) and the plot is shown on the next page.



This plot is consistent with the p-value result.

(d) The command *summary(fit)* returns the p-value for testing the significant of each predictor. Only the predictor "Wind Speed" is significant at level 0.05 (with a p-value of 0.026).

(e) The command *confint(fit)* returns the 95% confidence intervals for the intercept and GS, WS, and T are respectively (1.295, 19.620), (-0.0022, 0.0017), (-8.515, -0.925), and (-0.1707, 0.1774).

2. (a) Using the R commands *fit =lm(y∼x1+x2+x3)*, we can get the fitted model as

$$y = 17.5238 + 0.7156x1 + 1.2953x2 - 0.1521x3.$$

The adjusted R^2 value is reported as 0.8983 and the p-value for the model utility test is 3.016×10^{-9}, thus the model is useful for predicting stackloss.

(b) The p-value for testing the significance of *x3* is 0.34405, thus it is not a useful predictor in the model. We use command *fitR=lm(y∼x1+x2)* to fit the MLR model using only x1 and x2, with no polynomial or interaction terms. The adjusted R^2 value is reported as 0.8986, very close to the adjusted R^2 value in part (a). This is consistent with the conclusion from the p-value that x3 is not useful.

(c) To get the asked confidence interval, we use the following commands

```
y=stackloss$stack.loss; x1=stackloss$Air.Flow;
x2=stackloss$Water.Temp; x3=stackloss$Acid.Conc;
m1 = mean(x1); m2 = mean(x2); m3 = mean(x3);
x1=x1-m1; x2=x2-m2; x3=x3-m3; fitR=lm(y∼x1+x2);
predict(fitR, data.frame(x1=65-m1, x2=20-m2), interval="confidence").
```

The CI is obtained as (16.72191, 21.62454).

(d) To fit the MLR model based on second order polynomials for x1 and x2, as well as their interaction, we use the command

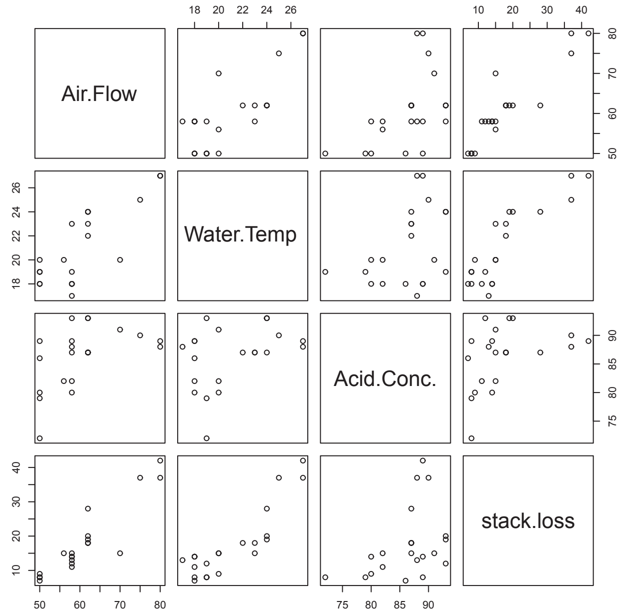
fitF=lm(y∼poly(x1,2,raw=T)+poly(x2,2,raw=T)+x1:x2).

The fitted model is

$$y = 17.19405+0.62793x1-0.03857x1^2+1.12312x2-0.06624x2^2+0.18757x1x2.$$

To test the joint (i.e., as a group) significance of the two quadratic terms and the interaction, we use the command *anova(fitR, fitF)* and it returns a p-value of 0.07238. Thus, the two quadratic terms and the interaction are significant at level 0.05.

(e) The scatterplot matrix for the data is given below.



From this figure, "Acid.Conc." appear correlated with "stack.loss." The high p-value for "Acid.Conc." in part (a) is because that the effect of "Acid.Conc." is explained by "Air.Flow" and "Water.Temp."

3. (a) The estimated regression model is

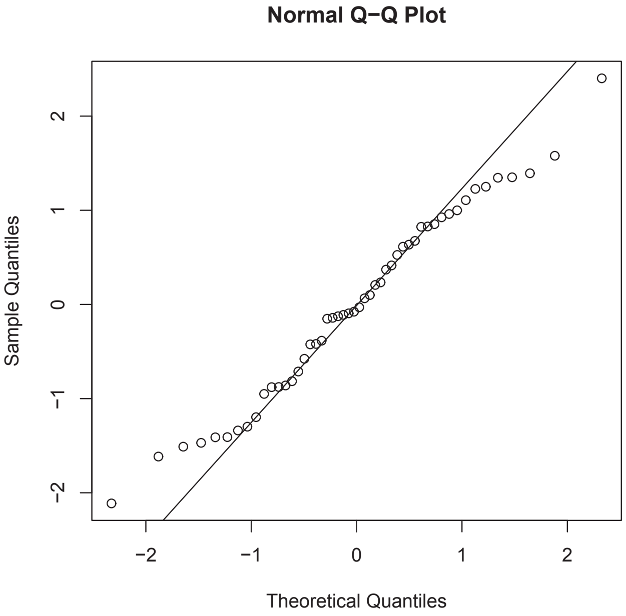
$$\hat{y} = 70.94 + 5.18 \times 10^{-5}x_1 - 2.18 \times 10^{-5}x_2 + 3.382 \times 10^{-2}x_3 \\ - 0.3011x_4 + 4.893 \times 10^{-2}x_5 - 5.735 \times 10^{-3}x_6 - 7.383 \times 10^{-8}x_7$$

and $R^2_{\text{adj}} = 0.6922$. The p-value for the model utility test is 2.534×10^{-10}. The model is useful for predicting life expectancy.

(b) Using the given command get *h2* and the command *anova(h1, h2)*, gives a p-value of 0.9993 for testing the joint significance. Thus, the variables “Income,” “Illiteracy,” and “Area” are not significant at level 0.05.

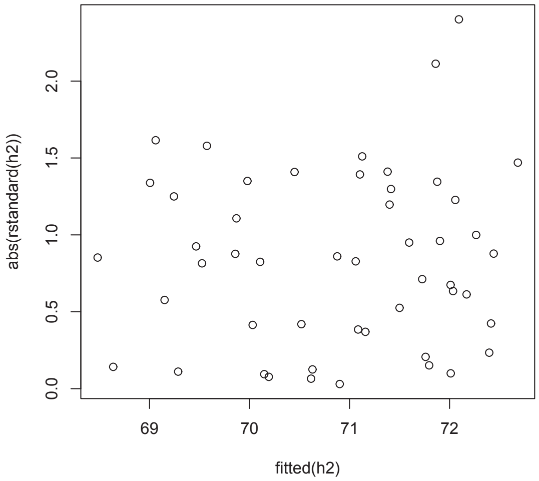
(c) The R^2 values for the full and reduced model are 0.7362 and 0.736, respectively. This is because the variables “Income,” “Illiteracy,” and “Area” are not significant. The R^2_{adj} values for the full and reduced model are 0.6922 and 0.7126, respectively. This is because in the reduced model, the almost identical R^2 is adjusted for fewer predictors.

(d) To test the normal assumption, we use the commands *shapiro.test(rstandard(h2))* which returns a p-value of 0.5606, and this suggests that the normal assumption is not contradicted by the data. The QQ plot is produced by the commands *qqnorm(rstandard(h2)); qqline(rstandard(h2))* and is given below.



The QQ-plot is consistent with the p-value result. To test the homoscedasticity assumption, we use the commands *r1=lm(abs(rstandard(h2))∼poly(fitted(h2),2)); summary(r1)* and we get a p-value of 0.7113, which suggests that the homoscedasticity assumption is not contradicted by the data. The standardized residual versus the fitted value is given by the command

plot(fitted(h2),abs(rstandard(h2))), and the plot is shown below.



This plot is consistent with the p-value result.

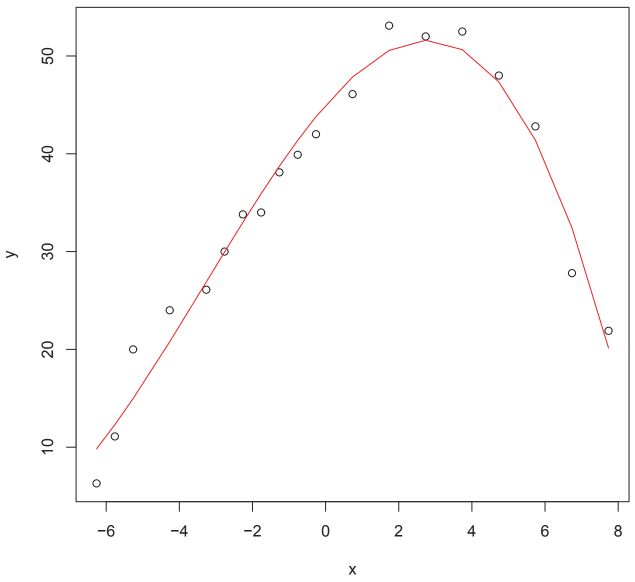
(e) The fitted value for CA is given by *fitted(h2)[5]*, which returns 71.796. To find a prediction for the life expectancy in the state of California with the murder rate reduced to 5, we use *predict(h2, data.frame(Population =21198, Income =5114 , Illiteracy= 1.1, Murder =5, HS.Grad =62.6, Frost= 20, Area =156361))* and the returned value is 73.386. To get a 95% prediction interval for life expectancy with the murder rate reduced to 5, we use *predict(h2, data.frame(Population =21198, Income =5114 , Illiteracy= 1.1, Murder =10.3, HS.Grad =62.6, Frost= 20, Area =156361), interval="prediction")* and the prediction interval is given as (71.62966, 75.14321).

4. (a) The fitted model is

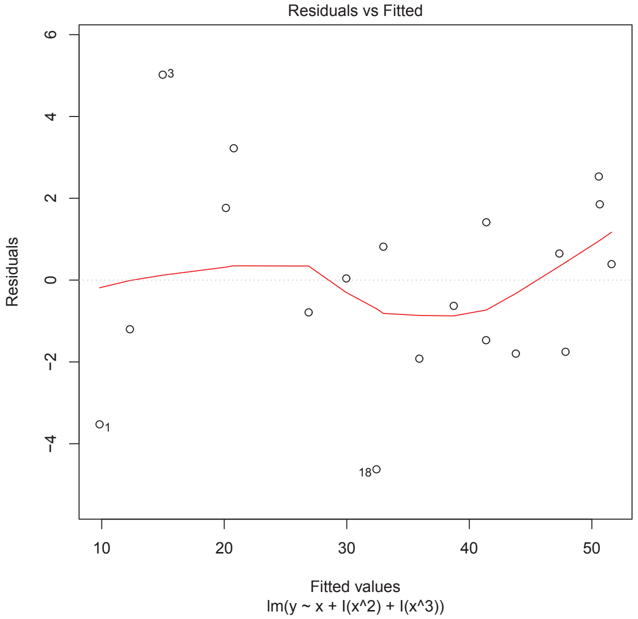
$$y = 44.97556 + 4.33939x - 0.54887x^2 - 0.05519x^3.$$

The adjusted R^2 is 0.9648 and the p-value for the model utility test is 1.025×10^{-11}, thus the model is useful. The p-value for testing the significance of x, x^2 and x^3 are 2.87×10^{-9}, 5.11×10^{-10}, and 4.72×10^{-5}, respectively. Thus they are all significant at level 0.01.

(b) The scatterplot of the data with the fitted curve superimposed is given below. This plot shows that the fit provided by the 3rd order polynomial is satisfactory.



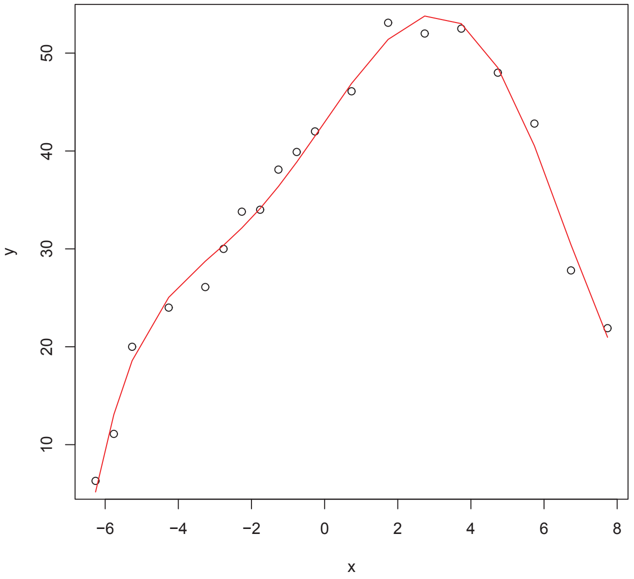
(c) The command *plot(hc3, which=1)* produces the plot below, which suggests that the fit can be improved.



For the 5th order polynomial model, the adjusted R^2 is 0.9847. When testing the significance of the coefficients at level 0.01, we found that only the coefficient of x^2 is not significant with a p-value of 0.216.

(d) Ignoring the term x^2, we fit a model by *hc52=lm(y∼x+I(*x^3*)+I(*x^4*)+I(*x^5*));*

summary(hc52). The p-values show that all the coefficients are significant at level 0.01. The adjusted R^2 is 0.984, which increased a little compared to part (a). The scatterplot of the data with the fitted curve superimposed is given below. Compared to the plot in part (b), the plot in part (d) shows that the fit increases at some points.

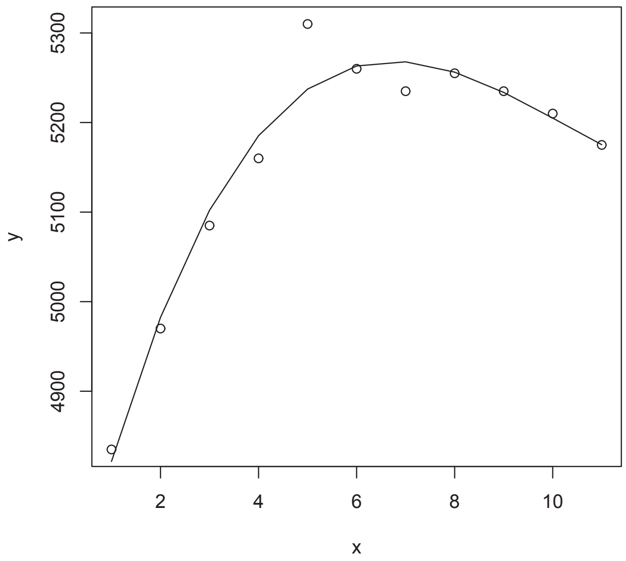


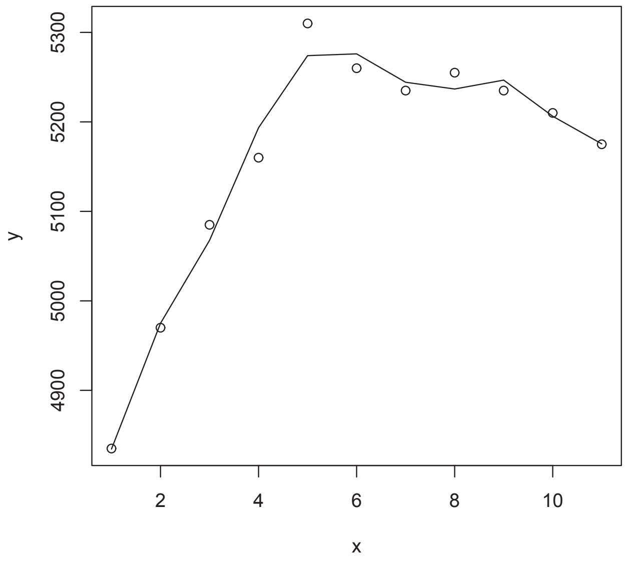
5. (a) The commands return the R^2 value of 0.962 and the adjusted R^2 value of 0.946. The p-value for the significant test is 2.403×10^{-5}, thus it is significant at level 0.01.

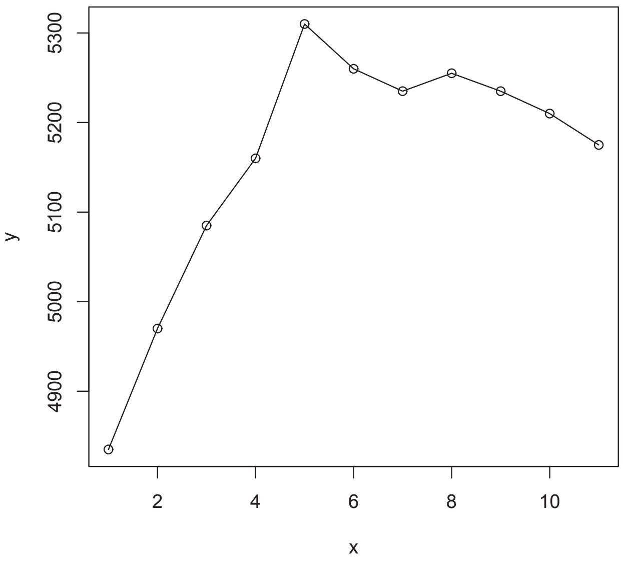
(b) To test the joint significance of the quadratic and cubic terms, we use the commands *prR = lm(y∼x); anova(pr3, prR)* and the returned p-value is 9.433×10^{-5}. Thus, the quadratic and cubic terms are jointly significant at level 0.01.

(c) To test the the joint significance of the polynomial terms of orders four through eight, we use the commands *pr8=lm(y∼poly(x, 8, raw=T)); anova(pr8, pr3)*. The commands return the p-value as 0.7917. Thus, the polynomial terms of orders four through eight are not jointly significant at level 0.01. From the fit *pr8*, we have the R^2 value 0.9824 and the adjusted R^2 value 0.912. Compared to those in (a), R^2 is somewhat bigger but the adjusted R^2 is somewhat smaller, consistent with the non-significance of the higher order polynomial term.

(d) The following figures show the scatterplot superimposed with the 3rd degree, 8th degree and 10th degree polynomial fits.


5300
5200
5100
5000
4900
y
2
4
6
8
10
x


5300
5200
5100
5000
4900
y
2
4
6
8
10
x



We can observe that as the degree of the polynomial increases, the curve fits the data better, but the curve becomes less and less smooth.

6. From (12.3.5), there is $SSR = R^2SST$, and $SSE = SST - SSR = (1 - R^2)SST$. Thus,

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{R^2SST/k}{(1-R^2)SST/(n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

as was to be proved.

12.4 Additional Topics

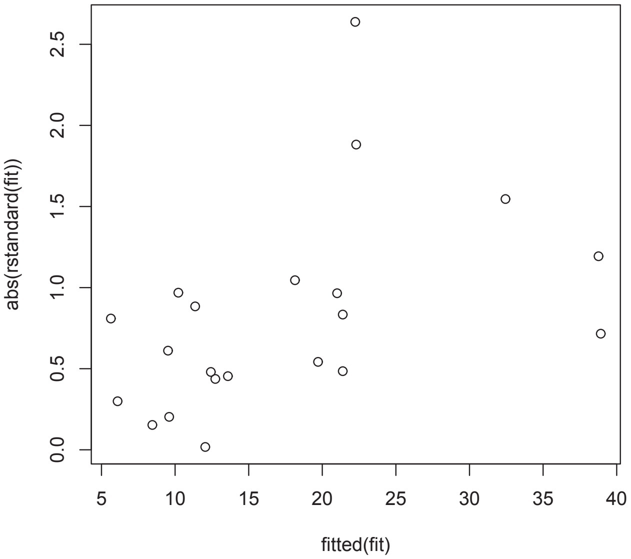
1. (a) By running the given commands, the standard errors of the slopes estimate obtained from the OLS and WLS are respectively 0.2209 and 0.1977.

   (b) By running the given commands, we have the standard deviation of the estimated slopes in OLS and WLS are respectively 0.5545655 and 0.1916188. Compare to the results in part (a), we see that OLS underestimates the variability of the estimated slope while WLS analysis estimates it correctly.

2. (a) We run the commands

y=stackloss$stack.loss; x1=stackloss$Air.Flow;
x2=stackloss$Water.Temp; x3=stackloss$Acid.Conc.;
x1=x1-mean(x1); x2=x2-mean(x2); x3=x3-mean(x3); fit = lm(y∼x1+x2+x3);
r1=lm(abs(rstandard(fit))∼poly(fitted(fit),2)); summary(r1).

The code gives us the p-value of 0.0377 for the model utility test, suggesting violation of the homoscedasticity assumption. The command *plot(fitted(fit), abs(rstandard(fit)))* gives us the plot below, which is consistent with the formal test.



(b) To fit the weighted least square model, we run the following commands

abse=abs(resid(fit)); yhat=fitted(fit); efit=lm(abse∼yhat);

*w=1/fitted(efit)**2; fitw = lm(y∼x1+x2+x3, weights=w); summary(fitw).*

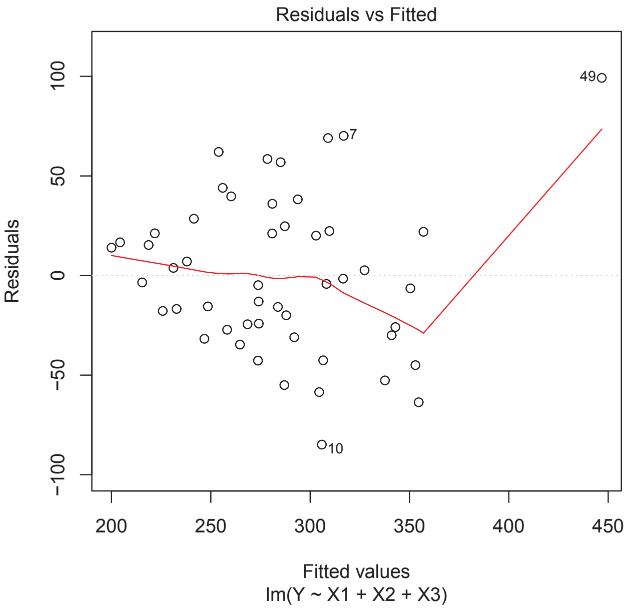
The p-value of the model utility test is reported as 1.868×10^{-8}.

(c) The 95% CI for the regression parameters are summarized in the table below.

Parameter	OLS	WLS
Intercept	(16.03, 19.02)	(16.04, 18.82)
x1	(0.43, 1.00)	(0.43, 0.94)
x2	(0.52, 2.07)	(0.41, 1.71)
x3	(-0.48, 0.18)	(-0.29, 0.19)

Clearly, WLS gives shorter CIs.

3. (a) The plot of the residuals versus the predicted values is given below.



On the basis of this plot, the homoscedasticity assumption is suspicious. To perform a formal test, we run the commands

r1=lm(abs(rstandard(edu.fit))∼poly(fitted(edu.fit),2)); summary(r1)

and the model utility test gives a p-value of 0.0002524, suggesting violation of the homoscedasticity assumption.

(b) To fit the weighted least square model, we run the following commands

abse=abs(resid(edu.fit)); yhat=fitted(edu.fit); efit=lm(abse∼yhat);
*w=1/fitted(efit)**2; fitw = lm(Y∼X1+X2+X3, weights=w, data=edu);*
summary(fitw).

(c) The 95% CI for the regression parameters are summarized in the table below.

Parameter	OLS	WLS
Intercept	(-804.5472, -308.5889)	(-564.0155, -153.4258)
X1	(0.0490, 0.0957)	(0.0426, 0.0858)
X2	(0.9187, 2.1855)	(0.4770, 1.5470)
X3	(-0.1077, 0.0992)	(-0.0634, 0.1028)

4. From the model (12.4.9), the cell mean for cell (i, j) is

$$\mu_{ij} = \mu_0 + \alpha_i + \beta_j, \quad \text{for} \quad i = 1, \cdots, a; \quad j = 1, \cdots, b.$$

According to the regression model (12.4.11), the cell mean can also be calculated as

$$\mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B, \quad \text{for} \quad i = 1, \cdots, a-1; \quad j = 1, \cdots, b-1$$

$$\mu_{aj} = \beta_0 - \beta_1^A - \cdots - \beta_{a-1}^A + \beta_j^B, \quad \text{for} \quad j = 1, \cdots, b-1$$
$$\mu_{ib} = \beta_0 + \beta_i^A - \beta_1^B - \cdots - \beta_{b-1}^B, \quad \text{for} \quad i = 1, \cdots, a-1$$

and

$$\mu_{ab} = \beta_0 - \beta_1^A - \cdots - \beta_{a-1}^A - \beta_1^B - \cdots - \beta_{b-1}^B.$$

Thus, we have a system of equations

$$\mu_0 + \alpha_i + \beta_j = \beta_0 + \beta_i^A + \beta_j^B, \quad \text{for} \quad i = 1, \cdots, a-1; \;\; j = 1, \cdots, b-1$$

$$\mu_0 + \alpha_a + \beta_j = \beta_0 - \beta_1^A - \cdots - \beta_{a-1}^A + \beta_j^B, \quad \text{for} \quad j = 1, \cdots, b-1$$
$$\mu_0 + \alpha_i + \beta_b = \beta_0 + \beta_i^A - \beta_1^B - \cdots - \beta_{b-1}^B, \quad \text{for} \quad i = 1, \cdots, a-1$$

and

$$\mu_0 + \alpha_a + \beta_b = \beta_0 - \beta_1^A - \cdots - \beta_{a-1}^A - \beta_1^B - \cdots - \beta_{b-1}^B.$$

Summing up the ab equations and using the condition that $\sum_{i=1}^a \alpha_i = 0$ and $\sum_{j=1}^b \beta_j = 0$, we have $ab\mu_0 = ab\beta_0$, that is $\beta_0 = \mu_0$.

Fixing $j \in \{1, \cdots, b-1\}$, summing up the $a-1$ equations

$$\mu_0 + \alpha_i + \beta_j = \beta_0 + \beta_i^A + \beta_j^B, \quad \text{for} \quad i = 1, \cdots, a-1$$

and

$$\mu_0 + \alpha_a + \beta_j = \beta_0 - \beta_1^A - \cdots - \beta_{a-1}^A + \beta_j^B,$$

and using $\sum_{i=1}^a \alpha_i = 0$, we have $a\beta_j = a\beta_j^B$, that is $\beta_j^B = \beta_j$, for $j = 1, \cdots, b-1$. Since $\sum_{j=1}^b \beta_j = 0$, we have

$$\beta_b = -\beta_1 - \cdots - \beta_{b-1} = -\beta_1^B - \cdots - \beta_{b-1}^B.$$

Similarly, it follows that $\beta_i^A = \alpha_i$, for $i = 1, \cdots, a-1$ and $\alpha_a = -\beta_1^A - \cdots - \beta_{a-1}^A$.

5. (a) We can define an indicator variable $X = 1$ or -1 depending on whether the observation comes from route 1 or 2.

   (b) The regression parameters β_0 and β_1 related to the population means μ_1 and μ_2 via the relations $\mu_1 = \beta_0 + \beta_1$ and $\mu_2 1 = \beta_0 - \beta_1$.

   (c) To compare the p-value from the model utility test with that from the two-sample t-test, we use the commands

   y=dd$duration; x=rep(1, length(y)); x[which(dd$route==2)]=-1;
   summary(lm(y∼x)); t.test(y∼dd$route, var.equal=T).

   Both give the p-value of 6.609×10^{-10}.

   (d) The Levene's test returns the p-value of 0.08981 and the regression test gives the p-value of 0.07751.

(e) For the WLS analysis, using the commands given in the hint, a p-value $< 2.2 \times 10^{-16}$ is returned; for the t-test without the equal variances assumption, we use *t.test(y~dd$route)* and it returns p-value of 1.626×10^{-10}. Thus, both methods suggest that the two population means are significant different at $\alpha = 0.05$.

6. We use the following code

```
R1=rep(0, length(edu$R)); R1[which(edu$R==1)]=1; R1[which(edu$R==4)]=-1;
R2=rep(0, length(edu$R)); R2[which(edu$R==2)]=1; R2[which(edu$R==4)]=-1;
R3=rep(0, length(edu$R)); R3[which(edu$R==3)]=1; R3[which(edu$R==4)]=-1;
fit = lm(Y~X1+X2+X3+R1+R2+R3, data=edu)
abse=abs(resid(fit)); yhat=fitted(fit); efit=lm(abse~yhat); w=1/fitted(efit)**2;
fitFw = lm(Y~X1+X2+X3+R1+R2+R3, weights=w, data=edu);
fitRw = lm(Y~X1+X2+X3, weights=w, data=edu); anova(fitFw, fitRw).
```

The given p-value is 0.07833, which suggests that the variable "Region" is significant at a 0.1 level of significance.

7. The commands show that the final model includes five predictors: mmax, cach, mmin, chmax, and syct. The p-values are listed as: mmax 1.18×10^{-15}, cach 5.11×10^{-6}, mmin 4.34×10^{-15}, chmax 3.05×10^{-11}, and syct 0.00539. The p-value for the model utility test is $< 2.2 \times 10^{-16}$.

8. Let Population, Income, Illiteracy, Murder, HS.Grad, Frost, and Area be x_1 to x_7, respectively.

(a) The full model is

$$\begin{aligned}\hat{y} = 70.94 + 5.18 \times 10^{-5} x_1 - 2.18 \times 10^{-5} x_2 + 3.382 \times 10^{-2} x_3\\ - 0.3011 x_4 + 4.893 \times 10^{-2} x_5 - 5.735 \times 10^{-3} x_6 - 7.383 \times 10^{-8} x_7,\end{aligned}$$

and the model without Area (x_7) is

$$\begin{aligned}\hat{y} = 70.99 + 5.188 \times 10^{-5} x_1 - 2.444 \times 10^{-5} x_2 + 2.846 \times 10^{-2} x_3\\ - 0.3018 x_4 + 4.847 \times 10^{-2} x_5 - 5.776 \times 10^{-3} x_6.\end{aligned}$$

(b) In the model from last step, the variable Illiteracy (x_3) has the largest p-value at 0.9340, thus we update the model by *h=update(h, . ~ . -Illiteracy)*, and get the model as

$$\begin{aligned}\hat{y} = 71.07 + 5.115 \times 10^{-5} x_1 - 2.477 \times 10^{-5} x_2 - 0.3 x_4\\ + 4.776 \times 10^{-2} x_5 - 5.910 \times 10^{-3} x_6.\end{aligned}$$

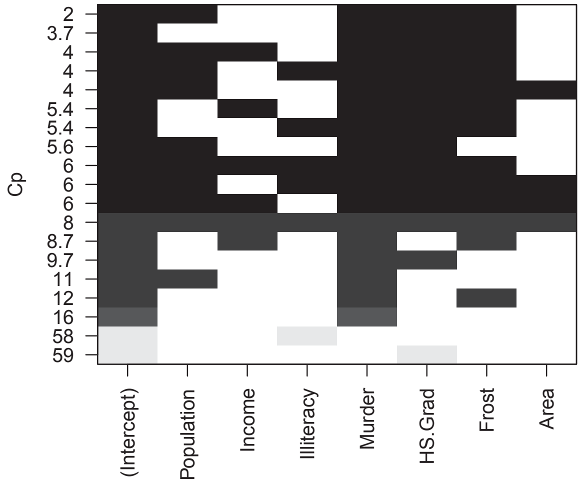
Now the variable Income (x_2) has the largest p-value at 0.9153. We update the model by *h=update(h, . ∼. -Income)* and the obtained model is

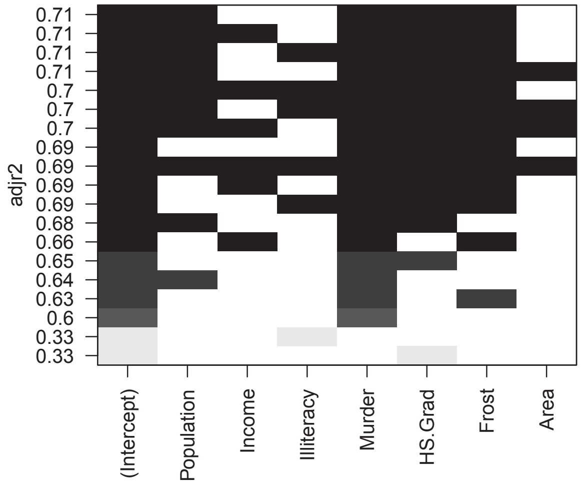
$$\hat{y} = 71.03 + 5.014 \times 10^{-5} x_1 - 0.3001 x_4 + 4.658 \times 10^{-2} x_5 - 5.943 \times 10^{-3} x_6.$$

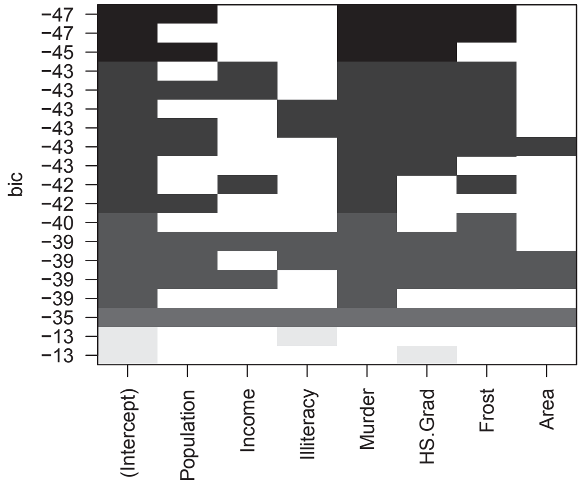
Now, every variable has p-value less than 0.1 and we stop removing variables.

The R^2 for the final model is 0.736, very close to that of the full model (0.7362). This verifies that the removed variables have very little contribution to the model.

9. We use the commands *library(leaps); vs.out = regsubsets(Life.Exp∼ . , nbest=3, data=st); plot(vs.out, scale="Cp"); plot(vs.out, scale="adjr2"); plot(vs.out, scale="bic")* to create the plots for ordering the models, using C_p, adjusted R^2, and BIC, respectively. The plots are given below.

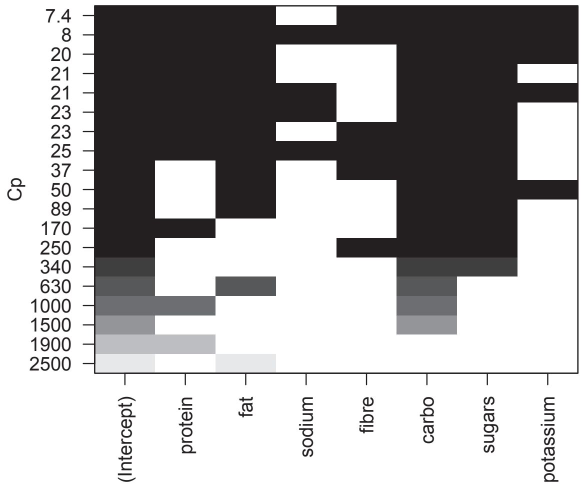


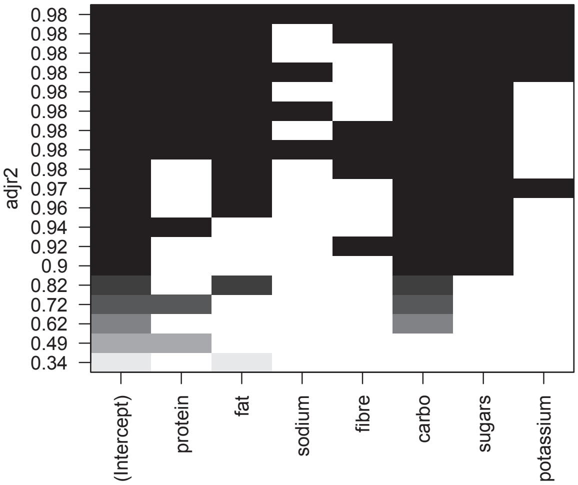


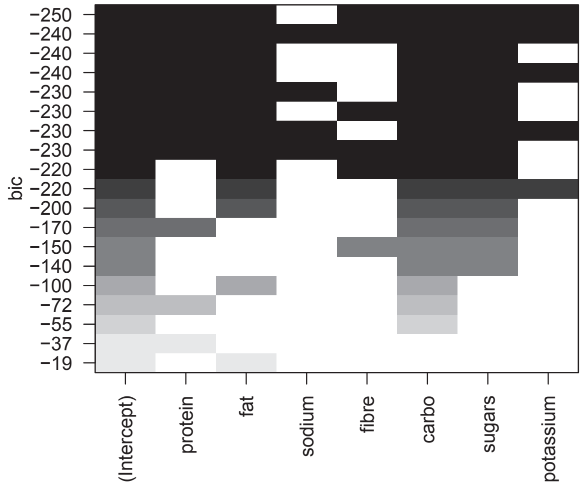


We see that all the three selection criteria give us the same best set of variables – that is, removing Income, Illiteracy, and Area. We also notice that this result is the same as that of Exercise 8.

10. (a) We use the commands *library(leaps); vs.out = regsubsets(calories∼ . , nbest=3, data=uscer); plot(vs.out, scale="Cp"); plot(vs.out, scale="adjr2"); plot(vs.out, scale="bic")* to create the plots for ordering the models, using C_p, adjusted R^2, and BIC, respectively. The plots are given below.

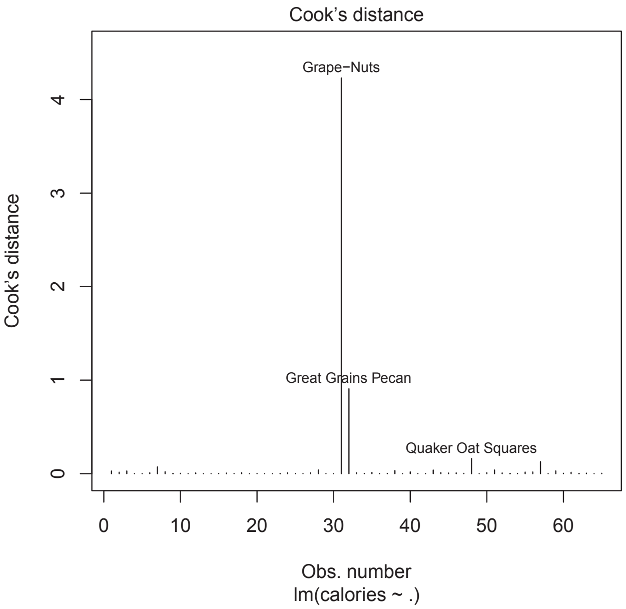

Cp
7.4
8
20
21
21
23
23
25
37
50
89
170
250
340
630
1000
1500
1900
2500
(Intercept)
protein
fat
sodium
fibre
carbo
sugars
potassium


adjr2
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.98
0.97
0.96
0.94
0.92
0.9
0.82
0.72
0.62
0.49
0.34
(Intercept)
protein
fat
sodium
fibre
carbo
sugars
potassium

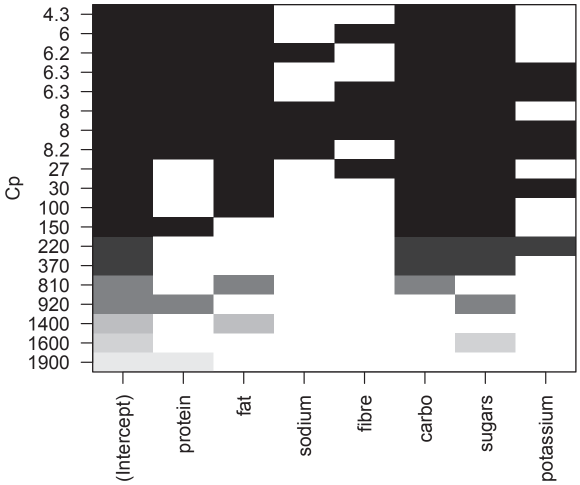


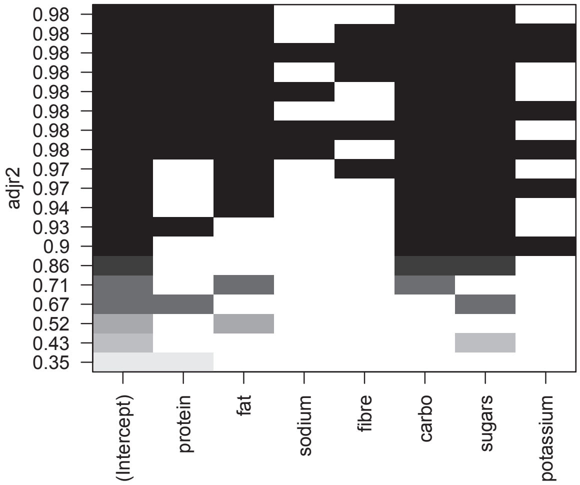
We see that the adjusted R^2 criteria gives the full model as the best, while C_p and BIC give the best model as the model without the "sodium" predictor.

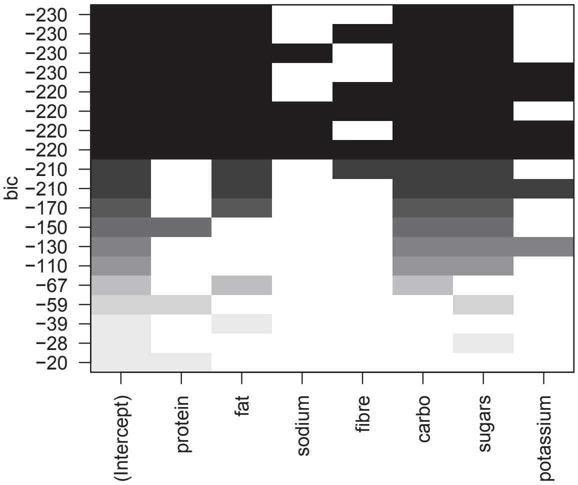
(b) The plot is given below. According to the rule of thumb, we identify observations 31 and 32 in the *uscer* data set as influential.



(c) We remove the influential observations by command *uscer1 = uscer[-c(31,32),]*, and then run the commands *vs.out = regsubsets(calories∼ . , nbest=3, data=uscer1); plot(vs.out, scale="Cp"); plot(vs.out, scale="adjr2"); plot(vs.out, scale="bic")* to create the plots for ordering the models, using C_p, adjusted R^2, and BIC, respectively. The plots are given below.







We see that all the three criterion give the same model as final model – that is, the model without "sodium," "fiber," and "potassium." The final model seems reasonable.

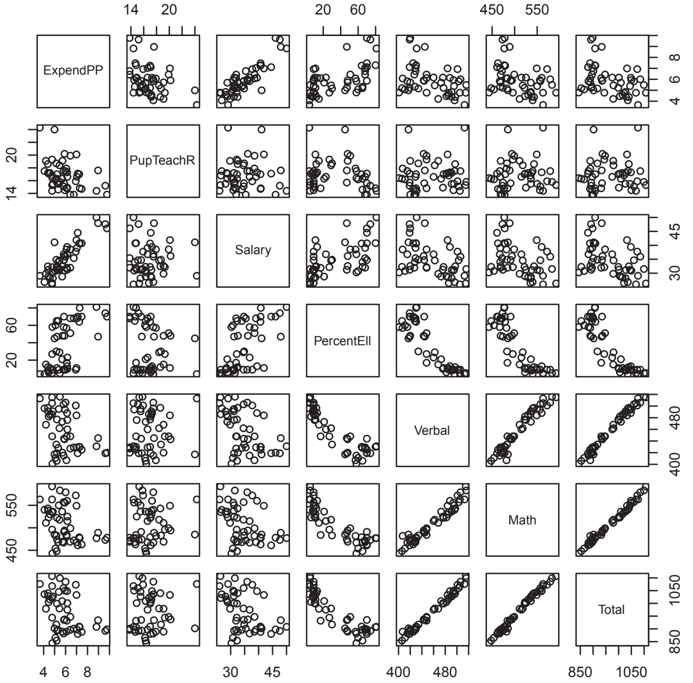
11. (a) The p-values for x_1, x_2, x_3 and x_4 are 0.0708, 0.5009, 0.8959, and 0.8441, respectively. Thus, no variable is significant at level 0.05. The R^2 value of 0.9824 and the p-value of 4.756×10^{-7} for the model utility test suggest that at least some of the variables should be significant. This is probably due to multicollinearity.

(b) Using command *vif(hc.out)*, the variance inflation factors for each variable are respectively 38.49621, 254.42317, 46.86839, and 282.51286 and they indicate that multicollinearity is an issue with this data.

(c) x_4 has the highest variance inflation factor and we remove it from the model by the command *hc1.out=update(hc.out, .∼. -x4)*. In the new model, $x1$ and $x2$ are both significant at level 0.05. The R^2 value and the adjusted R^2 are very close to those of the full model.

(d) Starting from the full model, we first remove $x3$ because it has the highest p-value. Then we remove $x4$, getting a model with only $x1$ and $x2$ – both variables have p-values less than 0.15. The final model has adjusted R^2 of 0.9744 and, compared to that of the full model (0.9824), there is not much loss.

12. (a) The scatterplot matrix is given below.



The Salary vs Total scatterplot suggests that increasing teacher salary will have a negative effect on student SAT scores.

Using the command *summary(lm(Total∼Salary, data=sat))*, we fit a least squares line through the scatterplot and the value of the slope is -5.540, verifying our observation from the scatterplot.

(b) Using command

summary(lm(formula = Total ∼ Salary + ExpendPP + PupTeachR + PercentEll, data = sat)),

we fit the MLR model for predicting the total SAT scores in terms of all available covariate. The coefficient for Salary now is 1.6379. This suggests that increasing teacher salary, while keeping all other predictor variables the same, appears to have a positive effect on student SAT scores. This is not compatible with the answer in part (a). The reason is because the predictor "Salary" has strong linearly dependencies with the predictor "ExpendPP" and "PercentEll," as demonstrated in the scatterplot matrix.

(c) Only the predictor "PercentEll" has a p-value of less than 0.05, thus it is the only significant predictor at 5% level.

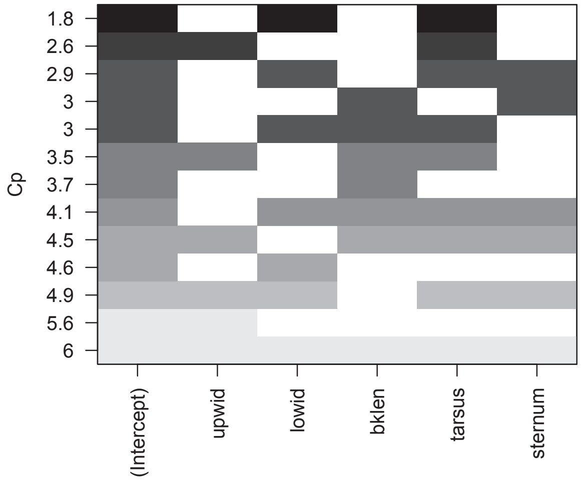
(d) $R^2 = 0.8246$, $R^2_{\text{adj}} = 0.809$, and the p-value for the model utility test is $< 2.2 \times 10^{-16}$. These values are consistent with finding a significant variable in part (c).

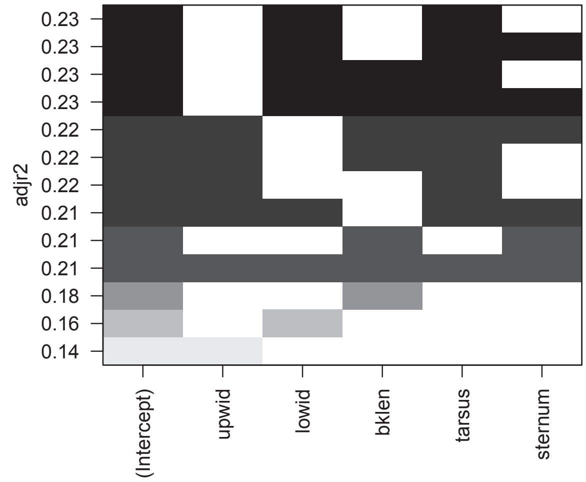
(e) Using the command *vif(lm(Total∼Salary+ExpendPP+PupTeachR+ PercentEll, data=sat))*, the variance inflation factor for the variables Salary, ExpendPP, PupTeachR, and PercentEll are respectively 9.217237, 9.465320, 2.433204, and 1.755090. They suggest multicollinearity of the predictors.

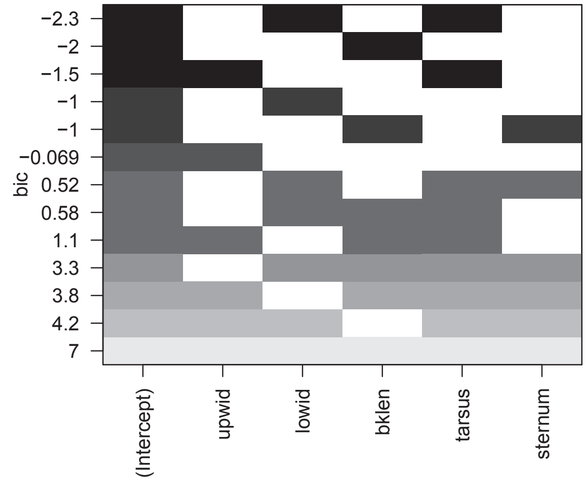
13. (a) The commands give $R^2 = 0.3045$, $R^2_{\text{adj}} = 0.2105$, and the p-value for the model utility test as 0.01589. Thus, the model utility is significant at 0.05 level. There is no predictor significant at 0.05 level.

(b) Using command *vif(lf.out)*, the variance inflation factors for upwid, lowid, bklen, tarsus, and sternum are respectively 13.951765, 14.188745, 1.924634, 1.326368, and 1.203584. These values suggest that multicollinearity is an issue with this data. The side effect of multicollinearity is the phenomenon of a significant model utility test when all predictors are not significant.

(c) We use the commands *library(leaps); vs.out = regsubsets(wt∼ . , nbest=3, data=lf); plot(vs.out, scale="Cp"); plot(vs.out, scale="adjr2"); plot(vs.out, scale="bic")* to create the plots for ordering the models, using C_p, adjusted R^2, and BIC, respectively. The plots are given below.







We see that all the three selection criteria give us the same best set of variables – that is, lowid and tarsus.

(d) Using the commands *lf.R=lm(wt∼ lowid + tarsus , data=lf); summary(lf.R)*, $R^2 = 0.2703$ and $R^2_{\text{adj}} = 0.2338$, and the p-value for the model utility test is 0.0018. Compared to the model obtained in part (a), the reduced model has a somewhat smaller R^2, somewhat larger R^2_{adj}, and smaller p-value for the model utility test. The new variance inflation factors are both equal to 1.0098.

Multicollinearity is not an issue now.

14. (a) We use the commands *fit =glm(y∼x,family=binomial()); summary(fit)* to get the fitted model as

$$\hat{y} = \frac{e^{-3.3367+0.8139x}}{1+e^{-3.3367+0.8139x}}.$$

The p-value for testing the significance of the stress variable is 0.00786, thus the variable is significant at level 0.05.

(b) To estimate the probability of failure at stress level 3.5, we use the command *predict(fit, list(x=3.5), type="response")*, which returns the probability as 0.380359.

(c) To fit the logistic regression model that includes a quadratic term of the stress variable, we use the command *fit =glm(y x+I(x**2),family=binomial()); summary(fit)* and the fitted model is returned as

$$\hat{y} = \frac{e^{-2.69440+0.46834x+0.04236x^2}}{1+e^{-2.69440+0.46834x+0.04236x^2}}.$$

Using the command *confint(fit)*, we get 95% CI for the interception, the coefficients for x and x^2 as (-12.108, 4.399), (-3.357, 5.068), and (-0.494, 0.526). According to the CIs, at 5% significant level, we cannot reject the hypothesis that the coefficient of the quadratic component is zero.

Chapter 13
Statistical Process Control

13.2 The $\bar{X}$ Chart

1. (a) The probability that a 2.7σ $\bar{X}$ chart issues a false alarm is

$$P_{\mu_0,\sigma_0}\left(\bar{X} < \mu_0 - 2.7\frac{\sigma_0}{\sqrt{n}}\right) + P_{\mu_0,\sigma_0}\left(\bar{X} > \mu_0 + 2.7\frac{\sigma_0}{\sqrt{n}}\right) = 2\Phi(-2.7).$$

The corresponding ARL is $1/(2\Phi(-2.7))$, which can be calculated by the command *1/(2*pnorm(-2.7))*, and the result is 144.2.

The probability that a 3.1σ $\bar{X}$ chart issues a false alarm is

$$P_{\mu_0,\sigma_0}\left(\bar{X} < \mu_0 - 3.1\frac{\sigma_0}{\sqrt{n}}\right) + P_{\mu_0,\sigma_0}\left(\bar{X} > \mu_0 + 3.1\frac{\sigma_0}{\sqrt{n}}\right) = 2\Phi(-3.1).$$

The corresponding ARL is $1/(2\Phi(-3.1))$, which can be calculated by the command *1/(2*pnorm(-3.1))*, and the result is 516.7.

(b) From the symmetry of the standard normal distribution, for any real value x, $\Phi(-x) = 1 - \Phi(x)$. Thus, for $\Delta = -|\Delta|$, (13.2.13) becomes

$$\begin{aligned}\Phi(-3 + \sqrt{n}|\Delta|) + 1 - \Phi(3 + \sqrt{n}|\Delta|) &= 1 - \Phi(3 - \sqrt{n}|\Delta|) + 1 - [1 - \Phi(-3 - \sqrt{n}|\Delta|)]\\ &= \Phi(-3 - \sqrt{n}|\Delta|) + 1 - \Phi(3 - \sqrt{n}|\Delta|),\end{aligned}$$

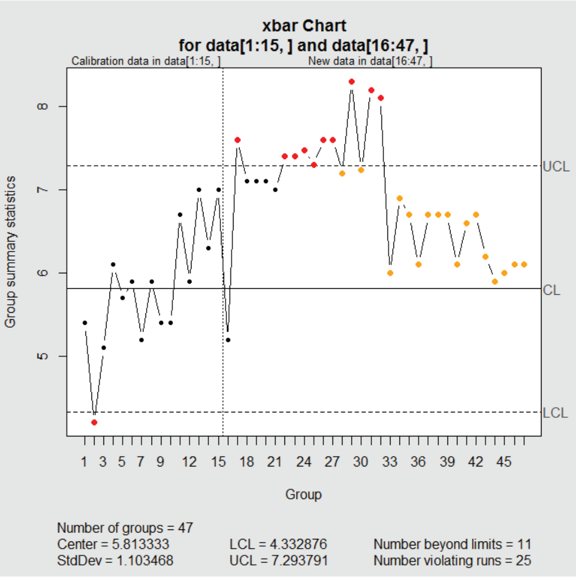
and this finishes the proof.

(c) We use the commands *Delta=1; n=c(3:7); p=1+pnorm(-3-sqrt(n)*Delta)-pnorm(3-sqrt(n)*Delta); p; 1/p*. The probability of an out-of-control signal and the corresponding ARL are listed as

n	3	4	5	6	7
Probability	0.1024092	0.1586555	0.2224540	0.2909847	0.3615763
ARL	9.764752	6.302963	4.495312	3.436606	2.765668

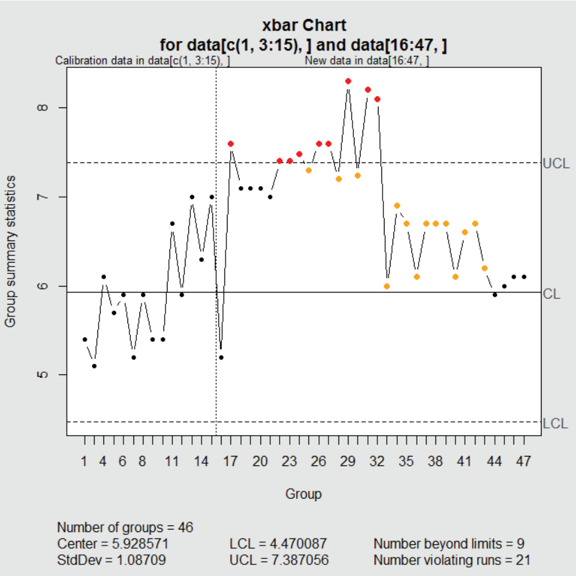
2. We load the data by commands *SSL=read.table("SqcSyringeL.txt", header =T); data = qcc.groups(SSL$x, SSL$sample)*.

(a) We use the commands *require(qcc); qcc(data[1:15,],type="xbar",newdata=data[16:47,])* to construct the 3σ $\bar{X}$ chart with the standard deviation estimated by $\hat{\sigma}_2$. The chart is given below.



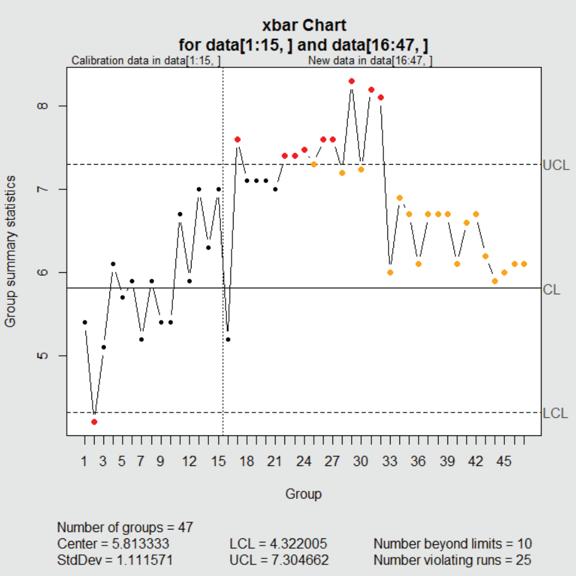
The chart suggests that the adjustment made after the 32nd sample brought the subgroup means within the control limits. Even so, the adjustment did not bring the process back in control since there are more than eight points on the same side of the central line.

(b) We use the commands *qcc(data[c(1, 3:15),], type="xbar", newdata=data[16:47,])* to construct the 3σ $\bar{X}$ chart with the standard deviation estimated by $\hat{\sigma}_2$, with the second point deleted. The chart is given below.

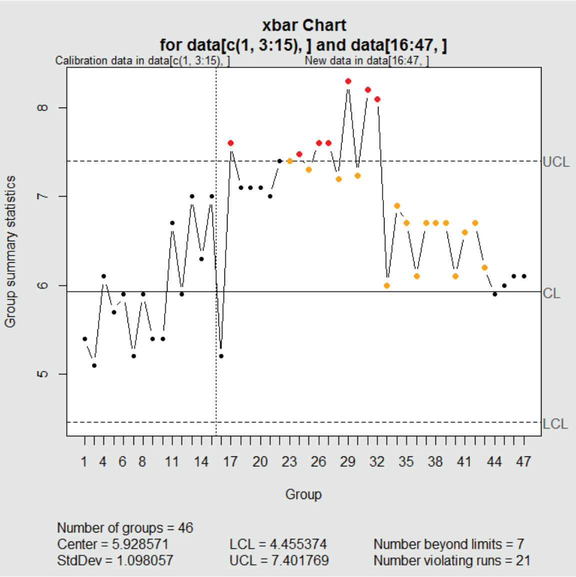


Comparing to part (a), the conclusions do not change.

(c) To construct the 3σ $\bar{X}$ chart with the standard deviation estimated by $\hat{\sigma}_1$, we use the following command *qcc(data[1:15,],type=“xbar”,newdata=data[16:47,], std.dev=“UWAVE-SD”)*, and the chart is given below.



With the second point deleted, the command is *qcc(data[c(1, 3:15),],type=“xbar”,newdata=data[16:47,], std.dev=“UWAVE-SD”)*, and the chart is given below.



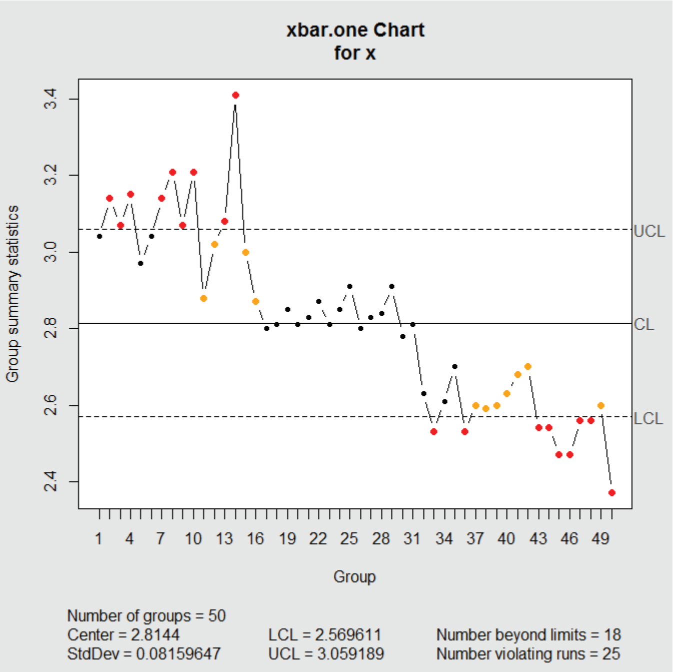
By comparing these charts to those in parts (a) and (b), we can see that all the conclusions are the same.

(d) We use the commands

x=SSL$x/100+4.9; sum(x[161:235]>=4.92 & x[161:235]<=4.98) to get the number of measurements within the limits. The commands returned 74 and, since there are 75 measurements after the adjustment, therefore the process yield after adjustment is 98.67%.

3. We load the data by commands *SCV=read.table("SqcCoolVisc.txt", header=T); x=SCV$x.*

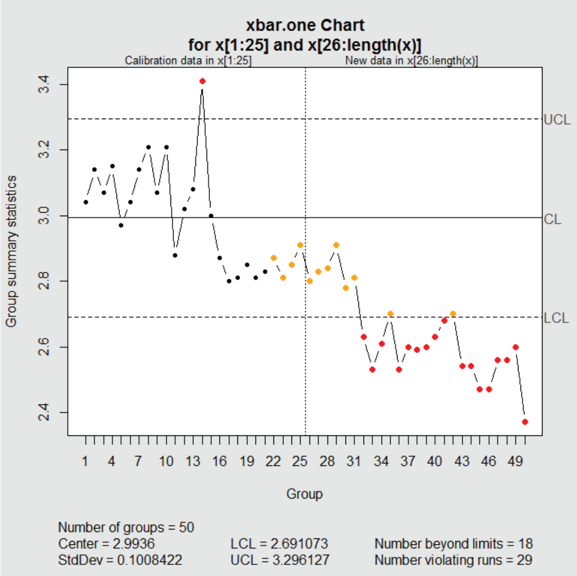
   (a) We use the command *qcc(x, type="xbar.one")* to get the 3σ X chart with the center and standard deviation computed from the entire data set, and it is shown below.



```
Number of groups = 50
Center = 2.8144            LCL = 2.569611        Number beyond limits = 18
StdDev = 0.08159647        UCL = 3.059189        Number violating runs = 25
```

   There are 18 points, marked in red, falling outside the control limits. There are 11 points, marked in yellow, suggesting an out-of-control state according to the General Electric supplemental rules.

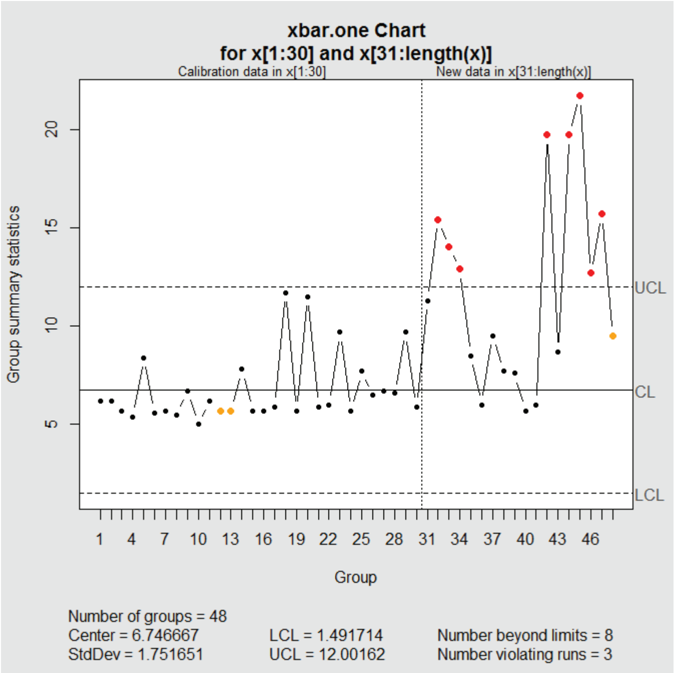
   (b) We use the command *qcc(x[1:25], type="xbar.one", newdata=x[26:length(x)])* to get the 3σ X chart with the center and standard deviation computed from the first 25 observations, taken when the process is believed to be in control, and it is shown on the next page.



There are 18 points, marked in red, falling outside the control limits. There are 12 points, marked in yellow, suggesting an out-of-control state according to the General Electric supplemental rules.

4. We load the data by commands *SLP=read.table("SqcLaborProd.txt", header=T); x=SLP$x.*

(a) We use the command *qcc(x[1:30], type="xbar.one", newdata=x[31:length(x)])* to get the 3σ X chart with the center and standard deviation computed from the first 30 observations, and it is shown below.

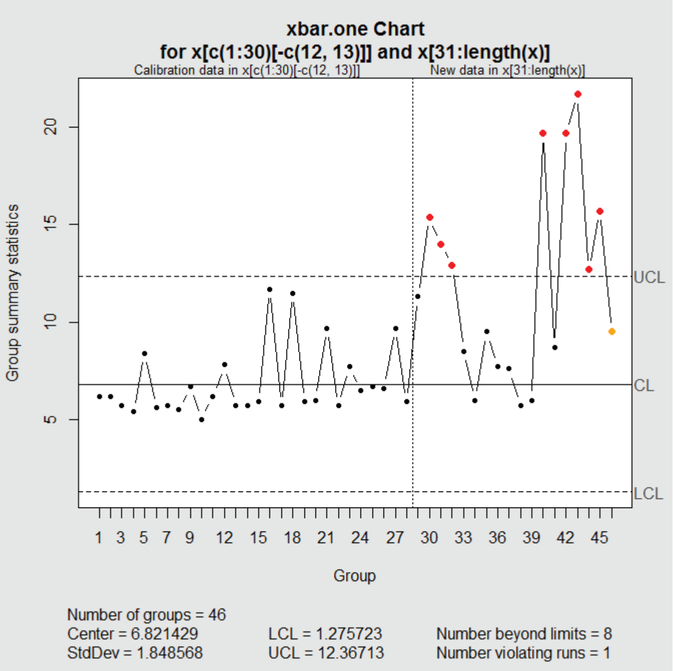


The graph shows that, during the calibration period, there are two points, marked in yellow, suggesting an out-of-control state according to the General Electric supplemental rules. After the calibration period, there are eight points, marked in red, falling outside the control limits. Thus, the process is not in control either during the calibration period or after the calibration period.

(b) We use the command

qcc(x[c(1:30)[-c(12,13)]], type="xbar.one", newdata=x[31:length(x)])

to get the 3σ X chart with the center and standard deviation computed from the first 30 observations, with the 12th and 13th daily measurements removed, and it is shown below.



The graph shows that, during the calibration period, the process is in control. After the calibration period, there are eight points, marked in red, falling outside the control limits. Thus, the process is not in control after the calibration period.

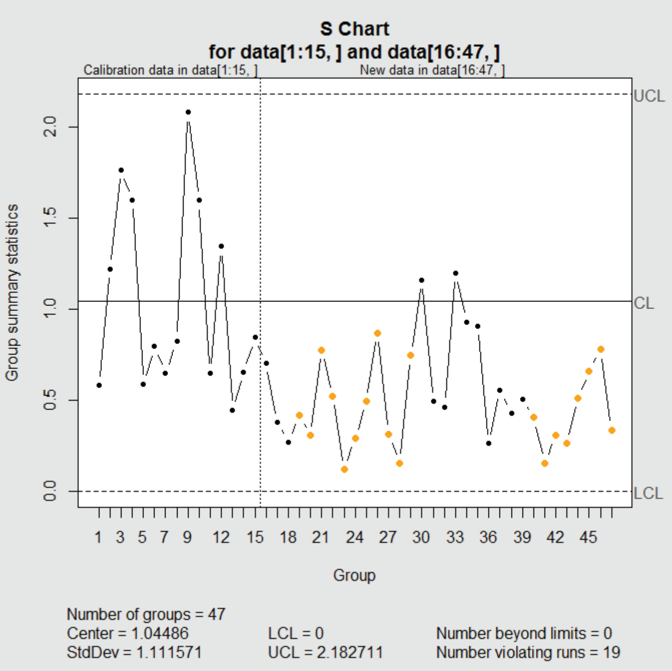
13.3 The S and R Charts

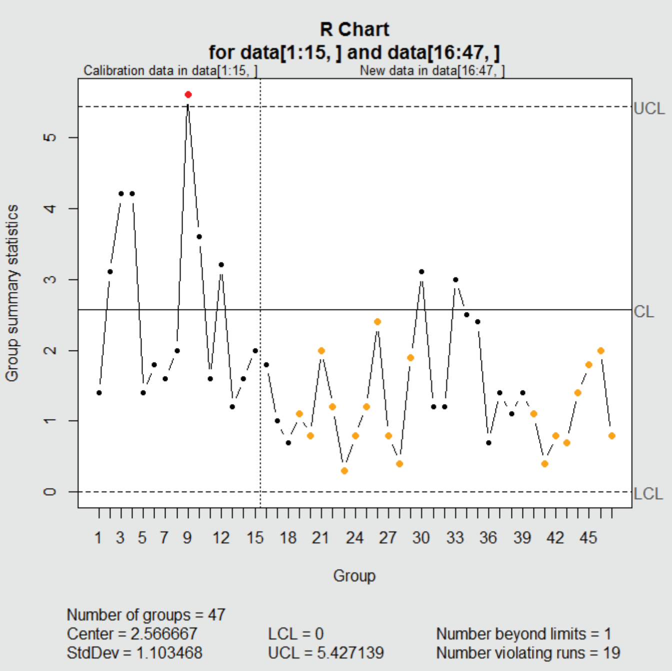
1. We load the data by commands *SSL=read.table("SqcSyringeL.txt", header =T); data = qcc.groups(SSL\$x, SSL\$sample)*. We use the commands

require(qcc); qcc(data[1:15,],type="S",newdata=data[16:47,])

to construct the S chart, and the command *qcc(data[1:15,],type="R",newdata=data[16:47,])*

to construct the R chart. These charts are given below.

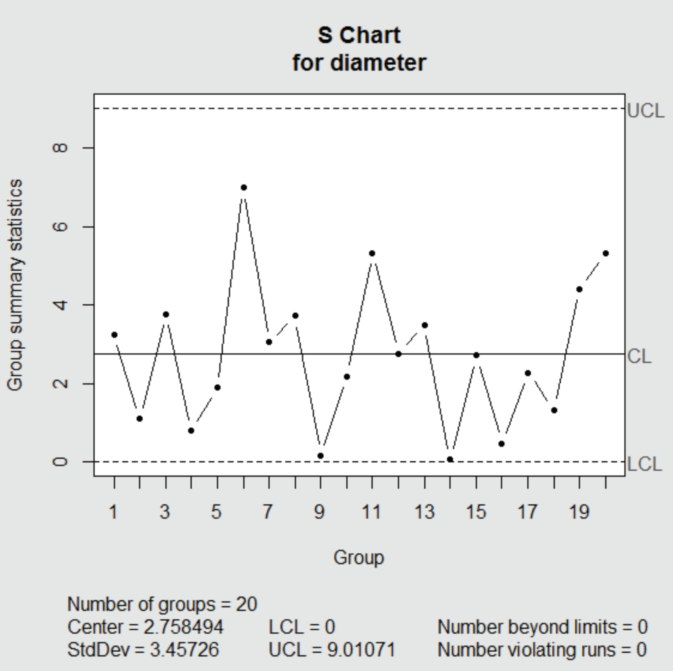




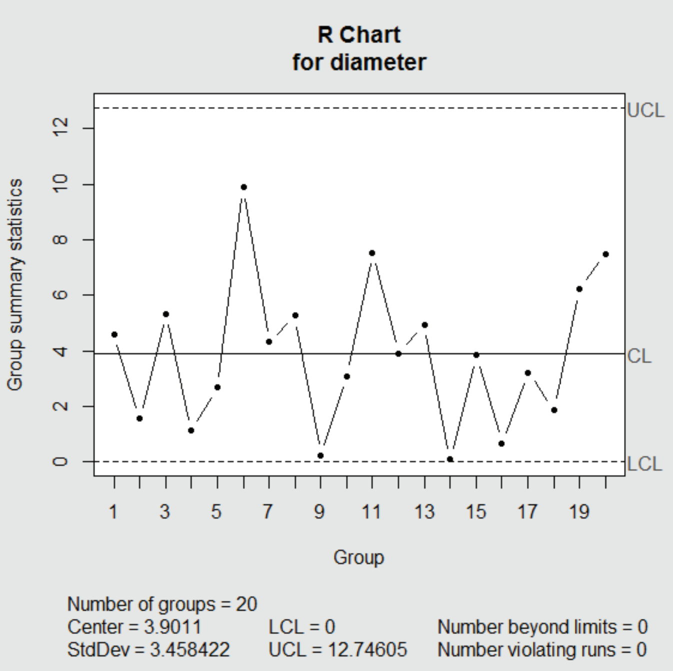
Both charts show most points after the calibration period to be below the center line. It appears that the adjustment had no effect on the process variability.

2. We load the data by commands *SSD=read.table("SqcSemicondDiam.txt", header=T); attach(SSD); diameter=qcc.groups(x,lot)*.

(a) We use the command *qcc(diameter, type =“S”)* to get the S chart and use *qcc(diameter, type =“R”)* to get the R chart, and they are shown below.



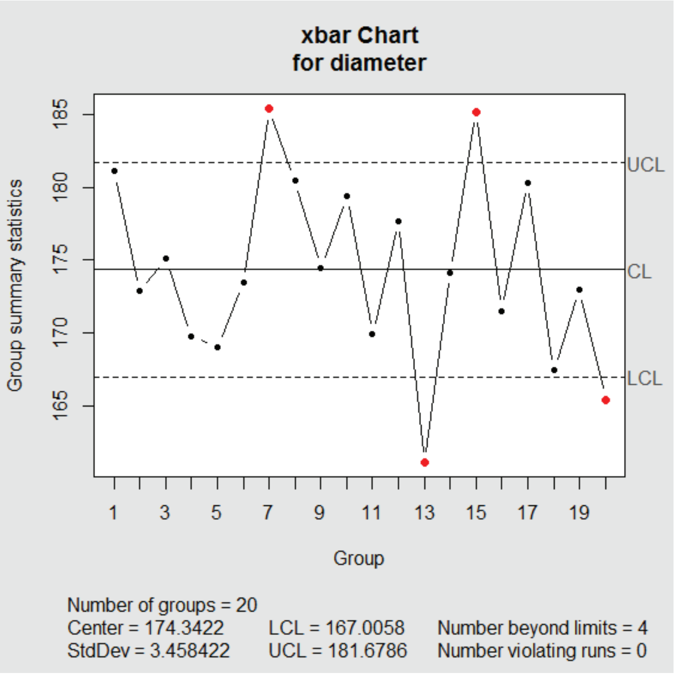
```
Number of groups = 20
Center = 2.758494      LCL = 0           Number beyond limits = 0
StdDev = 3.45726       UCL = 9.01071     Number violating runs = 0
```



```
Number of groups = 20
Center = 3.9011        LCL = 0           Number beyond limits = 0
StdDev = 3.458422      UCL = 12.74605    Number violating runs = 0
```

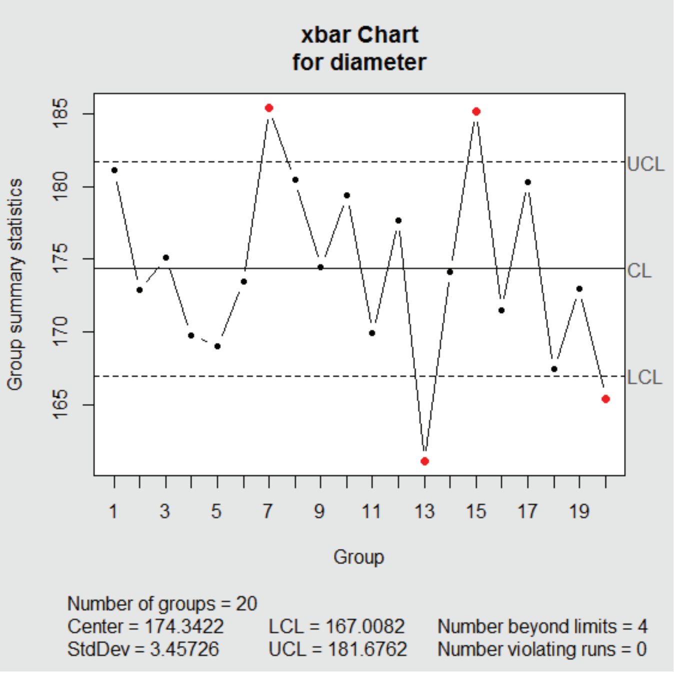
Clearly, these charts suggest that the process variability is in control.

(b) We use the command *qcc(diameter, type =“xbar”, std.dev=“UWAVE-R”)* to construct the 3σ $\bar{X}$ chart with the standard deviation estimated by $\hat{\sigma}_2$ given in (13.2.5). The chart is given on the next page.



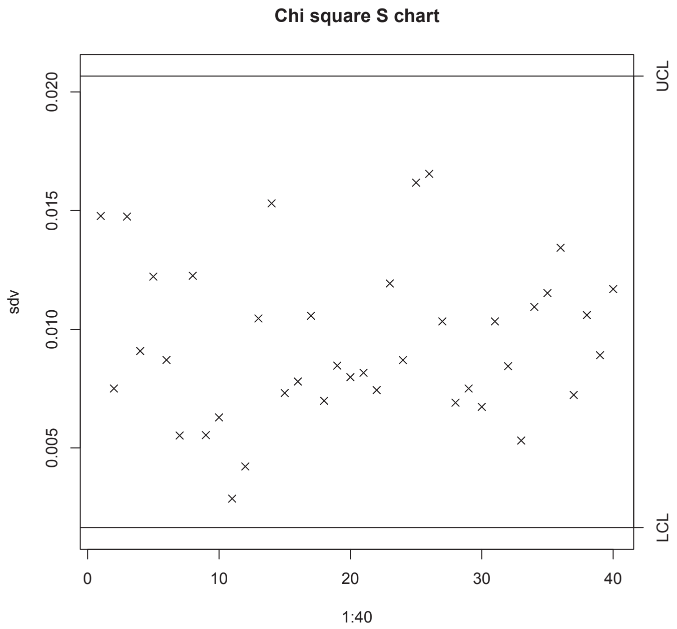
This chart shows that the process mean is out of control.

(c) We use the command *qcc(diameter, type =“xbar”, std.dev=“UWAVE-SD”)* to construct the 3σ $\bar{X}$ chart with the standard deviation estimated by $\hat{\sigma}_1$ given in (13.2.5). The chart is given below.



By comparing to the chart is part (b), the conclusion does not change.

3. Using the commands given in the book, the constructed χ^2-based S chart for the *pistonrings* data is shown below.

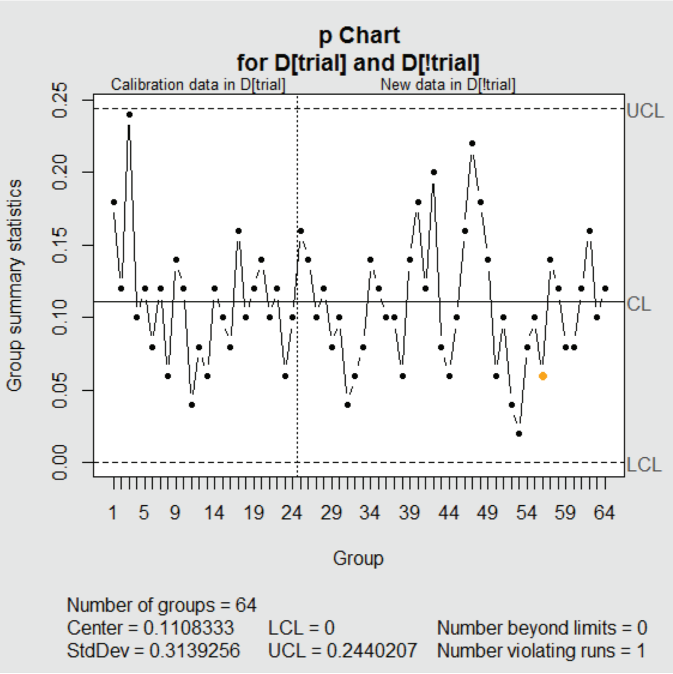


13.4 The p and c Charts

1. To construct a 3σ p chart using the first 24 samples in the *orangejuice2* data frame as the calibration data, we use the following commands:

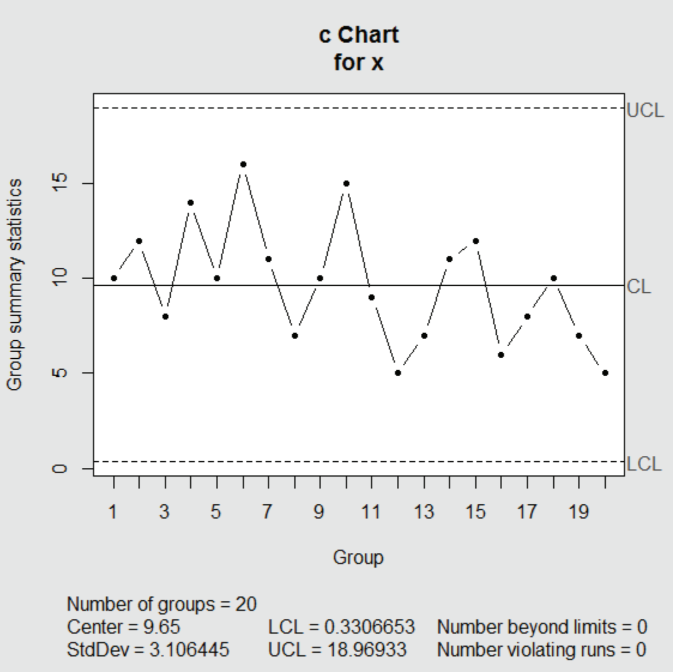
```
data(orangejuice2); attach(orangejuice2);

qcc(D[trial], sizes=size[trial], type="p", newdata=D[!trial], newsizes=size[!trial]).
```

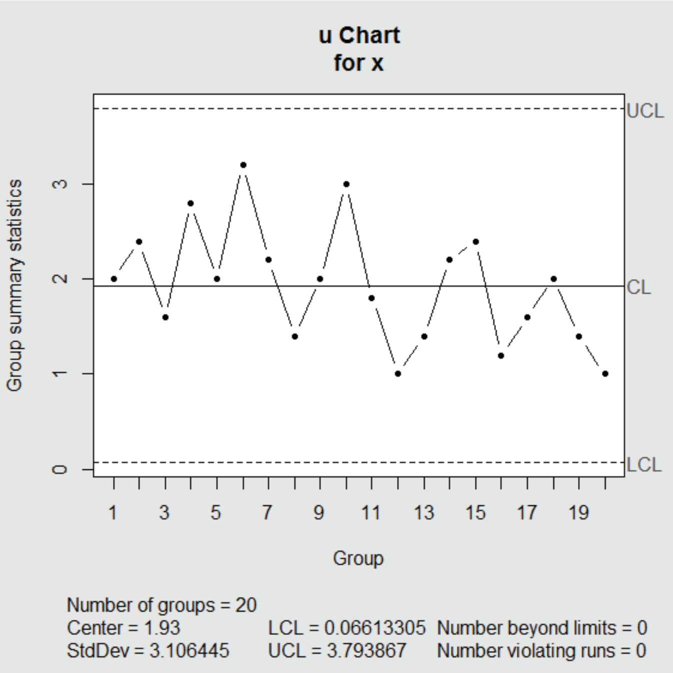
The p chart is given on the next page.



This chart suggests that the process remains in control.

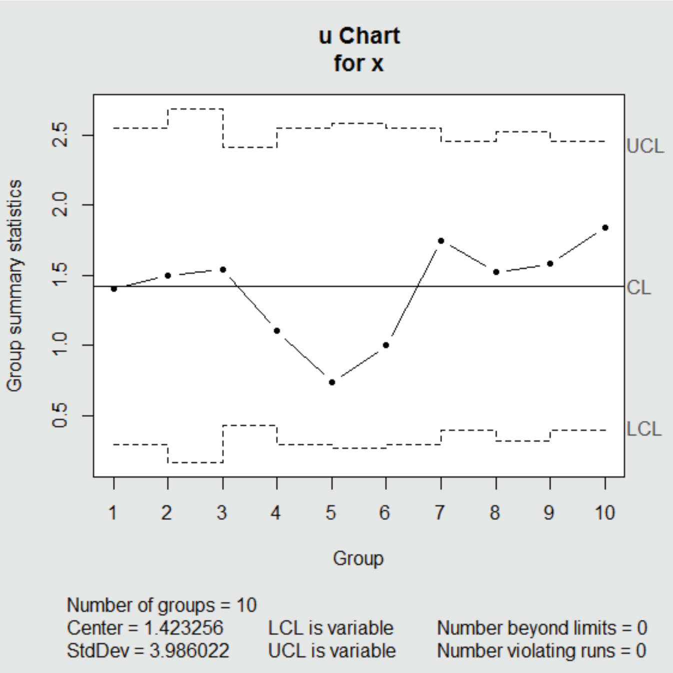
2. We use commands *data(pcmanufact); attach(pcmanufact)* to import data in the R session. We use commands *require(qcc); qcc(x, sizes=size, type="c")* to construct a 3σ c chart, with λ estimated from the entire data set, and the 3σ u chart is constructed by the command *qcc(x, sizes=size, type="u")*. The charts are given below.





When the Poisson count pertains to the total number of nonconformities in batches of items, a u chart plots the average number of nonconformities per unit. These charts show that the process is in control.

3. The u chart with unequal sample sizes could be obtained by the following commands: *data(dyedcloth); attach(dyedcloth); qcc(x,sizes=size,type="u")*. The chart is given below.

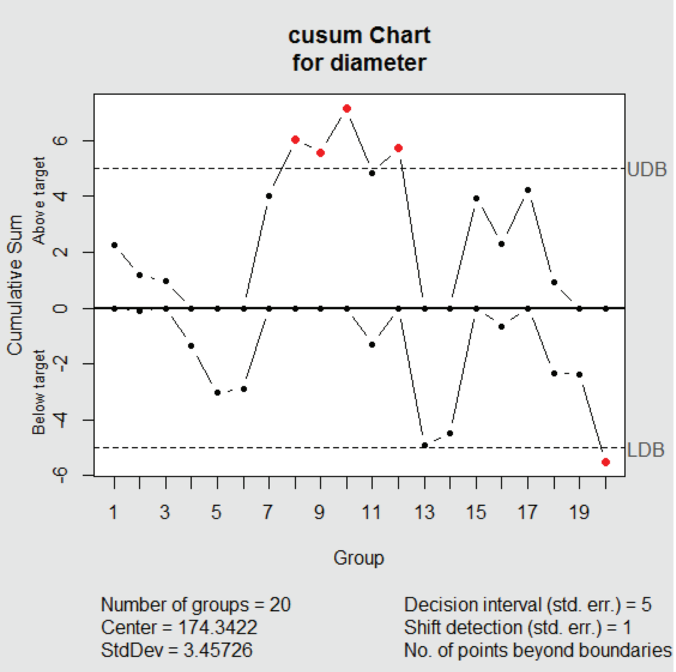


We notice that, due to the unequal sample sizes, the UCL and LCL are not constant anymore. This chart shows that the process is in control.

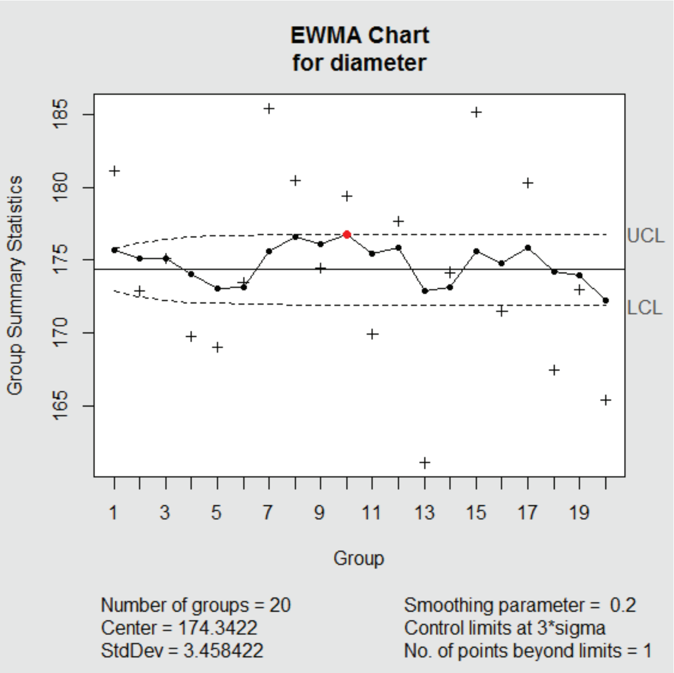
13.5 CUSUM and EWMA Charts

1. We load the data by commands *SSD=read.table("SqcSemicondDiam.txt", header=T); attach(SSD); diameter=qcc.groups(x,lot).*

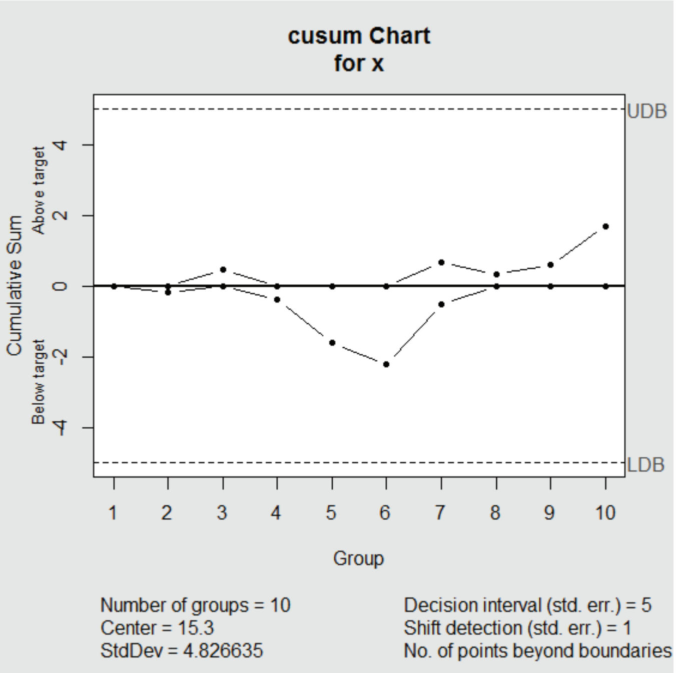
   (a) We use command *require(qcc); cusum(diameter, std.dev="UWAVE-SD")* to construct a CUSUM chart for the semiconductor wafer diameter data, which is shown below. The chart shows that the process is out of control since there are five points outside the control limits.

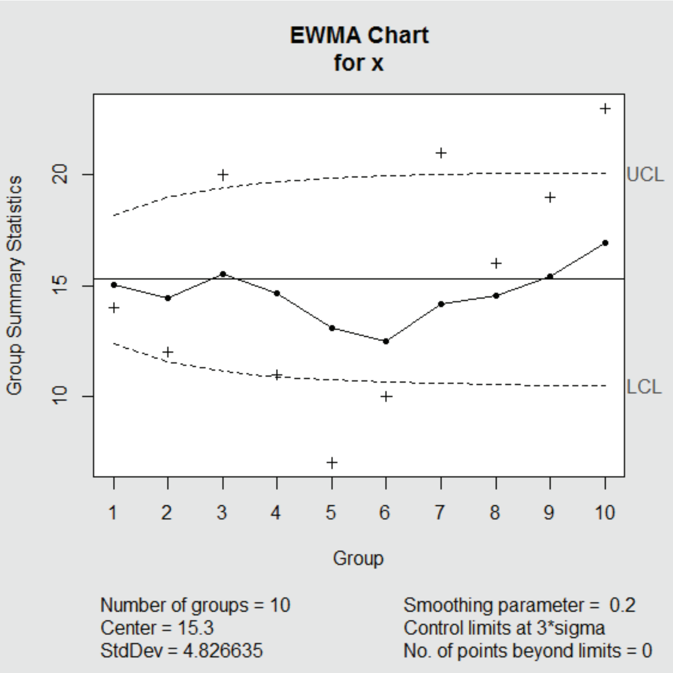


   (b) We use command *ewma(diameter)* to construct a EWMA chart for the semiconductor wafer diameter data, which is shown below. The chart shows that the process is out of control since there is one point outside the control limits.



2. We load the data by commands *SRP=read.table("SqcRedoxPotent.txt", header=T); attach(SRP); require(qcc)*. To construct the CUSUM and EWMA charts for the chlorine data, we use *cusum(x)* and *ewma(x)*, respectively, and the charts are given below.





The charts suggest the process mean is in control.