Chapter 1

Odd solutions

Solution for 1:

(a) 131.041

```
> round((7 - 8) + 5^3 - 5/6 + sqrt(62), 3)

[1] 131.041
```

(b) -18.987

```
> round(log(3) - sqrt(2) * sin(pi) - exp(3), 3)

[1] -18.987
```

(c) 94.551

```
> round(2 * (5 + 3) - sqrt(6) + 9^2, 3)

[1] 94.551
```

(d) 2.22

```
> round(log(5) - exp(2) + 2^3, 3)

[1] 2.22
```

(e) 13.911

```
> round(9/2 * 4 - sqrt(10) + log(6) - exp(1), 3)

[1] 13.911
```

Solution for 3:

```
> Treatment <- rep(c("Treatment One", "Treatment Two", "Treatment Three"),
+     c(20, 18, 22))
> xtabs(~Treatment)

Treatment
  Treatment One Treatment Three   Treatment Two
             20              22              18
```

Solution for 5:

(a) The valued stored in **z** is 78125.

```
> x <- 5
> y <- 7
> z <- x^y
> z

[1] 78125
```

(b)

```
> u <- c(1, 2, 5, 4)
> v <- c(2, 2, 1, 1)
```

(c)

```
> which(u == 5)

[1] 3
```

(d)

```
> which(v >= 2)

[1] 1 2
```

(e) Multiplication of vectors with R is element by element.

```
> uv <- u * v
> uv

[1] 2 4 5 4
```

(f) The values in the shorter vector are recycled until the two vectors are the same size. In this case, `u*c(u, v)` is the same as `c(u, u)*c(u, v)`.

```
> u * (c(u, v))

[1]  1  4 25 16  2  4  5  4

> c(u, u) * c(u, v)

[1]  1  4 25 16  2  4  5  4
```

(g)

```
> G <- 1:10
> G[1:3]

[1] 1 2 3
```

(h)

```
> J <- seq(from = 1, to = 30, by = 2)
> J[c(1, 3, 8)]

[1]  1  5 15
```

(i)

```
> q <- c(3, 0, 1, 6)
> r <- c(1, 0, 2, 4)
> q %*% r

     [,1]
[1,]   29
```

(j)

```
> X <- rbind(u, v)
> X

  [,1] [,2] [,3] [,4]
u    1    2    5    4
v    2    2    1    1
```

(k)

```
> Y <- cbind(u, v)
> Y

     u v
[1,] 1 2
[2,] 2 2
[3,] 5 1
[4,] 4 1
```

(l)

```
> W <- X %*% Y
> W

   u  v
u 46 15
v 15 10
```

(m)

```
> solve(W)

            u           v
u  0.04255319 -0.06382979
v -0.06382979  0.19574468

> t(solve(W))

            u           v
u  0.04255319 -0.06382979
v -0.06382979  0.19574468
```

Solution for 7:

(a)

```
> community <- c("Galicia", "Asturias", "Cantabria", "Pais Vasco",
+     "Navarra", "La Rioja", "Aragon", "Cataluna", "Islas Baleares",
+     "Castilla y Leon", "Madrid", "Castilla-La Mancha", "C. Valenciana",
+     "Region de Murcia", "Extremadura", "Andalucia", "Islas Canarias")
> wheat.surface <- c(18817, 65, 440, 25143, 66326, 34214,
+     311479, 74206, 7203, 619858, 13118, 263424, 6111, 9500,
+     143250, 558292, 100)
> wheat.spain <- data.frame(community, wheat.surface)
> rm(community, wheat.surface)
> head(wheat.spain)

   community wheat.surface
1    Galicia         18817
2   Asturias            65
3  Cantabria           440
4 Pais Vasco         25143
5    Navarra         66326
6   La Rioja         34214
```

(b) The maximum for `wheat.surface` is 619858 ha, the minimum for `wheat.surface` is 65 ha, and the range for `wheat.surface` is 619793 ha.

```
> max(wheat.spain$wheat.surface)

[1] 619858

> min(wheat.spain$wheat.surface)

[1] 65

> diff(range(wheat.spain$wheat.surface))

[1] 619793
```

(c)

```
> wheat.spain[wheat.spain$wheat.surface == max(wheat.spain$wheat.surface),]

         community wheat.surface
10 Castilla y Leon        619858
```

(d)

```
> IO <- wheat.spain[order(wheat.spain$wheat.surface, decreasing = FALSE), ]
> head(IO)

          community wheat.surface
2          Asturias            65
17   Islas Canarias           100
3         Cantabria           440
13    C. Valenciana          6111
9    Islas Baleares          7203
14 Region de Murcia          9500
```

(e)

```
> DO <- wheat.spain[order(wheat.spain$wheat.surface, decreasing = TRUE), ]
> head(DO)

            community wheat.surface
10    Castilla y Leon        619858
16          Andalucia        558292
7              Aragon        311479
12 Castilla-La Mancha        263424
15        Extremadura        143250
8            Cataluna         74206
```

(f)

```
> wheat.c <- wheat.spain[wheat.spain$community != "Asturias", ]
> head(wheat.c)

   community wheat.surface
1    Galicia         18817
3  Cantabria           440
4 Pais Vasco         25143
5    Navarra         66326
6   La Rioja         34214
7     Aragon        311479
```

(g)

```
> RM <- wheat.spain[wheat.spain$community == "Asturias", ]
> wheat.c <- rbind(wheat.c, RM)
> wheat.c

            community wheat.surface
1            Galicia         18817
3          Cantabria           440
4          Pais Vasco         25143
5             Navarra         66326
6            La Rioja         34214
7              Aragon        311479
8            Cataluna         74206
9      Islas Baleares          7203
10    Castilla y Leon        619858
11             Madrid         13118
12 Castilla-La Mancha        263424
13     C. Valenciana          6111
14   Region de Murcia          9500
15        Extremadura        143250
16          Andalucia        558292
17     Islas Canarias           100
2             Asturias            65
```

(h)

```
> wheat.c <- within(data = wheat.c, {
+   acre <- wheat.surface/0.40468564224
+ })
> head(wheat.c)

   community wheat.surface       acre
1    Galicia         18817  46497.820
3  Cantabria           440   1087.264
4 Pais Vasco         25143  62129.706
5    Navarra         66326 163895.115
6   La Rioja         34214  84544.635
7     Aragon        311479 769681.371
```

(i) The total harvested surface in Spain in 2004 is 2151546 hectares and 5316585.9507 acres.

```
> sum(wheat.c$wheat.surface)

[1] 2151546

> sum(wheat.c$acre)

[1] 5316586
```

(j)

```
> nc <- wheat.c[, -1]
> row.names(nc) <- wheat.c[, 1]
> wheat.c <- nc
> head(wheat.c)

           wheat.surface       acre
Galicia            18817  46497.820
Cantabria            440   1087.264
Pais Vasco         25143  62129.706
Navarra            66326 163895.115
La Rioja           34214  84544.635
Aragon            311479 769681.371
```

(k) 29.4118% of the autonomous communities have a harvested wheat surface greater than the mean wheat surface area.

```
> PA <- mean(wheat.c$wheat.surface > mean(wheat.c$wheat.surface))*100
> PA

[1] 29.41176
```

(l)

```
> AO <- wheat.c[order(row.names(wheat.c)), ]
> head(AO)

               wheat.surface         acre
Andalucia             558292 1379569.5763
```

```
Aragon                 311479  769681.3711
Asturias                   65     160.6185
C. Valenciana            6111   15100.6099
Cantabria                 440    1087.2637
Castilla y Leon        619858 1531702.4755
```

(m) The total harvested area is 36537 acres or 90284.8932 hectares.

```
> lessthan40k <- wheat.c[wheat.c$acre < 40000, ]
> lessthan40k

                wheat.surface       acre
Cantabria                 440  1087.2637
Islas Baleares           7203 17799.0006
Madrid                  13118 32415.2839
C. Valenciana            6111 15100.6099
Region de Murcia         9500 23475.0112
Islas Canarias            100   247.1054
Asturias                   65   160.6185

> apply(lessthan40k, 2, sum)

wheat.surface          acre
     36537.00      90284.89
```

(n)

```
> lt40 <- apply(lessthan40k, 2, sum)
> gt40 <- wheat.c[wheat.c$acre >= 40000, ]
> wheat.sum <- rbind(gt40, lt40)
> row.names(wheat.sum)[11] <- c("less than 40,000")
> wheat.sum

                  wheat.surface       acre
Galicia                   18817   46497.82
Pais Vasco                25143   62129.71
Navarra                   66326  163895.12
La Rioja                  34214   84544.64
Aragon                   311479  769681.37
Cataluna                  74206  183367.02
Castilla y Leon          619858 1531702.48
Castilla-La Mancha       263424  650934.88
Extremadura              143250  353978.46
Andalucia                558292 1379569.58
less than 40,000          36537   90284.89
```

(o)

```
> dump("wheat.c", file = "wheat.txt")
> rm("wheat.c")
> wheat.c  # no longer available

Error in eval(expr, envir, enclos):  object 'wheat.c' not found
```

```
> source("wheat.txt")
> head(wheat.c)

           wheat.surface       acre
Galicia            18817  46497.820
Cantabria            440   1087.264
Pais Vasco         25143  62129.706
Navarra            66326 163895.115
La Rioja           34214  84544.635
Aragon            311479 769681.371
```

(p) There are different values stored in each of `wheat.txt` and `wheat.dat`. Specifically, the values from part (m) are collapsed into one category `"less than 40,000"` in `wheat.dat`, whereas `wheat.txt` has all of the values.

```
> write.table(x = wheat.sum, file = "wheat.dat")
```

(q)

```
> tail(read.table(file = "wheat.dat"))

                   wheat.surface       acre
Cataluna                   74206  183367.02
Castilla y Leon           619858 1531702.48
Castilla-La Mancha        263424  650934.88
Extremadura               143250  353978.46
Andalucia                 558292 1379569.58
less than 40,000           36537   90284.89
```

Solution for 9:

(a)

```
> STATES <- WHEATUSA2004$states
> row.names(WHEATUSA2004) <- STATES
> head(WHEATUSA2004)

   states acres
AR     AR   620
CA     CA   320
CO     CO  1700
DE     DE    47
GA     GA   190
ID     ID   700
```

(b)

```
> WHEATUSA2004$ha <- WHEATUSA2004$acres * 0.40468564224
> head(WHEATUSA2004)

   states acres        ha
AR     AR   620 250.90510
CA     CA   320 129.49941
CO     CO  1700 687.96559
```

```
DE     DE    47  19.02023
GA     GA   190  76.89027
ID     ID   700 283.27995
```

(c)

```
> io <- WHEATUSA2004[order(WHEATUSA2004$acres), ]
> head(io)

   states acres       ha
DE     DE    47 19.02023
NY     NY   100 40.46856
MS     MS   135 54.63256
PA     PA   135 54.63256
MD     MD   145 58.67942
SC     SC   180 72.84342
```

(d) Kansas, Oklahoma, and Texas are in the top 10% of states for harvested surface area.

```
> top10 <- quantile(WHEATUSA2004$acres, prob = 0.9)
> ans <- WHEATUSA2004[WHEATUSA2004$acres > top10, ]
> row.names(ans)

[1] "KS" "OK" "TX"
```

(e)

```
> dump("WHEATUSA2004", "WHEATUSA.txt")
> rm(WHEATUSA2004)
> source("WHEATUSA.txt")
> head(WHEATUSA2004)

   states acres        ha
AR     AR   620 250.90510
CA     CA   320 129.49941
CO     CO  1700 687.96559
DE     DE    47  19.02023
GA     GA   190  76.89027
ID     ID   700 283.27995
```

(f) This question needs an answer!

```
> write.table(WHEATUSA2004, "WHEATUSA.dat")
```

(g) The total harvested area for the bottom 10% of states is 147 acres.

```
> bottom10 <- quantile(WHEATUSA2004$acres, prob = 0.1)
> ans <- WHEATUSA2004[WHEATUSA2004$acres < bottom10, ]
> ans

   states acres       ha
DE     DE    47 19.02023
NY     NY   100 40.46856
```

```
> THA <- sum(ans[, "acres"])
> THA  # Total Harvested Acres

[1] 147
```

Solution for 11:

(a) A total of 171 patients have been treated with hamstring stretch position.

```
> xtabs(~treatment, data = EPIDURALF)

treatment
  Hamstring Stretch Traditional Sitting
                171                 171

> xtabs(~treatment, data = EPIDURALF)[1]

Hamstring Stretch
              171
```

(b) The percent of patients treated with hamstring stretch that were classified as `Easy`, `Difficult`, and `Impossible` was 58.4795%, 36.8421%, and 4.6784%, respectively.

```
> T1 <- xtabs(~treatment + ease, data = EPIDURALF)
> T1

                     ease
treatment             Difficult Easy Impossible
  Hamstring Stretch          63  100          8
  Traditional Sitting        51  107         13

> prop.table(T1[1, ]) * 100

 Difficult       Easy Impossible
 36.842105  58.479532   4.678363
```

(c) 51.6908% of the patients classified as easy to palpate were assigned to the traditional sitting position.

```
> T1

                     ease
treatment             Difficult Easy Impossible
  Hamstring Stretch          63  100          8
  Traditional Sitting        51  107         13

> prop.table(T1[, "Easy"])[2] * 100

Traditional Sitting
           51.69082
```

(d)

```
> tapply(EPIDURALF$kg, list(EPIDURALF$ease, EPIDURALF$treatment),
+     mean)

           Hamstring Stretch Traditional Sitting
Difficult           92.66667            94.27451
Easy                78.67000            79.40187
Impossible         127.87500           113.61538
```

(e) 9.0643% of patients have a body mass index less than 25 and are classified as easy to palpate.

```
> EPIDURALF$BMI <- EPIDURALF$kg/(EPIDURALF$cm/100)^2
> EPIDURALF[1:5, 3:8]

   cm       ease           treatment oc complications      BMI
1 172  Difficult Traditional Sitting  0          None 39.21038
2 176       Easy   Hamstring Stretch  0          None 27.76343
3 157  Difficult Traditional Sitting  0          None 29.21011
4 169       Easy   Hamstring Stretch  2          None 22.05805
5 163 Impossible Traditional Sitting  0          None 42.90715

> mean(EPIDURALF$ease =="Easy" & EPIDURALF$BMI <25)*100

[1] 9.064327
```

Solution for 13:

```
> for (celsius in seq(from = 18, to = 28, by = 2)) {
+     print(c(celsius, 9/5 * celsius + 32))
+ }

[1] 18.0 64.4
[1] 20 68
[1] 22.0 71.6
[1] 24.0 75.2
[1] 26.0 78.8
[1] 28.0 82.4
```

Solution for 15:

```
> picker <- function(m = 4, n = 40, ...) {
+     presenters <- sample(n, m, replace = FALSE)
+     presenters
+ }
> set.seed(1)
> picker()

[1] 11 15 22 34

> ###
> pickerM <- function(m = 4, ...) {
+     if (m > 6)
```

```
+         stop("m must be less than 6")
+     n <- c("Joe", "Bill", "Mark", "Karen", "Anne", "Mary")
+     presenters <- sample(n, m, replace = FALSE)
+     presenters
+ }
> set.seed(1)
> pickerM()

[1] "Bill"  "Mary"  "Mark"  "Karen"
```

Solution for 17:

(a)

```
> ansA <- IRF(A = 3000, t = 3, i = 0.04, n = 365)
> ansA

[1] 2660.779
```

John needs to invest $2660.78 today to have $3000 to pay for his trip in 3 years.

(b)

```
> ansB <- IRF(P = 9000, t = 15, i = 0.10, n = 2)
> ansB

[1] 38897.48
```

Fred will have have $38897.48 in 15 years.

Solution for 19:

(a)

```
> ansA <- ARF(R = 200, i = 0.03, n = 12, t = 30)
> ansA

[1] 116547.4
```

Mary will have $116547.3769 at the end of 30 years.

(b)

```
> ansB <- ARF(A = 200000, i = 0.03, n = 12, t = 30)
> ansB

[1] 343.2081
```

Mary needs to increase her monthly savings by $143.21 over her current $200 monthly contributions to have $200,000 in 30 years.

Chapter 2

Odd solutions

Solution for 1:

(a)

```
> library(MASS)
> help(package = "MASS")
```

(b) The description file says `lqs()` fits a regression to the points in the data set, thereby achieving a regression estimator with a high breakdown point.

(c) The function `search()` provides a list of attached packages and the function `library()` shows all installed packages.

Solution for 3:

(a) The distribution of harvested wheat is unimodal and skewed to the right. This skew is seen in how much larger the mean (126561.5294) is versus the median (25143). The difference between Q_1 and Q_2 is also much smaller than the difference between Q_3 and Q_2. The total harvested area is 2151546 hectares.

```
> quantile(WHEATSPAIN$hectares)

    0%    25%    50%    75%   100%
    65   7203  25143 143250 619858

> quantile(WHEATSPAIN$hectares, probs = seq(from = 0.1, to = 1.0, by = 0.1))

     10%      20%      30%      40%      50%      60%      70%      80%
   304.0   6329.4   9040.6  15397.6  25143.0  53481.2  88014.8 239389.2
     90%     100%
410204.2 619858.0

> mean(WHEATSPAIN$hectares)

[1] 126561.5

> IQR(WHEATSPAIN$hectares)

[1] 136047

> var(WHEATSPAIN$hectares)

[1] 38934822657

> sd(WHEATSPAIN$hectares)

[1] 197319.1

> sum(WHEATSPAIN$hectares)

[1] 2151546
```

(b)

```
> describe <- function(x, ...){
+   Quantiles <- quantile(x)
+   Mean <- mean(x)
+   Var <- var(x)
+   SD <- sd(x)
+   Total <- sum(x)
+   Range <- diff(range(x))
+   print(c(Quantiles = Quantiles, Mean = Mean, Var = Var,
+           SD = SD, Total = Total, Range = Range))
+ }
> describe(WHEATSPAIN$hectares)

  Quantiles.0%  Quantiles.25%  Quantiles.50%  Quantiles.75% Quantiles.100%
  6.500000e+01   7.203000e+03   2.514300e+04   1.432500e+05   6.198580e+05
          Mean            Var             SD          Total          Range
  1.265615e+05   3.893482e+10   1.973191e+05   2.151546e+06   6.197930e+05
```

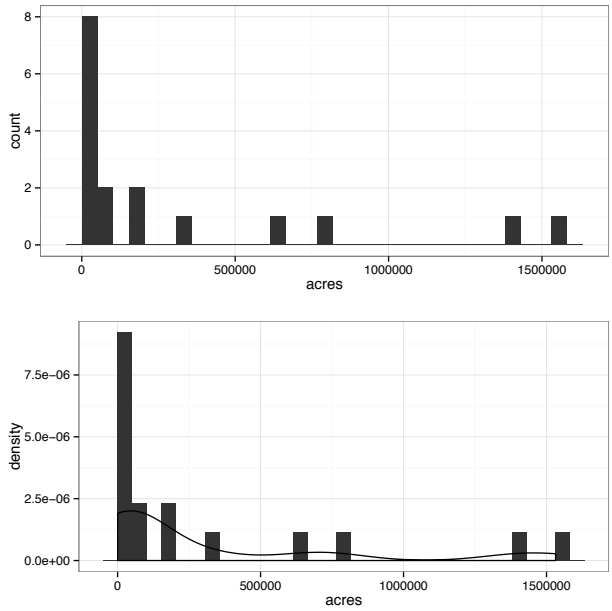
(c) Asturias and Canarias are below 304 hectares. Castilla-Leon and Andalucia are above 410204.2 hectares. Since Navarra is the eleventh out of seventeen areas, it corresponds to the $10/16 \times 100 = 62.5^{\text{th}}$ percentile.

```
> bottom10 <- quantile(WHEATSPAIN$hectares, probs = 0.10)
> bottom10

10%
304

> WHEATSPAIN[WHEATSPAIN$hectares < bottom10, ] # bottom communities

   community hectares acres
2   Asturias       65 160.6
17  Canarias      100 247.1

> top10 <- quantile(WHEATSPAIN$hectares, probs = 0.90)
> top10

     90%
410204.2

> WHEATSPAIN[WHEATSPAIN$hectares > top10, ]    # top communities

       community hectares   acres
10 Castilla-Leon   619858 1531702
16     Andalucia   558292 1379570

> WHEATSPAIN[order(WHEATSPAIN$hectares), ]

            community hectares     acres
2            Asturias       65     160.6
17           Canarias      100     247.1
3           Cantabria      440    1087.3
13       C.Valenciana     6111   15100.6
```

```
9            Baleares     7203   17799.0
14             Murcia     9500   23475.0
11             Madrid    13118   32415.3
1             Galicia    18817   46497.8
4             P.Vasco    25143   62129.7
6            La Rioja    34214   84544.6
5             Navarra    66326  163895.1
8            Cataluna    74206  183367.0
15       Extremadura   143250  353978.5
12 Castilla-La Mancha  263424  650934.9
7              Aragon  311479  769681.4
16          Andalucia  558292 1379569.6
10      Castilla-Leon  619858 1531702.5

> which(WHEATSPAIN[order(WHEATSPAIN$hectares), ]$community=="Navarra")

[1] 11

> pk <- (11 - 1)/(17 - 1)
> pk

[1] 0.625

> quantile(WHEATSPAIN$hectares, probs = pk)

62.5%
66326
```

(d)

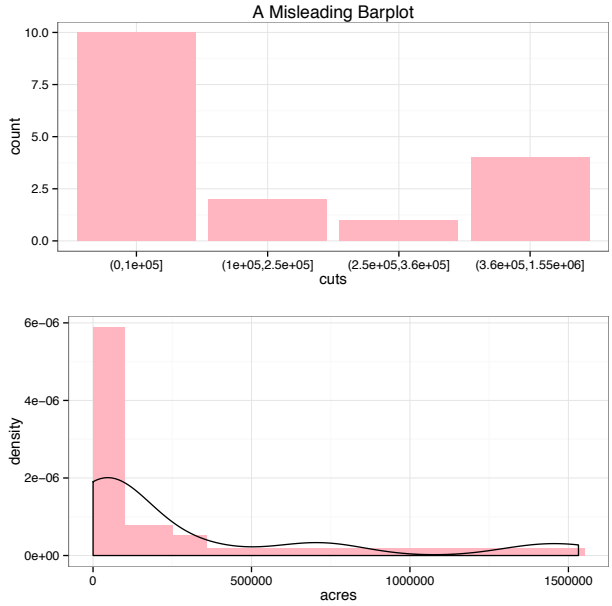
```
> p1 <- ggplot(data = WHEATSPAIN, aes(x = acres)) +
+   geom_histogram() + theme_bw()
> p2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) +
+   geom_histogram() + theme_bw() + geom_density()
> multiplot(p1, p2)
```



(e) If the breaks used in `hist()` are not equidistant, the default is to produce a density histogram.

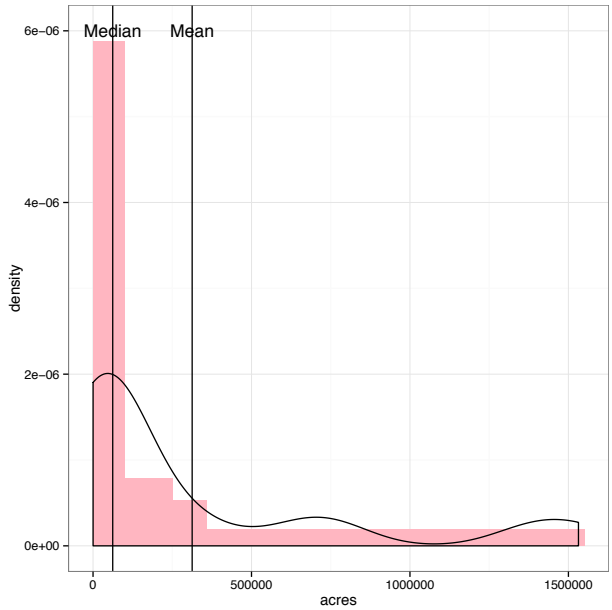
(f)

```
> bins <- c(0, 100000, 250000, 360000, 1550000)
> WHEATSPAIN$cuts <- cut(WHEATSPAIN$acres,
+                        breaks = bins)
> p1 <- ggplot(data = WHEATSPAIN, aes(x = cuts)) +
+   geom_bar(fill = "lightpink") +
+   theme_bw() +
+   labs(title = "A Misleading Barplot")
> p2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) +
+   geom_histogram(breaks = bins, fill = "lightpink") +
+   theme_bw() +
+   geom_density()
> multiplot(p1, p2, layout = matrix(c(1, 2), byrow = TRUE, ncol = 1))
```



(g)

```
> p2 <- ggplot(data = WHEATSPAIN, aes(x = acres, y = ..density..)) +
+   geom_histogram(breaks = bins, fill = "lightpink") +
+   theme_bw() +
+   geom_density()
> p2 + geom_vline(xintercept = c(median(WHEATSPAIN$acres),
+                                mean(WHEATSPAIN$acres))) +
+   annotate("text", label = "Median", x = median(WHEATSPAIN$acres),
+            y = 6e-06) +
+   annotate("text", label = "Mean", x = mean(WHEATSPAIN$acres), y = 6e-06)
```



(h) Note: click just to the right of the outliers using the following code.

```
> with(data = WHEATSPAIN, boxplot(hectares))
> with(data = WHEATSPAIN, identify(rep(1, length(hectares)),
+     hectares, labels = community))
```

(i) Based on the output from part (c), Castilla-Leon the tenth indexed position in the `WHEATSPAIN` data frame has the largest harvested wheat surface. The mean, median, and standard deviation are smaller than those computed in part (a) once Castilla-Leon is removed.

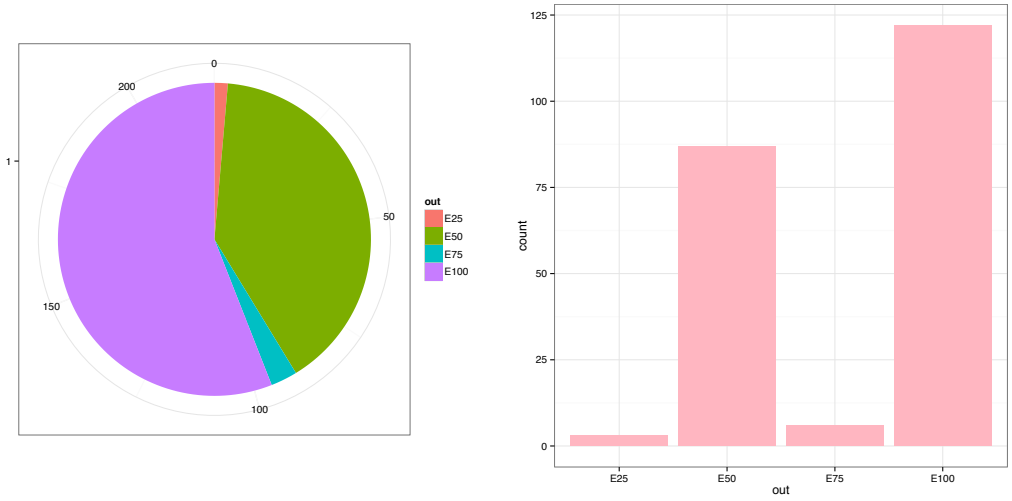
```
> noCL <- WHEATSPAIN[-10, ]
> mean(WHEATSPAIN$hectares)

[1] 126561.5

> mean(noCL$hectares)

[1] 95730.5

> median(WHEATSPAIN$hectares)

[1] 25143

> median(noCL$hectares)

[1] 21980

> sd(WHEATSPAIN$hectares)

[1] 197319.1

> sd(noCL$hectares)

[1] 155864.7
```

Solution for 5:

(a) The barplot is the easiest to read.

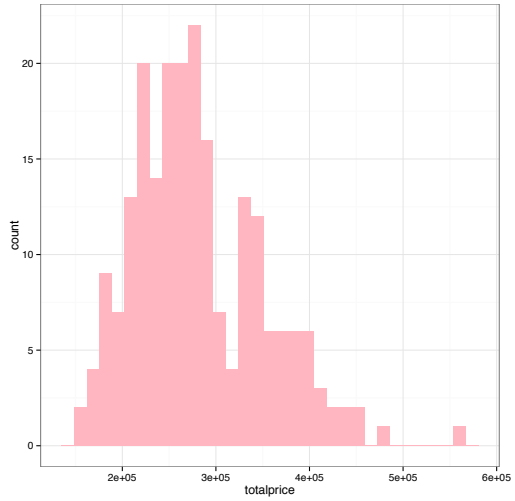
```
> VIT2005$out <- factor(VIT2005$out,
+                       levels = c("E25", "E50", "E75", "E100"))
> levels(VIT2005$out)

[1] "E25"  "E50"  "E75"  "E100"

> xtabs(~out, data = VIT2005)

out
 E25  E50  E75 E100
   3   87    6  122

> p1 <- ggplot(data = VIT2005, aes(x = factor(1), fill = out)) +
+   geom_bar(width = 1)
> p1 + coord_polar(theta = "y") +
```

```
+   theme_bw() +
+   labs(x = "", y = "")
> p2 <- ggplot(data = VIT2005, aes(x = out)) +
+   geom_bar(fill = "lightpink") +
+   theme_bw()
> p2
```



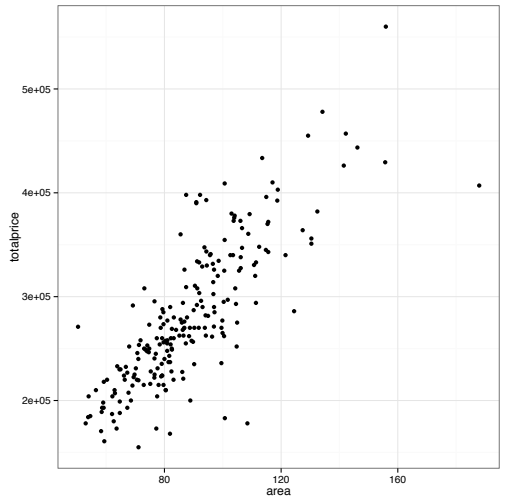
(b) The distribution of `totalprice` is skewed to the right with one outlier (€560000). The median `totalprice` is €269750 and the IQR for `totalprice` is €100125.

```
> ggplot(data = VIT2005, aes(x = totalprice)) +
+   geom_histogram(fill = "lightpink") +
+   theme_bw()
> max(VIT2005$totalprice)

[1] 560000

> median(VIT2005$totalprice)

[1] 269750

> IQR(VIT2005$totalprice)

[1] 100125
```



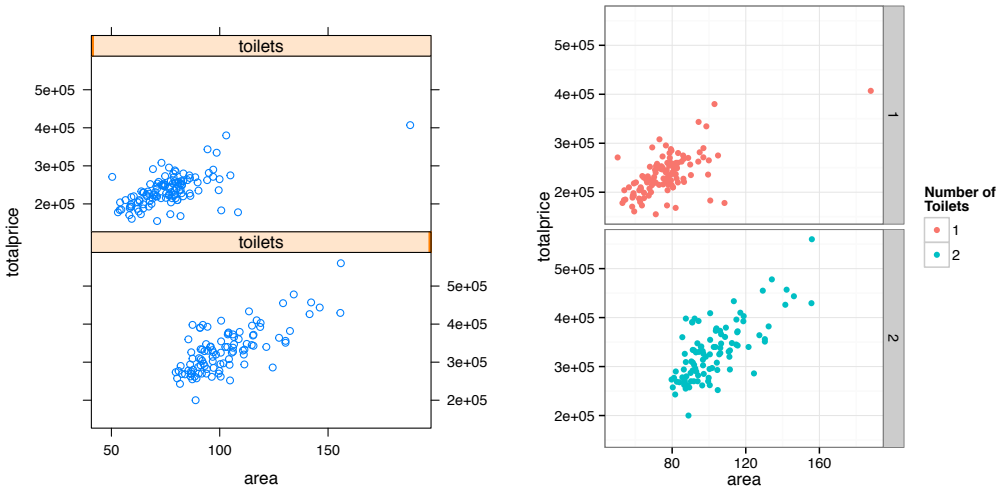
(c) There is a positive linear relationship between `totalprice` and `area`.

```
> ggplot(data = VIT2005, aes(x = area, y = totalprice)) +
+   geom_point() +
+   theme_bw()
```



(d) Apartments with one bathroom are generally between 50 and 100 m^2, while apartments with two bathrooms are generally between 80 and 160 m^2. The intersection of apartments with one and two bathrooms is roughly 80 to 100 m^2.

```
> xyplot(totalprice ~ area | toilets, data = VIT2005, layout = c(1, 2),
+        as.table = TRUE)
> TEXT <- "Number of\nToilets"
> ggplot(data = VIT2005, aes(x = area, y = totalprice,
+                            color = as.factor(toilets))) +
+   geom_point() +
+   facet_grid(toilets ~ .) +
+   theme_bw() +
+   guides(color = guide_legend(TEXT))
```



(e) The median increase in `totalprice` for a second bathroom for apartments between 80 and 100 m^2 is €36000. Answers will vary for answering whether readers would be willing to spend €36000 for an additional bathroom.

```
> bothbaths <- subset(VIT2005, subset = area >= 80 & area <= 100)
> ANS <- tapply(bothbaths$totalprice, bothbaths$toilets, median)
> ANS

     1      2
255000 291000

> diff(ANS)

    2
36000
```

Solution for 7:

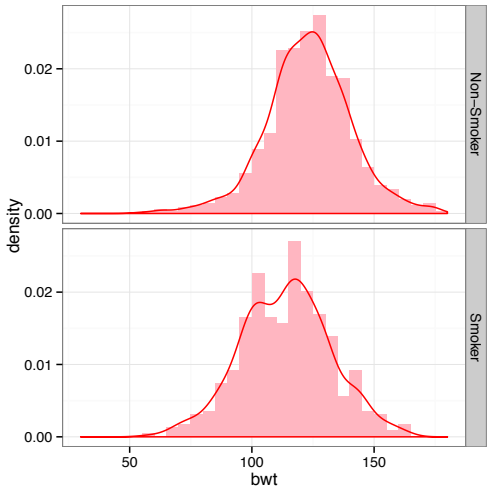
```
> site <- "http://www.stat.berkeley.edu/users/statlabs/data/babies.data"
> BABIES <- read.table(file = url(site), header = TRUE)
> head(BABIES)

  bwt gestation parity age height weight smoke
1 120       284      0  27     62    100     0
2 113       282      0  33     64    135     0
3 128       279      0  28     64    115     1
4 123       999      0  36     69    190     0
5 108       282      0  23     67    125     1
6 136       286      0  25     62     93     0
```

(a)

```
> CLEAN <- with(data = BABIES, BABIES[bwt != 999 & gestation !=
+     999 & parity != 9 & age != 99 & height != 99 & weight !=
+     999 & smoke != 9, ])
```

(b)

```
> CLEAN$smoke <- factor(CLEAN$smoke, levels = 0:1,
+                       labels =c("Non-Smoker", "Smoker"))
> ggplot(data = CLEAN, aes(x = bwt, y = ..density..)) +
+   geom_histogram(fill = "lightpink") +
+   geom_density(color = "red") +
+   facet_grid(smoke ~.) +
+   xlim(30, 180) +
+   theme_bw()
```



(c) Based on the density histograms in part (b), the distributions of birth weights for both smoking and non-smoking mothers are unimodal and symmetric. The mean and standard deviation for birth weights of non-smoking mothers are 123.0853 and 17.4237 ounces, respectively. The mean and standard deviation for birth weights of smoking mothers are 113.8192 and 18.295 ounces, respectively.

```
> mean(CLEAN$bwt[CLEAN$smoke == "Non-Smoker"])

[1] 123.0853

> sd(CLEAN$bwt[CLEAN$smoke == "Non-Smoker"])

[1] 17.4237

> mean(CLEAN$bwt[CLEAN$smoke == "Smoker"])

[1] 113.8192

> sd(CLEAN$bwt[CLEAN$smoke == "Smoker"])

[1] 18.29501
```

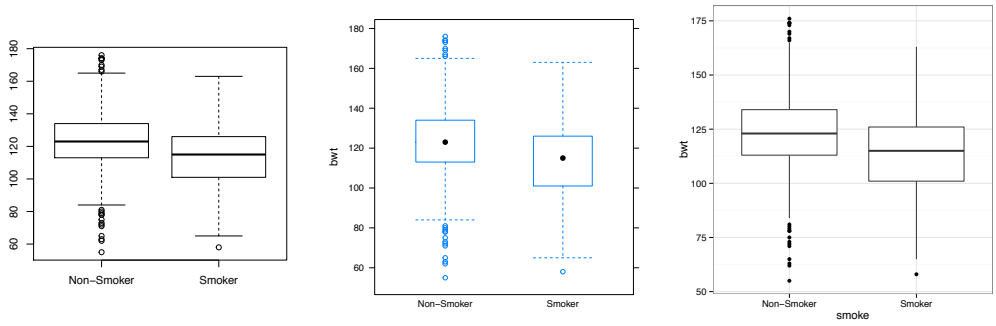
(d) The mean birth weight difference between non-smoking and smoking mother's birth weights is 9.2661 ounces.

```
> ANS <- tapply(CLEAN$bwt, CLEAN$smoke, mean)
> ANS

Non-Smoker     Smoker
```

```
  123.0853   113.8192

> DIFF <- ANS[1] - ANS[2]
> DIFF

Non-Smoker
  9.266143
```

(e)

```
> boxplot(bwt ~ smoke, data = CLEAN)
> bwplot(bwt ~ smoke, data = CLEAN)
> ggplot(data = CLEAN, aes(x = smoke, y = bwt)) +
+   geom_boxplot() +
+   theme_bw()
```

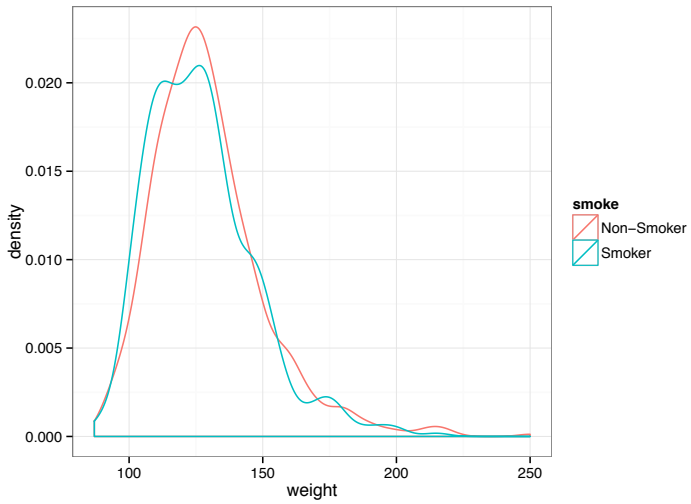


(f) The median birth weight difference between firstborn babies and those that are not firstborn is 2 ounces.

```
> ANS <- tapply(CLEAN$bwt, CLEAN$parity, median)
> ANS

  0   1
120 118

> DIFF <- ANS[1] - ANS[2]
> DIFF

0
2
```

(g)

```
> ggplot(data = CLEAN, aes(x = weight, color = smoke)) +
+   geom_density() +
+   theme_bw()
```



(h) The distribution of pre-pregnancy weight for both smokers and non-smokers is unimodal and skewed to the right. The median and IQR of pre-pregnancy weight for smokers are 125 and 24.5 pounds, respectively. The median and IQR of pre-pregnancy weight for non-smokers are 126 and 25 pounds, respectively.

```
> median(CLEAN$weight[CLEAN$smoke == "Smoker"])

[1] 125

> IQR(CLEAN$weight[CLEAN$smoke == "Smoker"])

[1] 24.5

> median(CLEAN$weight[CLEAN$smoke == "Non-Smoker"])

[1] 126

> IQR(CLEAN$weight[CLEAN$smoke == "Non-Smoker"])

[1] 25
```

(i) The mean pre-pregnancy weight difference between non-smokers and smokers is 2.5603 pounds. The mean should not be used as a measure of center in this problem since both distributions are skewed.

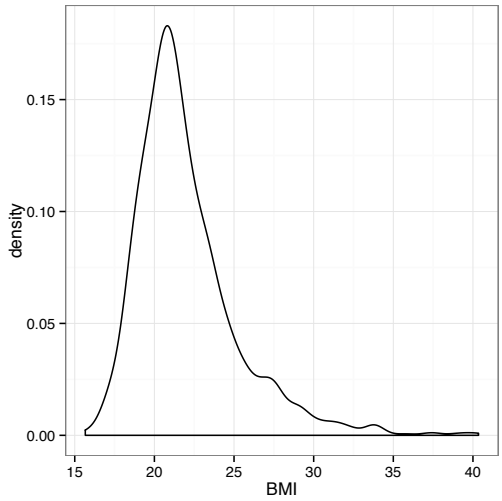
```
> ANS <- tapply(CLEAN$weight, CLEAN$smoke, mean)
> ANS

Non-Smoker     Smoker
  129.4797   126.9194

> DIFF <- ANS[1] - ANS[2]
> DIFF

Non-Smoker
   2.56033
```

(j)

```
> CLEANP <- within(data = CLEAN, expr = {
+     weightM <- 0.45359 * weight
+     heightM <- 0.0254 * height
+     BMI <- weightM/heightM^2
+ })
> head(CLEANP)

  bwt gestation parity age height weight      smoke      BMI heightM
1 120       284      0  27     62    100 Non-Smoker 18.28996  1.5748
2 113       282      0  33     64    135 Non-Smoker 23.17234  1.6256
3 128       279      0  28     64    115     Smoker 19.73940  1.6256
5 108       282      0  23     67    125     Smoker 19.57746  1.7018
6 136       286      0  25     62     93 Non-Smoker 17.00966  1.5748
7 138       244      0  33     62    178 Non-Smoker 32.55612  1.5748
   weightM
1 45.35900
2 61.23465
3 52.16285
5 56.69875
6 42.18387
7 80.73902
```

(k) The distribution of BMI is unimodal skewed to the right with a median of 21.2842 and an IQR of 3.4083 kg/m^2, respectively.

```
> median(CLEANP$BMI)

[1] 21.28422

> IQR(CLEANP$BMI)

[1] 3.408279

> ggplot(data = CLEANP, aes(x = BMI)) + geom_density() + theme_bw()
```

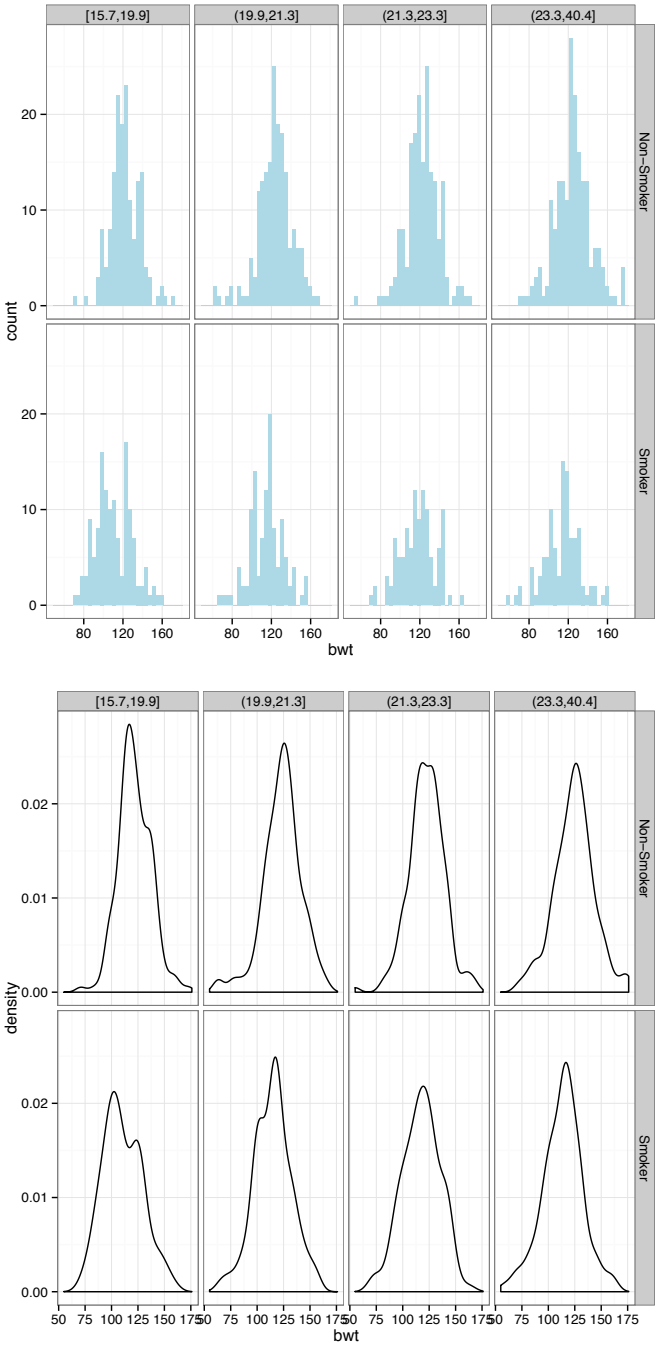


(l) The requested answers are computed in the following R Code 2.1. Based on the values, birth weight in each quartile appears to be symmetric regardless of the mother's smoking status.

R Code 2.1

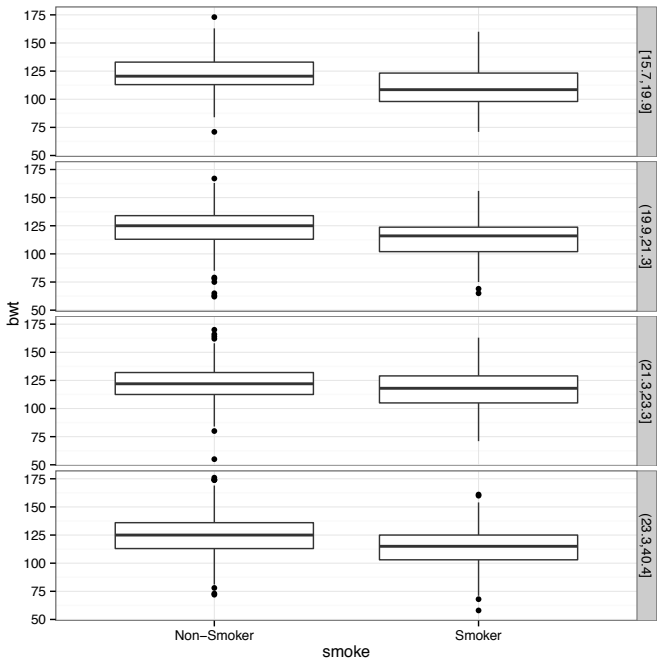
```
> values <- quantile(CLEANP$BMI)
> CLEANP <- within(data = CLEANP, expr = {
+     Quartiles <- cut(BMI, values, include.lowest = TRUE)
+ })
> tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke),
+     mean)

            Non-Smoker   Smoker
[15.7,19.9]   121.8125 110.5662
(19.9,21.3]   123.4696 114.6754
(21.3,23.3]   122.3552 117.4393
(23.3,40.4]   124.4869 113.4020

> tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke),
+     sd)

            Non-Smoker   Smoker
[15.7,19.9]   15.34434 18.44459
(19.9,21.3]   18.07778 17.66076
(21.3,23.3]   16.70788 17.74355
(23.3,40.4]   19.04755 18.82923

> tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke),
+     median)

            Non-Smoker Smoker
[15.7,19.9]      120.5  108.5
(19.9,21.3]      125.0  116.0
(21.3,23.3]      122.0  118.0
(23.3,40.4]      125.0  115.0

> tapply(CLEANP$bwt, list(CLEANP$Quartiles, CLEANP$smoke),
+     IQR)

            Non-Smoker Smoker
[15.7,19.9]       20.0  25.25
(19.9,21.3]       21.0  21.75
(21.3,23.3]       19.5  24.00
(23.3,40.4]       23.0  22.00
```

```
> ggplot(data = CLEANP, aes(x = bwt)) + geom_histogram(fill="lightblue") +
+   theme_bw() + facet_grid(smoke ~ Quartiles)
> ggplot(data = CLEANP, aes(x = bwt)) + geom_density() +
+   theme_bw() + facet_grid(smoke ~ Quartiles)
```



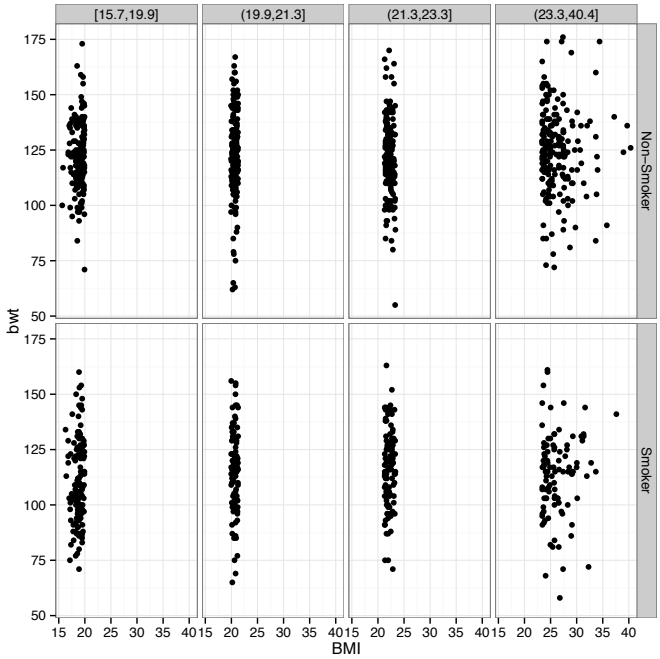
(m) The boxplots also suggest the distribution of `bwt` is symmetric for both smokers and non-smokers in each quartile.

```
> ggplot(data = CLEANP, aes(x = smoke, y = bwt)) +
+   geom_boxplot() +
+   facet_grid(Quartiles~.) +
+   theme_bw()
```



(n) There appears to be no association between birth weight and BMI.

```
> ggplot(data = CLEANP, aes(x = BMI, y = bwt)) +
+   geom_point() +
+   facet_grid(smoke ~ Quartiles) +
+   theme_bw()
```



(o) The mean, standard deviation, median, and IQR for gestation grouped according to BMI quartile and smoking status are computed in R Code 2.2 on the next page. Based on

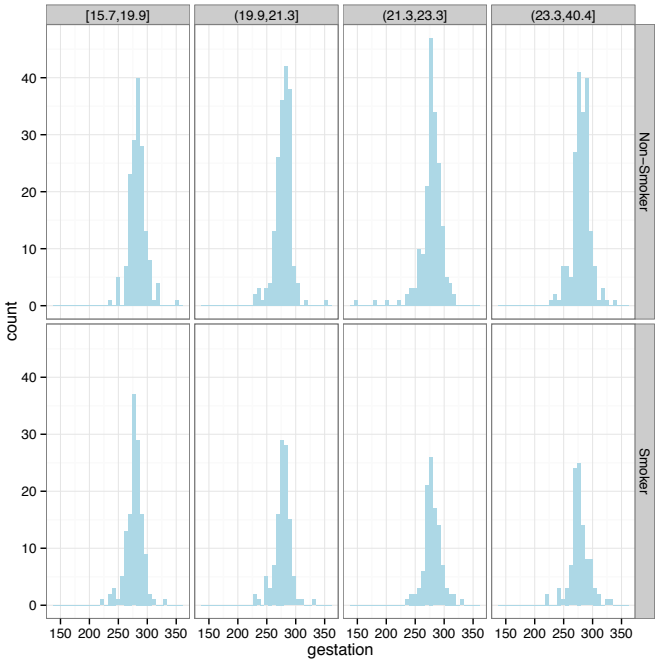
the values, gestation in each quartile appears to be symmetric regardless of the mother's smoking status.

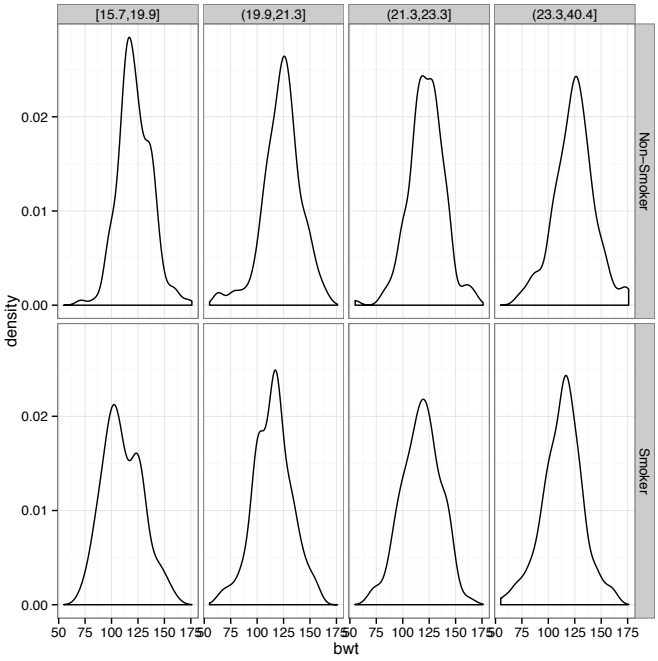
R Code 2.2

```
> tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke),
+     mean)

            Non-Smoker   Smoker
[15.7,19.9]   282.8938 277.2132
(19.9,21.3]   279.0331 277.4649
(21.3,23.3]   277.4372 279.6636
(23.3,40.4]   280.4764 277.4412

> tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke),
+     sd)

            Non-Smoker   Smoker
[15.7,19.9]   14.57214 14.55330
(19.9,21.3]   14.57810 14.40082
(21.3,23.3]   20.08376 15.33890
(23.3,40.4]   15.48780 16.77727

> tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke),
+     median)

            Non-Smoker Smoker
[15.7,19.9]        283    279
(19.9,21.3]        281    280
(21.3,23.3]        279    279
(23.3,40.4]        281    277

> tapply(CLEANP$gestation, list(CLEANP$Quartiles, CLEANP$smoke),
+     IQR)

            Non-Smoker Smoker
[15.7,19.9]      14.25   16.0
(19.9,21.3]      16.00   14.5
(21.3,23.3]      15.00   17.0
(23.3,40.4]      16.50   14.0
```

The histograms and density plots confirm a symmetric distribution of gestation regardless of BMI quartile or mother's smoking status.

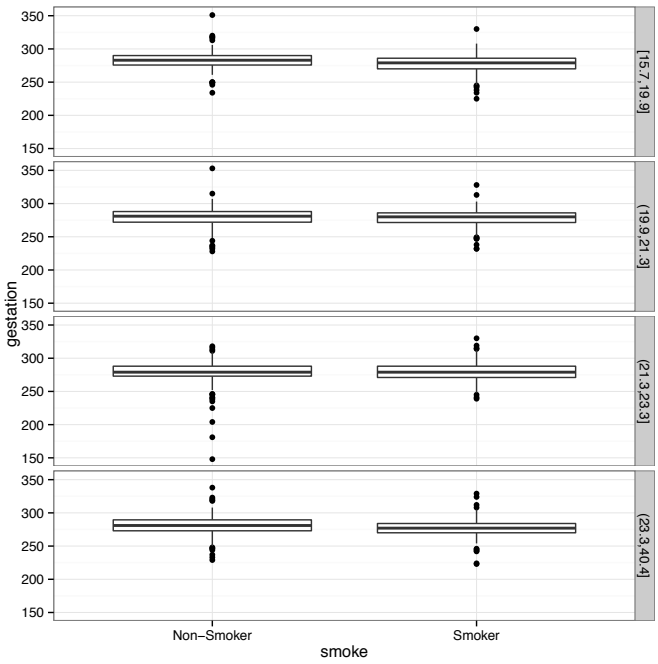
```
> ggplot(data = CLEANP, aes(x = gestation)) +
+   geom_histogram(fill = "lightblue") +
+   theme_bw() +
+   facet_grid(smoke ~ Quartiles)
> ggplot(data = CLEANP, aes(x = bwt)) +
+   geom_density() +
+   theme_bw() +
+   facet_grid(smoke ~ Quartiles)
```





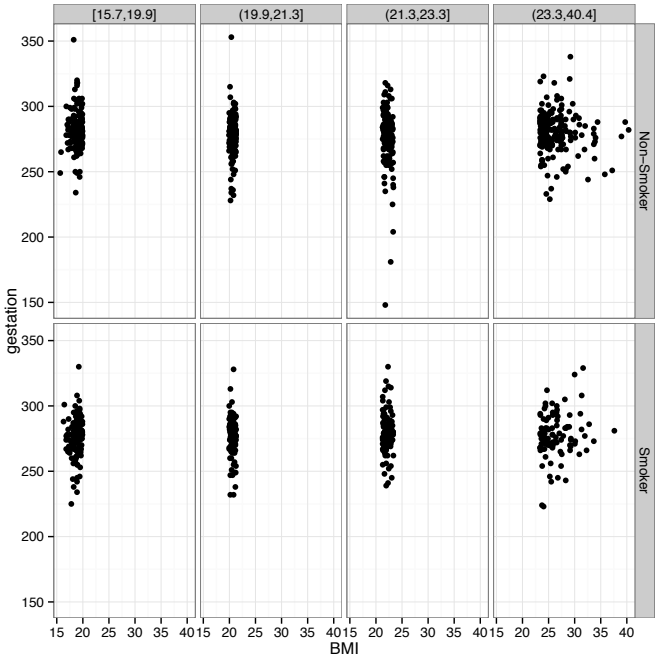
The boxplots also suggest the distribution of `gestation` is symmetric for both smokers and non-smokers in each quartile.

```
> ggplot(data = CLEANP, aes(x = smoke, y = gestation)) +
+   geom_boxplot() +
+   facet_grid(Quartiles~.) +
+   theme_bw()
```



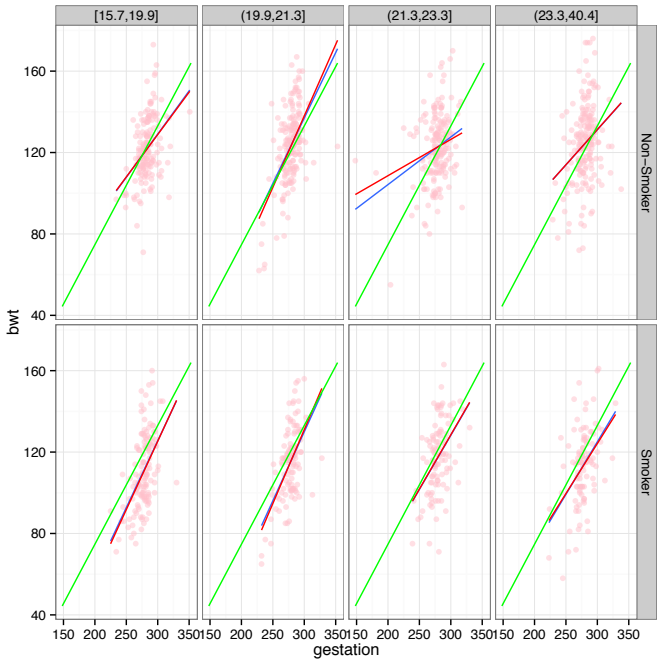
There appears to be no association between gestation and BMI.

```
> ggplot(data = CLEANP, aes(x = BMI, y = gestation)) +
+   geom_point() +
+   facet_grid(smoke ~ Quartiles) +
+   theme_bw()
```



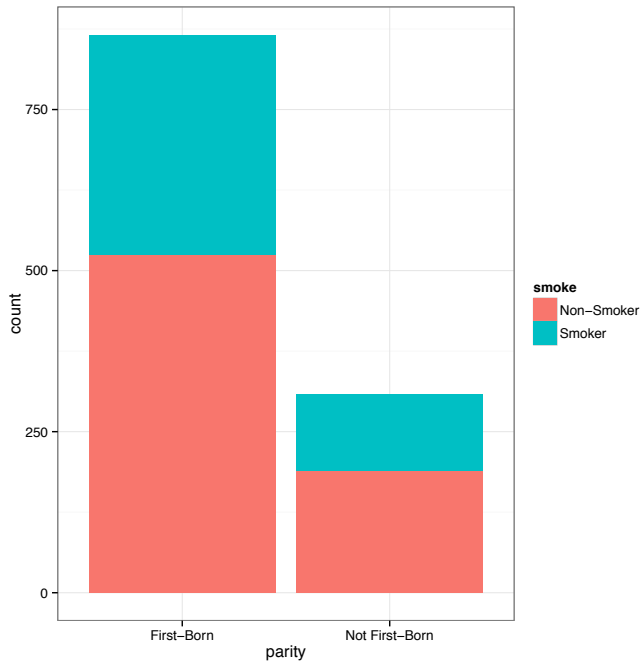
(p) There seems to be less variability among the three model fits for smoking mothers versus the non-smoking mothers.

```
> library(MASS)
> lqsmod <- lqs(bwt ~ gestation, data = CLEANP)
> xval <- seq(min(CLEANP$gestation), max(CLEANP$gestation),
+             length.out = 100)
> predata <- data.frame(gestation = xval)
> predata$bwt <- predict(lqsmod, newdata = predata)
> ggplot(data = CLEANP, aes(x = gestation, y = bwt)) +
+   geom_point(alpha = 0.5, color ="pink") +
+   facet_grid(smoke ~ Quartiles) +
+   theme_bw() +
+   stat_smooth(method = "lm", se = FALSE) +
+   stat_smooth(method = "rlm", se = FALSE, color = "red") +
+   geom_line(data = predata, color = "green")
```



(q) The percent of mothers that did not smoke during the birth of their first child is 60.6236%.

```
> CLEANP$parity <- factor(CLEAN$parity, levels = 0:1,
+                         labels =c("First-Born", "Not First-Born"))
> T1 <- xtabs(~smoke + parity, data = CLEANP)
> prop.table(T1, 2)

            parity
smoke        First-Born Not First-Born
  Non-Smoker  0.6062356      0.6168831
  Smoker      0.3937644      0.3831169

> ggplot(data = CLEANP, aes(x = parity, fill = smoke)) +
+   geom_bar() + theme_bw()
```



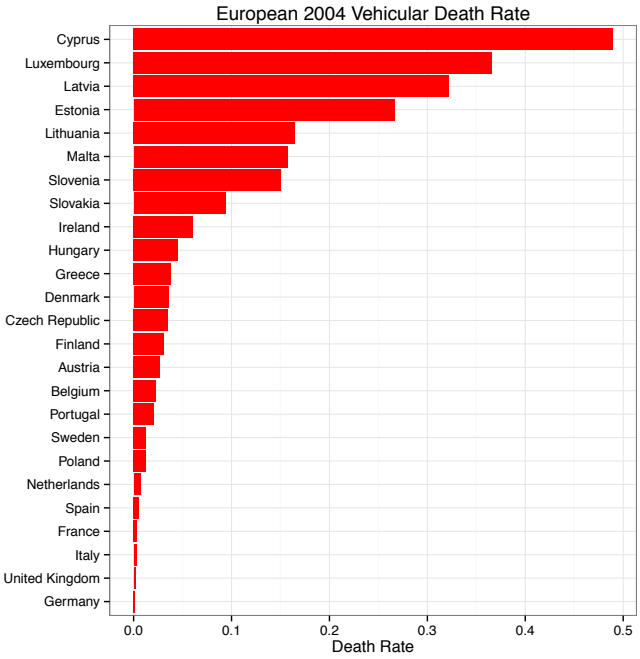
Solution for 9:

(a)

```
> CARS2004 <- within(data = CARS2004, expr = {
+     total.cars = cars * population/1000
+     death.rate = deaths/total.cars
+ })
> head(CARS2004)

         country cars deaths population death.rate total.cars
1        Belgium  467    112      10396 0.02306932   4854.932
2 Czech Republic  373    135      10212 0.03544167   3809.076
3        Denmark  354     68       5398 0.03558548   1910.892
4        Germany  546     71      82532 0.00157559  45062.472
5        Estonia  350    126       1351 0.26646928    472.850
6         Greece  348    147      11041 0.03825865   3842.268
```

(b)

```
> ggplot(data =CARS2004, aes(x = reorder(country, death.rate),
+                            y = death.rate)) +
+   geom_bar(stat = "identity", fill = "red") +
+   coord_flip() +
+   labs(x = "", y = "Death Rate",
+        title = "European 2004 Vehicular Death Rate") +
+   theme_bw()
```



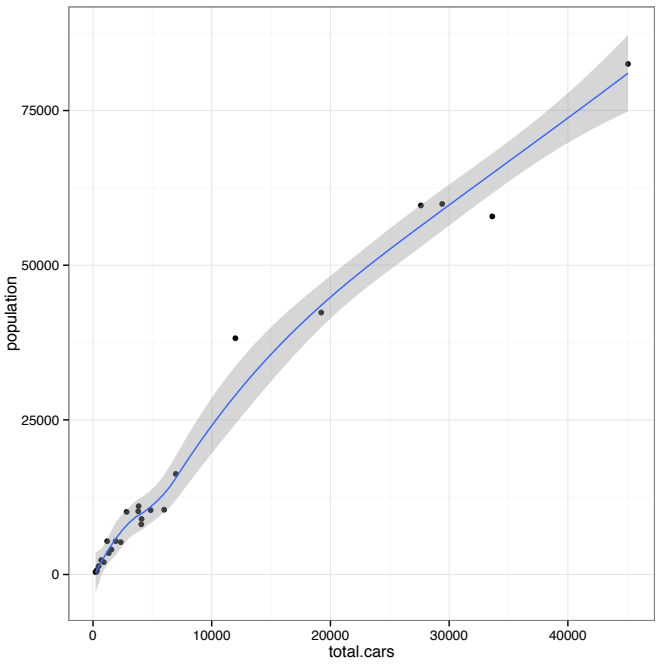
(c)

The country with the lowest automobile death rate is Germany while Cyprus has the highest automobile death rate.

(d)

There is a positive curvilinear relationship between `total.cars` and `population`.

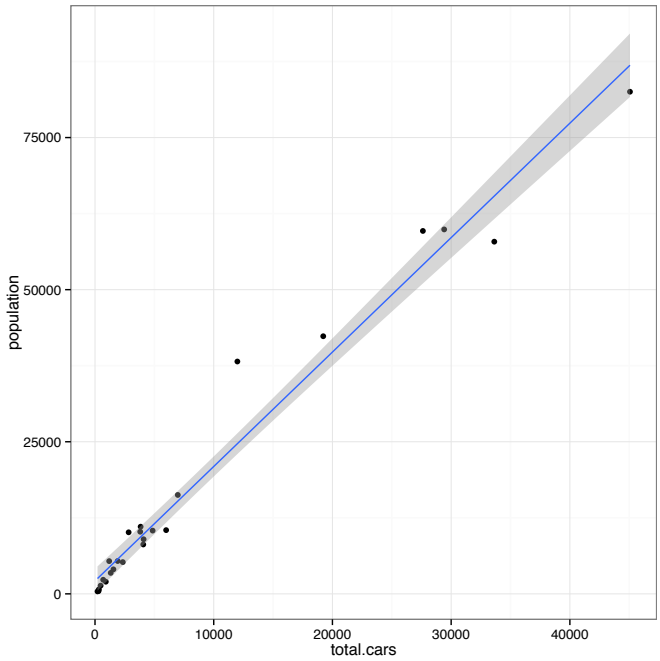
```
> ggplot(data = CARS2004, aes(x = total.cars, y = population)) +
+   geom_point() +
+   geom_smooth() +
+   theme_bw()
```



(e)

```
> mod.lm <- lm(population ~ total.cars, data = CARS2004)
> summary(mod.lm)


Call:
lm(formula = population ~ total.cars, data = CARS2004)

Residuals:
   Min     1Q Median     3Q    Max
 -7500  -1840  -1013   1015  13510

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.124e+03  9.731e+02   2.183   0.0395 *
total.cars  1.881e+00  6.561e-02  28.668   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3967 on 23 degrees of freedom
Multiple R-squared:  0.9728,Adjusted R-squared:  0.9716
F-statistic: 821.8 on 1 and 23 DF,  p-value: < 2.2e-16

> ggplot(data = CARS2004, aes(x = total.cars, y = population)) +
+   geom_point() +
+   geom_smooth(method = "lm") +
+   theme_bw()
> POP <- predict(mod.lm, newdata = data.frame(total.cars = 19224.630))*1000
> POP
```

```
       1
38285550

> resid(mod.lm)[7]*1000

      7
4059450
```

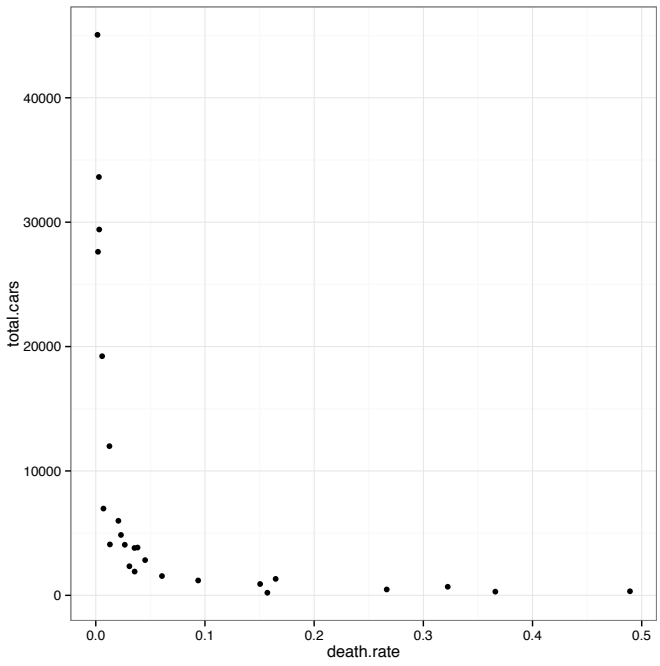


The least squares model predicts a population of 38285550.4948 people. Spain has a `total.cars` value of 19224.630 and a reported population of 42,345,000. The difference between Spain's actual population and the value predicted with least squares is the seventh residual $42,345,000 - 38,285,550 = 4,059,450$.

(f)

There is a decreasing monotonic relationship between `total.cars` and `death.rate`.

```
> ggplot(data = CARS2004, aes(x = death.rate, y = total.cars)) +
+   geom_point() +
+   theme_bw()
```



(g)

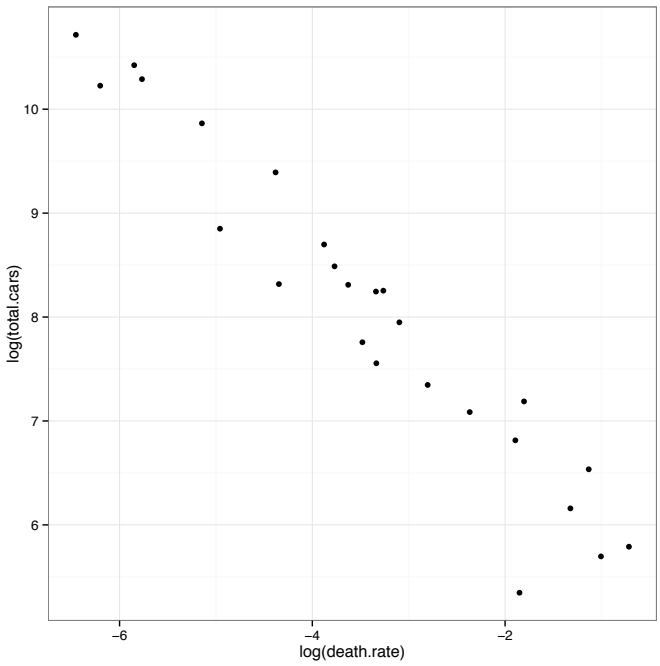
Spearman's rank correlation is a measure of the monotonic relationship between two variables.

```
> with(data = CARS2004, cor(total.cars, death.rate, method = "spearman"))

[1] -0.9676923
```

(h)

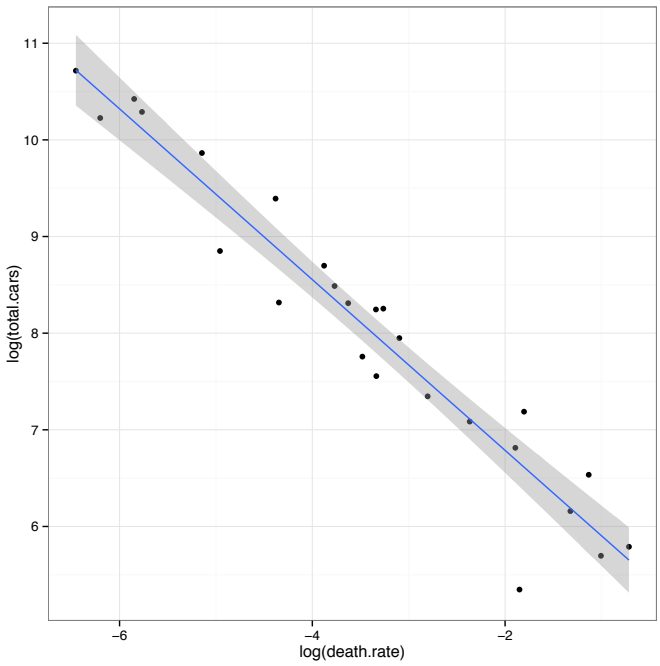
The relationship is strong, negative, and linear between the logarithm of `total.cars` and the logarithm of `death.rate`.

```
> ggplot(data = CARS2004, aes(x = log(death.rate), y = log(total.cars))) +
+   geom_point() +
+   theme_bw()
```



(i) The total number of cars predicted for a country with a `logdeath.rate = -3.769252` is 4231.018 cars.

```
> ggplot(data = CARS2004, aes(x = log(death.rate), y = log(total.cars))) +
+   geom_point() +
+   theme_bw() +
+   geom_smooth(method = "lm")
> modlm.log <- lm(log(total.cars) ~ log(death.rate), data = CARS2004)
> coef(summary(modlm.log))

                 Estimate Std. Error   t value     Pr(>|t|)
(Intercept)      5.0206666 0.19568324  25.65711 1.994256e-18
log(death.rate) -0.8833401 0.05142204 -17.17824 1.293676e-14

> TOTCARS <- exp(predict(modlm.log,
+                        newdata = data.frame(death.rate = exp(-3.769252))))
> TOTCARS

       1
4231.018
```


log(total.cars)
log(death.rate)
11
10
9
8
7
6
−6
−4
−2

Chapter 3

Odd solutions

Solution for 1:

```
> choose(90, 8)

[1] 77515521435
```

There are $\binom{90}{8} = 77,515,435$ ways to choose 8 people from 90.

Solution for 3:

```
> 26 * 25 * 10 * 9 * 8

[1] 468000
```

There are a total of 468000 possible license plates if repetition among letters and numbers is not permissible.

Solution for 5:

```
> choose(10, 2)

[1] 45
```

There is one way for the smallest number to be a 5 and the largest number to be a 16. This leaves two numbers to be drawn between the 6 and 15. There are 10 numbers between 6 and 15 inclusive, so the remaining two numbers can be selected $\binom{10}{2} = 45$ different ways.

Solution for 7:

```
> BA <- factorial(7)/factorial(2)
> PA <- factorial(11)/(factorial(2)*factorial(2))
> SA <- factorial(10)/(factorial(3)*factorial(3)*factorial(2))
> c(BA, PA, SA) # arrangements of BIOLOGY, PROBABILITY, and STATISTICS

[1]    2520 9979200   50400
```

Solution for 9:

```
> choose(3, 2)*choose(47, 2) + choose(3, 3)*choose(47, 1)

[1] 3290
```

Solution for 11:

```
> factorial(5)

[1] 120

> factorial(5) - 2 * factorial(4)

[1] 72
```

There are 120 ways five politicians can stand in line. If two of the politicians refuse to stand next to each other, there are 72 ways they may stand in line.

Solution for 13:

(a)

```
> 10 * 9 * 8

[1] 720
```

(b)

```
> 9 * 8 * 7 + 9 * 8

[1] 576
```

(c)

```
> 8 * 7 * 6 + 3 * 2 * 8 * 7

[1] 672
```

(d)

```
> 8 * 7 * 6 + 3 * 2 * 8

[1] 384
```

(e)

```
> 3 * 9 * 8

[1] 216
```

Solution for 15:

```
> choose(4, 2) * choose(6, 2)/choose(10, 4)

[1] 0.4285714
```

Solution for 17:

```
> choose(6, 6)/choose(54, 6)                # Pr(first prize)

[1] 3.871892e-08

> choose(6, 5)*choose(48, 1)/choose(54, 6)  # Pr(second prize)

[1] 1.115105e-05

> choose(6, 4)*choose(48, 2)/choose(54, 6)  # Pr(third prize)

[1] 0.0006551242
```

Solution for 19:

```
> fractions(choose(3, 2)*(1/3)^2*(2/3)^1)

[1] 2/9
```

Solution for 21:

(a) $6^4 = 1296$ (b) $5^4 = 625$ (c) $6 \cdot 5 \cdot 4 \cdot 3 = 360$ (d) $\frac{6 \cdot 5 \cdot 4 \cdot 3}{6^4} = \frac{360}{1296} = \frac{5}{18}$

Solution for 23:

(a) 0.8, (b) 0.3, (c) 0.9, (d) 0.3

Solution for 25:

It must be shown that $\mathbb{P}(F|E) = \dfrac{\mathbb{P}(F \cap E)}{\mathbb{P}(E)}$ satisfies

(1) $0 \leq \mathbb{P}(F|E) \leq 1$

(2) $\mathbb{P}(\Omega|E) = 1$

(3) $\mathbb{P}(\cup_{i=1}^{\infty} F_i|E) = \sum_{i=1}^{n} \mathbb{P}(F_i|E)$

(1) The left side is obvious since all probabilities must be greater than or equal to zero. Since $F \cap E \subset E$, it follows that $\mathbb{P}(F \cap E) \leq \mathbb{P}(E)$ which is less than or equal to one.

(2) $\mathbb{P}(\Omega|E) = \dfrac{\mathbb{P}(\Omega \cap E)}{\mathbb{P}(E)} = \dfrac{\mathbb{P}(E)}{\mathbb{P}(E)} = 1$

(3)

$$\begin{aligned}
\mathbb{P}(\cup_{i=1}^{\infty} F_i|E) &= \frac{\mathbb{P}\big(\cup_{i=1}^{\infty}(F_i \cap E)\big)}{\mathbb{P}(E)} \\
&= \sum_{i=1}^{\infty} \frac{\mathbb{P}(F_i \cap E)}{\mathbb{P}(E)} \\
&= \sum_{i=1}^{n} \mathbb{P}(F_i|E)
\end{aligned}$$

Solution for 27:

(a) If A and B are mutually exclusive, $\mathbb{P}(A|B) + \mathbb{P}(A^c|B^c) \neq 1$. Counterexample: Consider rolling a fair die and let the event A be rolling an even number and the event B be rolling an odd number. Then, $\mathbb{P}(A|B) + \mathbb{P}(A^c|B^c) = 0 + 0 = 0 \neq 1$.

(b) If A and B are independent, then $\mathbb{P}(A|B) + \mathbb{P}(A^c|B^c) = 1$ because $\mathbb{P}(A|B) + \mathbb{P}(A^c|B^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1$. Recall that if A and B are independent, A^c and B^c are also independent; so $\mathbb{P}(A|B) = \mathbb{P}(A)$ and $\mathbb{P}(A^c|B^c) = \mathbb{P}(A^c)$.

Solution for 29:

Let B = banned substance is present and $+$ = a positive test.

$$\begin{array}{ll}
\mathbb{P}(+|B) = 0.98 & \mathbb{P}(B) = 0.02 \\
\mathbb{P}(+|B^c) = 0.02 & \mathbb{P}(B^c) = 0.98
\end{array}$$

$$\begin{aligned}\mathbb{P}(B|+) &= \frac{\mathbb{P}(B \cap +)}{\mathbb{P}(+)} \\ &= \frac{\mathbb{P}(+|B) \cdot \mathbb{P}(B)}{\mathbb{P}(+|B) \cdot \mathbb{P}(B) + \mathbb{P}(+|B^c) \cdot \mathbb{P}(B^c)} \\ &= \frac{0.98 \times 0.02}{0.98 \times 0.02 + 0.02 \times 0.98} = 0.50\end{aligned}$$

Solution for 31:

(a) Let B_1 represent the lot obtained from passing two chipped beads from A to B. Let B_2 represent the lot obtained from passing one chipped bead and one non-chipped bead from A to B. Let B_3 represent the lot obtained from passing two non-chipped beads from A to B. Let C be the event selecting a chipped bead.

$$\mathbb{P}(B_1) = \frac{\binom{2}{2}\binom{2}{0}}{\binom{4}{2}} = \frac{1}{6}; \quad \mathbb{P}(B_2) = \frac{\binom{2}{1}\binom{2}{1}}{\binom{4}{2}} = \frac{4}{6}; \quad \mathbb{P}(B_3) = \frac{\binom{2}{0}\binom{2}{2}}{\binom{4}{2}} = \frac{1}{6}$$

$$\begin{aligned}\mathbb{P}(C) &= \mathbb{P}(C|B_1) \cdot \mathbb{P}(B_1) + \mathbb{P}(C|B_2) \cdot \mathbb{P}(B_2) + \mathbb{P}(C|B_3) \cdot \mathbb{P}(B_3) \\ &= \frac{4}{7} \cdot \frac{1}{6} + \frac{3}{7} \cdot \frac{4}{6} + \frac{2}{7} \cdot \frac{1}{6} = \frac{18}{42} = \frac{3}{7}\end{aligned}$$

(b)

$$\begin{aligned}\mathbb{P}(B_3|C^c) &= \frac{\mathbb{P}(C^c|B_3) \cdot \mathbb{P}(B_3)}{\mathbb{P}(C^c)} \\ &= \frac{\frac{5}{7} \cdot \frac{1}{6}}{\frac{4}{7}} = \frac{5}{24}\end{aligned}$$

Solution for 33:

Let I, II, and III be events associated with suppliers and D be the event associated with a defective appliance.

$$\begin{aligned}\mathbb{P}(D) &= \mathbb{P}(D|I) \cdot \mathbb{P}(I) + \mathbb{P}(D|II) \cdot \mathbb{P}(II) + \mathbb{P}(D|III) \cdot \mathbb{P}(III) \\ &= 0.02 \times 0.35 + 0.01 \times 0.25 + 0.03 \times 0.40 = 0.0215\end{aligned}$$

$$\mathbb{P}(III|D) = \frac{\mathbb{P}(D|III) \cdot \mathbb{P}(III)}{\mathbb{P}(D)} = \frac{0.03 \times 0.4}{0.0215} = 0.5581$$

Solution for 35:

Let W_1 = first draw is white and W_2 be second draw is white. Since there are only two colors of balls, W_1^c = first draw is black and W_2^c is the second draw is black.

$$\begin{aligned}
\mathbb{P}(W_1^c|W_2) &= \frac{\mathbb{P}(W_1^c \cap W_2)}{\mathbb{P}(W_2)} \\
&= \frac{\mathbb{P}(W_2|W_1^c)\cdot\mathbb{P}(W_1^c)}{\mathbb{P}(W_2)} \\
&= \frac{\mathbb{P}(W_2|W_1^c)\cdot\mathbb{P}(W_1^c)}{\mathbb{P}(W_2|W_1^c)\cdot\mathbb{P}(W_1^c)+\mathbb{P}(W_2|W_1)\cdot\mathbb{P}(W_1)} \\
&= \frac{\frac{3}{10}\cdot\frac{8}{14}}{\frac{3}{10}\cdot\frac{8}{14}+\frac{4}{10}\cdot\frac{6}{14}} \\
&= \frac{\frac{24}{140}}{\frac{24}{140}+\frac{24}{140}} = \frac{1}{2}
\end{aligned}$$

Solution for 37:

$\mathbb{P}(\text{John wins}) = p^2 + (1-p)^2$ and $\mathbb{P}(\text{Peter wins}) = 1 - \mathbb{P}(\text{John wins})$.

(a) When $\mathbb{P}(\text{John wins}) = \mathbb{P}(\text{Peter wins}) = 1/2$, the game is fair.

$$\begin{aligned}
\mathbb{P}(\text{John wins}) = p^2 + (1-p)^2 &= \frac{1}{2} \\
2p^2 + 2(1-p)^2 &= 1 \\
4p^2 - 4^p + 1 &= 0 \\
(2p-1)^2 &= 0 \\
\implies p &= \frac{1}{2}
\end{aligned}$$

If $p = 1/2$ both of them have the same probability of winning the game.

(b) Since $(2p-1)^2 > 0$ for all $p \neq 1/2$, John wins for any different answer than that in (a).

Solution for 39:

The probability density is

x	2	3	4	5	6	7	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

which implies $\mu_X = \sum_x x \cdot p(x) = 7$ and $\sigma_x^2 = \sum_x (x-\mu_x)^2 \cdot p(x) = 5.8\overline{3}$.

The bound given by Chebyshev's Inequality says

$$\begin{aligned}
\mathbb{P}(|X-\mu_x| \geq k) &\leq \frac{\sigma_x^2}{k^2} \\
\mathbb{P}(|X-7| \geq 4) &\leq \frac{5.8\overline{3}}{4^2} \\
\implies \mathbb{P}(|X-7| \geq 4) &\leq 0.3645833
\end{aligned}$$

```
> dicerolls <- expand.grid(1:6, 1:6)
> SDR <- apply(dicerolls, 1, sum)
> PDF <- fractions(table(SDR)/36)
> PDF
```

```
SDR
   2    3    4    5    6    7    8    9   10   11   12
1/36 1/18 1/12  1/9 5/36  1/6 5/36  1/9 1/12 1/18 1/36

> MX <- sum(2:12 * PDF)
> VX <- sum((2:12 - MX)^2 * PDF)
> SX <- sqrt(VX)
> c(MX, VX, SX)

[1] 7.000000 5.833333 2.415229

> UL <- VX/4^2
> UL

[1] 35/96
```

The exact probability is $\mathbb{P}(X = 2) + \mathbb{P}(X = 3) + \mathbb{P}(X = 11) + \mathbb{P}(X = 12) = 1/6$.

Solution for 41:

(a) Let A = Alvin buys a stamp. Let S_i = stamp is won on the i^{th} bid. Note that $(1-p) < 1$, so that the geometric series simplification holds.

$$\begin{aligned}
\mathbb{P}(S_i) &= (1-p)^{i-1} \cdot p \\
\mathbb{P}(A) &= \mathbb{P}(S_1 \cup S_3 \cup S_5 \cup \cdots) \\
&= p + (1-p)^2 \cdot p + (1-p)^4 \cdot p + \cdots \\
&= p\left[1 + (1-p)^2 + (1-p)^4 + (1-p)^6 + \cdots\right] \\
&= p \cdot \frac{1}{1-(1-p)^2} \\
&= \frac{p}{1-(1-2p+p^2)} \\
&= \frac{1}{2-p}
\end{aligned}$$

(b) The probability that Alvin wins two auctions is the probability he wins the first auction squared, $\left(\frac{1}{2-p}\right)^2$.

The probability that Bob wins two auctions is the probability he wins one auction squared. Let B = Bob buys a stamp. Let S_i = stamp is won on the i^{th} bid.

$$\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(S_2 \cup S_4 \cup S_6 \cup \cdots) \\
&= (1-p) \cdot p + (1-p)^3 \cdot p + (1-p)^5 \cdot p + \cdots \\
&= p(1-p)\left[1 + (1-p)^2 + (1-p)^4 + (1-p)^6 + \cdots\right] \\
&= p(1-p) \cdot \frac{1}{1-(1-p)^2} \\
&= \frac{p(1-p)}{1-(1-2p+p^2)} \\
&= \frac{1-p}{2-p}
\end{aligned}$$

The probability Bob wins both auctions is then $\left(\frac{1-p}{2-p}\right)^2$.

This means the probability that both stamps are purchased by the same bidder is

$$\begin{gathered}\mathbb{P}(\text{Alvin wins both stamps} \cup \text{Bob wins both stamps}) = \\ \mathbb{P}(\text{Alvin wins both stamps}) + \mathbb{P}(\text{Bob wins both stamps}) = \\ \left(\frac{1}{2-p}\right)^2 + \left(\frac{1-p}{2-p}\right)^2 = \frac{2-2p+p^2}{(2-p)^2}\end{gathered}$$

Solution for 43:

Given their current points, the probability that Louis will score next is $\frac{19}{37}$ and the probability that Joseph scores next is $\frac{18}{37}$. Assuming that the rest of the point scoring maintains these probabilities, then the games that can be played are LL, LJL, JJLL, JLL, JLJL, LJJL, which give Louis wins and JJJ, LJJJ, JLJJ, JJLJ which give Joseph wins. Multiplying and adding gives Louis a probability of winning of 0.707491512 and Joseph a probability of winning of 0.292508488. This means Louis receives €70.75 and Joseph receives €29.25.

For example, the probability that the game JJLL is played is $\left(\frac{18}{37}\right)^2\left(\frac{19}{37}\right)^2 = 0.0624$.

Solution for 45:

x	0	100	200	300	400
$p(x)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Solution for 47:

Let X = number of rounds a person plays until he loses once.

(a)

$$\begin{aligned}\mathbb{P}(X > 4) &= 1 - \mathbb{P}(X \leq 4) \\ &= 1 - [\mathbb{P}(X=4) + \mathbb{P}(X=3) + \mathbb{P}(X=2) + \mathbb{P}(X=1)] \\ &= 1 - \left[\frac{1}{2^4} + \frac{1}{2^3} + \frac{1}{2^2} + \frac{1}{2}\right] = 0.0625\end{aligned}$$

(b) To leave with exactly€600, he must win the first seven games and lose the eighth, so $\mathbb{P}(X=8) = \frac{1}{2^8} = 0.0039$.

(c) If you lose on the first round, you have lost the ante: -100 winnings. If win the first and lose the second, you break even. By the second round, if you win, you have positive winnings.

$$\begin{aligned}
E(\text{winnings}) &= \frac{1}{2}(-100) + \frac{1}{4}(100) + \frac{1}{8}(100) + \frac{1}{2^4}(200) + \dots \\
&= \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i [-100 + 100(i-1)] \\
&= \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i [-100 + 100i - 100] \\
&= \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i (100i) - 200 \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i \\
&= \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i (100i) - 200(1) \quad \text{infinite geometric sum}
\end{aligned}$$

Solving for the first term of previous expression:

Let $S = \sum_{i=1}^{\infty} \left(\frac{1}{2}\right)^i (100i)$. This means that $2S = \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i 100(i+1)$ (note i starts at 0 in the second sum). Then,

$$\begin{aligned}
2S &= 100 + 200\left(\frac{1}{2}\right) + 300\left(\frac{1}{4}\right) + 400\left(\frac{1}{8}\right) + \dots \\
S &= \qquad 100\left(\frac{1}{2}\right) + 200\left(\frac{1}{4}\right) + 300\left(\frac{1}{8}\right) + \dots \\
2S - S &= = 100 + 100\left(\frac{1}{2}\right) + 100\left(\frac{1}{4}\right) + 100\left(\frac{1}{8}\right) + \dots
\end{aligned}$$

Therefore

$$S = 100 \sum_{i=0}^{\infty} \left(\frac{1}{2}\right)^i = 100(2) = 200$$

This means that $E(\text{winnings}) = S - 200 = 200 - 200 = 0$.

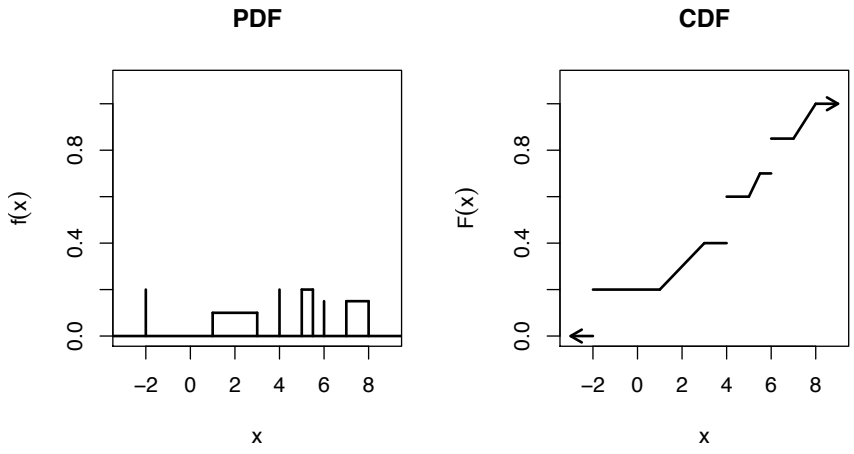
Solution for 49:

Verify that the probabilities of points and areas given add to 1:

$$\begin{aligned}
&\mathbb{P}(X = -2) + \mathbb{P}(1 < X \leq 3) + \mathbb{P}(X = 4) \\
&\quad + \mathbb{P}(5 < X \leq 5.5) + \mathbb{P}(X = 6) + \mathbb{P}(7 < X \leq 8) \stackrel{?}{=} 1 \\
&\quad 0.2 + 0.1(2) + 0.2 + 0.2(.5) + .15 + .15(1) \stackrel{?}{=} 1 \\
&\quad 1 = 1
\end{aligned}$$

```
> opar <- par(no.readonly = TRUE)
> par(mfrow=c(1, 2))
> x <- -3:9
```

```
> plot(x, type = "n", ylab = expression(f(x)), xlab = "x",
+ ylim=c(0, 1.1), xlim = c(-3, 9), main = "PDF")
> segments(-2, 0, -2, 0.2, lwd = 2)
> segments(1, 0, 1, 0.1, lwd = 2)
> segments(1, 0.1, 3, 0.1, lwd = 2)
> segments(3, 0.1, 3, 0, lwd = 2)
> segments(4, 0, 4, 0.2, lwd = 2)
> segments(5, 0, 5, 0.2, lwd = 2)
> segments(5, 0.2, 5.5, 0.2, lwd = 2)
> segments(5.5, 0.2, 5.5, 0, lwd = 2)
> segments(6, 0, 6, 0.15, lwd = 2)
> segments(7, 0, 7, 0.15, lwd = 2)
> segments(7, 0.15, 8, 0.15, lwd = 2)
> segments(8, 0.15, 8, 0, lwd = 2)
> abline(h = 0, lwd = 2)
> plot(x, type = "n", ylab = expression(F(x)), xlab = "x",
+ ylim = c(0, 1.1), xlim = c(-3, 9), main = "CDF")
> arrows(-2, 0, -3, 0, length = 0.1, lwd = 2)
> segments(-2, 0.2, 1, 0.2, lwd = 2)
> segments(1, 0.2, 3, 0.4, lwd = 2)
> segments(3, 0.4, 4, 0.4, lwd = 2)
> segments(4, 0.6, 5, 0.6, lwd = 2)
> segments(5, 0.6, 5.5, 0.7, lwd = 2)
> segments(5.5, 0.7, 6, 0.7, lwd = 2)
> segments(6, 0.85, 7, 0.85, lwd = 2)
> segments(7, 0.85, 8, 1, lwd = 2)
> arrows(8, 1, 9, 1, length = 0.1, lwd = 2)
> par(opar)
```



Solution for 51:

First, find k, $\int_{-1}^{1} k\,dx = kx|_{-1}^{1} = k(1-(-1)) = 2k \overset{\text{set}}{=} 1 \implies k = {}^1\!/_2$.

Skewness is $\gamma_1 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$, so expected value and variance must be calculated as must $E[X^3]$.

$$\mu = E[X] = \int_{-1}^{1} \frac{x}{2}\, dx = \left.\frac{x^2}{4}\right|_{-1}^{1} = 0$$

and

$$\sigma^2 = Var[X] = \int_{-1}^{1} \frac{(x-0)^2}{2}\, dx = \left.\frac{x^3}{6}\right|_{-1}^{1} = \frac{2}{6} = \frac{1}{3}$$

$$E[X^3] = \int_{-1}^{1} \frac{x^3}{2}\, dx = \left.\frac{x^4}{8}\right|_{-1}^{1} = 0$$

Since $Var[X] > 0$ and $E[(X-\mu)^3] = E[X^3] = 0$, the skewness is zero.

```
> f <- function(x){x/2}
> g <- function(x){x^2/2}
> h <- function(x){x^3/2}
> Mu <- integrate(f, -1, 1)$value
> Mu

[1] 0

> Va <- integrate(g, -1, 1)$value
> Va

[1] 0.3333333

> EX3 <- integrate(h, -1, 1)$value
> EX3

[1] 0
```

Solution for 53:

(a) From Markov's theorem, $\mathbb{P}\big(g(X) \geq K\big) \leq \frac{E[g(X)]}{K}$, if $g(X) = X$, then $\mathbb{P}\big(X \geq 112\big) \leq \frac{E[X]}{112} = \frac{100}{112} = 89.2857\%$.

(b) Chebyshev's inequality says that $\mathbb{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$.
Here, $\mathbb{P}(|X-100| \geq 12 = 2 \times 6) \leq \frac{1}{2^2} = \frac{1}{4} = 25\%$.

(c) $\mathbb{P}(88 < X < 112) = \mathbb{P}(|X-100| < 12) \geq 1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$.

(d) $\mathbb{P}(|\overline{X} - 100| < k\sigma_{\overline{X}}) \geq 1 - \frac{1}{k^2}$ to get within 6, $k\frac{\sigma}{\sqrt{n}} = 6 \Longrightarrow k = \sqrt{n}$. Since $1 - \frac{1}{k^2}$ is to be 0.9, $1 - \frac{1}{n} = .9 \Longrightarrow n = 10$.

Solution for 55:

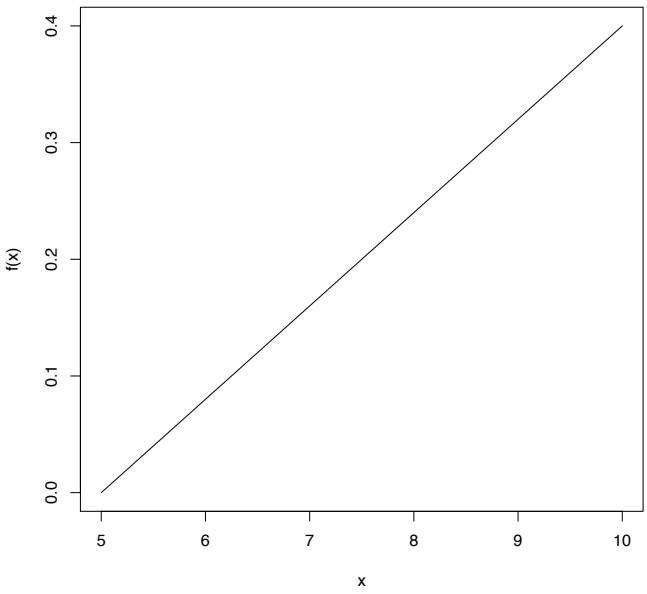
(a) Property 1 says $f(x) \geq 0$. Since $(x-5) \geq 0$ when $5 \leq x \leq 10$, this is satisfied. Property 2 says the integral over the domain is 1:

$$\int_5^{10} \frac{2}{25}(x-5)\, dx = \left.\frac{(x-5)^2}{25}\right|_5^{10} = \frac{25}{25} - 0 = 1.$$

```
> f <- function(x){2/25*(x - 5)}
> integrate(f, 5, 10)$value

[1] 1
```

(b)

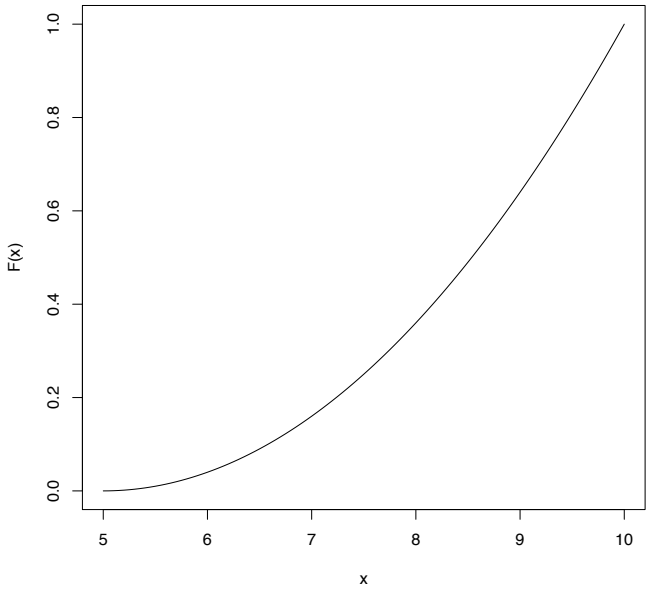
```
> curve(f, 5, 10)
```



(c)

$$F(x) = \int_5^x \frac{2}{25}(y-5)\,dy = \frac{(y-5)^2}{25}\bigg|_5^x = \frac{(x-5)^2}{25}$$

```
> F <- function(x){(x - 5)^2/25}
> curve(F, 5, 10)
```



(d) $\mathbb{P}(X \leq 8) = F(8) = \dfrac{(8-5)^2}{25} = \dfrac{9}{25} = 0.36$

$\mathbb{P}(X \geq 6) = 1 - F(6) = 1 - \dfrac{(6-5)^2}{25} = \dfrac{24}{25} = 0.96$

$\mathbb{P}(7 \leq X \leq 8) = F(8) - F(7) = \dfrac{(8-5)^2}{25} - \dfrac{(7-5)^2}{25} = \dfrac{9}{25} - \dfrac{4}{25} = \dfrac{1}{5} = 0.2$

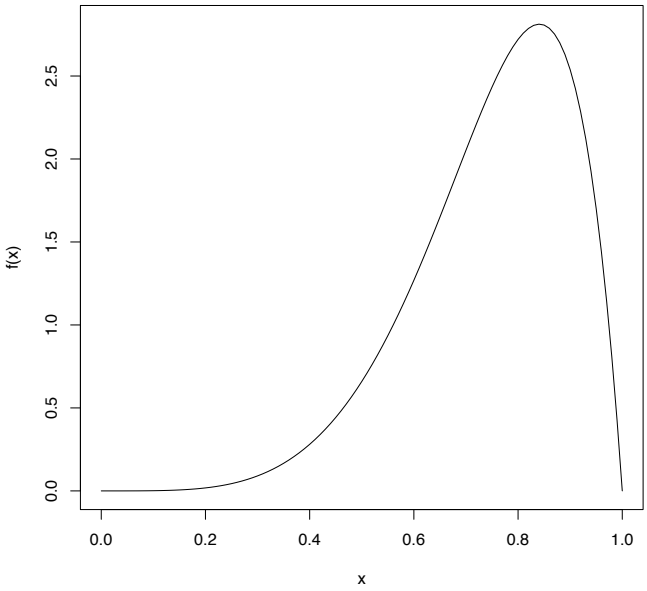
(e)

```
> integrate(f, 5, 8)$value  # P(X <=8)
[1] 0.36
> integrate(f, 6, 10)$value  # P(X >=6)
[1] 0.96
> integrate(f, 7, 8)$value  # P(7 < X < 8)
[1] 0.2
```

Solution for 57:

(a)

```
> f <- function(x){630/56*x^4*(1-x^4)}
> curve(f, 0, 1)
```



(b)

$$\begin{aligned} E[X] &= \int_0^1 x \cdot \frac{630}{56} x^4 \left(1 - x^4\right)\, dx \\ &= \frac{630}{56} \int_0^1 x^5 - x^9\, dx \\ &= \frac{630}{56} \left[\frac{x^6}{6} - \frac{x^{10}}{10} \Big|_0^1 \right] \\ &= \frac{630}{56} \left[\frac{1}{6} - \frac{1}{10} - 0 \right] \\ &= \frac{630}{56} \cdot \frac{2}{30} = \frac{3}{4} \end{aligned}$$

(c) First, find $E[X^2]$.

$$\begin{aligned} E[X^2] &= \int_0^1 x^2 \cdot \frac{630}{56} x^4 \left(1 - x^4\right) \, dx \\ &= \frac{630}{56} \int_0^1 x^6 - x^10 \, dx \\ &= \frac{630}{56} \left[\frac{x^7}{7} - \frac{x^{11}}{11} \Big|_0^1 \right] \\ &= \frac{630}{56} \left[\frac{1}{7} - \frac{1}{11} - 0 \right] \\ &= \frac{630}{56} \cdot \frac{4}{77} = \frac{45}{77} = 0.5844 \end{aligned}$$

$$Var[X] = E[X^2] - (E[X])^2 = \frac{45}{77} - \frac{9}{16} = \frac{27}{1232} = 0.0219$$

(d)

$$\begin{aligned} \mathbb{P}(0.2 < X < 0.8) &= \int_{0.2}^{0.8} \frac{630}{56} x^4 \left(1 - x^4\right) \, dx \\ &= \frac{630}{56} \left[\frac{x^5}{5} - \frac{x^9}{9} \Big|_{0.2}^{0.8} \right] \\ &= \frac{630}{56} \left[\left(\frac{(0.8)^5}{5} - \frac{(0.8)^9}{9} \right) - \left(\frac{(0.2)^5}{5} - \frac{(0.2)^9}{9} \right) \right] \\ &= \frac{222,183}{390,625} = 0.5688 \end{aligned}$$

(e)

```
> fxe <- function(x){x*f(x)}
> EX <- integrate(fxe, 0, 1)$value
> EX

[1] 0.75
```

(f)

```
> fxf <- function(x){x^2*f(x)}
> EX2 <- integrate(fxf, 0, 1)$value
> VX <- EX2 - EX^2
> VX

[1] 0.02191558
```

(g)

```
> integrate(f, 0.2, 0.8)$value  # P(0.2 < X < 0.8)

[1] 0.5687885
```

Solution for 59:

(a)

$$\mathbb{P}(X = 3) = p$$
$$\mathbb{P}(X = 5) = 2p$$
$$\mathbb{P}(X = 7) = 4p$$
$$\mathbb{P}(X = 8) = 8p$$

$$p + 2p + 4p + 8p \stackrel{set}{=} 1 \longrightarrow 15p = 1 \longrightarrow p = 1/15$$

The probability density function for X is

$X = x$	3	5	7	8
$\mathbb{P}(X = x)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{4}{15}$	$\frac{8}{15}$

(b) The cumulative distribution function for X is

$$F(X) = \begin{cases} 0 & x < 3 \\ \frac{1}{15} & 3 \leq x < 5 \\ \frac{3}{15} & 5 \leq x < 7 \\ \frac{7}{15} & 7 \leq x < 8 \\ 1 & x \geq 8 \end{cases}$$

(c) The mean and variance of X are

```
> x <- c(3, 5, 7, 8)
> px <- c(1/15, 2/15, 4/15, 8/15)
> EX <- sum(x * px)
> EX

[1] 7

> VX <- sum((x - EX)^2 * px)
> VX

[1] 2.133333

> EX2 <- sum(x^2 * px)
> EX2

[1] 51.13333

> VX <- EX2 - EX^2
> VX

[1] 2.133333
```

$$E[X] = \sum_x x \cdot \mathbb{P}(X = x) = 3 \cdot \frac{1}{15} + 5 \cdot \frac{2}{15} + 7 \cdot \frac{4}{15} + 8 \cdot \frac{8}{15} = 7$$

$$Var[X] = \sum_x (x - \mu)^2 \cdot \mathbb{P}(X = x) = 2.1333$$

Solution for 61:

(a) $\mathbb{P}(X > 60) = 0.3011942$

```
> fx <- function(x){1/50*exp(-x/50)}
> integrate(f = fx, lower = 60, upper = Inf)$value

[1] 0.3011942
```

(b) $E[X] = 50$, $Y = 0.5 + 0.03X$, $E[Y] = 0.5 + 0.03E[X] = 2$, and $E[1000Y] = 1000E[Y] = 2000$.

```
> Me <- function(x){x*fx(x)}
> EX <- integrate(f = Me, lower = 0, upper = Inf)$value
> EY <- 0.5 + 0.03*EX
> E1000Y <- 1000*EY
> c(EX, EY, E1000Y)

[1]   50    2 2000
```

(c) $Var[Y] = Var[0.5 + 0.03X] = 0.03^2\, Var[X] = 2.25$, $Var[X] = 2500$, and $\gamma_1 = E[(Y - \mu_Y)^3]/\sigma_Y^3 = 213508.74$.

```
> VX <- integrate(f = function(x){(x-EX)^2*fx(x)},
+                 lower = 0, upper = Inf)$value
> VY <- 0.03^2*VX
> SkewY <- integrate(f = function(x){(x-EY)^3*fx(x)},
+                    lower = 0, upper = Inf)$value / (VY^(3/2))
> c(VX, VY, SkewY)

[1]   2500.00      2.25 213508.74
```

Solution for 63:

(a) To be a **pdf**, the integral of $f(x)$ must equal 1.

$$1 = \int_0^1 k(1-x)\,dx = -\frac{k}{2}(1-x)^2\Big|_0^1 = \frac{k}{2}(1-0)^2 = \frac{k}{2}$$

The integral of $f(x)$ is 1 when $k = 2$.

(b) Since

$$\int_0^x 2(1-t)\,dt = -(1-t)^2\Big|_0^x = 1 - (1-x)^2 = x(2-x),$$

it follows that

$$F(x) = \begin{cases} 0 & x < 0 \\ x(2-x) & 0 \le x \le 1 \\ 1 & \text{otherwise.} \end{cases}$$

The probability the child practices more than 48 minutes on a Saturday is

$$\mathbb{P}\left(X > {}^{48}/_{60}\right) = 1 - \mathbb{P}(X \le 0.8) = 1 - F(0.8) = 1 - 0.8(2 - 0.8) = 0.04.$$

```
> f <- function(x){2*(1 - x)}
> integrate(f, 0.8, 1)$value

[1] 0.04
```

Chapter 4

Odd solutions

Solution for 1:
$\mathbb{P}(X = 0) = 0.1353$, $\mathbb{P}(X \geq 3) = 0.3233$, and $\mathbb{P}(X \leq k) \geq 0.70 \implies k = 3$.

```
> dpois(0, 2)

[1] 0.1353353

> ppois(2, 2, lower = FALSE)

[1] 0.3233236

> qpois(0.7, 2)

[1] 3
```

Solution for 3:
$\mathbb{P}(X > 7.1) = 0.4867$, and $k = 9.5249$ if $\mathbb{P}(X < k) = 0.8$.

```
> pnorm(7.1, 7, 3, lower = FALSE)

[1] 0.4867044

> qnorm(0.80, 7, 3)

[1] 9.524864
```

Solution for 5:
$a = 0.7906$ if $\mathbb{P}(X < a) = 0.95$

```
> qgamma(0.95, 2, 6)

[1] 0.7906441
```

Solution for 7:
$X \sim Bin(n = 60, \pi = \frac{1}{6})$

$$
\begin{aligned}
E[X] &= n\pi = 60(1/6) = 10 \\
Var[X] &= n\pi(1-\pi) = 60 \cdot \frac{1}{6} \cdot \frac{5}{6} = \frac{300}{36} \\
E[X^2] = Var[X] + (E[X])^2 &= \frac{50}{6} + 10^2 = \frac{325}{3} = 108.3333
\end{aligned}
$$

Solution for 9:

Let X = number of seeds that germinate. $X \sim Bin(n = 20, \pi = 0.97)$, $\mathbb{P}(X \leq 17) = 0.021$.

```
> pbinom(17, 20, 0.97)

[1] 0.02100836
```

Solution for 11:

(a) Let X = number of cars passing a certain point on the road in 30 seconds. $X \sim Pois(\lambda = 2)$. $\mathbb{P}(X > 3) = 1 - \mathbb{P}(X <= 3) = 0.1429$

```
> 1 - ppois(3, 2)

[1] 0.1428765
```

(b) Let Y = number of cars passing a certain point on the road in 3 minutes. $Y \sim Pois(\lambda = 12)$. $\mathbb{P}(Y > 10) = 1 - \mathbb{P}(X <= 10) = 0.6528$.

```
> 1 - ppois(10, 12)

[1] 0.6527706
```

Solution for 13:

If he fires at targets until the pistol is empty, the probability that he hits only one target out of the bullets shot in the first round of bullets in the pistol he is carrying that day is 0.0142.

Let X = number of shots that hit the target. $X|TRS \sim Bin(5, 0.7)$ and $X|MWFS \sim Bin(7, 0.7)$.

$$\begin{aligned}
\mathbb{P}(X = 1) &= \mathbb{P}(X = 1|TRS)P(TRS) + \mathbb{P}(X = 1|MWFS)P(MWFS) \\
&= \binom{5}{1}(0.7)^1(0.3)^4 \cdot \frac{3}{7} + \binom{7}{1}(0.7)^1(0.3)^6 \cdot \frac{4}{7} \\
&= 0.0142
\end{aligned}$$

```
> AnsA <- dbinom(1, 5, 0.7)*3/7 + dbinom(1, 7, 0.7)*4/7
> AnsA

[1] 0.0141912
```

The probability that he used the pistol with 7 bullets if he hits only one target is 0.1438.

$$\begin{aligned}
\mathbb{P}(MWFS|X = 1) &= \frac{\mathbb{P}(X = 1|MWFS)P(MWFS)}{\mathbb{P}(X = 1)} \\
&= \frac{\binom{7}{1}(0.7)^1(0.3)^6 \cdot \frac{4}{7}}{0.0142} \\
&= 0.1438
\end{aligned}$$

```
> AnsB <- (dbinom(1, 7, 0.7)*4/7)/AnsA
> AnsB

[1] 0.1438356
```

Solution for 15:

(a)

```
> z <- qnorm(0.13)
> z

[1] -1.126391

> sig <- (45 - 48)/z
> sig

[1] 2.663373
```

$$\begin{aligned} z_{0.13} &= \frac{x-\mu}{\sigma} \\ -1.1264 &= \frac{45-48}{\sigma} \\ -1.1264 \cdot \sigma &= -3 \\ \sigma &= \frac{-3}{-1.1264} = 2.6634 \end{aligned}$$

The standard deviation for the current variety of green peppers is 2.6634 grams.

(b)

```
> z <- qnorm(0.05)
> z

[1] -1.644854

> mu <- 45 - z*sig
> mu

[1] 49.38086
```

$$\begin{aligned} z_{0.05} &= \frac{x-\mu}{\sigma} \\ -1.6449 &= \frac{45-\mu}{2.6634} \\ -1.6449 \cdot 2.6634 = -4.3809 &= 45-\mu \\ \mu &= 45 - -4.3809 = 49.3809 \end{aligned}$$

The agronomists should attempt to create green peppers with a mean weight of 49.3809 grams.

(c)

```
> z <- qnorm(0.05)
> sig2 <- (45 - 50)/z
> sig2

[1] 3.039784
```

$$
\begin{aligned}
z_{0.05} &= \frac{x-\mu}{\sigma} \\
-1.6449 &= \frac{45-50}{\sigma} \\
\sigma &= \frac{45-50}{-1.6449} = 3.0398
\end{aligned}
$$

The standard deviation for the new variety of green peppers is 3.0398 grams.

(d) Since the standard deviation of the new variety (3.0398 grams) is greater than the standard deviation of the current variety (2.6634 grams), the new variety is less consistent with respect to weight than the current variety.

Solution for 17:

X is a discrete uniform, which means it takes on values in $\{1, 2, 3, \ldots, n\}$ with probability $\frac{1}{n}$ each.

$E[X] = \sum_{i=1}^{n} x_i \cdot \mathbb{P}(X = x_i) = \sum_{i=1}^{n} i \cdot \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$

$E[X^2] = \sum_{i=1}^{n} x_i^2 \cdot \mathbb{P}(X = x_i) = \sum_{i=1}^{n} i^2 \cdot \frac{1}{n} = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$

$$
\begin{aligned}
Var[X] = E[X^2] - (E[X])^2 &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\
&= \frac{2n^2+3n+1}{6} - \frac{n^2+2n+1}{4} \\
&= \frac{4n^2+6n+2-3n^2-6n-3}{12} \\
&= \frac{n^2-1}{12}
\end{aligned}
$$

Solution for 19:

(a) Given $X \sim N(60, 2)$, $\mathbb{P}(57 < X < 65) = 0.927$. The percent of the manufacturer's sheet metal that can be expected to fall within the specification is 92.6983%.

```
> p <- pnorm(65, 60, 2) - pnorm(57, 60, 2)
> p

[1] 0.9269831
```

(b) Let Y = number of sheets that test between 57 and 65 micra. Then, $Y \sim Bin(4, 0.927)$ and $\mathbb{P}(Y = 4) = 0.7384$, so there is a 0.7384 chance the contractor will purchase from the local manufacturer and pay a premium price.

```
> dbinom(4, 4, p)

[1] 0.7383926
```

(c) $\mathbb{P}(X < k) = 0.985 \implies k = 64.3402 \implies c = 4.3402$

```
> k <- qnorm(0.985, 60, 2)
> INC <- k - 60
> c(k, INC)

[1] 64.340181  4.340181
```

(d) Let W = number of sheets that have hardness greater than 60. Then, $W \sim Bin(20, 0.5)$, and $\mathbb{P}(W \geq 10) = 1 - \mathbb{P}(W \leq 9) = 0.5881$.

```
> 1 - pbinom(9, 20, 0.5)

[1] 0.5880985
```

Solution for 21:

Let X = number of accounts closed and refunded out of 50. Then, $X \sim Bin(50, 0.01)$. The minimum number of accounts closed where the probability is at least 0.95 occurs where $\mathbb{P}(X \leq k) \geq 0.95 \implies k = 2$. Therefore, the bank must have on hand $2 \times 25,000 =$ € 50,000.

```
> k <- qbinom(0.95, 50, 0.01)
> k

[1] 2

> euros <- k*25000
> euros

[1] 50000
```

Solution for 23:

(a) Let X = number of buckets received per minute. Then, $X \sim Pois(\lambda = 6)$. If 6 buckets arrive in 60 seconds, 9 buckets are expected to arrive in 90 seconds. Let Y = number of buckets that arrive in 90 seconds. It follows then that $Y \sim Pois(\lambda = 9)$. Assume an employee is hand transporting the buckets at the same rate given in the problem. $\mathbb{P}(Y > 8) = 1 - \mathbb{P}(Y \leq 8) = 0.5443$

```
> 1 - ppois(8, 9)

[1] 0.5443474
```

(b) Let T = time until the first bucket arrives. $T \sim \Gamma(1, 6)$ and $\mathbb{P}(T < 1/6) = 0.6321$.

```
> pgamma(1/6, 1, 6)

[1] 0.6321206
```

Solution for 25:

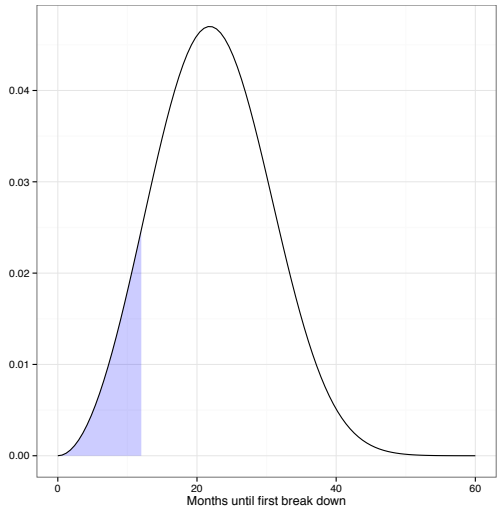
Let X = the number of months until the first breakdown. $X \sim Weib(3, 25)$. $\mathbb{P}(X \leq 12) = 0.1047$

```
> pweibull(12, 3, 25)

[1] 0.104696
```

Since there is a 10.4696% chance of break down in the first 12 months and there are 50 cars, we can expect 5.2348 cars to break down during the guarantee period. Consequently, the expected price of the guarantee is $4187.8417.

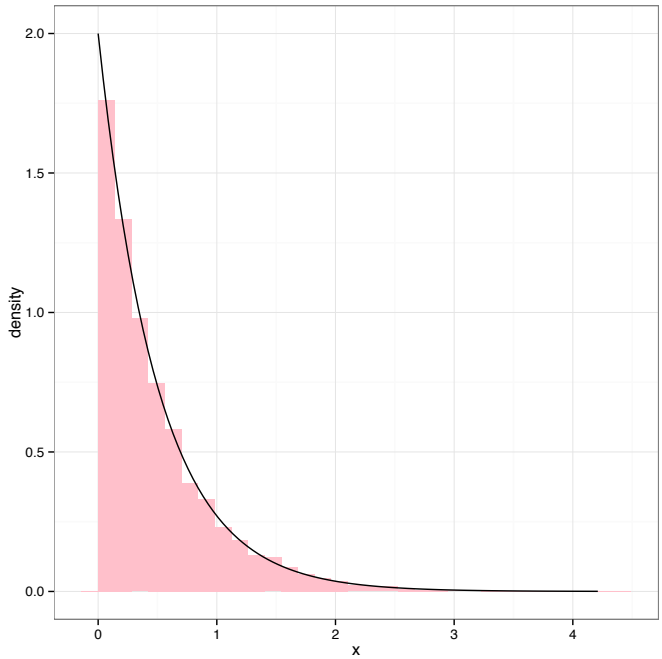
```
> EBD <- 50 * pweibull(12, 3, 25)  # E(Break Downs)
> EBD

[1] 5.234802

> EGP <- EBD*800                   # E(Price of Guarantee)
> EGP

[1] 4187.842
```

```
> limitRange <- function(fun, shape, scale, min, max){
+ function(x){
+ y <- fun(x, shape, scale)
+ y[x < min | x > max] <- NA
+ return(y)
+ }
+ }
> dlimit <- limitRange(dweibull, 3, 25, 0, 12)
> p <- ggplot(data = data.frame(x = c(0, 60)), aes(x = x))
> p + stat_function(fun = dlimit, geom = "area", fill = "blue",
+ alpha = 0.2) +
+ stat_function(fun = dweibull, arg = list(3, 25)) +
+ theme_bw() +
+ labs(x = "Months until first break down", y = "")
```



Solution for 27:

```
> set.seed(50)
> rs <- rexp(10000, 2)
> DF <- data.frame(x = rs)
> ggplot(data = DF, aes(x = x)) +
+ geom_histogram(aes(y = ..density..), fill = "pink") + theme_bw() +
+ stat_function(fun = dexp, arg = list(2))
> Mu <- mean(rs)
> Va <- var(rs)
> abs((Mu - 1/2)/(1/2))*100

[1] 0.2781262

> abs((Va - 1/4)/(1/4))*100

[1] 0.7784307
```



The mean and variance of the simulated values are 0.4986 and 0.2519, respectively. Note that both values are within 1% of the theoretical mean and variance of an $Exp(2)$ distribution. (That is, $\mu = 1/\lambda = 0.5$ and $\sigma^2 = 1/\lambda^2 = 0.25$.)

Solution for 29:

(a)

$$\begin{aligned} F_X(x) &= \int_1^x 3\left(\frac{1}{t}\right)^4 dt \\ &= -t^{-3}\Big|_1^x \\ &= -x^{-3} + 1 \end{aligned}$$

(b) To do the generation, find the relationship between a uniform and X.

$$\begin{aligned} F_X(x) = -x^{-3} + 1 &\stackrel{\text{set}}{=} u \\ -x^{-3} &= u - 1 \\ x^{-3} &= 1 - u \\ x &= (1-u)^{-1/3} \\ x &= 1/(1-u)^{1/3} \end{aligned}$$

```
> set.seed(98)
> n <- 100000
> u <- runif(n, 0, 1)
> x <- 1/(1 -u)^(1/3)
> ans <- c(mean(x), var(x), mean((x - mean(x))^3)/sd(x)^3)
> names(ans) <- c("Mean", "Variance", "Skewness")
> ans

      Mean   Variance   Skewness
 1.5001303  0.7145073 10.6328404
```

(c) Mean:

$$E[X] = \int_1^\infty x \cdot 3\left(\frac{1}{x}\right)^4 dx = \int_1^\infty 3x^{-3}\,dx = \left.\frac{-3x^{-2}}{2}\right|_1^\infty = \frac{3}{2}$$

Variance:

$$E[X^2] = \int_1^\infty x^2 \cdot 3\left(\frac{1}{x}\right)^4 dx = \int_1^\infty 3x^{-2}\,dx = \left.-3x^{-1}\right|_1^\infty = 3$$

$$Var[X] = E[X^2] - (E[X])^2 = 3 - \left(\frac{3}{2}\right)^2 = 3 - \frac{9}{4} = \frac{3}{4}$$

Coefficient of skewness:

$$\gamma_1 = \frac{E\left[(X-\mu)^3\right]}{\sigma^3}$$

$$\begin{aligned} E\left[(X-\mu)^3\right] &= \int_1^\infty (x-1.5)^3 \cdot 3\left(\frac{1}{x}\right)^4 dx \\ &= \int_1^\infty 3x^{-4}(x^3 - 4.5x^2 + 6.75x - 3.375)\,dx \\ &= \int_1^\infty (3x^{-1} - 13.5x^{-2} + 20.25x^{-3} - 10.125x^{-4})\,dx \\ &= 3\ln(x) + 13.5x^{-1} - 10.125x^{-2} + \left.\frac{10.125x^{-3}}{3}\right|_1^\infty = \infty \end{aligned}$$

$\mu_X = \frac{3}{2}$, $\sigma_X^2 = \frac{3}{4}$, and $\gamma_1 = \infty$.

```
> c(abs(ans[1] - 3/2)/(3/2)*100,
+ abs(ans[2] - 3/4)/(3/4)*100)

       Mean    Variance
0.008686381 4.732364625
```

The estimated mean and variance from (b) are both within 3% of their theoretical values. The skewness from (b) is not close to ∞.

Solution for 31:

(a) $\int_0^1 f(x)\,dx = \int_0^1 \frac{4}{3}x(2-x^2)\,dx = \int_0^1 \frac{8x}{3} - \frac{4x^3}{3}\,dx = \frac{4x^2}{3} - \frac{x^4}{3}\Big|_0^1 = \left(\frac{4}{3} - \frac{1}{3}\right) - (0-0) = 1$

```
> f <- function(x){4/3*x*(2 - x^2)}
> integrate(f, 0, 1)$value

[1] 1
```

(b) $F_X(x) = \int_0^x \frac{8t}{3} - \frac{4t^3}{3}\,dt = \frac{4t^2}{3} - \frac{t^4}{3}\Big|_0^x = \frac{4x^2}{3} - \frac{x^4}{3}, 0 \le x \le 1$

(c) $\mathbb{P}(X > 0.75) = 1 - F(.75) = 0.3555$

```
> 1 - (4*0.75^2/3 - 0.75^4/3)

[1] 0.3554688

> # or
> integrate(f, 0.75, 1)$value

[1] 0.3554688
```

(d) Set the **cdf** equal to u from a uniform.

$$\begin{aligned} u &\overset{\text{set}}{=} F_X(x) \\ u &= \frac{4x^2}{3} - \frac{x^4}{3} \\ x^4 - 4x^2 + 3u &= 0 \end{aligned}$$

Let $y = x^2$.

$$\begin{aligned} y^2 - 4y + 3u &= 0 \\ \implies y &= \frac{4 \pm \sqrt{16 - 4(1)(3u)}}{2} \\ y &= 2 \pm \sqrt{4 - 3u} \end{aligned}$$

Since x must fall between 0 and 1, only $2 - \sqrt{4-3u}$ is a viable solution for y. This means

$$x = \sqrt{2 - \sqrt{4 - 3u}}$$

```
> set.seed(13)
> n <- 1e+05
> u <- runif(n, 0, 1)
> x <- sqrt(2 - sqrt(4 - 3 * u))
> ans <- c(mean(x), var(x))
> names(ans) <- c("Mean", "Variance")
> ans

      Mean   Variance
0.62316807 0.05768694
```

The mean and variance of the random sample are 0.6232 and 0.0577, respectively.

(e) Mean:

$$\begin{aligned} E[X] &= \int_0^1 x \cdot \frac{4}{3}x(2-x^2)\,dx \\ &= \int_0^1 \frac{8x^2}{3} - \frac{4x^4}{3}\,dx \\ &= \frac{8x^3}{9} - \frac{4x^4}{15}\Big|_0^1 \\ &= \frac{8}{9} - \frac{4}{15} - 0 \\ &= \frac{28}{45} = 0.6222 \end{aligned}$$

Variance:

$$\begin{aligned} E[X^2] &= \int_0^1 x^2 \cdot \frac{4}{3}x(2-x^2)\,dx \\ &= \int_0^1 \frac{8x^3}{3} - \frac{4x^5}{3}\,dx \\ &= \frac{2x^4}{3} - \frac{2x^6}{9}\Big|_0^1 \\ &= \frac{2}{3} - \frac{2}{9} - 0 \\ &= \frac{4}{9} = 0.4444 \end{aligned}$$

$Var[X] = E[X^2] - (E[X])^2 = \frac{4}{9} - \left(\frac{28}{45}\right)^2 = \frac{116}{2025} = 0.0573$

(f) The simulated mean (0.6232) and the simulated variance (0.0577) are both within 1% of their theoretical values.

```
> c(abs(ans[1] - 28/45)/(28/45) * 100, abs(ans[2] - 116/2025)/(116/2025) *
+     100)

     Mean  Variance
0.1520114 0.7034922
```

Solution for 33:

(a) For $x, \theta > 0$, $\int_0^\infty 3\pi\theta x^2 e^{-\theta\pi x^3}\, dx = -e^{-\theta\pi x^3}\Big|_0^\infty = 0 + 1 = 1$

(b) $F_X(x) = \int_0^x 3\pi\theta t^2 e^{-\theta\pi t^3}\, dt = -e^{-\theta\pi t^3}\Big|_0^x = 1 - e^{-\theta\pi x^3},\ x \geq 0,\ \theta > 0$

(c) $\mathbb{P}(X > 1|\theta = 5) = 1 - P(X \leq 1|\theta = 5) = 0$

```
> Fx <- function(x) {
+     1 - exp(-5 * pi * x^3)
+ }
> 1 - Fx(1)

[1] 1.507017e-07

> # Or
> f <- function(x, theta = 5) {
+     3 * pi * theta * x^2 * exp(-theta * pi * x^3)
+ }
> integrate(f, 1, Inf)$value

[1] 1.507017e-07
```

(d) Set $u = F_X(x)$, and solve for x.

$$\begin{aligned} u &\overset{\text{set}}{=} F_X(x) \\ u &= 1 - e^{-\theta\pi x^3} \\ e^{-\theta\pi x^3} &= 1 - u \\ -\theta\pi x^3 &= \ln(1-u) \\ x^3 &= \frac{\ln(1-u)}{-\theta\pi} \\ x &= \left(\frac{\ln(1-u)}{-\theta\pi}\right)^{1/3} \end{aligned}$$

Use $\theta = 5$ for the simulation.

```
> set.seed(201)
> n <- 100000
> u <- runif(n, 0, 1)
> x <- (log(1 - u)/(-5*pi))^(1/3)
> ans <- c(mean(x), var(x))
> names(ans) <- c("Mean", "Variance")
> ans

      Mean   Variance
0.35618738 0.01687483
```

The mean and the variance of the random sample are 0.3562 and 0.0169, respectively.

(e) Theoretical mean and variance given that $\theta = 5$ follow:

Mean:

$$E[X] = \int_0^\infty x \cdot 3\pi\theta x^2 e^{-\theta\pi x^3}\, dx$$
$$= \int_0^\infty 3\pi\theta x^3 e^{-\theta\pi x^3}\, dx$$

Recall $\int_0^\infty x^{\alpha-1}e^{-x}\, dx = \Gamma(\alpha)$ and use the substitution $u = \theta\pi x^3$.

Note that $du = 3\theta\pi x^2\, dx$ and $x = \left(\frac{u}{\theta\pi}\right)^{1/3}$.

$$= \int_0^\infty \left(\frac{u}{\theta\pi}\right)^{1/3} e^{-u}\, du$$
$$= \left(\frac{1}{\theta\pi}\right)^{1/3} \Gamma\left(\frac{4}{3}\right)$$

So, $E[X|\theta = 5] = \left(\frac{1}{5\pi}\right)^{1/3} \Gamma\left(\frac{4}{3}\right) = 0.3565618456$.
Variance:

$$E[X^2] = \int_0^\infty x^2 \cdot 3\pi\theta x^2 e^{-\theta\pi x^3}\, dx$$
$$= \int_0^\infty 3\pi\theta x^4 e^{-\theta\pi x^3}\, dx$$

Recall $\int_0^\infty x^{\alpha-1}e^{-x}\, dx = \Gamma(\alpha)$ and use the substitution $u = \theta\pi x^3$.

Note that $du = 3\theta\pi x^2\, dx$ and $x = \left(\frac{u}{\theta\pi}\right)^{1/3}$.

$$= \int_0^\infty \left(\frac{u}{\theta\pi}\right)^{2/3} e^{-u}\, du$$
$$= \left(\frac{1}{\theta\pi}\right)^{2/3} \Gamma\left(\frac{5}{3}\right)$$

$Var[X] = E[X^2] - (E[X])^2 = \frac{\Gamma(5/3)}{(\theta\pi)^{2/3}} - \left(\frac{\Gamma(4/3)}{(\theta\pi)^{1/3}}\right)^2$
So, $Var[X|\theta = 5] = \frac{\Gamma(5/3) - \Gamma(4/3)^2}{(5\pi)^{2/3}} = 0.0167938676$

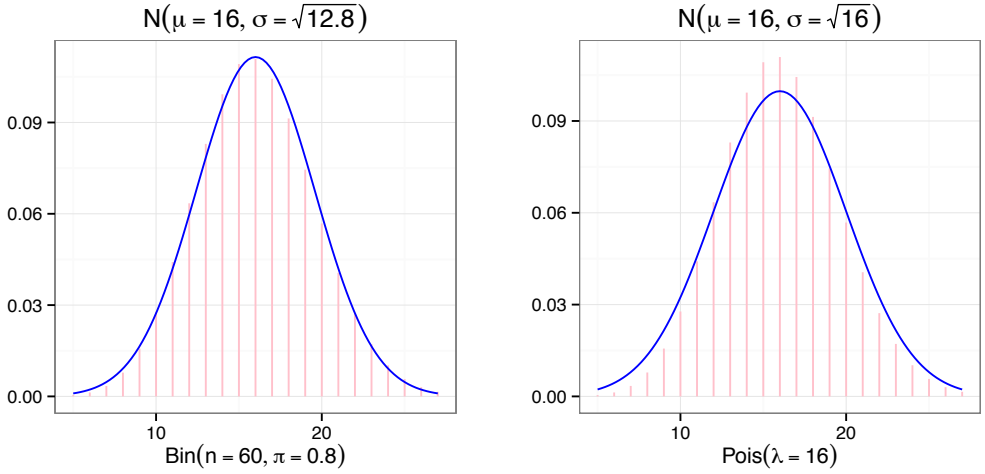
(f) The simulated mean (0.3562) and the simulated variance (0.0169) are both within 1% of their theoretical means.

```
> EX <- (1/(5*pi))^(1/3)*gamma(4/3)
> VX <- (gamma(5/3) - gamma(4/3)^2)/(5*pi)^(2/3)
> c(EX, VX)

[1] 0.35656185 0.01679387

> c(abs(ans[1] - EX)/EX*100, abs(ans[2] - VX)/VX*100)

     Mean  Variance
0.1050226 0.4820777
```

Solution for 35:

```
> n <- 80
> PI <- 0.20
> Lambda <- 16
> k <- 0.5  # correction factor
> x <- 5:27
> pxBINa <- pbinom(x, n, PI)
> pxNORa <- pnorm(x + k, n*PI, sqrt(n*PI*(1 - PI)))
> round(rbind(pxBINa, pxNORa), 3)

        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
pxBINa 0.001 0.002 0.005 0.013 0.029 0.056 0.101 0.164 0.247 0.346 0.455
pxNORa 0.002 0.004 0.009 0.018 0.035 0.062 0.104 0.164 0.242 0.338 0.444
       [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
pxBINa 0.566 0.671 0.762 0.837 0.893 0.934 0.961 0.978 0.989 0.994 0.997
pxNORa 0.556 0.662 0.758 0.836 0.896 0.938 0.965 0.982 0.991 0.996 0.998
       [,23]
pxBINa 0.999
pxNORa 0.999

> pxPOIb <- ppois(x, Lambda)
> pxNORb <- pnorm(x + k, Lambda, sqrt(Lambda))
> round(rbind(pxPOIb, pxNORb), 3)

        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11]
pxPOIb 0.001 0.004 0.010 0.022 0.043 0.077 0.127 0.193 0.275 0.368 0.467
pxNORb 0.004 0.009 0.017 0.030 0.052 0.085 0.130 0.191 0.266 0.354 0.450
       [,12] [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22]
pxPOIb 0.566 0.659 0.742 0.812 0.868 0.911 0.942 0.963 0.978 0.987 0.993
pxNORb 0.550 0.646 0.734 0.809 0.870 0.915 0.948 0.970 0.983 0.991 0.996
       [,23]
pxPOIb 0.996
pxNORb 0.998
```

```
> DF1 <- data.frame(x = x, y = dbinom(x, n, PI))
> ggplot(data = DF1, aes(x = x, y = y)) +
+ geom_linerange(aes(x = x, ymin = 0, ymax = y), color = "pink") +
+ stat_function(fun = dnorm, arg = list(n*PI, sqrt(n*PI*(1 - PI))),
+ color = "blue") +
+ labs(x = expression(Bin(n == 60, pi == 0.80)), y = "",
+ title = expression(N(mu == 16, sigma == sqrt(12.8)))) +
+ theme_bw()
> DF2 <- data.frame(x = x, y = dpois(x, Lambda))
> ggplot(data = DF1, aes(x = x, y = y)) +
+ geom_linerange(aes(x = x, ymin = 0, ymax = y), color = "pink") +
+ stat_function(fun = dnorm, arg = list(Lambda, sqrt(Lambda)),
+ color = "blue") +
+ labs(x = expression(Pois(lambda == 16)), y = "",
+ title = expression(N(mu == 16, sigma == sqrt(16)))) +
+ theme_bw()
```



Solution for 37:

Let X = number of cups where milk has been added before tea. Then $X \sim Hyper(4, 4, 4)$, and $\mathbb{P}(X = 3) = 0.2286$ or $\mathbb{P}(X = 3) = \binom{4}{3} \cdot \binom{4}{1} / \binom{8}{4} = 0.2286$.

```
> dhyper(3, 4, 4, 4)

[1] 0.2285714

> choose(4, 3)*choose(4, 1)/choose(8, 4)

[1] 0.2285714
```

Solution for 39:

(a) For $X \sim Weib(\alpha, \beta)$, $f(x) = \alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^\alpha}$ if $x \geq 0$, and $f(x) = 0$ if $x < 0$.

$$\begin{aligned} F_X(x) &= \int_0^x \alpha\beta^{-\alpha}t^{\alpha-1}e^{-(t/\beta)^\alpha}\,dt \\ &= -e^{-(t/\beta)^\alpha}\Big|_0^x \\ &= 1 - e^{-(x/\beta)^\alpha} \text{ for } x \geq 0 \end{aligned}$$

(b)

$$\begin{aligned} h(t) &= \frac{f(t)}{1 - F(t)} \\ &= \frac{\alpha\beta^{-\alpha}t^{\alpha-1}e^{-(t/\beta)^\alpha}}{1 - (1 - e^{-(t/\beta)^\alpha})} \\ &= \frac{\alpha\beta^{-\alpha}t^{\alpha-1}e^{-(t/\beta)^\alpha}}{e^{-(t/\beta)^\alpha}} \\ &= \alpha\beta^{-\alpha}t^{\alpha-1} = \frac{\alpha t^{\alpha-1}}{\beta^\alpha} \end{aligned}$$

■

Solution for 41:

(a)

$$E[V] = \int_0^\infty v \cdot \sqrt{\frac{2}{\pi}} \left(\frac{M}{RT}\right)^{\frac{3}{2}} v^2 e^{-\frac{Mv^2}{2RT}}\, dv$$

Let $y^2 = \frac{Mv^2}{2RT} \Rightarrow v = y\sqrt{\frac{2RT}{M}}$ and $dv = \sqrt{\frac{2RT}{M}}\, dy$

$$\begin{aligned}
&= \int_0^\infty \sqrt{\frac{2}{\pi}} \left(\frac{M}{RT}\right)^{\frac{3}{2}} \left(y\sqrt{\frac{2RT}{M}}\right)^3 e^{-y^2} \sqrt{\frac{2RT}{M}}\, dy \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{M}{RT}\right)^{\frac{3}{2}} \left(\frac{2RT}{M}\right)^{\frac{3}{2}} \left(\frac{2RT}{M}\right)^{\frac{1}{2}} \int_0^\infty y^3 e^{-y^2}\, dy \\
&= \sqrt{\frac{2}{\pi}} \left(\frac{RT}{M}\right)^{-\frac{3}{2}} \left(\frac{2RT}{M}\right)^{2} \cdot \frac{1}{2}\Gamma\left(\frac{3+1}{2}\right) \\
&= 4\sqrt{\frac{2}{\pi}} \left(\frac{RT}{M}\right)^{\frac{1}{2}} \cdot \frac{1}{2}\Gamma(2) \\
E[V] &= 2\sqrt{\frac{2RT}{\pi M}}
\end{aligned}$$

(b) Finding units for $2\sqrt{\frac{2RT}{\pi M}}$:

R is ${}^{\mathrm{J}}/_{\mathrm{mol}\cdot\mathrm{K}}$; T is in K; M is in $\frac{\mathrm{kg}}{\mathrm{mol}}$ and a J is ${}^{\mathrm{kg}\cdot\mathrm{m}^2}/_{\mathrm{s}^2}$.

$$\text{Units ONLY of } \sqrt{\frac{2RT}{\pi M}} \text{ are } \sqrt{\frac{\frac{\mathrm{J}}{\mathrm{mol}\cdot\mathrm{K}} \cdot \mathrm{K}}{\frac{\mathrm{kg}}{\mathrm{mol}}}} = \sqrt{\frac{\frac{\mathrm{kg}\cdot\mathrm{m}^2}{\mathrm{s}^2}}{\mathrm{kg}}} = \sqrt{\frac{\mathrm{m}^2}{\mathrm{s}^2}} = \frac{\mathrm{m}}{\mathrm{s}}$$

(c)

$$\begin{aligned}
E[E_k] &= \int_0^\infty \frac{Mv^2}{2} \cdot \sqrt{\frac{2}{\pi}} \left(\frac{M}{RT}\right)^{\frac{3}{2}} v^2 e^{-\frac{Mv^2}{2RT}}\, dv \\
&= \frac{M}{2} \left(\frac{M}{RT}\right)^{\frac{3}{2}} \sqrt{\frac{2}{\pi}} \int_0^\infty v^4 e^{-\frac{Mv^2}{2RT}}\, dv
\end{aligned}$$

Let $y^2 = \frac{Mv^2}{2RT} \Rightarrow v = y\sqrt{\frac{2RT}{M}}$ and $dv = \sqrt{\frac{2RT}{M}}\, dy$

$$\begin{aligned}
&= \frac{M}{2} \left(\frac{M}{RT}\right)^{\frac{3}{2}} \sqrt{\frac{2}{\pi}} \int_0^\infty \left(\sqrt{\frac{2RT}{M}}\right)^4 y^4 e^{-y^2} \sqrt{\frac{2RT}{M}}\, dy \\
&= \frac{4RT}{\sqrt{\pi}} \cdot \frac{1}{2}\, \Gamma\left(\frac{4+1}{2}\right) \\
&= \frac{2RT}{\sqrt{\pi}} \cdot \frac{3}{2} \cdot \frac{\sqrt{\pi}}{2} \\
&= \frac{3RT}{2}
\end{aligned}$$

(d)

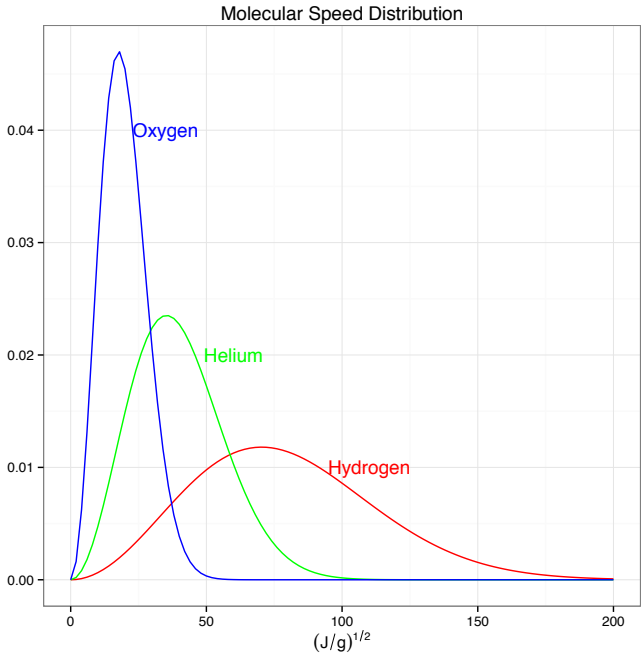
$$\begin{aligned} E[V] &= 2\sqrt{\frac{2RT}{\pi M}} \\ &= 2\sqrt{\frac{2(8.3145\frac{\text{J}}{\text{mol}\cdot\text{K}})\cdot 300\text{K}}{\pi\cdot 1.008\frac{\text{g}}{\text{mol}}}} \\ &= 2\sqrt{\frac{2(8.3145)\cdot 300}{\pi\cdot 1.008}\cdot\frac{\text{J}}{\text{g}}} \\ &= 2\sqrt{\frac{2(8.3145)\cdot 300}{\pi\cdot 1.008}\cdot\frac{1000\text{g}\cdot\text{m}^2}{\text{s}^2\cdot\text{g}}} \\ &= 2510.259\frac{\text{m}}{\text{s}} \end{aligned}$$

(e) Numerical integration for $E[V]$ gives an answer in $(\text{J/g})^{1/2}$. To convert to m/s, multiply by $\sqrt{1000}$.

```
> M <- 1.008
> R <- 8.3145
> Temp <- 300
> fv <- function(x, M=1.008)
+ {
+ sqrt(2/pi)*(M/(R*Temp))^(3/2)*x^3*exp(-(M*x^2)/(2*R*Temp))
+ }
> ans <- integrate(fv, 0, Inf)$value
> ans

[1] 79.38135

> ans*sqrt(1000)

[1] 2510.259
```

(f)

```
> M <- 1.008
> R <- 8.3145
> Temp <- 300
> f <- function(x, M = 1.008)
+ {
+ sqrt(2/pi)*(M/(R*Temp))^(3/2)*x^2*exp(-(M*x^2)/(2*R*Temp))
+ }
> p <- ggplot(data = data.frame(x = c(0, 200)), aes(x = x))
> p + stat_function(fun = f, args = list(M = 1.008), color = "red") +
+ stat_function(fun = f, args = list(M = 4.003), color = "green") +
+ stat_function(fun = f, args = list(M = 16), color = "blue") +
+ theme_bw() +
+ labs(x = expression((J/g)^{1/2}), y = "",
+ title = "Molecular Speed Distribution") +
+ annotate("text", x = 35, y = 0.04, label = "Oxygen", color = "blue") +
+ annotate("text", x = 60, y = 0.02, label = "Helium", color = "green") +
+ annotate("text", x = 110, y = 0.01, label = "Hydrogen", color = "red")
```



Solution for 43:

Let A = friend one buys the first ticket and wins, then $\mathbb{P}(A) = \frac{2}{n}$.

Let B = friend two buys the first ticket after the first prize is awarded and wins.

Let B_i = friend two buys the first ticket after the first prize is awarded and wins with the i^{th} ticket where $i > 1$.

Note that $\mathbb{P}(B) = \sum_{i=2}^{n} \mathbb{P}(B_i)$.

Now

$$\begin{aligned}
\mathbb{P}(B_2) &= \frac{2}{n} \cdot \frac{1}{n-1} \\
\mathbb{P}(B_3) &= \frac{n-2}{n} \cdot \frac{2}{n-1} \cdot \frac{1}{n-2} = \frac{2}{n} \cdot \frac{1}{n-1} \\
\mathbb{P}(B_4) &= \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdot \frac{2}{n-2} \cdot \frac{1}{n-3} = \frac{2}{n} \cdot \frac{1}{n-1} \\
\mathbb{P}(B_5) &= \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdot \frac{n-4}{n-2} \cdot \frac{2}{n-3} \cdot \frac{1}{n-4} = \frac{2}{n} \cdot \frac{1}{n-1} \\
\mathbb{P}(B_6) &= \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdot \frac{n-4}{n-2} \cdot \frac{n-5}{n-3} \cdot \frac{2}{n-4} \cdot \frac{1}{n-5} = \frac{2}{n} \cdot \frac{1}{n-1} \\
&\vdots \\
&\vdots \\
\mathbb{P}(B_i) &= \frac{n-2}{n} \cdot \frac{n-3}{n-1} \cdots \frac{n-(i-1)}{n-(i-3)} \cdot \frac{2}{n-(i-2)} \cdot \frac{1}{n-(i-1)} = \frac{2}{n} \cdot \frac{1}{n-1}.
\end{aligned}$$

Thus

$$\mathbb{P}(B) = \sum_{i=2}^{n} P(B_i) = \sum_{i=2}^{n} \frac{2}{n} \cdot \frac{1}{n-1} = (n-1) \cdot \frac{2}{n} \cdot \frac{1}{n-1} = \frac{2}{n}.$$

$\mathbb{P}(B) = \mathbb{P}(A) = \frac{2}{n}$

Let X = the number of winning tickets purchase by friend one. $X \sim Bin(m, \pi = 2/n)$

$$\begin{aligned}\mathbb{P}(X > 1) &= 1 - \mathbb{P}(X \leq 1) = 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) = \\ &= 1 - \binom{m}{0}\left(\frac{2}{n}\right)^0\left(\frac{n-2}{n}\right)^m - \binom{m}{1}\left(\frac{2}{n}\right)^1\left(\frac{n-2}{n}\right)^{m-1} \\ &= 1 - \left(\frac{n-2}{n}\right)^m - m\left(\frac{2}{n}\right)\left(\frac{n-2}{n}\right)^{m-1} \\ &= 1 - \frac{(n-2)^{m-1}(n+2m-2)}{n^m}.\end{aligned}$$

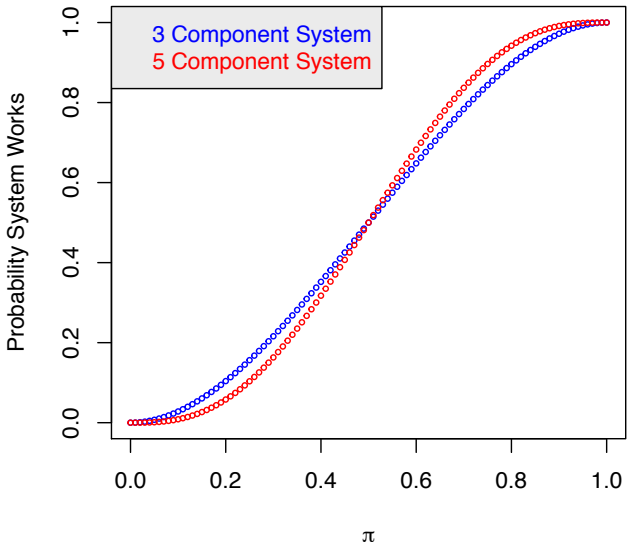
Since $\mathbb{P}(A) = \mathbb{P}(B)$, the probability that each wins more than one prize is

$$2 \cdot \left[1 - \frac{(n-2)^{m-1}(n+2m-2)}{n^m}\right].$$

Solution for 45:

For any value of π less than 0.5 the system with 3 components has a higher probability of functioning. For any value of π greater than 0.5, the system with 5 components has a higher probability of functioning.

```
> p <- seq(from = 0, to = 1, by = 0.01)
> p3 <- 1 - pbinom(1, 3, p)
> p5 <- 1 - pbinom(2, 5, p)
> plot(p, p3, col = "blue", cex = 0.5, ylab = "Probability System Works",
+      xlab = expression(pi))
> points(p, p5, col = "red", cex = 0.5)
> legend(x = "topleft", legend = c("3 Component System",
+                                  "5 Component System"),
+        text.col = c("blue", "red"), bg = "gray92")
```


3 Component System
5 Component System
Probability System Works
0.0
0.2
0.4
0.6
0.8
1.0
π

Chapter 5

Odd solutions

Solution for 1:

(a) $Cov[X,Y] = E[XY] - E[X] \cdot E[Y] = 0 - 0 \cdot \frac{-2}{6} = 0$

$$
\begin{aligned}
E[X] &= -1 \cdot \frac{2}{6} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{2}{6} = 0 \\
E[Y] &= -1 \cdot \frac{4}{6} + 0(0) + 1 \cdot \frac{2}{6} = \frac{-2}{6} \\
E[XY] &= (-1)(-1) \cdot \frac{1}{6} + (-1)(0)(0) + (-1)(1) \cdot \frac{1}{6} + (0)(-1) \cdot \frac{1}{3} \\
&\quad + (0)(0)(0) + (0)(1)(0) + (1)(-1) \cdot \frac{1}{6} + (1)(0)(0) + (1)(1) \cdot \frac{1}{6} \\
&= \frac{1}{6} - \frac{1}{6} - \frac{1}{6} + \frac{1}{6} = 0
\end{aligned}
$$

(b)

$$\mathbb{P}(X = -1 | Y = 1) = \frac{\mathbb{P}(X = -1, Y = 1)}{\mathbb{P}(Y = 1)} = \frac{1/6}{2/6} = \frac{1}{2}.$$

(c) X and Y must be dependent because

$$\mathbb{P}(X = -1, Y = -1) = \frac{1}{6} \neq \mathbb{P}(X = -1) \cdot \mathbb{P}(Y = -1) = \frac{2}{6} \cdot \frac{4}{6} = \frac{2}{9}.$$

Solution for 3:

(a)

		W			
		0	1	2	
Z	0	$4/9$	$2/9$	0	$6/9$
	1	0	$2/9$	$1/9$	$3/9$
		$4/9$	$4/9$	$1/9$	1

(b) $Cov[Z,W] = E[ZW] - E[Z] \cdot E[W] = \frac{4}{9} - \frac{1}{3} \cdot \frac{2}{3} = \frac{4-2}{9} = \frac{2}{9}$

$$
\begin{aligned}
E[Z] &= 0 \cdot \frac{6}{9} + 1 \cdot \frac{3}{9} = \frac{3}{9} = \frac{1}{3} \\
E[W] &= 0 \cdot \frac{4}{9} + 1 \cdot \frac{4}{9} + 2 \cdot \frac{1}{9} = \frac{6}{9} = \frac{2}{3} \\
E[ZW] &= (0)(0) \cdot \frac{4}{9} + (0)(1) \cdot \frac{2}{9} + (0)(2) \cdot 0 + (1)(0) \cdot 0 + (1)(1) \cdot \frac{2}{9} + (1)(2) \cdot \frac{1}{9} = \frac{4}{9}
\end{aligned}
$$

Note that Z and W are not independent because $Cov[Z, W] = \frac{2}{9} \neq 0$.
(c) $Cov[Z, W] = 0 \not\Longrightarrow$ independence.

Solution for 5:
(a) Note that $Z = {}^{X}/_{Y}$. Consequently, the probability distribution of Z is

z	18	24	36	90	120	180	270	360	540
$p(z)$	0.0025	0.05	0.3	0.15	0.15	0.025	0.1	0.2	0

(b)

$$\begin{aligned} E[Z] &= \sum_z z \cdot p(z) \\ &= (18)(0.0025) + (24)(0.05) + (36)(0.3) + (90)(0.15) + (120)(0.15) \\ &\quad + (180)(0.025) + (270)(0.1) + (360)(0.2) + (540)(0) = 147.45 \end{aligned}$$

```
> z <- c(18, 24, 36, 90, 120, 180, 270, 360, 540)
> pz <- c(0.025, 0.05, 0.3, 0.15, 0.15, 0.025, 0.1, 0.2, 0)
> EZ <- sum(z*pz)
> EZ

[1] 147.45
```

(c)

$$\begin{aligned} E[X] &= 64800(0.375) + 324000(0.325) + 972000(0.3) = 421,200 \\ E[X^2] &= 64800^2(0.375) + 324000^2(0.325) + 972000^2(0.3) = 319,127,040,000 \\ E[Y] &= 1800(0.325) + 2700(0.4) + 3600(0.275) = 2655 \\ E[Y^2] &= 1800^2(0.325) + 2700^2(0.4) + 3600^2(0.275) = 7,533,000 \\ Var[X] &= E[X^2] - (E[X])^2 = 319,127,040,000 - 421,200^2 = 141,717,600,000 \\ Var[Y] &= E[Y^2] - (E[Y])^2 = 7,533,000 - 2655^2 = 483975 \end{aligned}$$

$$\begin{aligned} E[XY] &= 1,245,132,000 \\ Cov[X, Y] &= E[XY] - E[X] \cdot E[Y] = 126,846,000 \end{aligned}$$

Note that the positive covariance indicates internet users transmitting more information get a faster server.

```
> x <- c(64800, 324000, 972000)
> px <- c(0.375, 0.325, 0.3)
> EX <- sum(x*px)
> EX2 <- sum(x^2*px)
> y <- c(1800, 2700, 3600)
> py <- c(0.325, 0.4, 0.275)
> EY <- sum(y*py)
> EY2 <- sum(y^2*py)
> VX <- EX2 - EX^2
> VY <- EY2 - EY^2
> c(EX, EX2, VX)
```

```
[1]       421200 319127040000 141717600000

> c(EY, EY2, VY)

[1]    2655 7533000  483975

> EXY <- 64800*1800*0.3 + 64800*2700*0.05 + 64800*3600*0.025 +
+        324000*1800*0.025 + 324000*2700*0.15 + 324000*3600*0.15 +
+        972000*1800*0 + 972000*2700*0.2 + 972000*3600*0.1
> CXY <- EXY - EX*EY
> c(EXY, CXY)

[1] 1245132000  126846000
```

Solution for 7:

Let X = interior diameter of a test tube and Y = thickness of a test tube. Then Z = exterior diameter is $X + 2Y$.

$$\begin{aligned} E[Z] &= E[X] + 2E[Y] = 5 + 2(0.5) = 6 \text{ cm} \\ Var[Z] &= Var[X] + 4\,Var[Y] = 0.03^2 + 4(0.001)^2 = 0.000904 \\ \sigma_Z &= 0.03006659 \text{ cm} \end{aligned}$$

```
> VZ <- 0.03^2 + 4*0.001^2
> SZ <- sqrt(VZ)
> c(VZ, SZ)

[1] 0.00090400 0.03006659
```

Solution for 9:

Let X = the amount of flour put into one multiple of the recipe, and let Y = the amount of milk put into one multiple of the recipe. Let TA = the total amount of flour and milk in the recipe ($TA = X + Y$).

(a)

$$\begin{aligned} E[4TA] &= E[4X + 4Y] = 4 \times 1 + 4 \times {}^3/_4 = 7 \\ \sigma_{4TA} &= \sqrt{4^2 \times ({}^1/_{16})^2 + 4^2 \times ({}^1/_8)^2} = 0.3125 \end{aligned}$$

(b) $\mathbb{P}(4TA > 7.5) = 0.1855$.

```
> ans <- pnorm(7.5, 7, sqrt(4^2*(1/16)^2 + 4^2*(1/8)^2), lower = FALSE)
> ans

[1] 0.1855467
```

Solution for 11:

Let X = Emily's monthly sales and Y = Albert's monthly sales, then total sales, $TS = X + Y \sim N(3000 + 2000, \sqrt{500^2 + 1000^2 + 2 \times 10000})$.

$$\mathbb{P}(0.30 \times TS \leq 1350) = \mathbb{P}(TS \leq {}^{1350}/_{0.30} = 4500) = 0.3286.$$

```
> ans <- pnorm(1350/0.30, 5000, sqrt(500^2 + 1000^2 + 2*10000))
> ans

[1] 0.3286376
```

32.86% of the time Emily and Albert will be able to spend less than 30% of their total income on their house payment.

Solution for 13:

$$\begin{aligned}\rho_{X,Y} &= \frac{Cov[X,Y]}{\sigma_X\sigma_Y} = \frac{E[XY]-E[X]E[Y]}{\sigma_X\sigma_Y} \\ &= \frac{E[X(aX+b)]-E[X][aE[X]+b]}{\sigma_X|a|\sigma_X} = \frac{aE[X^2]-aE[X]^2}{|a|\sigma_X^2} \\ &= \frac{aCov[X,X]}{|a|\sigma_X^2} = \frac{a\sigma_X^2}{|a|\sigma_X^2} = \frac{a}{|a|}\end{aligned}$$

If $a > 0$, then $\rho_{X,Y} = \frac{a}{|a|} = 1$. If $a < 0$, then $\rho_{X,Y} = \frac{a}{|a|} = -1$.

Solution for 15:

(a)

$$p_{X,Y}(x,y) = \frac{\binom{4}{x}\binom{4}{y}\binom{44}{5-x-y}}{\binom{52}{5}} \text{ for } x = 0,1,2,3,4;\ y = 0,1,2,3,4;\ 0 \le x+y \le 5.$$

(b)

$$p_X(x) = \sum_y f(x,y) = \frac{\binom{4}{x}}{\binom{52}{5}} \underbrace{\sum_y \binom{4}{y}\binom{44}{5-x-y}}_{\binom{48}{5-x}} = \frac{\binom{4}{x}\binom{48}{5-x}}{\binom{52}{5}}$$

(c)

$$p_Y(y) = \sum_y f(x,y) = \frac{\binom{4}{y}}{\binom{52}{5}} \underbrace{\sum_x \binom{4}{x}\binom{44}{5-x-y}}_{\binom{48}{5-y}} = \frac{\binom{4}{y}\binom{48}{5-y}}{\binom{52}{5}}$$

Solution for 17:

$$\mathbb{P}\left(X+3>Y \,\middle|\, X>\frac{1}{3}\right) = \frac{\mathbb{P}\left(X+3>Y,\ X>\frac{1}{3}\right)}{\mathbb{P}\left(X>\frac{1}{3}\right)}$$

$$\begin{aligned}
\mathbb{P}\left(X+3>Y,\, X>\frac{1}{3}\right) &= \int_{\frac{1}{3}}^{\infty}\int_{0}^{x+3} e^{-(x+y)}\,dy\,dx \\
&= \int_{\frac{1}{3}}^{\infty} -e^{-(x+y)}\Big|_{0}^{x+3}\,dx \\
&= \int_{\frac{1}{3}}^{\infty} -e^{-(2x+3)}+e^{-x}\,dx \\
&= \frac{e^{-(2x+3)}}{2}-e^{-x}\Big|_{\frac{1}{3}}^{\infty} \\
&= -\frac{e^{-11/3}}{2}+e^{-1/3}
\end{aligned}$$

$f_X(x) = \int_0^{\infty} e^{-(x+y)}\,dy = e^{-x}$

$$\begin{aligned}
\mathbb{P}\left(X>\frac{1}{3}\right) &= \int_{\frac{1}{3}}^{\infty} e^{-x}\,dx \\
&= -e^{-x}\Big|_{\frac{1}{3}}^{\infty} \\
&= e^{-1/3}
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}\left(X+3>Y \,\middle|\, X>\frac{1}{3}\right) &= \frac{\mathbb{P}\left(X+3>Y,\, X>\frac{1}{3}\right)}{\mathbb{P}\left(X>\frac{1}{3}\right)} \\
&= \frac{-\frac{e^{-11/3}}{2}+e^{-1/3}}{e^{-1/3}} \\
&= 1-\frac{1}{2}e^{-10/3} = 0.9822
\end{aligned}$$

Solution for 19:

k must have a value so that $\int_0^1\int_0^1 f(x,y)\,dx\,dy = 1$

$$\begin{aligned}
\int_0^1 \int_0^1 f(x,y)\,dx\,dy &= k\int_0^1 \int_{x^2}^1 (y-2x)\,dy\,dx \\
&= k\int_0^1 \left(\frac{y^2}{2} - 2xy\right)\Big|_{x^2}^1 dx \\
&= k\int_0^1 \left(\frac{1}{2} - 2x\right) - \left(\frac{x^4}{2} - 2x^3\right) dx \\
&= k\left[\left(\frac{x}{2} - x^2\right) - \left(\frac{x^5}{10} - \frac{x^4}{2}\right)\right]_0^1 \\
&= k\left(\frac{1}{2} - 1 - \frac{1}{10} + \frac{1}{2}\right) \\
&= k\left(-\frac{1}{10}\right) \\
\implies k &= -10
\end{aligned}$$

Solution for 21:

(a)

$$\begin{aligned}
\mathbb{P}\left(X < \frac{1}{2} \,\middle|\, Y < \frac{1}{4}\right) &= \frac{\mathbb{P}\left(X < \frac{1}{2}, Y < \frac{1}{4}\right)}{\mathbb{P}\left(Y < \frac{1}{4}\right)} \\
&= \frac{\int_0^{\frac{1}{2}} \int_0^{\frac{1}{4}} 6(x-y)^2\,dy\,dx}{\int_0^1 \int_0^{\frac{1}{4}} 6(x-y)^2\,dy\,dx} \\
&= \frac{\int_0^{\frac{1}{2}} \frac{(x-y)^3}{-3}\Big|_0^{\frac{1}{4}} dx}{\int_0^1 \frac{(x-y)^3}{-3}\Big|_0^{\frac{1}{4}} dx} \\
&= \frac{\int_0^{\frac{1}{2}} \frac{x^3}{3} - \frac{(x-\frac{1}{4})^3}{3}\,dx}{\int_0^1 \frac{x^3}{3} - \frac{(x-\frac{1}{4})^3}{3}\,dx} \\
&= \frac{x^4 - (x-\frac{1}{4})^4\big|_0^{\frac{1}{2}}}{x^4 - (x-\frac{1}{4})^4\big|_0^1} \\
&= \frac{\left(\frac{1}{2}\right)^4 - (\frac{1}{2}-\frac{1}{4})^4 - \left[0^4 - (0-\frac{1}{4})^4\right]}{1^4 - (1-\frac{1}{4})^4 - \left[0^4 - (0-\frac{1}{4})^4\right]} \\
&= \frac{\frac{1}{16} - \frac{1}{256} + \frac{1}{256}}{1 - \frac{81}{256} + \frac{1}{256}} = \frac{16}{176} = \frac{1}{11}
\end{aligned}$$

(b)

$$\begin{aligned}
\mathbb{P}\left(X < \frac{1}{2} \,\middle|\, Y = \frac{1}{4}\right) &= \frac{\int_0^{1/2} 6\left(x - \frac{1}{4}\right)^2 dx}{\int_0^1 6\left(x - \frac{1}{4}\right)^2} dx \\
&= \frac{2\left(x - \frac{1}{4}\right)^3\Big|_0^{1/2}}{2\left(x - \frac{1}{4}\right)^3\Big|_0^1} \\
&= \frac{2\left[\left(\frac{1}{2} - \frac{1}{4}\right)^3 - \left(0 - \frac{1}{4}\right)^3\right]}{2\left[\left(1 - \frac{1}{4}\right)^3 - \left(0 - \frac{1}{4}\right)^3\right]} \\
&= \frac{2\left[\frac{1}{64} + \frac{1}{64}\right]}{2\left[\frac{27}{64} + \frac{1}{64}\right]} \\
&= \frac{\frac{2}{64}}{\frac{28}{64}} \\
&= \frac{1}{14}
\end{aligned}$$

Solution for 23:

Recall that $E[Y|X] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and $Var[Y|X] = \sigma^2_{Y|x} = \sigma^2_Y(1 - \rho^2)$.

(a) $E\left[\,Y \,|\, X = 170\right] = 400 + 0.9\frac{50}{10}(170 - 180) = 355$

```
> EYgX170 <- 400 + 0.9*50/10*(170 - 180)
> EYgX170

[1] 355
```

(b) $E\left[\,Y \,|\, X = 200\right] = 400 + 0.9\frac{50}{10}(200 - 180) = 490$

```
> EYgX200 <- 400 + 0.9*50/10*(200 - 180)
> EYgX200

[1] 490
```

(c) $Var\left[\,Y \,|\, X = 170\right] = 50^2(1 - 0.9^2) = 475$

```
> VYgX170 <-  50^2*(1 - 0.9^2)
> VYgX170

[1] 475
```

(d) $Var\left[\,Y \,|\, X = 200\right] = 50^2(1 - 0.9^2) = 475$

```
> VYgX200 <-  50^2*(1 - 0.9^2)
> VYgX200

[1] 475
```

(e) $\mathbb{P}\left(Y \leq 380 \,|\, X = 170\right) = 0.8743$

```
> pnorm(380, EYgX170, sqrt(VYgX170))

[1] 0.8743254
```

(f)$\mathbb{P}(Y \geq 450 \,|\, X = 200) = 0.9668$

```
> 1 - pnorm(450, EYgX200, sqrt(VYgX200))

[1] 0.9667713
```

Solution for 25:

(a) $Y = \mu + X - 20$

(b) $E[Y] = E[\mu + X - 20] = E[\mu] + E[X] - E[20] = \mu + 0 - 20 = \mu - 20$ and $Var[Y] = Var[\mu + X - 20] = Var[X] = 100$, since $X \sim N(0, 10)$ it follows that $Y \sim N(\mu - 20, 10)$.

(c)

$$\begin{aligned} P(Y \geq 400) &= 0.98 \\ 1 - P(Y \leq 400) &= 0.98 \\ P(Y \leq 400) &= 0.02 \\ P(Z \leq \big(400 - (\mu - 20)\big)/10) &= 0.02 \\ \implies (420 - \mu)/10 &= Z_{0.02} \\ \mu &= 420 - Z_{0.02} \times 10 = 440.5375 \end{aligned}$$

```
> mu <- 420 - qnorm(0.02)*10
> mu

[1] 440.5375
```

μ must be at least 440.5375 grams to be 98% confident that the tins contain at least 400 grams of peppers.

(d) $Y = \mu + X - W$, where $X \sim N(0, 10)$ and $W \sim N(20, 5)$.

$$\begin{aligned} E[Y] &= \mu + 0 - 20 = \mu - 20 \\ Var[Y] &= Var[X] + Var[W] = 100 + 25 = 125 \end{aligned}$$

$Y \sim N(\mu - 20, \sqrt{125}\,)$.

$$\begin{aligned} \mathbb{P}(Y \geq 400) &= 0.98 \\ 1 - \mathbb{P}(Y \leq 400) &= 0.98 \\ \mathbb{P}(Y \leq 400) &= 0.02 \\ \mathbb{P}\left(Z \leq \big(400 - (\mu - 20)\big)/\sqrt{125}\right) &= 0.02 \\ \implies (420 - \mu)/\sqrt{125} &= Z_{0.02} \\ \mu &= 420 - Z_{0.02} \times \sqrt{125} = 442.9616 \end{aligned}$$

```
> mu <- 420 - qnorm(0.02) * sqrt(125)
> mu

[1] 442.9616
```

μ must be at least 442.9616 grams to be 98% confident that the tins contain at least 400 grams of peppers.

Solution for 27: Note that

$$\int_0^\infty e^{-x}\,dx = -e^{-x}\big|_0^\infty = 1$$

and

$$\int_0^\infty xe^{-x}\,dx = -xe^{-x}\big|_0^\infty + \int_0^\infty e^{-x}\,dx = 1.$$

(a)

$$\begin{aligned}
1 &\stackrel{?}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_0^\infty\int_0^\infty \frac{1+x+y+cxy}{(c+3)}\exp(-(x+y))\,dx\,dy \\
&= \frac{1}{c+3}\left[\int_0^\infty\int_0^\infty \exp(-(x+y))\,dx\,dy + \int_0^\infty\int_0^\infty x\exp(-(x+y))\,dx\,dy\right. \\
&\qquad \left. + \int_0^\infty\int_0^\infty y\exp(-(x+y))\,dx\,dy + c\int_0^\infty\int_0^\infty xy\exp(-(x+y))\,dx\,dy\right] \\
&= \frac{1}{c+3}\left[\int_0^\infty e^{-y}\int_0^\infty e^{-x}\,dx\,dy + \int_0^\infty e^{-y}\int_0^\infty xe^{-x}\,dx\,dy\right. \\
&\qquad \left. + \int_0^\infty e^{-x}\int_0^\infty ye^{-y}\,dy\,dx + c\int_0^\infty ye^{-y}\int_0^\infty xe^{-x}\,dx\,dy\right] \\
&= \frac{1}{c+3}\left[1+1+1+c\right] \\
1 &= 1
\end{aligned}$$

(b)

$$\begin{aligned}
f_X(x) &= \int_0^\infty f_{X,Y}(x,y)\,dy \\
&= \int_0^\infty \frac{1+x+y+cxy}{(c+3)}\exp(-(x+y))\,dy \\
&= \frac{1}{c+3}\left[\int_0^\infty \exp(-(x+y))\,dy + \int_0^\infty x\exp(-(x+y))\,dy\right. \\
&\qquad \left. + \int_0^\infty y\exp(-(x+y))\,dy + \int_0^\infty cxy\exp(-(x+y))\,dy\right] \\
&= \frac{1}{c+3}\left[e^{-x} + xe^{-x} + e^{-x} + cxe^{-x}\right] \\
f_X(x) &= \frac{e^{-x}\left[2+x+cx\right]}{c+3} \text{ for } x \geq 0
\end{aligned}$$

(c) If X and Y are to be independent, $f_X(x)\cdot f_Y(y) = f_{X,Y}(x,y)$.

$$
\begin{aligned}
f_X(x) \cdot f_Y(y) &\overset{\text{set}}{=} f_{X,Y}(x,y) \\
\frac{e^{-x}\,[2+x+cx]}{c+3} \cdot \frac{e^{-y}\,[2+y+cy]}{c+3} &= \frac{1+x+y+cxy}{(c+3)} e^{-x} e^{-y} \\
[2+x+cx] \cdot [2+y+cy] &= (1+x+y+cxy)(c+3) \\
4+2y+2cy+2x+xy+cxy+2cx+cxy+c^2xy &= \\
c+cx+cy+c^2xy+3+3x+3y+3cxy & \\
1-y-x+xy &= c(1-x-y+xy) \\
c &= 1 \quad \forall (x,y).
\end{aligned}
$$

If $c = 1$, X and Y are independent.

Solution for 29:

(a)

$$
\begin{aligned}
1 &\overset{\text{set}}{=} \int_4^6 \int_2^4 Kxy\,dx\,dy \\
&= \int_4^6 \frac{Kx^2y}{2}\Big|_2^4 dy \\
&= \int_4^6 \frac{16Ky}{2} - \frac{4Ky}{2}\,dy \\
&= \int_4^6 6Ky\,dy \\
&= 3Ky^2|_4^6 \\
&= 3K(36-16) \\
1 &= 60K \\
\Longrightarrow K &= \frac{1}{60}
\end{aligned}
$$

```
> f <- function(x){x[1]*x[2]}
> K <- 1 / adaptIntegrate(f, lowerLimit = c(2, 4),
+                         upperLimit = c(4, 6))$integral #
> MASS::fractions(K)

[1] 1/60
```

(b)

$$
f_X(x) = \int_4^6 \frac{1}{60} xy\,dy = \frac{xy^2}{120}\Big|_4^6 = \frac{x}{120}[36-16] = \frac{x}{6} \text{ for } 2 \le x \le 4
$$

$$
f_Y(y) = \int_2^4 \frac{1}{60} xy\,dx = \frac{x^2y}{120}\Big|_2^4 = \frac{y}{120}[16-4] = \frac{y}{10} \text{ for } 4 \le y \le 6
$$

(c) X and Y are independent since $f_X(x) \cdot f_Y(y) = \frac{x}{6} \cdot \frac{y}{10} = \frac{xy}{60} = f_{X,Y}(x,y)$.

Solution for 31:

(a) Note that $-2 \le x \le 2$ and $1 \le y \le x^2$ implies that $1 \le x^2$, which means that $(x \le -1) \cup (x \ge 1)$ is implied in the ranges of the variables.

$$
\begin{aligned}
1 &\overset{\text{set}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-2}^{-1}\int_{1}^{x^2} Ky\,dy\,dx + \int_{1}^{2}\int_{1}^{x^2} Ky\,dy\,dx \\
&= K\left[\int_{-2}^{-1}\frac{x^4-1}{2}\,dx + \int_{1}^{2}\frac{x^4-1}{2}\,dx\right] \\
&= K\left[\left(\frac{x^5}{10}-\frac{x}{2}\right)\Big|_{-2}^{-1} + \left(\frac{x^5}{10}-\frac{x}{2}\right)\Big|_{1}^{2}\right] \\
&= K\left[\left(\frac{-1}{10}+\frac{1}{2}\right) - \left(\frac{-32}{10}+1\right) + \left(\frac{32}{10}-1\right) - \left(\frac{1}{10}-\frac{1}{2}\right)\right] \\
&= K\left[\frac{4}{10}+\frac{22}{10}+\frac{22}{10}+\frac{4}{10}\right] \\
1 &= K\left[\frac{26}{5}\right] \\
\implies K &= \frac{5}{26}
\end{aligned}
$$

(b)

$$
\begin{aligned}
f_X(x) &= \int_y f_{X,Y}(x,y)\,dy \\
&= \int_1^{x^2}\frac{5}{26}y\,dy \\
&= \frac{5}{52}y^2\Big|_1^{x^2} \\
&= \frac{5}{52}(x^4-1) \\
\implies f_X(x) &= \begin{cases} \frac{5}{52}(x^4-1), & (-2 \le x \le -1) \cup (1 \le x \le 2) \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

$$
\begin{aligned}
f_Y(y) &= \int_x f_{X,Y}(x,y)\,dx \\
&= \int_{-2}^{-\sqrt{y}} \frac{5}{26} y\,dx + \int_{\sqrt{y}}^{2} \frac{5}{26} y\,dx \\
&= \frac{5}{26} xy\Big|_{-2}^{-\sqrt{y}} + \frac{5}{26} xy\Big|_{\sqrt{y}}^{2} \\
&= \frac{5}{26}\left[\left(-y^{3/2}+2y\right)+\left(2y-y^{3/2}\right)\right] \\
&= \frac{5}{26}\left[4y-2y^{3/2}\right] \\
\implies f_Y(y) &= \begin{cases} \frac{5}{13}(2y-y^{3/2}), & 1 \le y \le 4 \\ 0, & \text{otherwise} \end{cases}
\end{aligned}
$$

(c)

$$
\begin{aligned}
\mathbb{P}\left(Y > \frac{3}{2} \,\middle|\, X < \frac{1}{2}\right) &= \frac{\mathbb{P}\left(Y > \frac{3}{2}, X < \frac{1}{2}\right)}{\mathbb{P}\left(X < \frac{1}{2}\right)} \\
&= \frac{\int_{-2}^{-\sqrt{3/2}} \int_{3/2}^{x^2} \frac{5}{26} y\,dy\,dx}{\int_{-2}^{-1} \frac{5}{52}(x^4-1)\,dx} \\
&= \frac{\int_{-2}^{-\sqrt{3/2}} \frac{y^2}{2}\Big|_{3/2}^{x^2} dx}{\frac{1}{2}\left(\frac{x^5}{5}-x\right)\Big|_{-2}^{-1}} \\
&= \frac{\int_{-2}^{-\sqrt{3/2}} x^4 - \frac{9}{4}\,dx}{\left(\frac{-1}{5}+1\right)-\left(\frac{-32}{5}+2\right)} \\
&= \frac{\frac{x^5}{5} - \frac{9}{4}x\Big|_{-2}^{-\sqrt{3/2}}}{\frac{4}{5}+\frac{22}{5}} \\
&= \frac{\left(-\frac{9}{4}\frac{\sqrt{3/2}}{5} + \frac{9}{4}\sqrt{\frac{3}{2}}\right) - \left(\frac{-32}{5}+\frac{9}{2}\right)}{\frac{26}{5}} \\
&= \frac{\frac{36}{20}\sqrt{\frac{3}{2}} + \frac{19}{10}}{\frac{26}{5}} \\
&= \frac{9\sqrt{\frac{3}{2}} + \frac{19}{2}}{26} \\
&= \frac{9\sqrt{6}+19}{52} = 0.7893
\end{aligned}
$$

Solution for 33:

(a)

$$\begin{aligned}
1 &\overset{\text{set}}{=} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_0^1 \int_0^{1-y} kx^2y\,dx\,dy \\
&= \int_0^1 ky \cdot \frac{x^3}{3}\Big|_0^{1-y}\,dy \\
&= \frac{k}{3}\int_0^1 y - 3y^2 + 3y^3 - y^4\,dy \\
&= \frac{k}{3}\left(\frac{y^2}{2} - y^3 + \frac{3}{4}y^4 - \frac{y^5}{5}\right)\Big|_0^1 \\
&= \frac{k}{3}\left(\frac{1}{2} - 1 + \frac{3}{4} - \frac{1}{5}\right) \\
1 &= \frac{k}{60} \\
\Longrightarrow k &= 60
\end{aligned}$$

(b)

$$f_X(x) = \int_0^{1-x} 60x^2y\,dy = 30x^2y^2\big|_0^{1-x} = 30x^2(1-x)^2 \text{ for } 0 \le x \le 1$$
$$f_Y(y) = \int_0^{1-y} 60x^2y\,dx = 20x^3y\big|_0^{1-y} = 20y(1-y)^3 \text{ for } 0 \le y \le 1$$

(c)

$$\begin{aligned}
E[X] &= \int_0^1 x \cdot 30x^2(1-x)^2\,dx = \int_0^1 30x^3 - 60x^4 + 30x^5\,dx \\
&= \frac{15x^4}{2} - 12x^5 + 5x^6\Big|_0^1 = \frac{15}{2} - 12 + 5 = \frac{1}{2} \\
E[Y] &= \int_0^1 y \cdot 20y(1-y)^3\,dy = \int_0^1 20y^2(1 - 3y + 3y^2 - y^3)\,dy \\
&= \frac{20y^3}{3} - 15y^4 + 12y^5 - \frac{10y^6}{3}\Big|_0^1 = \frac{20}{3} - 15 + 12 - \frac{10}{3} = \frac{1}{3}
\end{aligned}$$

(d)

$$\begin{aligned}
\mathbb{P}\left(Y > \frac{1}{3} \,\middle|\, X > \frac{1}{2}\right) &= \frac{\mathbb{P}\left(Y > \frac{1}{3}, X > \frac{1}{2}\right)}{\mathbb{P}\left(X > \frac{1}{2}\right)} \\
&= \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} \int_{\frac{1}{2}}^{1-y} 60x^2 y \, dx \, dy}{\int_{\frac{1}{2}}^{1} 30x^2 - 60x^3 + 30x^4 \, dx} \\
&= \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} \int_{\frac{1}{2}}^{1-y} 2x^2 y \, dx \, dy}{\int_{\frac{1}{2}}^{1} x^2 - 2x^3 + x^4 \, dx} \\
&= \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} 2y \, \frac{x^3}{3}\Big|_{\frac{1}{2}}^{1-y} \, dy}{\frac{x^3}{3} - \frac{x^4}{2} + \frac{x^5}{5}\Big|_{\frac{1}{2}}^{1}} \\
&= \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} \frac{2y}{3}\left[\left(1 - 3y + 3y^2 - y^3\right) - \frac{1}{8}\right] \, dy}{\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{5}\right) - \left(\frac{1}{24} - \frac{1}{32} + \frac{1}{160}\right)} \\
&= \frac{\int_{\frac{1}{3}}^{\frac{1}{2}} \frac{7}{4}y - 6y^2 + 6y^3 - 2y^4 \, dy}{3\left(\frac{1}{30} - \frac{1}{60}\right)} \\
&= 20\left[\frac{7y^2}{8} - 2y^3 + \frac{3y^4}{2} - \frac{2y^5}{5}\right]_{\frac{1}{3}}^{\frac{1}{2}} \\
&= 20\left[\left(\frac{7}{32} - \frac{1}{4} + \frac{3}{32} - \frac{1}{80}\right) - \left(\frac{7}{72} - \frac{2}{27} + \frac{1}{54} - \frac{2}{1215}\right)\right] \\
&= 20\left[\frac{1}{20} - \frac{389}{9720}\right] = \frac{97}{486} = 0.1996
\end{aligned}$$

Solution for 35:

(a) Assuming that the day a person shops is uniformly distributed across the week and letting X = wait time in minutes that a shopper spend in a local supermarket's check out line and Y = weekend indicator where 0 is a weekday and 1 is a weekend,

$$\begin{aligned}
\mathbb{P}(X < 1) &= \mathbb{P}(X < 1 \cap Y = 0) + \mathbb{P}(X < 1 \cap Y = 1) \\
&= \mathbb{P}(X < 1 \,|\, Y = 0)\mathbb{P}(Y = 0) + \mathbb{P}(X < 1 \,|\, Y = 1)\mathbb{P}(Y = 1) \\
&= \frac{5}{7}\int_0^1 \frac{e^{-x/2}}{2} \, dx + \frac{2}{7}\int_0^1 \frac{e^{-x/3}}{3} \, dx \\
&= \frac{5}{7}\left(1 - e^{-1/2}\right) + \frac{2}{7}\left(1 - e^{-1/3}\right) = 0.362
\end{aligned}$$

```
> (1 - exp(-1/2))*5/7 + (1 - exp(-1/3))*2/7
[1] 0.3620406
> # or
> pexp(1, 1/2)*5/7 + pexp(1, 1/3)*2/7
[1] 0.3620406
```

(b) $\mathbb{P}(Y = 1|X < 2) = 0.2354$

```
> ((1 - exp(-2/3))*2/7) / ((1 - exp(-1))*5/7 + (1 - exp(-2/3))*2/7)

[1] 0.2354185
```

(c) $\mathbb{P}(X < 2) = 0.5905$
Same as (a) with integrals from 0 to 2 rather than 1.

```
> (1 - exp(-1))*5/7 + (1 - exp(-2/3))*2/7

[1] 0.5905384

> # or
> pexp(2, 1/2)*5/7 + pexp(2, 1/3)*2/7

[1] 0.5905384
```

Solution for 37:
(a) Let $X =$ time maintenance supervisor opens the sluice gate. Then $X \sim Unif(9, 10)$ and $f_X(x) = \frac{1}{10-9}$ for $9 \leq x \leq 10$.

$$E[X] = \int_9^{10} x \cdot f_X(x)\, dx = \int_9^{10} \frac{x}{10-9}\, dx = \frac{x^2}{2}\Big|_9^{10} = \frac{100-81}{2} = 9.5$$

```
> fx <- function(x){x}
> integrate(fx, lower = 9, upper = 10)$value

[1] 9.5
```

On the average, the sluice gate opens at 9:30 p.m.

(b) $\mathbb{P}(X < 9:30) = \mathbb{P}(X < 9.5) = \int_9^{9.5} 1\, dx = x|_9^{9.5} = 0.5$

```
> fx <- function(x){1}
> integrate(Vectorize(fx), lower=9, upper=9.5)$value

[1] 0.5
```

The probability that the sluice gate opens before 9:30 p.m. is 0.5.

Solution for 39:
Solving directly:

$$E[Y] = E\left[\sum_{i=1}^{n} c_i X_i\right] = \sum_{i=1}^{n} c_i E\left[X_i\right] = \sum_{i=1}^{n} c_i \mu_i$$

$$
\begin{aligned}
Var[Y] &= E\left[(Y - E[Y])^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} c_i X_i - E\left[\sum_{i=1}^{n} c_i X_i\right]\right)^2\right] \\
&= E\left[\left(\sum_{i=1}^{n} c_i (X_i - \mu_i)\right)^2\right] \\
&= \sum_{i=1}^{n} c_i^2 E\left[(X_i - \mu_i)^2\right] + 2\sum\sum_{i<j} c_i c_j E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\
&= \sum_{i=1}^{n} c_i^2 \sigma_i^2 + 2\sum\sum_{i<j} c_i c_j \, Cov[X_i, X_j]
\end{aligned}
$$

Note that if the X_is are independent, then $Cov[X_i, X_j] = 0$.

$$
Var[Y] = \sum_{i=1}^{n} c_i^2 \sigma_i^2
$$

Chapter 6

Odd solutions

Solution for 1:
There are $\binom{90}{8} = 77515521435$ ways to choose 8 people from 90.

```
> choose(90, 8)

[1] 77515521435
```

Solution for 3:
$M_X(t) = (1-2t)^{-5} \Longrightarrow X \sim \chi^2_{10}$, so $\mathbb{P}(X < 15.99) = 0.9001$

```
> pchisq(15.99, 10)

[1] 0.900081
```

Solution for 5:
$\mathbb{P}(X < 8) = 0.3712$ and $\mathbb{P}(X > 6) = 0.8153$. If $\mathbb{P}(X < a) = 0.15$, then $a = 5.5701$. The population mean and population variance of a χ^2_{10} random variable are 10 and 20, respectively.

```
> pchisq(8, 10)

[1] 0.3711631

> pchisq(6, 10, lower = FALSE)

[1] 0.8152632

> qchisq(0.15, 10)

[1] 5.570059
```

Solution for 7:
(a) The population mean and population variance are 2 and 2, respectively.

```
> pop <- c(0, 1, 2, 3, 4)
> Mu <- mean(pop)
> Ppop <- 1/length(pop)
> Va <- sum((pop - Mu)^2*Ppop)
> c(Mu, Va)

[1] 2 2
```

(b) The following code is used to verify that the mean of $\overline{X}$ is 2 and that the variance of $\overline{X}$ is $\frac{1}{3}$.

```
> SRS <- srs(pop, 3)
> xbar <- apply(SRS, 1, mean)
> Pxbar <- 1/length(xbar)
> MUxbar <- mean(xbar)
> VAxbar <- sum((xbar - MUxbar)^2*Pxbar)
> MASS::fractions(c(MUxbar, VAxbar))

[1]   2 1/3
```

(c) The following code is used to verify that the mean of $\overline{X}$ is 2 and that the variance of $\overline{X}$ is $\frac{2}{3}$.

```
> RS <- expand.grid(pop, pop, pop)
> xbar <- apply(RS, 1, mean)
> Pxbar <- 1/length(xbar)
> MUxbar <- mean(xbar)
> VAxbar <- sum((xbar - MUxbar)^2*Pxbar)
> MASS::fractions(c(MUxbar, VAxbar))

[1]   2 2/3
```

Solution for 9:

A statistic is a numerical summary from a sample that cannot be a function of a parameter. This means that all of the above except (c) are statistics.

Solution for 11:

Note that $\dfrac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$.

(a) $\mathbb{P}\left(\frac{\overline{X}}{S} > 2\right) = \mathbb{P}(t_5 > 2\sqrt{6}\,) = 0.0022$

```
> 1 - pt(sqrt(6)*2, 5)

[1] 0.002239216
```

(b) $\dfrac{\overline{X}-0}{\frac{S_u\sqrt{\frac{n}{n-1}}}{\sqrt{n}}} = \dfrac{\overline{X}}{\frac{S_u}{\sqrt{n-1}}} = \dfrac{\overline{X}\sqrt{n-1}}{S_u} \sim t_{n-1}$.

$\mathbb{P}\left(\left|\frac{\overline{X}}{S_u}\right| \le 4\right) = \mathbb{P}\left(-4\sqrt{5} \le \frac{\overline{X}}{S_u}\sqrt{5} \le 4\sqrt{5}\right) = \mathbb{P}\left(-4\sqrt{5} \le t_5 \le 4\sqrt{5}\right) = 0.9997.$

```
> pt(4*sqrt(5), 5) - pt(-4*sqrt(5), 5)

[1] 0.9997089
```

Solution for 13:

$$\begin{aligned}\mathbb{P}\left(0.5 < \frac{S^2}{\sigma^2} < 1.2\right) &= \mathbb{P}\left(10(0.5) \le 10\frac{S^2}{\sigma^2} \le 10(1.2)\right)\\ &= \mathbb{P}(5 \le \chi^2_{10} \le 12)\\ &= \mathbb{P}(\chi^2_{10} \le 12) - \mathbb{P}(\chi^2_{10} \le 5) = 0.6061\end{aligned}$$

```
> pchisq(12, 10) - pchisq(5, 10)

[1] 0.6061215
```

Solution for 15:

(a) Let X = weight of a particular vitamin. $X \sim N(0.6, 0.015)$. The therapy is not effective if more than $0.20 \times 125 = 25$ pills are under 0.58 grams. The probability of a vitamin being underweight is $\mathbb{P}(X \leq 0.58) = p \implies p = 0.0912$. Let W = number of underweight pills. $W \sim Bin(125, p)$. The probability the therapy is not effective is $\mathbb{P}(W > 25) = 1 - P(W \leq 25) = 1e - 04$.

```
> p <- pnorm(0.58, 0.60, 0.015)
> p

[1] 0.09121122

> 1 - pbinom(25, 125, p)

[1] 5.515032e-05
```

(b) Let Y = weight of a bottle of vitamins. $Y \sim N(125 \times 0.6, \sqrt{125 \times 0.015^2})$. That is $Y \sim N(75, 0.1677)$. $\mathbb{P}(Y > 74.7) = 1 - P(Y \leq 74.7) = 0.9632$.

```
> ans <- 1 - pnorm(74.7, 125*0.6, sqrt(125*0.015^2)) # P(Y > 74.7)
> ans

[1] 0.9631809
```

Let V = number of bottles that weigh in excess of 74.7 grams. $V \sim Bin(50, 0.9632)$ and $\mathbb{P}(V \leq 46) = 0.1117$. In other words, the probability a randomly selected box does not meet the manufacturers' guarantee is 0.1117.

```
> pbinom(46, 50, ans)  # P(V <= 46)

[1] 0.1117434
```

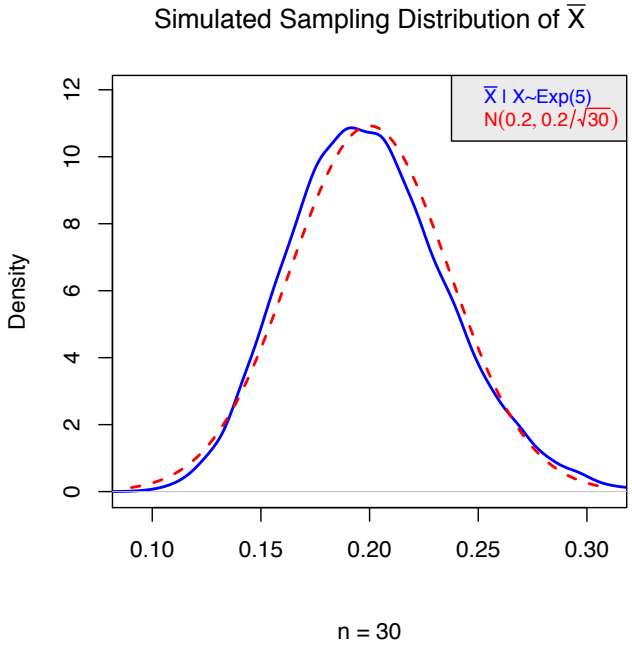
Solution for 17:

The answer to this question depends on how close you want to be to the normal distribution. What one can observe from the graphs is that there is still a positive skew in the sampling distribution of $\overline{X}$ even for samples of size 300 when sampling from an exponential.

$n = 30$

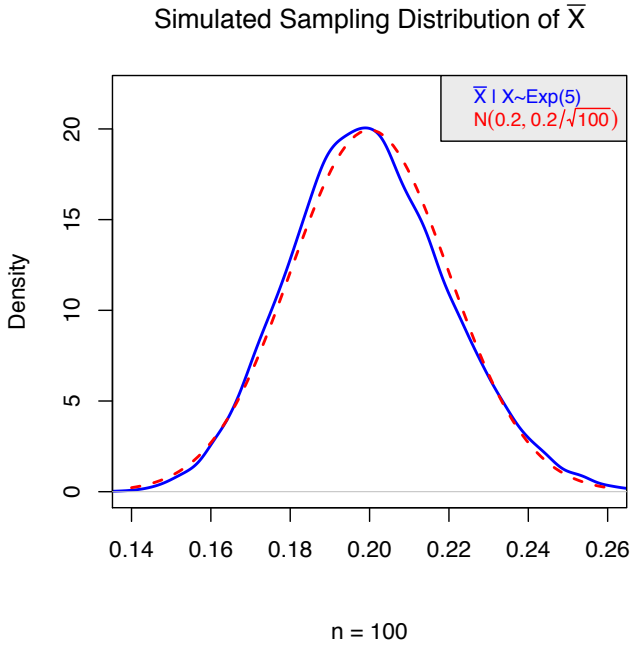
```
> sims <- 20000
> xbar30 <- numeric(sims)  # storage space
> n <- 30
> for(i in 1:sims){
+   xbar30[i] <- mean(rexp(n, 5))
+ }
> plot(density(xbar30), col = "blue", lwd = 2, xlab = "",
+ ylim=c(0, 1.1*max(density(xbar30)$y)),
+ main = substitute(paste("Simulated Sampling Distribution of ", bar(X))),
```

```
+ sub = "n = 30", xlim = c(0.2 - 3*0.2/sqrt(30), 0.2 + 3*0.2/sqrt(30)))
> curve(dnorm(x, 0.2, 0.2/sqrt(30)), add = TRUE, lwd = 2, lty = 2,
+       col = "red")
> legend(x = "topright",
+        legend = c(substitute(paste(bar(X), paste(" | X~Exp(5)"))),
+ expression(N(0.2, 0.2/sqrt(30)))), text.col = c("blue", "red"),
+ bg = "gray92", cex = 0.80)
```



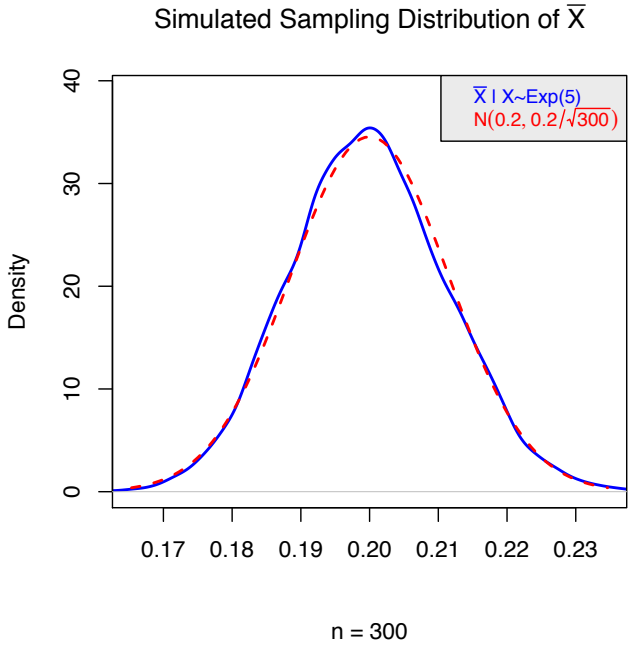
$n = 100$

```
> xbar100 <- numeric(sims)  # storage space
> n <- 100
> for(i in 1:sims){
+   xbar100[i] <- mean(rexp(n, 5))
+ }
> plot(density(xbar100), col = "blue", lwd = 2, xlab = "",
+ ylim=c(0, 1.1*max(density(xbar100)$y)),
+ main = substitute(paste("Simulated Sampling Distribution of ", bar(X))),
+ sub = "n = 100", xlim = c(0.2 - 3*0.2/sqrt(100), 0.2 + 3*0.2/sqrt(100)))
> curve(dnorm(x, 0.2, 0.2/sqrt(100)), add = TRUE, lwd = 2, lty = 2,
+       col = "red")
> legend(x = "topright",
+        legend = c(substitute(paste(bar(X), paste(" | X~Exp(5)"))),
+ expression(N(0.2, 0.2/sqrt(100)))), text.col = c("blue", "red"),
+ bg = "gray92", cex = 0.80)
```



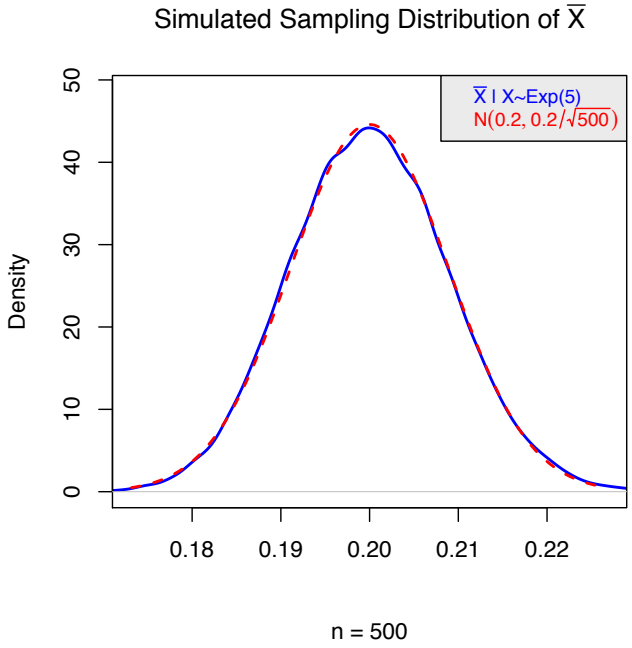
$n = 300$

```
> xbar300 <- numeric(sims)  # storage space
> n <- 300
> for(i in 1:sims){
+   xbar300[i] <- mean(rexp(n, 5))
+ }
> plot(density(xbar300), col = "blue", lwd = 2, xlab = "",
+ ylim=c(0, 1.1*max(density(xbar300)$y)),
+ main = substitute(paste("Simulated Sampling Distribution of ", bar(X))),
+ sub = "n = 300", xlim = c(0.2 - 3*0.2/sqrt(300), 0.2 + 3*0.2/sqrt(300)))
> curve(dnorm(x, 0.2, 0.2/sqrt(300)), add = TRUE, lwd = 2, lty = 2,
+       col = "red")
> legend(x = "topright",
+        legend = c(substitute(paste(bar(X), paste(" | X~Exp(5)"))),
+ expression(N(0.2, 0.2/sqrt(300)))), text.col = c("blue", "red"),
+ bg = "gray92", cex = 0.80)
```



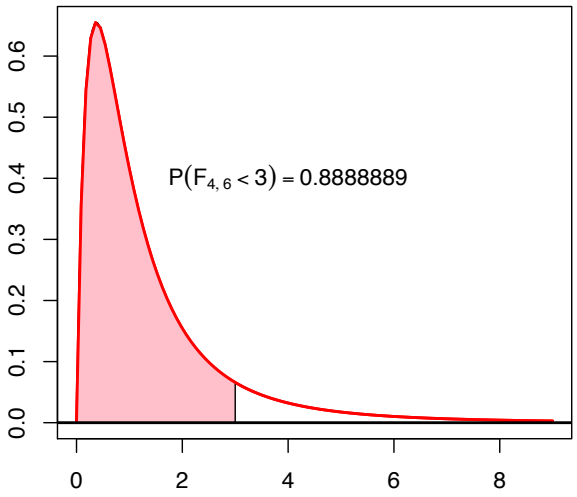
$n = 500$

```
> xbar500 <- numeric(sims)  # storage space
> n <- 500
> for(i in 1:sims){
+   xbar500[i] <- mean(rexp(n, 5))
+ }
> plot(density(xbar500), col = "blue", lwd = 2, xlab = "",
+ ylim=c(0, 1.1*max(density(xbar500)$y)),
+ main = substitute(paste("Simulated Sampling Distribution of ", bar(X))),
+ sub = "n = 500", xlim = c(0.2 - 3*0.2/sqrt(500), 0.2 + 3*0.2/sqrt(500)))
> curve(dnorm(x, 0.2, 0.2/sqrt(500)), add = TRUE, lwd = 2, lty = 2,
+       col = "red")
> legend(x = "topright",
+        legend = c(substitute(paste(bar(X), paste(" | X~Exp(5)"))),
+ expression(N(0.2, 0.2/sqrt(500)))), text.col = c("blue", "red"),
+ bg = "gray92", cex = 0.80)
> par(opar)  # restore original settings
```



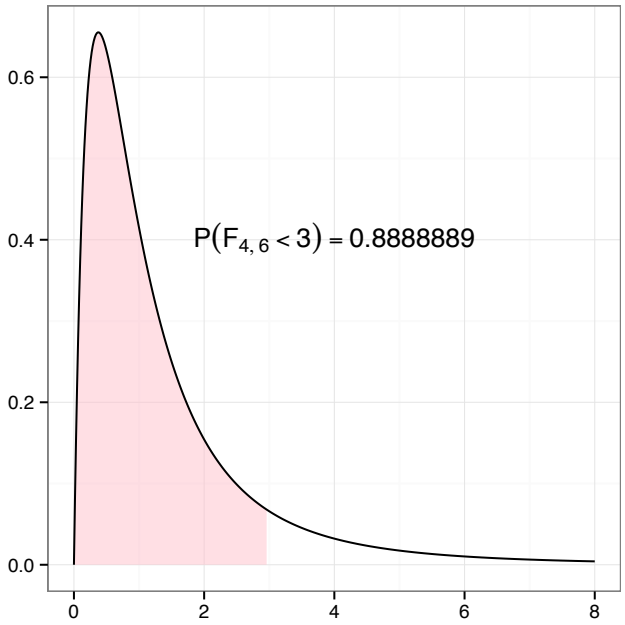
Solution for 19:

```
> curve(df(x, 4, 6), from = 0, to = 9, lwd = 2,
+ col = "red", ylab = "",xlab = "")
> x <- seq(0, 3, 0.05)
> y <- df(x, 4, 6)
> xs <- c(0, x, 3)
> ys <- c(0, y, 0)
> polygon(xs, ys, density = -1, col = "pink")
> curve(df(x, 4, 6), from = 0, to = 9, lwd = 2,
+ col = "red", add = TRUE, ylab = "",xlab = "")
> abline(h = 0, lwd = 2)
> text(4, 0.4, expression(P(F[list(4,6)] < 3)== 0.8888889))
```



Similar graph with `ggplot`:

```
> limitRange <- function(fun, dfn, dfd, min, max){
+   function(x){
+       y <- fun(x, dfn, dfd)
+       y[x < min | x > max] <- NA
+       return(y)
+   }
+ }
> dlimit <- limitRange(df, 4, 6, 0, 3)
> p <- ggplot(data = data.frame(x = c(0, 8)), aes(x = x))
> p + stat_function(fun = dlimit, geom = "area", fill = "pink",
+                   alpha = 0.5) +
+   stat_function(fun = df, arg = list(4, 6), n = 500) +
+   theme_bw() +
+   labs(x = "", y = "") +
+   annotate("text", x = 4, y = 0.4,
+            label = "P(F[list(4,6)] < 3)== 0.8888889",  parse = TRUE)
```



Solution for 21:

The numerator has a $N\left(0,\ \sqrt{c^2\sigma^2+\sigma^2+\sigma^2}\right)$ distribution, which means $\frac{cX_1+X_2+X_3-0}{\sigma\sqrt{c^2+2}} \sim N(0,1)$.

The expression $\frac{1}{\sigma^2}(X_4^2+X_5^2+X_6^2) \sim \chi_3^2$. A t distribution is constructed as $\frac{N(0,1)}{\sqrt{\chi_\nu^2/\nu}}$.

$$\begin{aligned}
\frac{cX_1+X_2+X_3}{\sqrt{X_4^2+X_5^2+X_6^2}} &= \frac{\frac{cX_1+X_2+X_3-0}{\sigma\sqrt{c^2+2}}}{\frac{\sqrt{X_4^2+X_5^2+X_6^2}}{\sigma\sqrt{c^2+2}}} \\
&= \frac{N(0,1)}{\sqrt{\frac{1}{c^2+2}\cdot\frac{1}{\sigma^2}(X_4^2+X_5^2+X_6^2)}} \\
&= \frac{N(0,1)}{\sqrt{\frac{1}{c^2+2}\cdot\chi_3^2}} \\
\implies c^2+2 &= 3 \\
c &= \pm 1
\end{aligned}$$

Solution for 23:

For $X \sim Exp(\lambda)$, $M_X(t) = \left(1-\frac{t}{\lambda}\right)^{-1}$. Also, the moment generating function of a $Y \sim \Gamma(n, \lambda n)$ is $M_Y(t) = \left(1-\frac{t}{\lambda n}\right)^{-n}$.

Since $\overline{X} = \sum_{i=1}^n \frac{X_i}{n}$, $M_{\overline{X}}(t) = E\left[e^{t\overline{X}}\right] = E\left[e^{t\left(\sum_{i=1}^n \frac{X_i}{n}\right)}\right] = E\left[\prod_{i=1}^n e^{t\cdot\frac{X_i}{n}}\right]$

Because the X_is are independent and identically distributed,

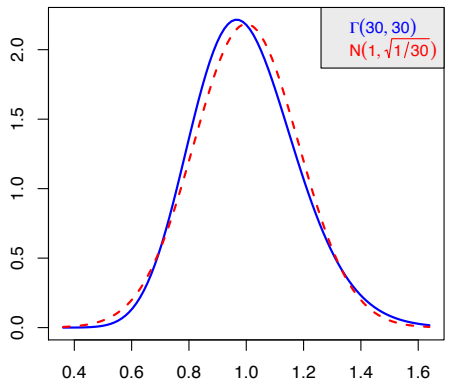
$$M_{\overline{X}}(t) = \prod_{i=1}^n E\left[e^{t\cdot\frac{X_i}{n}}\right] = \prod_{i=1}^n M_{X_i}\left(\frac{t}{n}\right) = \prod_{i=1}^n \left(1-\frac{t}{n\lambda}\right)^{-1} = \left(1-\frac{t}{n\lambda}\right)^{-n} = M_Y(t).$$

Note that the sampling distributions of the sample mean are a $\Gamma(30,30)$, $\Gamma(100,100)$, $\Gamma(300,300)$, and $\Gamma(500,500)$ for the sample sizes $n = 30$, 100, 300, and 500, respectively.

The normal distribution superimposed over the gamma distributions are $N(1, \sqrt{1/30})$, $N(1, \sqrt{1/100})$, $N(1, \sqrt{1/300})$, and $N(1, \sqrt{1/500})$, respectively, since the mean of the gamma is α/λ and the variance is α/λ^2.

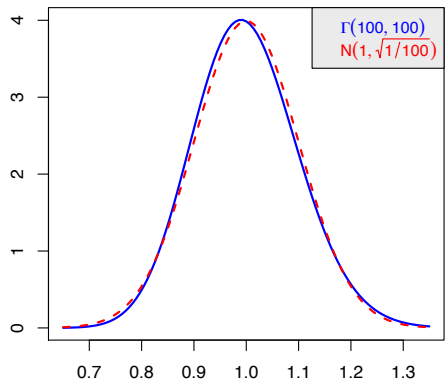
Code for the graph with $\Gamma(30, 30)$:

```
> curve(dgamma(x, 30, 30), from = 1 - 3.5*sqrt(1/30),
+   to = 1 + 3.5*sqrt(1/30), ylab = "", lwd = 2, col = "blue", xlab = "")
> curve(dnorm(x, 1, sqrt(1/30)), from = 1 - 3.5*sqrt(1/30),
+   to = 1 + 3.5*sqrt(1/30), ylab = "", lwd = 2, lty = 2, col = "red",
+   add = TRUE, xlab = "")
> legend(x = "topright", legend = c(expression(Gamma(list(30, 30))),
+                                   expression(N(list(1, sqrt(1/30))))),
+   text.col=c("blue", "red"), bg = "gray92", cex = 0.90)
```



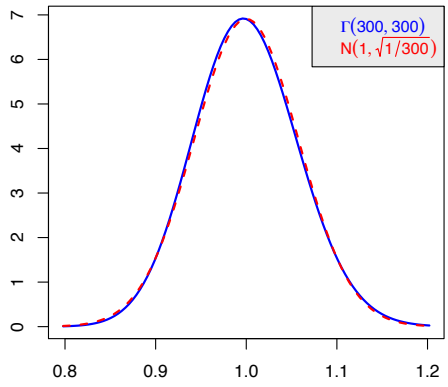
Code for the graph with $\Gamma(100, 100)$:

```
> curve(dgamma(x, 100, 100), from = 1 - 3.5*sqrt(1/100),
+   to = 1 + 3.5*sqrt(1/100), ylab = "", lwd = 2, col = "blue", xlab = "")
> curve(dnorm(x, 1, sqrt(1/100)), from = 1 - 3.5*sqrt(1/100),
+   to = 1 + 3.5*sqrt(1/100), ylab = "", lwd = 2, lty = 2, col = "red",
+   add = TRUE, xlab = "")
> legend(x = "topright", legend = c(expression(Gamma(list(100, 100))),
+                                   expression(N(list(1, sqrt(1/100))))),
+   text.col=c("blue", "red"), bg = "gray92", cex = 0.90)
```



Code for the graph with $\Gamma(300, 300)$:

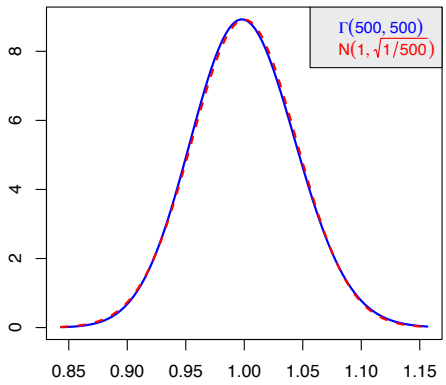
```
> curve(dgamma(x, 300, 300), from = 1 - 3.5*sqrt(1/300),
+   to = 1 + 3.5*sqrt(1/300), ylab = "", lwd = 2, col = "blue", xlab = "")
> curve(dnorm(x, 1, sqrt(1/300)), from = 1 - 3.5*sqrt(1/300),
+   to = 1 + 3.5*sqrt(1/300), ylab = "", lwd = 2, lty = 2, col = "red",
+   add = TRUE, xlab = "")
> legend(x = "topright", legend = c(expression(Gamma(list(300, 300))),
+                                   expression(N(list(1, sqrt(1/300))))),
+   text.col=c("blue", "red"), bg = "gray92", cex = 0.90)
```



Code for the graph with $\Gamma(500, 500)$:

```
> curve(dgamma(x, 500, 500), from = 1 - 3.5*sqrt(1/500),
+   to = 1 + 3.5*sqrt(1/500), ylab = "", lwd = 2, col = "blue", xlab = "")
> curve(dnorm(x, 1, sqrt(1/500)), from = 1 - 3.5*sqrt(1/500),
+   to = 1 + 3.5*sqrt(1/500), ylab = "", lwd = 2, lty = 2, col = "red",
+   add = TRUE, xlab = "")
> legend(x = "topright", legend = c(expression(Gamma(list(500, 500))),
```

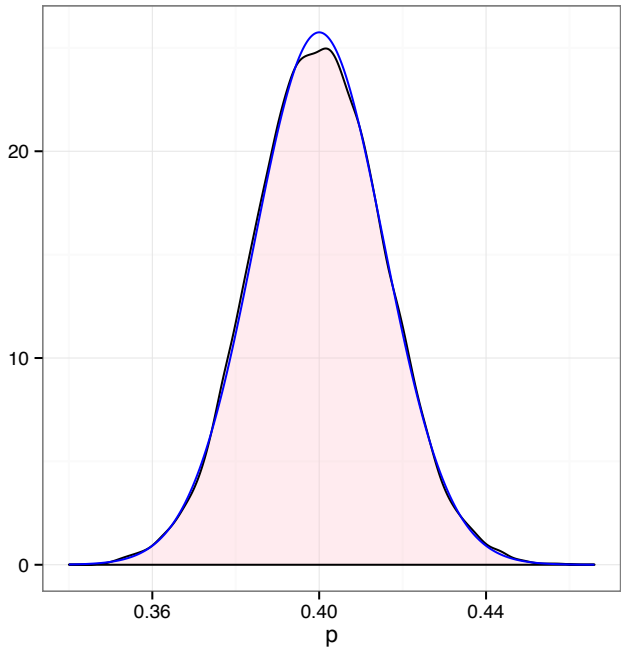
```
+                                    expression(N(list(1, sqrt(1/500))))),
+   text.col=c("blue", "red"), bg = "gray92", cex = 0.90)
```



Solution for 25:

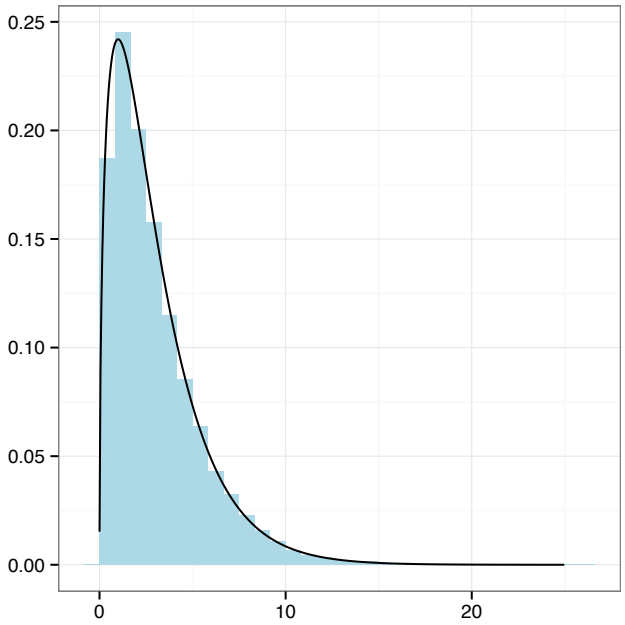
Note that the mean and standard deviation of the simulated sampling distribution are 0.4 and 0.0155, respectively. The pink simulated distribution is fairly close to the blue theoretical distribution.

```
> set.seed(95)
> sims <- 20000
> n <- 1000
> p <- rbinom(sims, n, 0.4)/n
> c(mean(p), sd(p))

[1] 0.39995085 0.01550862

> ggplot(data = data.frame(x = p), aes(x = x)) +
+   geom_density(fill = "pink", alpha = .3) +
+   stat_function(fun = dnorm, args = list(0.4, sqrt(0.4*0.6/n)),
+                 col = "blue", n = 500) +
+   labs(x = "p", y = "") +
+   theme_bw()
```



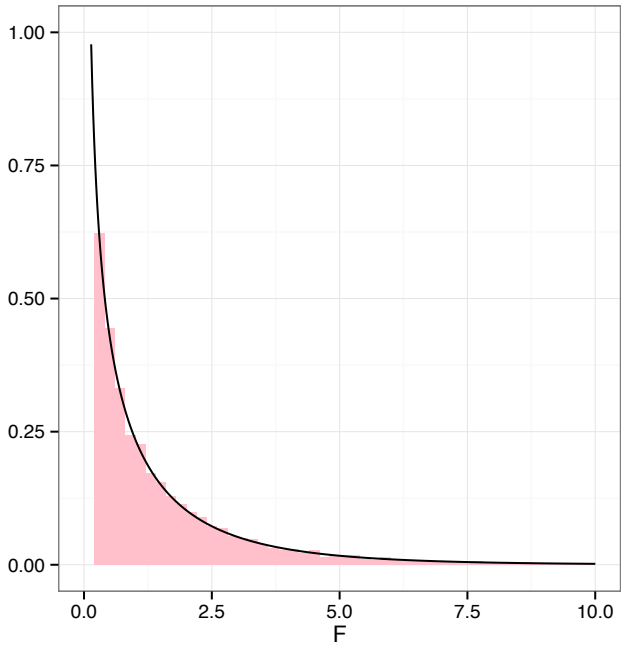
Solution for 27:

```
> set.seed(48)
> sims <- 20000
> Z1 <- rnorm(sims, 0, 1)
> Z2 <- rnorm(sims, 0, 1)
> Z3 <- rnorm(sims, 0, 1)
> chi2obs3 <- Z1^2 + Z2^2 + Z3^2
> c(mean(chi2obs3), var(chi2obs3))

[1] 3.004454 6.097796

> ggplot(data = data.frame(x = chi2obs3), aes(x = x)) +
+   geom_histogram(aes(x = x, y = ..density..), fill = "lightblue") +
+   stat_function(fun = dchisq, args = list(3), n = 500) +
+   theme_bw() +
+   labs(x = "", y = "")
```



Solution for 29:

```
> set.seed(37)
> sims <- 20000
> Z <- rnorm(sims, 0, 1)
> X2 <- rchisq(sims, 20)
> F_emp <- Z^2/(X2/20)
> ggplot(data = data.frame(x = F_emp), aes(x = x)) +
+   geom_histogram(aes(y = ..density..), binwidth = 0.2, fill = "pink") +
+   theme_bw() + ylim(0, 1) + xlim(0, 10) +
+   stat_function(fun = df, args = list(1, 20), n = 500) +
+   labs(x = "F", y = "")
```



(a)

```
> PE <- mean(F_emp <= 2 & F_emp >= 1.5)  # empirical probability
> PE

[1] 0.063

> PT <- pf(2, 1, 20) - pf(1.5, 1, 20)    # theoretical probability
> PT

[1] 0.0622232

> PDa <- abs(PE - PT)/PT*100
> PDa

[1] 1.248416
```

The empirical probability $\mathbb{P}(1.5 < F < 2) = 0.063$, while the theoretical probability $\mathbb{P}(1.5 < F_{1,20} < 2) = 0.0622$. The percent difference between the empirical and theoretical answers is 1.2484%.

(b)

```
> PEb <- mean(F_emp <= 2 & F_emp >= 1.5)/mean(F_emp >= 1.5)
> PEb

[1] 0.2664411

> PTb <- (pf(2, 1, 20) - pf(1.5, 1, 20))/pf(1.5, 1, 20, lower = FALSE)
> PTb

[1] 0.2648906
```

```
> PDb <- abs(PEb - PTb)/PTb*100
> PDb

[1] 0.585345
```

The empirical probability $\mathbb{P}(F < 2|F > 1.5) = 0.2664$, while the theoretical probability $\mathbb{P}(F_{1,20} < 2|F_{1,20} > 1.5) = 0.2649$. The percent difference between the empirical and theoretical answers is 0.5853%.

Solution for 31:

(a) $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{500}{\sqrt{n}} \leq 100 \Longrightarrow \sqrt{n} \geq 5 \Longrightarrow n \geq 25$, so at least 25 chickens must be sampled to ensure the a standard deviation of the mean is no more than 100 grams with a confidence level of 0.95.

(b) Let $\overline{C}_i$ = mean weight from chicken coop i in grams where $i = 1, 2, 3$ and C_i be the event chickens coming from coop i.

Let W = overall average weight.

Then, $\overline{C}_1 \sim N(1800, 500/\sqrt{50})$, $\overline{C}_2 \sim N(1900, 500/\sqrt{50})$, and ; $\overline{C}_3 \sim N(2000, 500/\sqrt{50})$.

The problem wishes to discover $\mathbb{P}(C_1 \,|\, W > 1975)$

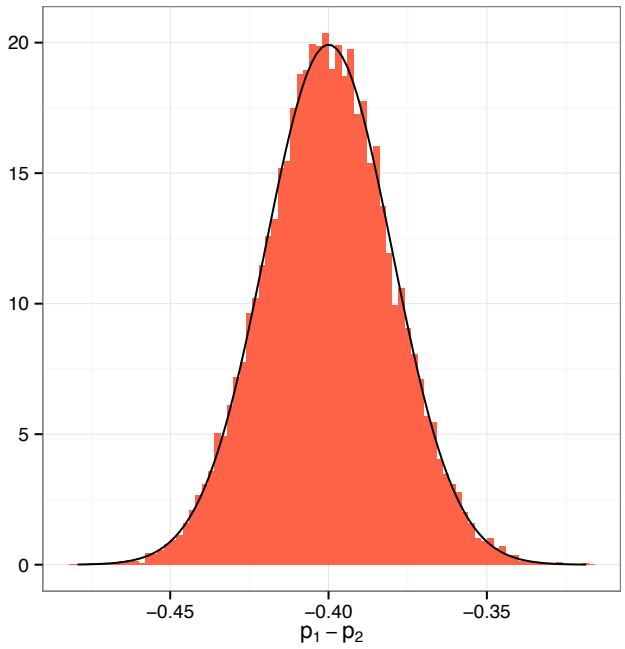
$$\begin{aligned}
\mathbb{P}(C_1 \,|\, W > 1975) &= \frac{\mathbb{P}(C_1, W > 1975)}{\mathbb{P}(W > 1975)} \\
&= \frac{\mathbb{P}(W > 1975 \,|\, C_1)\, \mathbb{P}(C_1)}{\sum_{i=1}^{3} \mathbb{P}(W > 1975 \,|\, C_i)\, \mathbb{P}(C_i)} \\
&= \frac{0.0022}{0.2631} = 0.0084
\end{aligned}$$

```
> num <- (1 - pnorm(1975, 1800, 500/sqrt(50)))*1/3
> num

[1] 0.002221388

> den <- (1 - pnorm(1975, 1800, 500/sqrt(50)))*1/3 +
+        (1 - pnorm(1975, 1900, 500/sqrt(50)))*1/3 +
+        (1 - pnorm(1975, 2000, 500/sqrt(50)))*1/3
> den

[1] 0.2630832

> PC1givnW <- num/den
> PC1givnW

[1] 0.008443672
```

Solution for 33:

Based on the graph, the shape of the sampling distribution of $p_1 - p_2$ is clearly approximately normal. Further, the mean and standard deviation of the simulated sampling distribution are -0.4001 and 0.02 which compare favorably with the theoretical answers of -0.4 and 0.02.

```
> set.seed(86)
> sims <- 20000
> n1 <- 1000
> pie1 <- 0.3
> n2 <- 1100
> pie2 <- 0.7
> p1 <- rbinom(sims, n1, pie1)/n1
> p2 <- rbinom(sims, n2, pie2)/n2
> p1Mp2 <- p1 - p2
> c(mean(p1Mp2), pie1 - pie2, sd(p1Mp2),
+   sqrt(pie1*(1 - pie1)/n1 + pie2*(1 - pie2)/n2))

[1] -0.40011982 -0.40000000  0.02004456  0.02002271

> ggplot(data = data.frame(x = p1Mp2), aes(x = x)) +
+   geom_histogram(aes(x = x, y = ..density..), fill = "tomato",
+                  binwidth = 0.002) +
+   stat_function(fun = dnorm,
+                 args = list(pie1 - pie2,
+                 sqrt(pie1*(1 - pie1)/n1 + pie2*(1 - pie2)/n2))) +
+   theme_bw() +
+   labs(x = expression(p[1] - p[2]), y = "")
```



Solution for 35:

$$\mathbb{P}\left(\left|\overline{X} - {}^{1}/_{2}\right|\right) \geq 0.90$$
$$\mathbb{P}\left(-0.10 < \overline{X} - {}^{1}/_{2} < 0.10\right) \geq 0.90$$

Since $X \sim Unif(0, 1)$, $E[X] = \frac{1}{2}$, $\text{Var}[X] = \frac{1}{12}$, this means $E[\overline{X}] = \frac{1}{2}$, $\text{Var}[\overline{X}] = \frac{\sigma_X^2}{n} = \frac{1/12}{n} = \frac{1}{12n} \Longrightarrow \sigma^2_{\overline{X}} = \frac{1}{12n}$, and $\sigma_{\overline{X}} = \frac{1}{\sqrt{12n}}$.

$$\begin{aligned} \mathbb{P}\left(-0.10 < \bar{X} - {}^1/_2 < 0.10\right) &\geq 0.90 \\ \mathbb{P}\left(\frac{-0.10}{\sigma_{\overline{X}}} < \frac{\overline{X} - {}^1/_2}{\sigma_{\overline{X}}} < \frac{0.10}{\sigma_{\overline{X}}}\right) &\geq 0.90 \\ \mathbb{P}\left(\frac{-0.10}{\sigma_{\overline{X}}} < Z < \frac{0.10}{\sigma_{\overline{X}}}\right) &\geq 0.90 \end{aligned}$$

$Z_{0.95} = \frac{0.10}{\sigma_{\bar{X}}} \Longrightarrow 1.645 = \frac{0.10}{1/\sqrt{12n}}$. Solving for n gives $n = \frac{(1.645)^2}{(0.10)^2 \cdot 12} = 22.5462$, so $n \geq 23$.

Chapter 7

Odd solutions

Solution for 1:

(a)

```
> with(data = WHEATSPAIN, c(mean(hectares), median(hectares),
+                           mad(hectares, constant = 1),
+                           sd(hectares), IQR(hectares)))

[1] 126561.5  25143.0  25043.0 197319.1 136047.0
```

(b)

```
> NCL <- subset(WHEATSPAIN, subset = community != "Castilla-Leon")
> with(data = NCL, c(mean(hectares), median(hectares),
+                           mad(hectares, constant = 1),
+                           sd(hectares), IQR(hectares)))

[1]  95730.5  21980.0  21710.0 155864.7  84537.0
```

Since the distribution is positively skewed, the preferred measures for the center of `hectares` with and without Castilla-Leon are the median and MAD. The median and MAD with Castilla-Leon are 25143 and 25043; and, without Castilla-Leon are 21980 and 21710 hectares, respectively. The preferred measure of spread with and without Castilla-Leon is the IQR.

Solution for 3:

$$eff[m, \overline{X}] = \frac{Var\left[\overline{X}\right]}{Var[m]} = \frac{\sigma^2/n}{\pi\sigma^2/2n} = 2/\pi = 0.6366$$

Solution for 5:

Since $X \sim Bin(n, \pi)$, it follows that $E[X] = n\pi$ and $Var[X] = n\pi(1-\pi)$.

(a) Given $T_1 = \frac{X}{n}$,

$$E[T_1] = E\left[\frac{X}{n}\right] = \frac{E[X]}{n} = \frac{n\pi}{n} = \pi$$

and

$$Var[T_1] = Var\left[\frac{X}{n}\right] = \frac{Var[X]}{n^2} = \frac{n\pi(1-\pi)}{n^2} = \frac{\pi(1-\pi)}{n}.$$

The $MSE[T_1]$ is $Var[T_1] + \big(Bias[T_1]\big)^2$.

$$\begin{aligned} MSE[T_1] &= Var[T_1] + \big(Bias[T_1]\big)^2 \\ &= \frac{\pi(1-\pi)}{n} + (E[T_1] - \pi)^2 \\ &= \frac{\pi(1-\pi)}{n} + (\pi - \pi)^2 \\ &= \frac{\pi(1-\pi)}{n} \end{aligned}$$

For $T_2 = \frac{X+2}{n+4}$,

$$E[T_2] = E\left[\frac{X+2}{n+4}\right] = \frac{n\pi + 2}{n+4},$$

$$Var[T_2] = Var\left[\frac{X+2}{n+4}\right] = \frac{n\pi(1-\pi)}{(n+4)^2}, \text{ and}$$

$$Bias[T_2] = (E[T_2] - \pi) = \frac{n\pi+2}{n+4} - \pi = \frac{(2-4\pi)}{n+4}.$$

It follows that

$$\begin{aligned} MSE[T_2] &= Var[T_2] + \big(Bias[T_2]\big)^2 \\ &= \frac{n\pi(1-\pi)}{(n+4)^2} + \frac{(2-4\pi)^2}{(n+4)^2} \\ &= \frac{n\pi(1-\pi) + (2-4\pi)^2}{(n+4)^2}. \end{aligned}$$

Thus, $MSE[T_1] = \frac{\pi(1-\pi)}{n}$ and $MSE[T_2] = \frac{n\pi(1-\pi)+(2-4\pi)^2}{(n+4)^2}$.

(b) For $n = 20$ and $\pi = 0.4$,

$$MSE(T_1) = \frac{\pi(1-\pi)}{n} = \frac{(0.4)(1-0.4)}{20} = 0.012, \text{ and}$$

$$MSE[T_2] = \frac{n\pi(1-\pi) + (2-4\pi)^2}{(n+4)^2} = \frac{20(0.4)(1-0.4) + [2-4(0.4)]^2}{(20+4)^2} = 0.0086,$$

so T_2 has the smaller MSE.

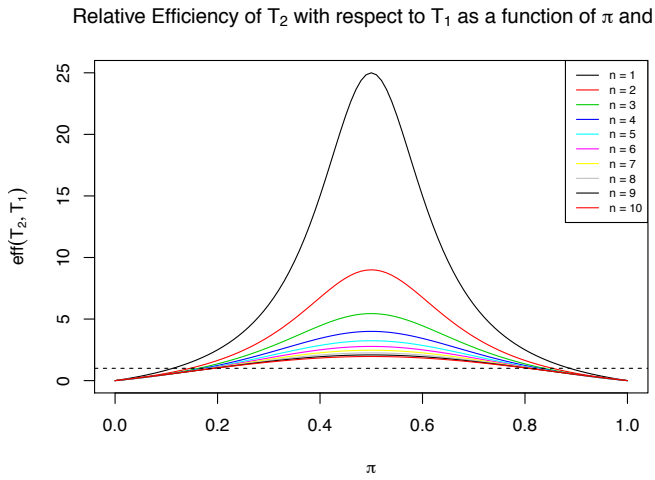
(c) The efficiency of T_2 relative to T_1 is

$$eff(T_2, T_1) = \frac{MSE[T_1]}{MSE[T_2]} = \frac{\dfrac{\pi(1-\pi)}{n}}{\dfrac{n\pi(1-\pi) + (2-4\pi)^2}{(n+4)^2}}.$$

The following code is used to graph the efficiency of T_2 relative to T_1 for values of π in $(0, 1)$ and n values from 1 to 10.

```
> p <- seq(from = 0, to = 1, by = 0.01)
> rel.ef <- function(p, n){
+   (p*(1 - p)/n) / ((n*p*(1 - p)+(2 - 4*p)^2)/(n + 4)^2)
+ }
```

```
> plot(p, rel.ef(p, 1), type = "n", xlab = expression(pi),
+ ylab = expression(eff(T[2], T[1])), ylim = c(0, 25))
> for(i in 1:10){
+   lines(p, rel.ef(p, i), col = i)
+ }
> abline(h = 1, lty = 2)
> names <- numeric(10)
> for(i in 1:10){
+   names[i] <- paste("n =", i)
+ }
> legend("topright", legend = names, lty = 1, col = c(1:10), cex = 0.7)
> title(expression(paste(plain("Relative Efficiency of "),T[2],
+ plain(" with respect to "),T[1],plain(" as a function of ")*pi,
+ plain(" and " )*n)))
```



Based on the graph, the relative efficiencies of T_2 relative to T_1 are very similar for π values close to 0 and 1. For values of π between 0.2 and 0.8, the graph illustrates the superiority of the T_2 estimator, which for $n = 1$ and $\pi = 0.5$ is roughly 25 times more efficient than T_1.

Solution for 7:

(a) For each estimator, expected value, variance, bias, and mean squared error are determined as intermediate steps to the solution of relative efficiency.

For T_1:

$E[T_1] = E\left[\overline{X}\right] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n\left(\frac{1}{\lambda}\right)}{n} = \frac{1}{\lambda}$

$Var[T_1] = Var\left[\overline{X}\right] = Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = \frac{n\left(\frac{1}{\lambda^2}\right)}{n^2} = \frac{1}{n\lambda^2}$

$Bias[T_1] = E[T_1] - \frac{1}{\lambda} = \frac{1}{\lambda} - \frac{1}{\lambda} = 0$

$MSE[T_1] = Var[T_1] + (Bias[T_1])^2 = \frac{1}{n\lambda^2} + 0 = \frac{1}{n\lambda^2}$

For T_2:

$E[T_2] = E\left[\frac{\sum_{i=1}^{n} X_i + 1}{n+2}\right] = \frac{\sum_{i=1}^{n} E[X_i] + 1}{n+2} = \frac{\frac{n}{\lambda} + 1}{n+2}$

$Var[T_2] = Var\left[\frac{\sum_{i=1}^{n} X_i + 1}{n+2}\right] \frac{n\,Var[X]}{(n+2)^2} = \frac{\frac{n}{\lambda^2}}{(n+2)^2}$

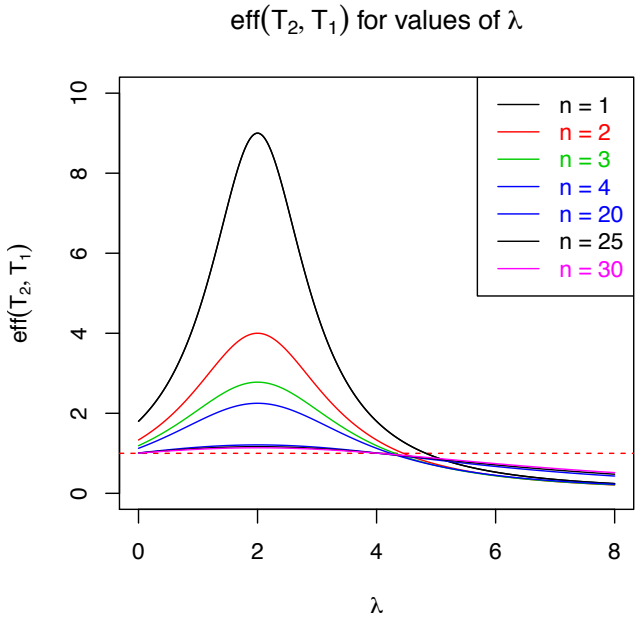
$Bias[T_2] = E[T_2] - \frac{1}{\lambda} = \frac{\frac{n}{\lambda} + 1}{n+2} - \frac{1}{\lambda} = \frac{\lambda - 2}{\lambda(n+2)}$

$$MSE[T_2] = Var[T_2] + (Bias[T_2])^2 = \frac{\frac{n}{\lambda^2}}{(n+2)^2} + \left(\frac{\lambda-2}{\lambda(n+2)}\right)^2 = \frac{n+(\lambda-2)^2}{\lambda^2(n+2)^2}$$

Consequently, $eff[T_2, T_1] = \frac{MSE[T_1]}{MSE[T_2]} = \frac{\frac{1}{n\lambda^2}}{\frac{n+(\lambda-2)^2}{\lambda^2(n+2)^2}} = \frac{(n+2)^2}{n(n+\lambda^2-4\lambda+4)}$

(b) The code to plot $eff(T_2, T_1)$ versus n values of 1, 2, 3, 4, 20, 25, and 30 is

```
> eff <- function(param, n){
+   (n + 2)^2/(n*(n + param^2 - 4*param + 4))
+ }
> lamb <- seq(from = 8, to = 0, by = -0.001)
> ns <- c(1, 2, 3, 4, 20, 25, 30)
> plot(lamb, eff(lamb, 1), type = "l", ylab = expression(eff(T[2],T[1])),
+      xlab = expression(lambda), main = expression(paste(eff(T[2],T[1]),
+      plain(" for values of ")*lambda)), col = 1, ylim = c(0, 10))
> abline(h = 1, lty = "dashed", col = "red")
> for(i in ns){
+   lines(lamb, eff(lamb, i), col = i)
+ }
> names <- numeric(length(ns))
> for(i in 1:7){
+    names[i] <- paste("n =", ns[i])
+ }
> legend("topright", legend = names, lty = 1, col = ns, text.col = ns)
```



(c) For $\lambda \leq 4$, T_2 is always more efficient than T_1 for all n. For $\lambda \geq 2+2\sqrt{2}$, T_1 is always more efficient than T_2 for all n. For $4 < \lambda < 2+2\sqrt{2}$, the efficiency of T_2 with respect to T_1 varies with n.

Solution for 9:

$$Var\left[\frac{\partial \ln f(X|\theta)}{\partial \theta}\right] = E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right] - E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)\right]^2$$

So, it is sufficient to show that $E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)\right] = 0$.

$$\begin{aligned} E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)\right] &= E\left[\left(\frac{\frac{\partial}{\partial \theta} f(X|\theta)}{f(X|\theta)}\right)\right] \\ &= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta} f(X|\theta)}{f(X|\theta)} \cdot f(X|\theta) dx \\ &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(X|\theta) dx \\ &= \frac{\partial}{\partial \theta}(1) = 0 \quad \blacksquare \end{aligned}$$

So, $Var\left[\frac{\partial \ln f(X|\theta)}{\partial \theta}\right] = E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]$.

Solution for 11:

Since $\mathbb{P}(T) = \pi$, the likelihood function is

$$L(\pi \,|\, \mathbf{x}) = L(\pi \,|\, HTT) = (1-\pi)\pi^2.$$

π	$(1-\pi)\pi^2$
$1/2$	$1/8$
$1/3$	$2/27$
$2/3$	$4/27$

The value $\pi = 2/3$ maximizes the likelihood function; therefore, $\hat{\pi}(\mathbf{x}) = 2/3$.

Solution for 13:

(a) The first sample moment, m_1 is $\overline{X}$, and the first population moment about zero is $\alpha_1(\pi) = E[X] = \frac{1-\pi}{\pi}$. By equating the first population moment to the first sample moment,

$$\begin{aligned} \alpha_1(\pi) = \frac{1-\pi}{\pi} &\overset{\text{set}}{=} \overline{X} = m_1 \\ 1 - \pi &= \pi \overline{X} \\ 1 &= \pi(\overline{X} + 1) \\ \pi &= \frac{1}{1+\overline{X}} \\ \Longrightarrow \tilde{\pi} &= \frac{1}{1+\overline{X}} \quad \text{is the method of moments estimator of } \pi \end{aligned}$$

(b) The likelihood function $L(\pi|\mathbf{x}) = \prod_{i=1}^{n} \pi(1-\pi)^{x_i}$ for a geometric distribution. To maximize this function in π, the logarithm and partial derivative will be calculated.

$$\begin{aligned}\ln L(\pi|\mathbf{x}) &= \ln\left(\prod_{i=1}^{n}\pi(1-\pi)^{x_i}\right)\\ &= \sum_{i=1}^{n}\left[\ln(\pi)+x_i\ln(1-\pi)\right]\\ &= n\ln(\pi)+\ln(1-\pi)\sum_{i=1}^{n}x_i\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln L(\pi|\mathbf{x})}{\partial\pi} = \frac{n}{\pi}-\frac{\sum_{i=1}^{n}x_i}{1-\pi} &\stackrel{\text{set}}{=} 0\\ n(1-\pi) &= \pi\sum_{i=1}^{n}x_i\\ n &= \pi\left(\sum_{i=1}^{n}x_i+n\right)\\ \pi &= \frac{n}{\sum_{i=1}^{n}x_i+n} = \frac{1}{1+\bar{x}}\\ \Longrightarrow \hat{\pi}\left(\mathbf{X}\right) &= \frac{1}{1+\overline{X}}\end{aligned}$$

(c)

```
> X <- c(8, 1, 2, 0, 0, 0, 2, 1, 3, 3)
> mean(X)

[1] 2
```

The mean of the sample given is 2 ($\bar{x}=2$). Since $\tilde{\pi}=\frac{1}{1+\overline{X}}=\hat{\pi}\left(\mathbf{X}\right)$, both estimates of π are $^1/_3$.

Solution for 15:

(a) First, calculate the maximum likelihood estimator

$$\begin{aligned}L(\mu|\mathbf{x}) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\mu}}\,e^{-\frac{(x_i-\mu)^2}{2\mu}}\\ \ln L(\mu|\mathbf{x}) &= \sum_{i=1}^{n}\left[-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\ln(\mu)-\frac{1}{2\mu}(x_i-\mu)^2\right]\\ &= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\mu)-\frac{1}{2\mu}\left(\sum_{i=1}^{n}x_i^2+n\mu^2-2\mu\sum_{i=1}^{n}x_i\right)\\ &= -\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\mu)-\frac{1}{2\mu}\sum_{i=1}^{n}x_i^2-\frac{n\mu}{2}+\sum_{i=1}^{n}x_i.\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ln L(\mu|\mathbf{x})}{\partial \mu} &= -\frac{n}{2\mu} + \frac{\sum_{i=1}^{n} x_i^2}{2\mu^2} - \frac{n}{2} \overset{\text{set}}{=} 0 \\
-2\mu^2 \left(-\frac{n}{2\mu} + \frac{\sum_{i=1}^{n} x_i^2}{2\mu^2} - \frac{n}{2} \right) &= 0 \\
n\mu - \sum_{i=1}^{n} x_i^2 + n\mu^2 &= 0 \\
\implies \mu &= \frac{-n \pm \sqrt{n^2 + 4n \sum_{i=1}^{n} x_i^2}}{2n}
\end{aligned}$$

Since μ must be greater than zero, $\hat{\mu}(\mathbf{X}) = \frac{-n+\sqrt{n^2+4n\sum_{i=1}^{n} X_i^2}}{2n}$.

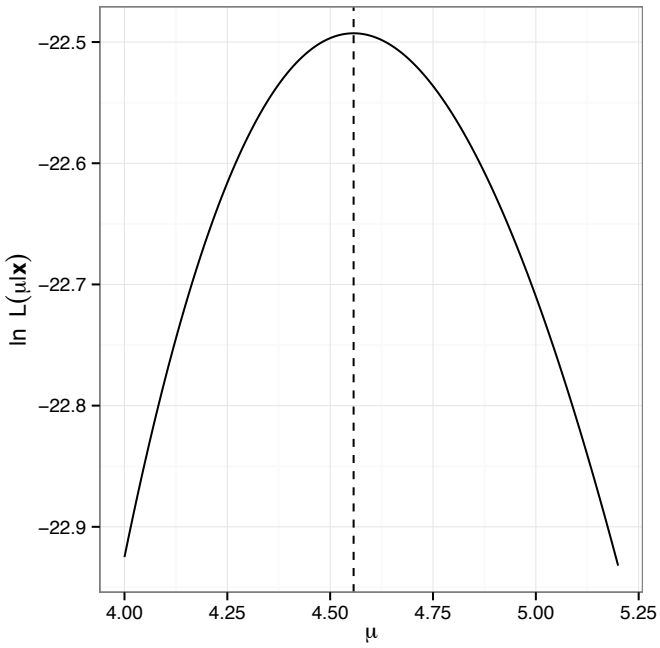
```
> xs <- c(4.37, 9.30, 1.67, 1.25, 4.30, 6.97, 2.68, 5.49, 4.36, 4.46)
> n <- length(xs)
> mle <- (-n + sqrt(n^2 + 4*n*sum(xs^2)))/(2*n)
> mle

[1] 4.557004

> loglike <- function(mu){
+   -n/2*log(2*pi) - n/2*log(mu) - 1/(2*mu)*sum(xs^2) - n*mu/2 + sum(xs)
+ }
> negloglike <- function(mu){
+   n/2*log(2*pi) + n/2*log(mu) + 1/(2*mu)*sum(xs^2) + n*mu/2 - sum(xs)
+ }
> EM <- nlm(f = negloglike, p = 4)
> EM$estimate

[1] 4.557004
```

The maximum likelihood estimate is $\hat{\mu} = \frac{-10+\sqrt{10^2+4(10)(253.2329)}}{2(10)} = 4.557$

(b) R Code 7.1 creates a plot of the log likelihood function from 4 to 5.2.

R Code 7.1

```
> ggplot(data = data.frame(x = c(4, 5.2)), aes(x = x)) +
+   stat_function(fun = loglike, n = 500) +
+   theme_bw() +
+   labs(x = expression(mu),
+        y = expression(textstyle(ln)~~L(mu*"|"*bold(x)))) +
+   geom_vline(xintercept = 4.557, lty = "dashed")
```



Solution for 17:

(a)

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n}(\theta+1)(1-x_i)^{\theta}$$

$$\ln L(\theta|\mathbf{x}) = \sum_{i=1}^{n}\ln(\theta+1) + \theta\sum_{i=1}^{n}\ln(1-x_i)$$

$$\frac{\partial \ln L(\theta|\mathbf{x})}{\partial\theta} = \frac{n}{\theta+1} + \sum_{i=1}^{n}\ln(1-x_i) \stackrel{\text{set}}{=} 0$$

Solve for θ:

$$\begin{aligned}
\frac{n}{\theta+1} + \sum_{i=1}^{n} \ln(1-x_i) &= 0 \\
n + (\theta+1)\sum_{i=1}^{n} \ln(1-x_i) &= 0 \\
\theta \sum_{i=1}^{n} \ln(1-x_i) &= -\left(n + \sum_{i=1}^{n} \ln(1-x_i)\right) \\
\theta &= \frac{-n}{\sum_{i=1}^{n} \ln(1-x_i)} - 1 \\
\Longrightarrow \hat{\theta}(\mathbf{X}) &= \frac{-n}{\sum_{i=1}^{n} \ln(1-X_i)} - 1
\end{aligned}$$

(b) To generate values from $f(x)$, an expression for x in terms of a uniform distribution must be calculated.

$$\begin{aligned}
F_X(x) &= \int_0^x (\theta+1)(1-t)^\theta \, dt \\
&= -(1-t)^{\theta+1}\Big|_0^x \\
&= -(1-x)^{\theta+1} + 1
\end{aligned}$$

Next, solve for x in terms of u where $u = F_X(x)$:

$$\begin{aligned}
u &= -(1-x)^{\theta+1} + 1 \\
(1-x)^{\theta+1} &= 1-u \\
1-x &= (1-u)^{\frac{1}{\theta+1}} \\
x &= 1-(1-u)^{\frac{1}{\theta+1}}
\end{aligned}$$

The code to perform the simulation is

```
> set.seed(3)
> n <- 20000
> u <- runif(n, 0, 1)
> theta <- 5
> x <- 1 - (1 - u)^(1/(theta + 1))
> mle <- -n/sum(log(1-x)) - 1
> mle

[1] 4.97824

> PE <- abs(mle - theta)/theta*100
> PE

[1] 0.4351948
```

(c) The mle $= 4.9782$ is within 0.4352% of the value of $\theta = 5$.

Solution for 19:

If $X \sim Exp(\theta)$, $f(x) = \frac{1}{\theta}e^{-x/\theta}$, $E[X] = \theta$, and $Var[X] = \theta^2$.

(a)

$$
\begin{array}{llll}
E\left[\hat{\theta}_1\right] = E[X_1] & = \theta & Var\left[\hat{\theta}_1\right] = Var[X_1] & = \theta^2 \\
E\left[\hat{\theta}_2\right] = E\left[\frac{X_1+X_2}{2}\right] & = \theta & Var\left[\hat{\theta}_2\right] = Var\left[\frac{X_1+X_2}{2}\right] & = \frac{2\theta^2}{4} = \frac{\theta^2}{2} \\
E\left[\hat{\theta}_3\right] = E\left[\frac{\sum_{i=1}^n X_i}{n}\right] & = \theta & Var\left[\hat{\theta}_3\right] = Var\left[\frac{\sum_{i=1}^n X_i}{n}\right] & = \frac{\theta^2}{n} \\
E\left[\hat{\theta}_4\right] = E[\min\{X_1, X_2, X_3\}] & = \frac{\theta}{3} & Var\left[\hat{\theta}_4\right] = Var[\min\{X_1, X_2, X_3\}] & = \frac{\theta^2}{9}
\end{array}
$$

Calculations for $\hat{\theta}_4 = \min\{X_1, X_2, X_3\}$:

$$
\begin{aligned}
F(x) = \mathbb{P}(X \leq x) &= \mathbb{P}\left(\min\{X_1, X_2, X_3\} \leq x\right) \\
&= 1 - \mathbb{P}\left(\min\{X_1, X_2, X_3\} > x\right) \\
&= 1 - \mathbb{P}\left(X_1 > x, X_2 > x, X_3 > x\right) \\
&= 1 - \prod_{i=1}^{3} \mathbb{P}\left(X_i > x\right) \\
&= 1 - \prod_{i=1}^{3} \left(1 - F(x)\right) \\
&= 1 - \prod_{i=1}^{3} \left(1 - \left(1 - e^{-x/\theta}\right)\right) \\
&= 1 - \prod_{i=1}^{3} e^{-x/\theta} = 1 - e^{-3x/\theta} \\
\implies f(x) &= \frac{3}{\theta}e^{-3x/\theta}, \text{ which is an exponential density with parameter } \frac{3}{\theta}
\end{aligned}
$$

(b) For an estimator to be considered efficient, its variance must equal the CRLB. That is

$$
\begin{aligned}
Var\left[\hat{\theta}(\mathbf{X})\right] &\stackrel{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]} \\
&= \frac{1}{n \cdot E\left[\left(\frac{\partial \ln\left(\frac{1}{\theta} e^{-X/\theta}\right)}{\partial \theta}\right)^2\right]} \\
&= \frac{1}{n \cdot E\left[\left(\frac{\partial}{\partial \theta}\left(-\ln(\theta) - \frac{X}{\theta}\right)\right)^2\right]} \\
&= \frac{1}{n \cdot E\left[\left(\left(-\frac{1}{\theta} + \frac{X}{\theta^2}\right)\right)^2\right]} \\
&= \frac{\theta^4}{n \cdot E\left[(X-\theta)^2\right]} \\
&= \frac{\theta^4}{n \cdot Var[X]} \\
&= \frac{\theta^4}{n \cdot \theta^2} = \frac{\theta^2}{n}
\end{aligned}
$$

Since $Var\left[\hat{\theta}_3(\mathbf{X})\right] = \frac{\theta^2}{n}$, it is efficient.

(c) $\hat{\theta}_3(\mathbf{X})$ is the MLE because it is efficient.

(d) If $X \sim Exp(\theta + 2)$, $E[X] = \theta + 2$. To create an unbiased estimator of θ, a statistic which has an expected value of $\theta - 2$ can be used. Any of θ_1, θ_2, or θ_3 minus 2 will yield an unbiased estimator of θ.

Solution for 21:

(a) Since $f(x) = \lambda e^{-\lambda x}$, the likelihood function is

$$L(\lambda|\mathbf{x}) = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

The likelihood for different values of λ is computed in R Code 7.2.

R Code 7.2

```
> lifetimes <- c(3.70, 1.76, 3.63, 15.73, 5.85, 0.20, 9.87,
+                14.55, 0.43, 2.46, 0.45, 5.09, 10.53, 12.41,
+                3.19, 3.41, 3.80, 1.66, 0.40, 1.10)
> like <- function(lifetimes, lambda){
+   prod(dexp(lifetimes, lambda))
+ }
> like(lifetimes, 1/6)

[1] 1.523384e-23

> like(lifetimes, 1/4)

[1] 1.195506e-23
```

$$L\left(\lambda = \frac{1}{6}\middle|\,\mathbf{x}\right) = \left(\frac{1}{6}\right)^{20} e^{-\frac{1}{6}(100.22)} = 0.$$

$$L\left(\lambda = \frac{1}{4}\middle|\,\mathbf{x}\right) = \left(\frac{1}{4}\right)^{20} e^{-\frac{1}{4}(100.22)} = 0.$$

Given the observed values from the accelerated test, it is more likely that the mean life of the resistor is 6 years.

(b) The maximum likelihood estimator is calculated and used to determine $\hat{\lambda}(\mathbf{x})$'s value for this sample.

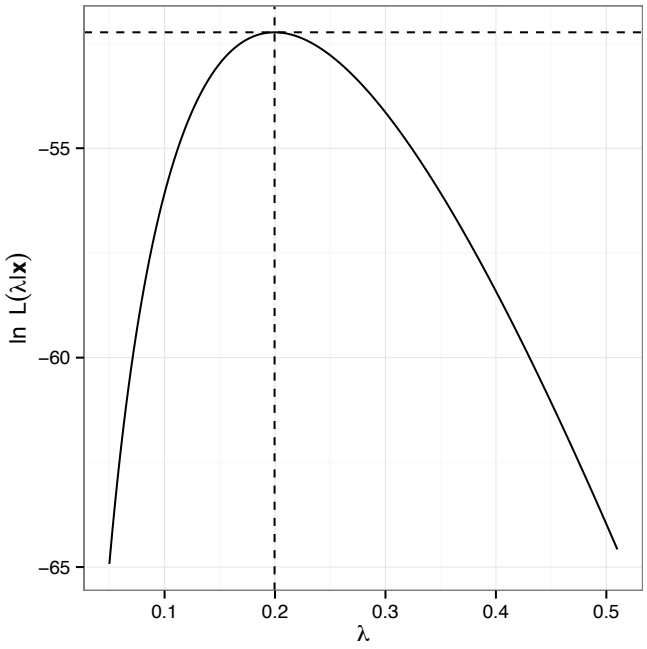
$$\begin{aligned} L(\lambda|\mathbf{x}) &= \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \\ \ln L(\lambda|\mathbf{x}) &= n\ln(\lambda) - \lambda \sum_{i=1}^n x_i \\ \frac{\partial}{\partial\lambda} \ln L(\lambda|\mathbf{x}) &= \frac{n}{\lambda} - \sum_{i=1}^n x_i \stackrel{\text{set}}{=} 0 \\ \Longrightarrow n &= \lambda \sum_{i=1}^n x_i \\ \lambda &= \frac{n}{\sum_{i=1}^n x_i} \\ \Longrightarrow \hat{\lambda}(\mathbf{X}) &= \frac{1}{\overline{X}} \end{aligned}$$

```
> mle <- 1/mean(lifetimes)
> mle

[1] 0.199561
```

For this sample, $\hat{\lambda}(\mathbf{x}) = 0.1996$.

(c)

```
> loglike <- function(lambda){
+   n <- length(lifetimes)
+   n*log(lambda) - lambda*sum(lifetimes)
+ }
> MLL <- optimize(loglike, c(0.05, 0.51), maximum = TRUE)$objective
> MLL

[1] -52.23271

> ggplot(data = data.frame(x = c(0.05, 0.51)), aes(x = x)) +
+   stat_function(fun = loglike, n = 500) +
+   labs(x = expression(lambda),
+        y = expression(textstyle(ln)~~L(lambda*"|"*bold(x))) ) +
+   theme_bw() +
+   geom_vline(xintercept = 1/mean(lifetimes), lty = "dashed") +
+   geom_hline(yintercept = MLL, lty = "dashed")
```



Solution for 23:

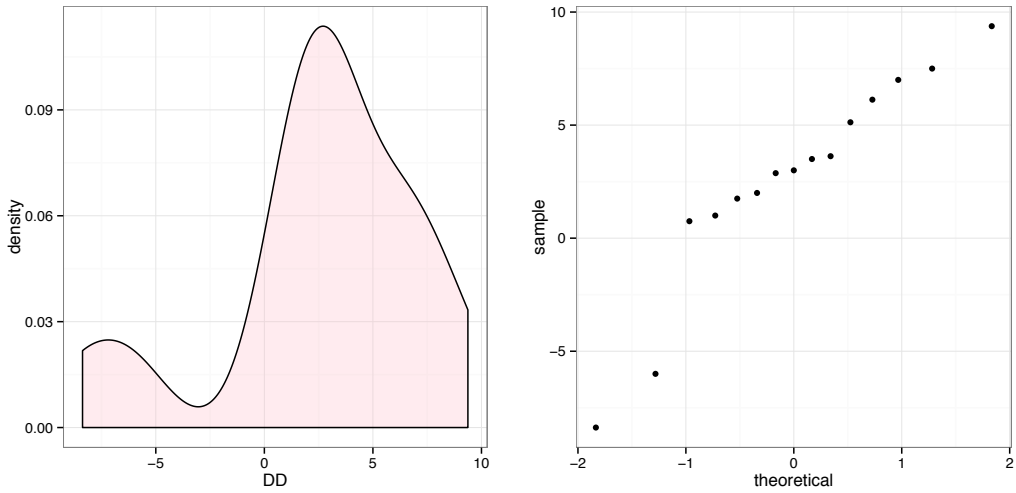
(a)

```
> TwoColumns <- unstack(FERTILIZE, height~fertilization)
> ThreeColumns <- within(data = TwoColumns, expr ={
+   DD <- cross - self
+ })
> head(ThreeColumns)

   cross   self     DD
1 23.500 17.375  6.125
2 12.000 20.375 -8.375
3 21.000 20.000  1.000
4 22.000 20.000  2.000
5 19.125 18.375  0.750
6 21.500 18.625  2.875
```

(b)

```
> ggplot(data = ThreeColumns, aes(x = DD)) +
+   geom_density(fill = "pink", alpha = 0.3) +
+   theme_bw()
> ggplot(data = ThreeColumns, aes(sample = DD)) +
+   stat_qq() +
+   theme_bw()
> shapiro.test(ThreeColumns$DD)


Shapiro-Wilk normality test
```

```
data:  ThreeColumns$DD
W = 0.90079, p-value = 0.09785
```



Based on the density estimate, the quantile-quantile plot, and the Shapiro-Wilk normality test (using an α level of 0.05) one might assume normality for the distribution of `DD`.

(c)

```
> library(MASS)
> ans <- fitdistr(x = ThreeColumns$DD, densfun = "normal")
> ans

     mean         sd
  2.6166667   4.5580667
 (1.1768878) (0.8321853)

> MU <- ans$estimate[1]
> MU

    mean
2.616667

> SIGMA <- ans$estimate[2]
> SIGMA

      sd
4.558067
```

The maximum likelihood estimates of μ and σ using the function `fitdistr()` assuming a normal distribution are: 2.6167 and 4.5581, respectively.

(d) The following code verifies that the numbers 2.6167 and 4.5581 are the sample mean and the uncorrected sample standard deviation of `DD`.

```
> with(data = ThreeColumns,
+      c(mean(DD), sqrt(var(DD)*(length(DD) - 1)/length(DD)) )
+ )

[1] 2.616667 4.558067
```

Solution for 25:

(a) The MLEs for the mean and variance of W are

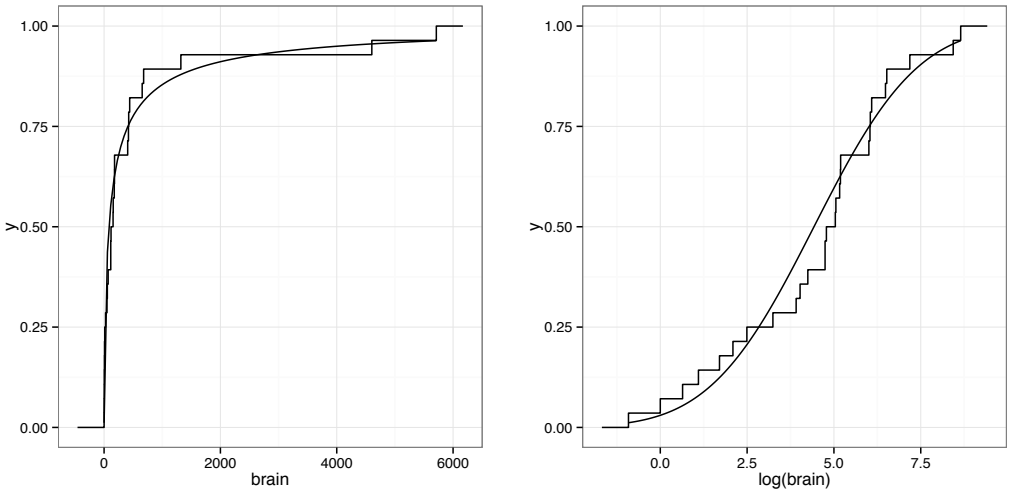
$$\widehat{E}[w] = e^{\overline{X}+S_u^2/2} \quad \text{and} \quad \widehat{Var}[w] = e^{2\overline{X}+S_u^2}(e^{S_u^2} - 1).$$

```
> library(MASS)
> mle <- fitdistr(x = Animals$brain, densfun = "lognormal")
> mle

    meanlog      sdlog
  4.4254457   2.3560485
 (0.4452513) (0.3148402)

> EW <- exp(mle$estimate[1] + mle$estimate[2]^2/2)
> VW <- exp(2*mle$estimate[1] + mle$estimate[2]^2)*
+   (exp(mle$estimate[2]^2) - 1)
> c(EW, VW)

     meanlog      meanlog
1.340674e+03 4.610096e+08
```

The maximum likelihood estimates of the mean and standard deviation for the logarithm of the variable `brain` are 4.4254 and 2.356, respectively. It follows using invariance properties of MLEs that the estimates of the mean and variance of `brain` are 1340.6744 kg and 461009623.032 kg^2, respectively.

(b)

```
> ggplot(data = Animals, aes(x = brain)) + stat_ecdf() +
+   stat_function(fun = plnorm,
+                 args = list(mle$estimate[1], mle$estimate[2])) +
+   theme_bw()
> #
> ggplot(data = Animals, aes(x = log(brain))) + stat_ecdf() +
+   stat_function(fun = pnorm,
+                 args = list(mle$estimate[1], mle$estimate[2])) +
+   theme_bw()
```



It seems reasonable to assume **brain** follows a lognormal distribution based on the graphs.

(c) In agreement with part (a), the estimated mean and variance of the variable **brain** are 1340.6744 kg and 461009623.032 kg^2, respectively.

```
> ans <- with(data = Animals, c(mean(log(brain)), var(log(brain))))
> ans

[1] 4.425446 5.756556

> xbar <- ans[1]
> V <- ans[2]
> c(xbar, V)

[1] 4.425446 5.756556

> VU <- V*27/28
> EW <- exp(xbar + VU/2)
> VW <- exp(2*xbar + VU)*(exp(VU) - 1)
> c(EW, VW)

[1] 1.340674e+03 4.610096e+08
```

(d) Repeating parts (a)-(c) after removing the dinosaurs.

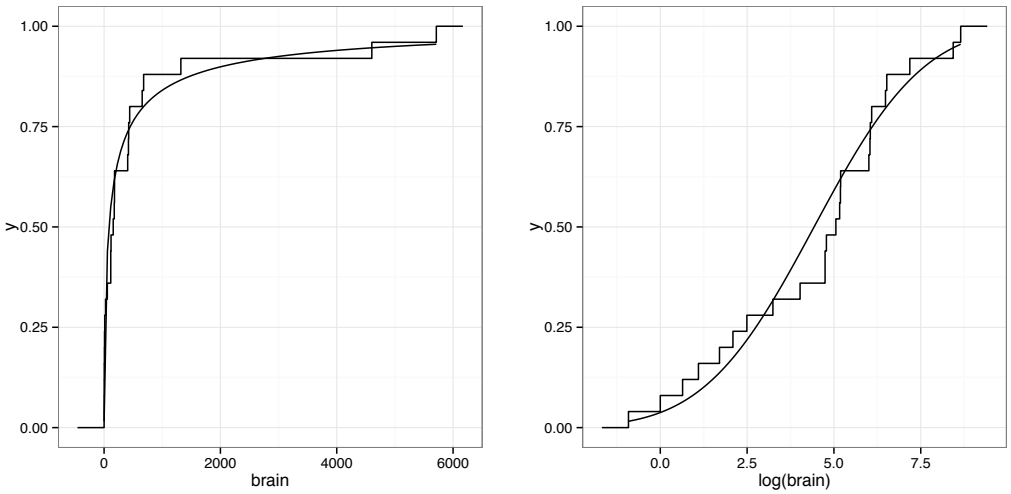
R Code 7.3

```
> NoDinos <- subset(x = Animals, subset = body < 9400)
> mle <- fitdistr(x = NoDinos$brain, densfun = "lognormal")
> mle

    meanlog       sdlog
  4.4284706   2.4880028
 (0.4976006) (0.3518567)

> EW <- exp(mle$estimate[1] + mle$estimate[2]^2/2)
```

```
> VW <- exp(2*mle$estimate[1] + mle$estimate[2]^2)*
+   (exp(mle$estimate[2]^2) - 1)
> c(EW, VW)

     meanlog      meanlog
1.851127e+03 1.668525e+09
```

After removing the three dinosaurs, the estimated mean and variance for the variable **brain** are 1851.1266 kg and 1668525179.2198 kg^2, respectively.

```
> ggplot(data = NoDinos, aes(x = brain)) + stat_ecdf() +
+   stat_function(fun = plnorm,
+                 args = list(mle$estimate[1], mle$estimate[2])) +
+   theme_bw()
> #
> ggplot(data = NoDinos, aes(x = log(brain))) + stat_ecdf() +
+   stat_function(fun = pnorm,
+                 args = list(mle$estimate[1], mle$estimate[2])) +
+   theme_bw()
```



It seems reasonable to assume **brain** still follows a lognormal distribution after removing the three dinosaurs based on the graphs. In agreement with R Code 7.3 on the facing page, the estimated mean and variance of the variable **brain** are 1851.1266 kg and 1668525179.2198 kg^2, respectively.

```
> ans <- with(data = NoDinos, c(mean(log(brain)), var(log(brain))))
> ans

[1] 4.428471 6.448081

> xbar <- ans[1]
> V <- ans[2]
> c(xbar, V)

[1] 4.428471 6.448081
```

```
> VU <- V*(27 - 3)/(28 - 3)
> EW <- exp(xbar + VU/2)
> VW <- exp(2*xbar + VU)*(exp(VU) - 1)
> c(EW, VW)

[1] 1.851127e+03 1.668525e+09
```

After removing the three dinosaurs, the estimate of the mean brain weight has increased while the estimate of the variance of the brain weight has decreased.

Solution for 27:

(a)

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{2\theta^3} x_i^2 e^{-\frac{x_i}{\theta}} \\ &= \frac{1}{2^n \theta^{3n}} \prod_{i=1}^{n} x_i^2 e^{-\frac{x_i}{\theta}} \\ \ln L(\theta|\mathbf{x}) &= -n \ln 2 - 3n \ln \theta + 2 \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i}{\theta} \\ \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} &= -\frac{3n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} \stackrel{\text{set}}{=} 0 \end{aligned}$$

$$\begin{aligned} \implies -3n\theta &= -\sum_{i=1}^{n} x_i \\ \theta &= \frac{\sum_{i=1}^{n} x_i}{3n} \\ \implies \hat{\theta}(\mathbf{X}) &= \frac{\overline{X}}{3} \end{aligned}$$

To verify this is a maximum value, take the second partial and show it is less than zero.

$$\begin{aligned} \frac{\partial^2 \ln L(\theta|\mathbf{x})}{\partial \theta^2} &= \frac{3n}{\theta^2} - 2\frac{\sum_{i=1}^{n} x_i}{\theta^3}\bigg|_{\hat{\theta}=\frac{\bar{x}}{3}} \\ &= \frac{27n}{\bar{x}^2} - \frac{2n\bar{x}}{\bar{x}^3/27} \\ &= \frac{27n}{\bar{x}^2} - \frac{54n}{\bar{x}^2} < 0 \text{ for all } n. \end{aligned}$$

(b) Show that $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{3}$ is unbiased and efficient.
Unbiased:

$$E[X] = \int_0^\infty x \cdot \frac{1}{2\theta^3} x^2 e^{-x/\theta}\, dx$$

Let $u = \frac{x}{\theta}$, and note that $du = \frac{dx}{\theta}$

$$\begin{aligned} &= \int_0^\infty \frac{u^3}{2} e^{-u} \theta \, du \\ &= \frac{\theta}{2}\Gamma(4) \\ &= 3!\frac{\theta}{2} \\ E[X] &= 3\theta \end{aligned}$$

$$E\left[\hat{\theta}(\mathbf{X})\right] = E\left[\frac{\overline{X}}{3}\right] = \frac{\sum_{i=1}^n E[X_i]}{3n} = \frac{n \cdot 3\theta}{3n} = \theta$$

Therefore, $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{3}$ is an unbiased estimator for θ.

For later calculations, $E[X^2]$ is also needed.

$$E[X^2] = \int_0^\infty x^2 \cdot \frac{1}{2\theta^3} x^2 e^{-x/\theta} \, dx$$

Let $u = \frac{x}{\theta}$, and note that $du = \frac{dx}{\theta}$

$$\begin{aligned} &= \theta \int_0^\infty \frac{u^4}{2} e^{-u} \theta \, du \\ &= \frac{\theta^2}{2}\Gamma(5) \\ &= 4!\frac{\theta^2}{2} \\ E[X^2] &= 12\theta^2 \end{aligned}$$

$Var[X] = E[X^2] - (E[X])^2 = 12\theta^2 - (3\theta)^2 = 3\theta^2$

For an estimator to be considered efficient, its variance must equal the CRLB. That is

$$Var\left[\hat{\theta}(\mathbf{X})\right] \stackrel{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial \ln f(X|\theta)}{\partial \theta}\right)^2\right]}$$

$$Var\left[\frac{\overline{X}}{3}\right] \stackrel{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial \ln\left(\frac{1}{2\theta^3} X^2 e^{-X/\theta}\right)}{\partial \theta}\right)^2\right]}$$

$$\frac{1}{9} \cdot Var\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] \stackrel{?}{=} \frac{1}{n \cdot E\left[\left(\frac{\partial}{\partial \theta}\left(-\ln 2 - 3\ln\theta + 2\ln X - \frac{X}{\theta}\right)\right)^2\right]}$$

$$\frac{1}{9n^2} \cdot Var\left[\sum_{i=1}^{n} X_i\right] \stackrel{?}{=} \frac{1}{n \cdot E\left[\left(-\frac{3}{\theta} + \frac{X}{\theta^2}\right)^2\right]}$$

$$\frac{1}{9n^2} \cdot \sum_{i=1}^{n} Var[X_i] \stackrel{?}{=} \frac{\theta^4}{n \cdot E\left[(X - 3\theta)^2\right]}$$

$$\frac{1}{9n^2} \cdot n \cdot 3\theta^2 \stackrel{?}{=} \frac{\theta^4}{n \cdot Var[X]}$$

$$\frac{\theta^2}{3n} = \frac{\theta^4}{n \cdot 3\theta^2} = \frac{\theta^2}{3n}$$

So, $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{3}$ is an efficient estimator of θ.

(c) To derive the method of moments estimator, solve

$$\alpha_1 = E[X] = 3\theta \stackrel{\text{set}}{=} \overline{X} = m_1$$

$$\theta = \frac{\overline{X}}{3}$$

$$\Longrightarrow \tilde{\theta} = \frac{\overline{X}}{3}$$

(d)

```
> waiting <- c(6, 12, 15, 14, 12, 10, 8, 9, 10, 9, 8, 7, 10, 7, 3)
> xbar <- mean(waiting)
> mle <- xbar/3
> c(xbar, mle)

[1] 9.333333 3.111111
```

When $\hat{\theta}(\mathbf{X}) = \frac{\overline{X}}{3}$, $\hat{\theta}(\mathbf{x}) = \frac{\bar{x}}{3} = \frac{9.3333}{3} = 3.1111$

Solution for 29:

(a) To be a density, $\int_{-\infty}^{\infty} f(x)\,dx$ must equal 1.

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{\theta}^{\infty} \frac{\theta}{x^2}\,dx = -\theta x^{-1}\Big|_{\theta}^{\infty} = 0 - \frac{-\theta}{\theta} = 1$$

So, $f(x)$ is a density function.

(b)

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{\theta}{x_i^2} \\ &= \theta^n \prod_{i=1}^{n} \frac{1}{x_i^2} \\ \ln L(\theta|\mathbf{X}) &= n\ln(\theta) - 2\sum_{i=1}^{n} \ln(x_i) \\ \frac{\partial L(\theta|\mathbf{X})}{\partial \theta} &= \frac{n}{\theta} \stackrel{\text{set}}{=} 0 \end{aligned}$$

Solving for θ directly cannot be accomplished.

(c) No.

(d) Since $E[X] = \int_\theta^\infty x \cdot \frac{\theta}{x^2}\, dx = \theta \ln(x)|_\theta^\infty = \infty$, the method of moments cannot be used to find an estimator of θ.

Solution for 31:

(a) Deriving the maximum likelihood estimator of θ:

$$\begin{aligned} L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \frac{1}{\theta} x_i^{\frac{1-\theta}{\theta}} \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} x_i^{\frac{1-\theta}{\theta}} \\ \ln L(\theta|\mathbf{x}) &= -n\ln\theta + \frac{1-\theta}{\theta}\sum_{i=1}^{n} \ln(x_i) \\ \frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} &= -\frac{n}{\theta} + \frac{(-1)\cdot\theta - 1(1-\theta)}{\theta^2}\sum_{i=1}^{n} \ln(x_i) \stackrel{\text{set}}{=} 0 \end{aligned}$$

$$\begin{aligned} -\frac{n}{\theta} + \frac{(-1)\cdot\theta - 1(1-\theta)}{\theta^2}\sum_{i=1}^{n} \ln(x_i) &= 0 \\ \frac{-\theta - 1 + \theta}{\theta^2}\sum_{i=1}^{n} \ln(x_i) &= \frac{n}{\theta} \\ -\sum_{i=1}^{n} \ln(x_i) &= n\theta \\ \Longrightarrow \theta &= \frac{-\sum_{i=1}^{n} \ln(x_i)}{n} \\ \Longrightarrow \hat{\theta}(\mathbf{X}) &= \frac{-\sum_{i=1}^{n} \ln(X_i)}{n} \end{aligned}$$

(b) Deriving the method of moments estimator of θ:

$$\begin{aligned}
\alpha_1(\theta) = E[X] &\stackrel{\text{set}}{=} \overline{X} = m_1 \\
\int_0^1 x \cdot \frac{1}{\theta} x^{\frac{1-\theta}{\theta}}\, dx &= \overline{X} \\
\frac{1}{\theta} \int_0^1 x^{\frac{1}{\theta}} &= \overline{X} \\
\frac{1}{\theta} \cdot \frac{\theta}{1+\theta} x^{\frac{1+\theta}{\theta}} \Big|_0^1 &= \overline{X} \\
\frac{1}{1+\theta} &= \overline{X} \\
\frac{1}{\overline{X}} &= 1+\theta \\
\theta &= \frac{1}{\overline{X}} - 1 \\
\Longrightarrow \tilde{\theta}(\mathbf{X}) &= \frac{1-\overline{X}}{\overline{X}}
\end{aligned}$$

(c) If $E\big[\hat{\theta}(\mathbf{X})\big] = \theta$, the MLE is unbiased.

$$\begin{aligned}
E\big[\hat{\theta}(\mathbf{X})\big] &= E\left[\frac{-\sum_{i=1}^n \ln(X_i)}{n}\right] \\
&= \frac{-1}{n} \sum_{i=1}^n E[\ln(X_i)] \\
&= -1E[\ln(X)]
\end{aligned}$$

Recall that to integrate by parts, $\int_a^b u\, dv = uv|_a^b - \int_a^b v\, du$.
Calculating $E[\ln(X)]$:

$$E[\ln(X)] = \int_0^1 \ln(x) \frac{1}{\theta} x^{\frac{1-\theta}{\theta}}\, dx$$

Let $u = \ln(x)$ and $dv = \frac{1}{\theta} x^{\frac{1-\theta}{\theta}}\, dx$, so $du = \frac{1}{x} dx$ and $v = x^{\frac{1}{\theta}}$

$$\begin{aligned}
&= \ln(x) \cdot x^{\frac{1}{\theta}} \Big|_0^1 - \int_0^1 x^{\frac{1}{\theta}} \cdot \frac{1}{x} dx \\
&= 0 - \theta x^{\frac{1}{\theta}} \Big|_0^1 = -\theta
\end{aligned}$$

So, $E\big[\hat{\theta}(\mathbf{X})\big] = -1E[\ln(X)] = -1(-\theta) = \theta$, and the MLE is unbiased.

Solution for 33:

(a) To find the MLE of θ, take the partial of the log likelihood function, set it equal to zero, and solve for θ.

$$
\begin{aligned}
L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} \theta \left(\frac{1}{x}\right)^{\theta+1} \\
&= \theta^n \prod_{i=1}^{n} \frac{1}{x_i^{\theta+1}} \\
\ln L(\theta|\mathbf{x}) &= n \ln \theta - (\theta+1) \sum_{i=1}^{n} \ln x_i \\
\frac{\partial \ln L(\theta|\mathbf{x})}{\partial \theta} &= \frac{n}{\theta} - \sum_{i=1}^{n} \ln x_i \overset{\text{set}}{=} 0 \\
\implies \theta &= \frac{n}{\sum_{i=1}^{n} \ln x_i} \\
\implies \hat{\theta}(\mathbf{X}) &= \frac{n}{\sum_{i=1}^{n} \ln X_i}
\end{aligned}
$$

(b) Calculating the MOM estimator:

$$
\begin{aligned}
\alpha_1 = E[X] &\overset{\text{set}}{=} \overline{X} = m_1 \\
\int_1^{\infty} x \cdot \theta \left(\frac{1}{x}\right)^{\theta+1} &= \overline{X} \\
\theta \int_1^{\infty} x^{-\theta} &= \overline{X} \\
\theta \frac{x^{-\theta+1}}{1-\theta}\Big|_1^{\infty} &= \overline{X} \\
\frac{\theta}{\theta-1} &= \overline{X} \\
\theta &= \theta\overline{X} - \overline{X} \\
\theta &= \frac{-\overline{X}}{1-\overline{X}} = \frac{\overline{X}}{\overline{X}-1} \\
\implies \tilde{\theta} &= \frac{\overline{X}}{\overline{X}-1}
\end{aligned}
$$

(c)

```
> x <- c(2, 3, 2, 2.5, 1, 2, 2, 3, 1, 4, 6, 3, 4.4)
> mle <- length(x)/sum(log(x))
> mom <- mean(x)/(mean(x) - 1)
> c(mle, mom)

[1] 1.116419 1.567686
```

(d) The mean of the distribution is $E[X] = \frac{\theta}{\theta-1}$ from part (b).

(e)

$$
\widehat{E[X]} = \frac{\hat{\theta}(\mathbf{x})}{\hat{\theta}(\mathbf{x}) - 1} = \frac{1.1164}{1.1164 - 1} = 9.5897
$$

Solution for 35:

(a) To generate a random sample from the distribution $F_X(x)$, the relationship between $X \sim F_X(x)$ and $U \sim Unif[0,1]$ must be determined.

$$F_X(x) = \int_0^x 3\pi\theta x^2 e^{-\theta\pi x^3}\, dx = -e^{-\theta\pi x^3}\Big|_0^x = -e^{-\theta\pi x^3} - (-1)$$

Solve for x in terms of u:

$$\begin{aligned} u &\stackrel{\text{set}}{=} 1 - e^{-\theta\pi x^3} \\ e^{-\theta\pi x^3} &= 1 - u \\ -\theta\pi x^3 &= \ln(1-u) \\ x &= \left[\frac{\ln(1-u)}{-\theta\pi}\right]^{1/3} \end{aligned}$$ Note that $\theta = 5$ in the code.

R Code 7.4

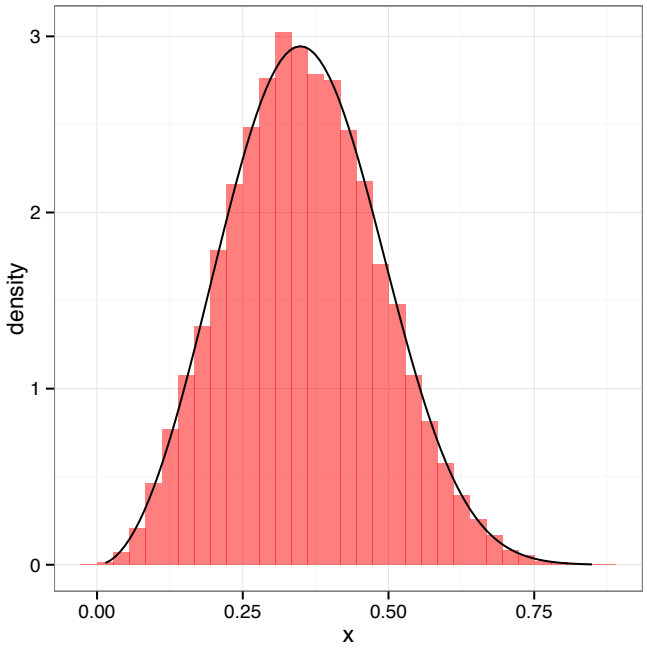
```
> set.seed(102)
> u <- runif(20000, 0, 1)
> x <- (log(1 - u)/(-5*pi))^(1/3)
```

(b) The mean and variance for the values generated in R Code 7.4 are 0.3554 and 0.0168, respectively.

```
> M <- mean(x)
> V <- var(x)
> c(M, V)

[1] 0.3554149 0.0167516
```

(c)

```
> f <- function(x){3*pi*5*x^2*exp(-5*pi*x^3)}
> ggplot(data = data.frame(x = x), aes(x = x)) +
+   geom_histogram(aes(x = x, y = ..density..), fill = "red",
+                  alpha = 0.5) +
+   theme_bw() +
+   stat_function(fun = f)
```



(d) To calculate the MLE of θ, take the partial of the log likelihood, set it equal to zero, and solve for θ.

$$
\begin{aligned}
L(\theta|\mathbf{x}) &= \prod_{i=1}^{n} 3\pi\theta x_i^2 e^{-\theta\pi x_i^3} \\
L(\theta|\mathbf{x}) &= (3\pi\theta)^n e^{-\theta\pi\sum_{i=1}^{n} x_i^3} \prod_{i=1}^{n} x_i^2 \\
\ln L(\theta|\mathbf{x}) &= n\ln(3\pi) + n\ln(\theta) - \theta\pi\sum_{i=1}^{n} x_i^3 + \sum_{i=1}^{n} 2\ln(x_i) \\
\frac{\partial \ln L(\theta|\mathbf{x})}{\partial\theta} &= \frac{n}{\theta} - \pi\sum_{i=1}^{n} x_i^3 \overset{\text{set}}{=} 0
\end{aligned}
$$

$$
\begin{aligned}
\frac{n}{\theta} - \pi\sum_{i=1}^{n} x_i^3 &= 0 \\
n &= \theta\pi\sum_{i=1}^{n} x_i^3 \\
\implies \theta &= \frac{n}{\pi\sum_{i=1}^{n} x_i^3} \\
\implies \hat{\theta}(\mathbf{X}) &= \frac{n}{\pi\sum_{i=1}^{n} X_i^3}
\end{aligned}
$$

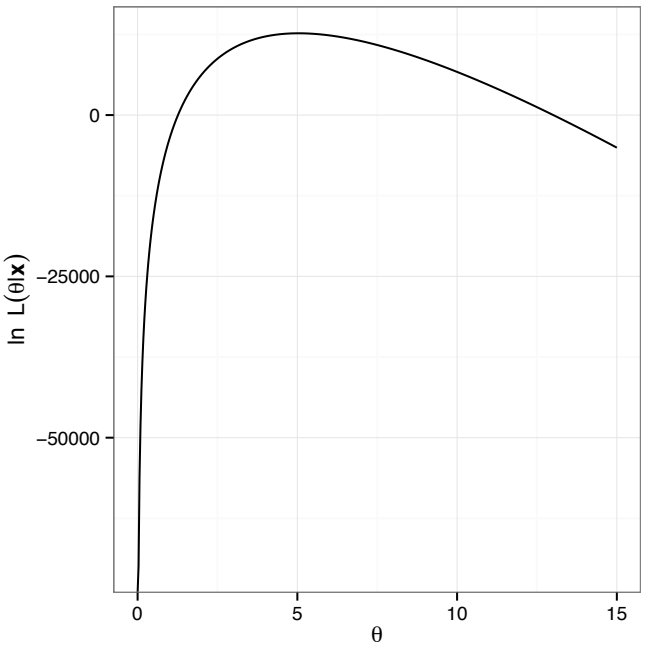
For the generated sample,

```
> n <- length(x)
> mle <- n/(pi*sum(x^3))
> mle

[1] 5.038159
```

$$\hat{\theta}(\mathbf{X}) = \frac{n}{\pi \sum_{i=1}^{n} X_i^3} = \frac{500}{(3.1416)(1263.5961)} = 5.0382.$$

(e)

```
> loglike <- function(theta){
+   n <- length(x)
+   n*log(3*pi*theta) - theta*pi*sum(x^3) + sum(log(x^2))
+ }
> ggplot(data = data.frame(x = c(0, 15)), aes(x = x)) +
+   stat_function(fun = loglike, n = 500) +
+   theme_bw() +
+   labs(x = expression(theta),
+        y = expression(textstyle(ln)~~L(theta*"|"*bold(x))))
```



Solution for 37:

The negative of the log-likelihood function needs to be determined:

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad 0 \le x \le 1$$

$$L(\alpha,\beta|\mathbf{x}) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n \prod_{i=1}^{n} x_i^{\alpha-1}(1-x_i)^{\beta-1}$$

$$\ln L(\alpha,\beta|\mathbf{x}) = n \ln\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right) + (\alpha-1)\sum_{i=1}^{n}\ln(x_i) + (\beta-1)\sum_{i=1}^{n}\ln(1-x_i)$$

The negative of $\ln L(\alpha,\beta|\mathbf{x})$ is what is stored in `negloglike()`.

```
> set.seed(8675)
> n <- 309
> x <- rbeta(n, 3, 2)
> negloglike <- function(p){
+   (-n*log((gamma(p[1] + p[2]))/(gamma(p[1])*gamma(p[2])))
+    - (p[1] - 1)*sum(log(x)) - (p[2] - 1)*sum(log(1 - x)))
+ }
> nlm(f = negloglike, p = c(1, 1))

$minimum
[1] -75.13128

$estimate
[1] 3.081033 2.149784

$gradient
[1]  1.955643e-06 -6.425273e-06

$code
[1] 1

$iterations
[1] 12

> nlm(f = negloglike, p = c(1, 1))$estimate

[1] 3.081033 2.149784

> # Also
> library(MASS)
> fitdistr(x = x, densfun = "beta", start = list(shape1 = 1 , shape2 = 1))

Warning in densfun(x, parm[1], parm[2], ...):  NaNs produced
Warning in densfun(x, parm[1], parm[2], ...):  NaNs produced

    shape1      shape2
  3.0810430   2.1497909
 (0.2422334) (0.1633135)
```

Chapter 8

Odd solutions

Solution for 1:
The interval $[\bar{x}-3,\ \bar{x}+3]$ is a confidence interval since it takes the form of point estimate minus and plus the margin of error.

Solution for 3:
(a) The confidence level is 0.96.

```
> 1 - pnorm(-2.053749)*2

[1] 0.96
```

(b) The confidence level is 0.84.

```
> 1 - pnorm(-1.405072)*2

[1] 0.8400001
```

(c) To answer the problem, one must find the value of $z_{0.005} = -2.5758$.

```
> qnorm(0.005)

[1] -2.575829
```

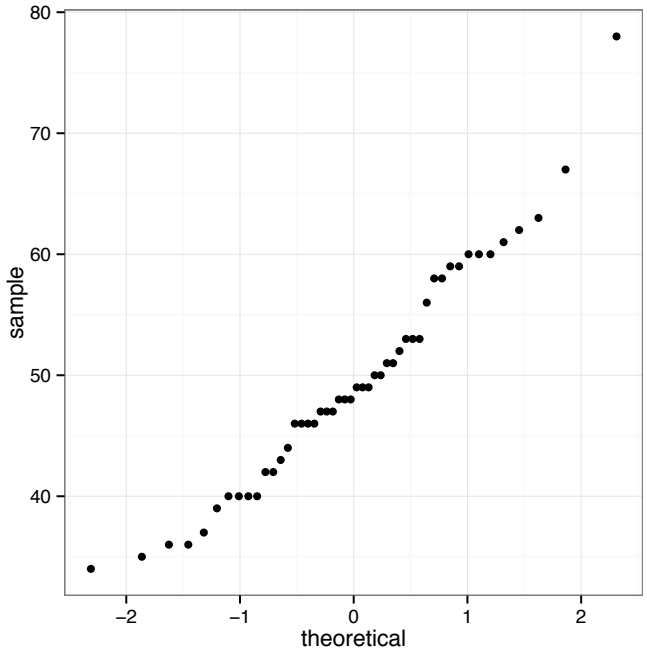
Solution for 5:
The length of a confidence interval for the mean given a normal population with known variance σ^2 is $2 \cdot z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. Reducing the length by a factor of k implies that the ratio of the original length to the new length will equal k.

$$\begin{aligned}
\frac{2 \cdot z_{1-\alpha/2}\frac{\sigma}{\sqrt{n_{\text{original}}}}}{2 \cdot z_{1-\alpha/2}\frac{\sigma}{\sqrt{n_{\text{new}}}}} &= k \\
\frac{\sqrt{n_{\text{original}}}}{\sqrt{n_{\text{new}}}} &= k \\
\implies n_{\text{new}} &= k^2 \cdot n_{\text{original}}
\end{aligned}$$

Therefore, to reduce the length of a confidence interval by a factor of k requires a sample size of $k^2 n_{\text{original}}$. For example, to reduce the length of a confidence interval by a factor of 3 requires a sample size of $9n$.

Solution for 7:
(a) The quantile-quantile plot does not rule out normality.

```
> ggplot(data = SALINITY, aes(sample = salinity)) +
+   stat_qq() +
+   theme_bw()
```



(b)

```
> CI <- t.test(SALINITY$salinity, conf.level = 0.95)$conf
> CI

[1] 46.85025 52.23308
attr(,"conf.level")
[1] 0.95
```

A 95% confidence interval for the mean salinity variation is $[46.8502, 52.2331]$.

Solution for 9:

(a)

```
> p <- 0.5
> alpha <- 0.04
> B <- 0.03
> n <- ceiling(p*(1 - p)*(qnorm(1 - alpha/2)/B)^2)
> n

[1] 1172

> # Or
> nsize(b = B, p = p, conf.level = 0.96, type = "pi")


The required sample size (n) to estimate the population
proportion of successes with a 0.96 confidence interval
so that the margin of error is no more than 0.03 is 1172 ^^H.
```

The required sample size (n) to estimate the population proportion of successes with a 0.96 confidence interval based on an asymptotic confidence interval so that the margin of error is no more than 0.03 is 1172.

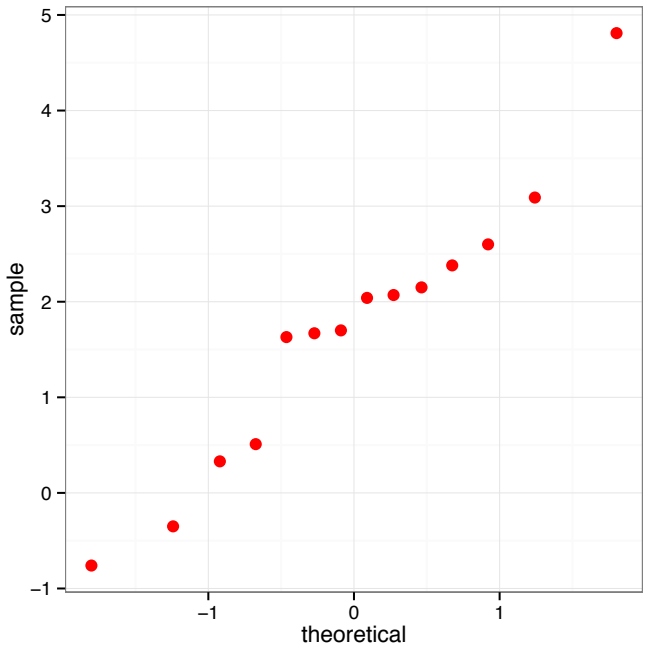
(b) The four 95% confidence intervals for the true proportion of accounts that are paid on time are very similar due to the sample size ($n = 800$).

```
> library(binom)
> binom.confint(x = 650, n = 800, conf.level = 0.95, methods = "asymptotic")

      method   x   n   mean     lower     upper
1 asymptotic 650 800 0.8125 0.7854532 0.8395468

> x <- 650
> n <- 800
> p <- x/n
> z <- qnorm(0.975)
> CIasy <- p + c(-1, 1)*z*sqrt(p*(1 -p)/n)
> CIasy

[1] 0.7854532 0.8395468

> binom.confint(x = 650, n = 800, conf.level = 0.95, methods = "wilson")

  method   x   n   mean     lower   upper
1 wilson 650 800 0.8125 0.7839832 0.83803

> prop.test(x = 650, n = 800, conf.level = 0.95, correct = FALSE)$conf

[1] 0.7839832 0.8380300
attr(,"conf.level")
[1] 0.95

> binom.confint(x = 650, n = 800, conf.level = 0.95, methods = "ac")

         method   x   n   mean     lower    upper
1 agresti-coull 650 800 0.8125 0.7839422 0.838071

> ntilde <- n + qnorm(0.975)^2
> ptilde <- 1/ntilde*(x + 1/2*qnorm(0.975)^2)
> CIac <- ptilde + c(-1, 1)*qnorm(0.975)*sqrt(ptilde*(1 - ptilde)/ntilde)
> CIac

[1] 0.7839422 0.8380710

> binom.confint(x = 650, n = 800, conf.level = 0.95, methods = "exact")

  method   x   n   mean     lower   upper
1  exact 650 800 0.8125 0.7836947 0.83898

> CIcp <- c(qbeta(0.025, x, n - x + 1), qbeta(0.975, x + 1, n - x))
> CIcp

[1] 0.7836947 0.8389800
```

Solution for 11:

(a) Based on the quantile-quantile plot of the differences between the Cosmed and Amatek VO2 values, there is no reason to rule out normality.

```
> COSAMA <- within(data = COSAMA, expr ={
+   DIFF <- cosmed - amatek
+ })
> head(COSAMA)

  subject cosmed amatek DIFF
1       1  31.71  31.20 0.51
2       2  33.96  29.15 4.81
3       3  30.03  27.88 2.15
4       4  24.42  22.79 1.63
5       5  29.07  27.00 2.07
6       6  28.42  28.09 0.33

> ggplot(data = COSAMA, aes(sample = DIFF)) +
+   stat_qq(color = "red", size = 3) +
+   theme_bw()
```



(b) The reported Cosmed and Amatek values are dependent as each subject in the study has both a Cosmed and an Amatek VO2 score.

(c)

```
> CI <- t.test(COSAMA$DIFF)$conf
> CI

[1] 0.8746899 2.5353101
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the average VO2 differenceis $[0.8747, 2.5353]$.

Solution for 13:

(a) The value of the sample mean is 460 liters/day.

```
> xbar1 <- (374.209 + 545.791)/2    # 95%
> xbar2 <- (340.926 + 579.074)/2    # 99%
> xbar3 <- (389.548 + 530.452)/2    # 90%
> c(xbar1, xbar2, xbar3)

[1] 460 460 460
```

(b) The first confidence interval is the 95% confidence interval; the second confidence interval is the 99% confidence interval; while the third confidence interval is the 90% confidence interval.

Solution for 15:

```
> CI <- t.test(TOE$energy)$conf
> CI

[1]  3756.425 12788.075
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the 2003 mean European TOE is $[3756.4249, 12788.0751]$.

Since the sample size is more than 10% of the population, a finite population correction factor should be applied $\left(\sqrt{(N-n)/(N-1)}\right)$ to the margin of error.

```
> N <- 15
> n <- 12
> fpc <- sqrt((N - n)/(N - 1))
> xbar <- sum(CI)/2 # sample mean
> ME <- diff(CI)/2  # margin of error
> CIfpc <- xbar + c(-1, 1)*ME*fpc
> CIfpc

[1]  6181.829 10362.671
```

The resulting confidence interval after applying the finite population correction factor to the margin of error is $[6181.8292, 10362.6708]$.

Solution for 17:

(a)

```
> CI <- with(data = EURD, t.test(rd2003, rd2002, paired = TRUE)$conf)
> CI

[1] -0.3403423 21.7788756
attr(,"conf.level")
[1] 0.95
```

A 95% confidence interval for the mean investment difference between 2003 and 2002 is $[-0.3403, 21.7789]$, which appears to suggest the mean difference is zero.

(b)

```
> N <- 28
> n <- 15
> fpc <- sqrt((N - n)/(N - 1))
> xbar <- sum(CI)/2 # sample mean
> ME <- diff(CI)/2  # margin of error
> CIfpc <- xbar + c(-1, 1)*ME*fpc
> CIfpc

[1]  3.045129 18.393404
```

Once a finite population correction factor is applied, the 95% confidence interval becomes $[3.0451, 18.3934]$, which suggests the mean investment difference is greater than zero. That is, the new policies are increasing investments.

Solution for 19:

If $t_{0.975,n-1}/\sqrt{n} \leq 1$, the confidence level is at least 0.95. Find the smallest value of n such that $t_{0.975,n-1}/\sqrt{n} \leq 1$. When $n = 7$, the confidence level is at least 95%.

```
> nfinder <- function(n, alpha, ...){
+   qt(1-alpha/2, n-1)/sqrt(n)
+ }
> nfinder(n = 2:8, alpha = 0.05)

[1] 8.9846435 2.4841377 1.5912232 1.2416640 1.0494356 0.9248457 0.8360209
```

Solution for 21:

(a) The assumptions to construct a confidence interval for the population variance are that you have a random sample from a normal distribution.

(b)

```
> values <- c(25.3, 23.8, 27.5, 23.2, 24.5, 25.3, 24.6, 26.8, 25.9, 29.2)
> n <- length(values)
> s2 <- var(values)
> chiU <- qchisq(0.975, n-1)
> chiL <- qchisq(0.025, n-1)
> CI <- sqrt(c((n-1)*s2/chiU, (n-1)*s2/chiL))
> CI

[1] 1.245068 3.304584
```

The 95% confidence interval for σ is $[1.2451, 3.3046]$.

(c) Recall that if $\{X_1, X_2, \ldots, X_n\}$ are independent random variables with $N(\mu_i, \sigma_i)$ distributions, respectively, then

$$Y = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_i^2} \sim \chi_n^2.$$

This means that

$$\mathbb{P}\left[\chi^2_{\alpha/2;\,n} \leq \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma_i^2} \leq \chi^2_{1-\alpha/2;\,n}\right] = 1 - \alpha$$

For this problem, $X_i \sim N(\mu = 25, \sigma)$, $n = 10$, and a 95% confidence interval is desired, so

$$\mathbb{P}\left[\chi^2_{0.025;\,10} \leq \sum_{i=1}^{10} \frac{(X_i - 25)^2}{\sigma^2} \leq \chi^2_{0.975;\,10}\right] = 0.95$$

$$\mathbb{P}\left[\frac{\chi^2_{0.025;\,10}}{\sum_{i=1}^{10}(X_i - 25)^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{0.975;\,10}}{\sum_{i=1}^{10}(X_i - 25)^2}\right] = 0.95$$

$$\mathbb{P}\left[\frac{\sum_{i=1}^{10}(X_i - 25)^2}{\chi^2_{0.975;\,10}} \leq \sigma^2 \leq \frac{\sum_{i=1}^{10}(X_i - 25)^2}{\chi^2_{0.025;\,10}}\right] = 0.95$$

$$\mathbb{P}\left[\sqrt{\frac{\sum_{i=1}^{10}(X_i - 25)^2}{\chi^2_{0.975;\,10}}} \leq \sigma \leq \sqrt{\frac{\sum_{i=1}^{10}(X_i - 25)^2}{\chi^2_{0.025;\,10}}}\right] = 0.95$$

$$\implies CI_{0.95}(\sigma) = \left[\sqrt{\frac{\sum_{i=1}^{10}(x_i - 25)^2}{\chi^2_{0.975;\,10}}}, \sqrt{\frac{\sum_{i=1}^{10}(x_i - 25)^2}{\chi^2_{0.025;\,10}}}\right]$$

```
> Num <- sum((values - 25)^2)
> chiU <- qchisq(0.975, 10)
> chiL <- qchisq(0.025, 10)
> CI <- sqrt(c(Num/chiU, Num/chiL))
> CI

[1] 1.273315 3.198123
```

The 95% confidence interval for the standard deviation is $[1.2733, 3.1981]$ when μ is known to be 25.

Solution for 23:

It is known that the quantity ${}^{(n-1)s^2}/_{\sigma^2}$ is distributed as a χ^2_{n-1}. Recall that the sum of r χ^2 random variables is also distributed as a χ^2 with $s = \sum_{i=1}^{r} n_i$ degrees of freedom. The probability statement that follows can then be rearranged to derive a confidence interval for σ.

$$\mathbb{P}\left[\chi^2_{\alpha/2;\,n_1+n_2-2} \leq \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \leq \chi^2_{1-\alpha/2;\,n_1+n_2-2}\right] = 1 - \alpha$$

$$\mathbb{P}\left[\frac{\chi^2_{\alpha/2;\,n_1+n_2-2}}{(n_1-1)S_1^2 + (n_2-1)S_2^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{1-\alpha/2;\,n_1+n_2-2}}{(n_1-1)S_1^2 + (n_2-1)S_2^2}\right] = 1 - \alpha$$

$$\mathbb{P}\left[\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\chi^2_{1-\alpha/2;\,n_1+n_2-2}} \leq \sigma^2 \leq \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\chi^2_{\alpha/2;\,n_1+n_2-2}}\right] = 1 - \alpha$$

$$\implies CI_{0.95}(\sigma) = \left[\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{\chi^2_{0.975;\,n_1+n_2-2}}}, \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{\chi^2_{0.025;\,n_1+n_2-2}}}\right]$$

In the context of this problem,

$$CI_{0.95}(\sigma) = \left[\sqrt{\frac{(22-1)38.7+(45-1)45.6}{89.17714}}, \sqrt{\frac{(22-1)38.7+(45-1)45.6}{44.60299}}\right]$$

```
> n1 <- 22
> n2 <- 45
> s12 <- 38.7
> s22 <- 45.6
> CI <- c( sqrt(((n1 - 1)*s12 +(n2 - 1)*s22)/qchisq(0.975, n1 + n2 - 2)),
+          sqrt(((n1 - 1)*s12 +(n2 - 1)*s22)/qchisq(0.025, n1 + n2 - 2)) )
> CI

[1] 5.622487 7.950112
```

The 95% confidence interval for σ is $[5.6225, 7.9501]$.

Solution for 25:

The minimum n is 21.

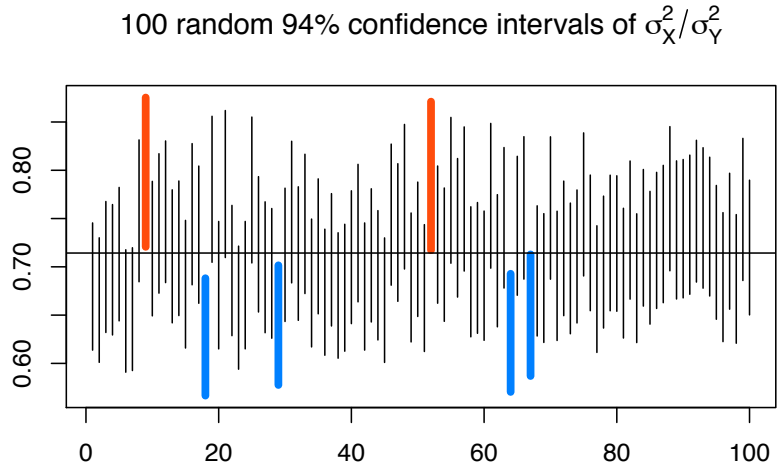
```
> f <- function(n){0.59*qchisq(0.97, n - 1) - 2*qchisq(0.03, n - 1)}
> n <- ceiling(uniroot(f, c(2, 100))$root)
> n

[1] 21
```

```
> n <- 2:100
> conf <- pchisq((n - 1)/0.59, n - 1) - pchisq((n - 1)/2, n - 1)
> results <- data.frame(n, conf)
> with(data = results, min(n[conf >= 0.94]))

[1] 21
```

Solution for 27:

```
> CIsimRV <- function (samples = 100, n1 = 1500, mu1 = 3, sigma1 = sqrt(5),
+                      n2 = 1500, mu2 = 6, sigma2 = sqrt(7),
+                      conf.level = 0.94)
+ {
+   Adkblue <- "#0080FF"
+   Aorange <- "#FF4C0C"
+   alpha <- 1 - conf.level
+   CL <- conf.level * 100
+   N <- samples
+   if (!missing(conf.level) && (length(conf.level) != 1 ||
+                                 !is.finite(conf.level) ||
+                                 conf.level <= 0 ||
+                                 conf.level >= 1))
+     stop("'conf.level' must be a single number between 0 and 1")
```

```
+   if (sigma1 <= 0 && sigma2 <= 0)
+     stop("Sigma1 and Sigma2 must be a positive values")
+   ll <- numeric(samples)
+   ul <- numeric(samples)
+     for(i in 1:samples){
+       xs <- rnorm(n1, mu1, sigma1)
+       ys <- rnorm(n2, mu2, sigma2)
+       sx2 <- var(xs)
+       sy2 <- var(ys)
+       ll[i] <- qf(alpha/2, n2 - 1, n1 - 1)*sx2/sy2
+       ul[i] <- qf(1 - alpha/2, n2 - 1, n1 - 1)*sx2/sy2
+     }
+   TR <- sigma1^2/sigma2^2
+     notin <- sum((ll > TR) + (ul < TR))
+     percentage <- round((notin/samples) * 100, 2)
+     plot(ll, type = "n", ylim = c(min(ll), max(ul)), xlab = " ",
+          ylab = " ")
+     title(sub = bquote(paste("Note: ", .(percentage),
+     "% of the random confidence intervals do not contain ",
+     sigma[X]^2/sigma[Y]^2, " = ", .(TR))))
+     title(main = bquote(paste(.(samples), " random ", .(CL),
+     "% confidence intervals of ", sigma[X]^2/sigma[Y]^2)))
+     for (i in 1:samples) {
+       low <- ll[i]
+       high <- ul[i]
+       if (low < TR & high > TR) {
+         segments(i, low, i, high)
+       }
+       else if (low > TR & high > TR) {
+         segments(i, low, i, high, col = Aorange, lwd = 5)
+       }
+       else {
+         segments(i, low, i, high, col = Adkblue, lwd = 5)
+       }
+     }
+     abline(h = TR)
+     cat(percentage,
+     "\b% of the random confidence intervals do not a variance ratio =",
+     TR, "\b.", "\n")
+   }
> set.seed(121)
> CIsimRV()

6 ^^H% of the random confidence intervals do not a variance ratio = 0.7142857 ^^H.
```



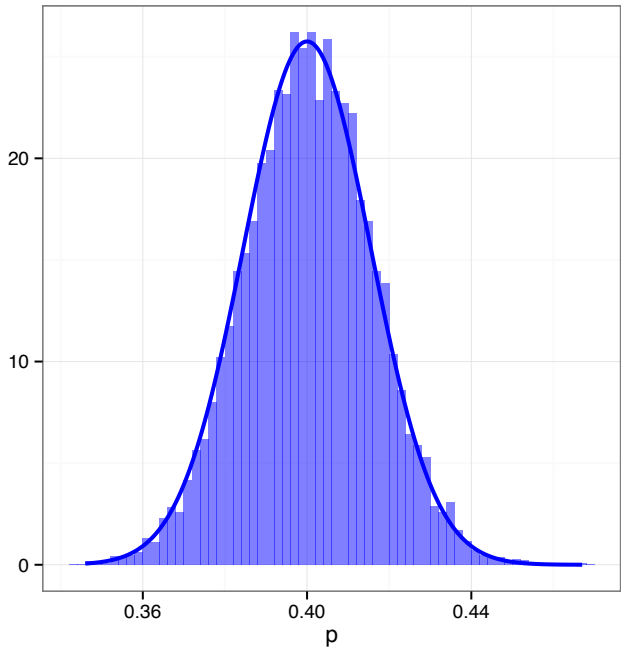
Solution for 29:

```
> CI <- prop.test(x = 300, n = 2000, correct = FALSE,
+                 conf.level = 0.98)$conf
> CI

[1] 0.1323712 0.1695179
attr(,"conf.level")
[1] 0.98
```

The 98% confidence interval for the proportion of ".org" domains registered in a particular country during the last few years is $[0.1324, 0.1695]$.

Solution for 31:

```
> set.seed(6)
> samples <- 10000
> n <- 1000
> PI <- 0.4
> p <- rbinom(samples, n, PI)/n
> c(mean(p), 0.4, sd(p), sqrt(PI*(1 - PI)/n))

[1] 0.40001290 0.40000000 0.01551069 0.01549193

> ggplot(data = data.frame(x = p), aes(x = x)) +
+   geom_histogram(aes(x = x, y = ..density..),
+                  binwidth = 0.002, fill = "blue", alpha = 0.5) +
+   stat_function(fun = dnorm, args = list(PI, sqrt(PI*(1 - PI)/n)),
+                 color = "blue", n = 500, size = 1) +
+   theme_bw() +
+   labs(x = "p", y = "")
```



Solution for 33:

(a)

```
> p <- 0.5
> alpha <- 0.05
> B <- 0.04
> n <- ceiling(p*(1 - p)*(qnorm(1 - alpha/2)/B)^2)
> n

[1] 601

> # Or
> nsize(b = B, conf.level = 1 - alpha, type = "pi")


The required sample size (n) to estimate the population
proportion of successes with a 0.95 confidence interval
so that the margin of error is no more than 0.04 is 601 ^^H.
```

The required sample size (n) to estimate the population proportion of successes with a 0.95 confidence interval so that the margin of error is no more than 0.04 is 601.

(b)

```
> CI <- prop.test(x = 650, n = 800, correct = FALSE,
+                 conf.level = 0.98)$conf
> CI

[1] 0.7783367 0.8424637
attr(,"conf.level")
[1] 0.98
```

A 98% confidence interval for the proportion of accounts paid on time is $[0.7783, 0.8425]$.

Solution for 35:

The confidence intervals for the various diagnoses are score/Wilson confidence intervals.

```
> CI <- prop.test(x = 358, n = 660, correct = FALSE)$conf
> CI

[1] 0.5042799 0.5800776
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of schizophrenic patients admitted to Virgen del Camino is $[0.5043, 0.5801]$.

```
> CI <- prop.test(x = 61, n = 660, correct = FALSE)$conf
> CI

[1] 0.07262507 0.11694046
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of schizophreniform patients admitted to Virgen del Camino is $[0.0726, 0.1169]$.

```
> CI <- prop.test(x = 37, n = 660, correct = FALSE)$conf
> CI

[1] 0.04094285 0.07631626
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of schizoaffective patients admitted to Virgen del Camino is $[0.0409, 0.0763]$.

```
> CI <- prop.test(x = 64, n = 660, correct = FALSE)$conf
> CI

[1] 0.07667093 0.12193291
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of bipolar patients admitted to Virgen del Camino is $[0.0767, 0.1219]$.

```
> CI <- prop.test(x = 24, n = 660, correct = FALSE)$conf
> CI

[1] 0.02455615 0.05353698
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of delusional patients admitted to Virgen del Camino is $[0.0246, 0.0535]$.

```
> CI <- prop.test(x = 54, n = 660, correct = FALSE)$conf
> CI

[1] 0.06324816 0.10522800
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of psychotic patients admitted to Virgen del Camino is $[0.0632, 0.1052]$.

```
> CI <- prop.test(x = 32, n = 660, correct = FALSE)$conf
> CI

[1] 0.03455100 0.06764427
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for the proportion of atypical psychosis patients admitted to Virgen del Camino is $[0.0346, 0.0676]$.

Chapter 9

Odd solutions

Solution for 1:

The probability of making a type I error (rejecting the null hypothesis when it is true) is α, while β is the probability of making a type II error (failing to reject the null hypothesis when it it false). The probability of rejecting the null hypothesis when it is false is $1 - \beta$, which is known as the power of the test.

Solution for 3:

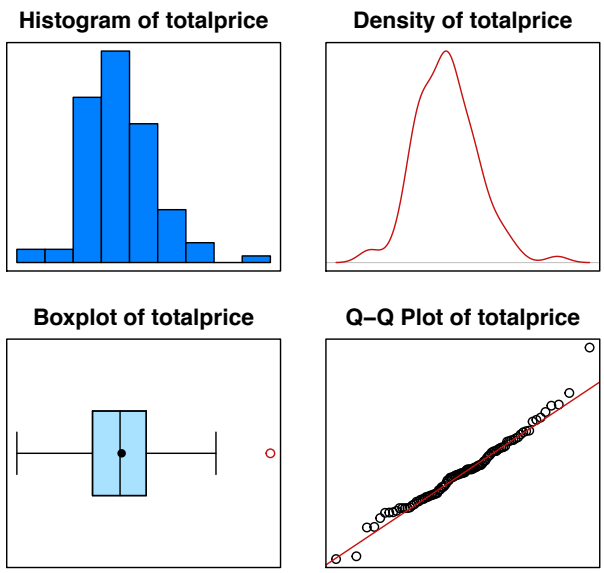
```
> alpha <- 0.05
> cvLower <- qnorm(alpha/2, 100, 50/sqrt(36))
> cvUpper <- qnorm(1 - alpha/2, 100, 50/sqrt(36))
> LeftPower <- pnorm(cvLower, 120, 50/sqrt(36))
> RightPower <- pnorm(cvUpper, 120, 50/sqrt(36), lower = FALSE)
> POWER <- LeftPower + RightPower
> POWER

[1] 0.670051
```

The power of the test is 0.6701.

Solution for 5:

Prior to using a test that is very sensitive to departures in normality, as a minimum, create a quantile-quantile plot to verify the assumption of normality.

```
> Greater90 <- subset(x = VIT2005, subset = area >= 90)
> with(data = Greater90,
+      eda(totalprice)
+      )
> n <- dim(Greater90)[1]
```



The results from applying `eda()`, suggest the appraised price for 90m^2 or larger pisos follows a normal distribution. Now, continue with the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the price for 90m^2 or larger pisos variability is greater than $60,000^2$€^{2}, are

$$H_0 : \sigma^2 = 60,000^2 \text{ versus } H_1 : \sigma^2 > 60,000^2.$$

Step 2: **Test Statistic** — The test statistic chosen is S^2 because $E\left[S^2\right] = \sigma^2$.

```
> TS <- var(Greater90$totalprice)
> TS

[1] 3822980710
```

The value of this test statistic is $s^2 = 3822980710.3638$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed χ^2_{93} and H_1 is an upper one-sided hypothesis, the rejection region is $\chi^2_{\text{obs}} > \chi^2_{0.95;\,94-1} = 116.511$. The value of the standardized test statistic is $\chi^2_{\text{obs}} = 98.7603$.

```
> RR <- qchisq(0.95, n - 1)
> RR

[1] 116.511
```

```
> STS <- (n - 1)*TS/60000^2
> STS

[1] 98.76034
```

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(\chi^2_{93} \geq 98.7603) = 0.3218$.

```
> pvalue <- pchisq(STS, n - 1, lower = FALSE)
> pvalue

[1] 0.3218218
```

I. From the rejection region, fail to reject H_0 because $\chi^2_{\text{obs}} = 98.7603$ is less than 116.511.

II. From the $\wp$-value, fail to reject H_0 because the $\wp$-value 0.3218 is greater than 0.05.

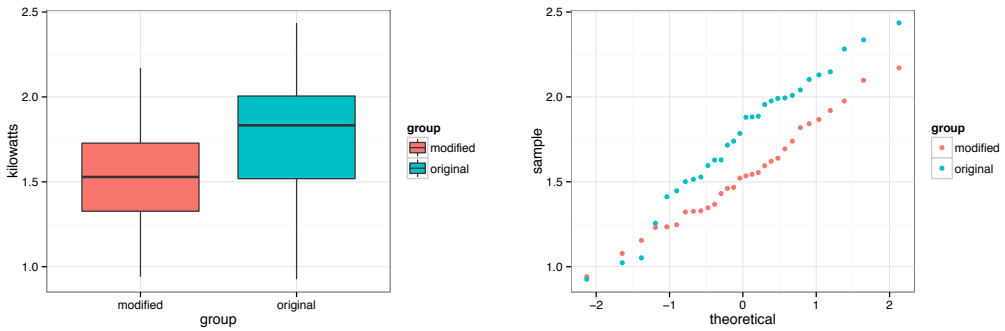
Fail to reject $\mathbf{H_0}$.

Step 5: **English Conclusion** — There is insufficient evidence to suggest the variance for the appraised price of 90m^2 or larger pisos is greater than $60,000^2 €^2$.

Solution for 7:

To solve this problem, start by verifying the reasonableness of the normality assumption.

```
> ggplot(data = REFRIGERATOR, aes(x = group, y = kilowatts, fill = group)) +
+   geom_boxplot() +
+   theme_bw()
> ggplot(data = REFRIGERATOR, aes(sample = kilowatts, color = group)) +
+   stat_qq() +
+   theme_bw()
```



The side-by-side boxplots and normal quantile-quantile plots suggest it is reasonable to assume the energy consumption for both models follows a normal distribution. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — Since the problem wants to test to see if the mean energy consumption for modified refrigerators is less than the mean energy consumption for original refrigerators, use a lower one-sided alternative hypothesis.

$$H_0 : \mu_{\text{modified}} - \mu_{\text{original}} = 0 \text{ versus } H_1 : \mu_{\text{modified}} - \mu_{\text{original}} < 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{X} - \overline{Y}$ because $E\left[\overline{X} - \overline{Y}\right] = \mu_X - \mu_Y$.

```
> Means <- with(data = REFRIGERATOR,
+               tapply(kilowatts, group, mean)
+ )
> Means

modified original
1.535800 1.760067
```

The value of this test statistic is $1.5358 - 1.7601 = -0.2243$. The standardized test statistic under the assumption that H_0 is true and its approximate distribution are

$$\frac{\left[\left(\overline{X} - \overline{Y}\right) - \delta_0\right]}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \overset{\bullet}{\sim} t_\nu.$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately t_ν and H_1 is a lower one-sided hypothesis, the rejection region is $t_{\text{obs}} < t_{0.05;\,54.7888} = -1.6731$.

```
> TR <- t.test(kilowatts ~ group, data = REFRIGERATOR,
+              alternative = "less")
> TR


Welch Two Sample t-test

data:  kilowatts by group
t = -2.5128, df = 54.789, p-value = 0.007475
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
        -Inf -0.07494116
sample estimates:
mean in group modified mean in group original
             1.535800              1.760067

> RR <- qt(0.05, TR$parameter)
> RR

[1] -1.673144
```

The degrees of freedom are

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} = 54.7888,$$

and the value of the standardized test statistic is

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} = -2.5128.$$

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(t_{54.7888} \leq -2.5128) = 0.0075$.

I. From the rejection region, reject H_0 because $t_{\text{obs}} = -2.5128$ is less than -1.6731.

II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 0.0075$ is less than 0.05.

Reject $\mathbf{H_0}$.

Step 5: **English Conclusion** — There is evidence to suggest the average energy consumption for modified refrigerators is less than the average energy consumption for unmodified (original) refrigerators.

Solution for 9:

To solve this problem, use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of women who favor low-fat yogurt is the same as the proportion of men who favor low-fat yogurt are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X \neq \pi_Y.$$

In this case, let the random variable X represent the number of females favoring low-fat yogurt, and let the random variable Y represent the number of males favoring low-fat yogurt.

Step 2: **Test Statistic** — The test statistic chosen is $P_X - P_Y$ since $E[P_X - P_Y] = \pi_X - \pi_Y$. The standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{P_X - P_Y}{\sqrt{P(1-P)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0,1)$ distribution and H_1 is a two-sided hypothesis, the rejection region is $|z_{\text{obs}}| > z_{0.995} = 2.5758$.

```
> RR <- qnorm(0.995)
> RR

[1] 2.575829
```

```
> x <- 825
> m <- 1200
> y <- 525
> n <- 1150
> p <- (x + y)/(m + n)
> p

[1] 0.5744681

> TRwoCC <- prop.test(x = c(x, y), n = c(m, n), correct = FALSE)
> TRwoCC


2-sample test for equality of proportions without continuity
correction

data:  c(x, y) out of c(m, n)
X-squared = 128.16, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1920354 0.2699211
sample estimates:
   prop 1    prop 2
0.6875000 0.4565217

> STSwoCC <- sqrt(TRwoCC$statistic)
> STSwoCC

X-squared
 11.32082

> TRwiCC <- prop.test(x = c(x, y), n = c(m, n), correct = TRUE)
> TRwiCC


2-sample test for equality of proportions with continuity
correction

data:  c(x, y) out of c(m, n)
X-squared = 127.22, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1911840 0.2707726
sample estimates:
   prop 1    prop 2
0.6875000 0.4565217

> STSwiCC <- sqrt(TRwiCC$statistic)
> STSwiCC

X-squared
 11.27908
```

The pooled estimate of π is $p = \frac{x+y}{m+n} = \frac{1350}{2350}$. The value of the standardized test statistic is

Without Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y}{\sqrt{p(1-p)\left(\frac{1}{m}+\frac{1}{n}\right)}} \\ &= \frac{\frac{825}{1200} - \frac{525}{1150}}{\sqrt{\frac{1350}{2350}\left(1-\frac{1350}{2350}\right)\left(\frac{1}{1200}+\frac{1}{1150}\right)}} \\ &= 11.3208 \end{aligned}$$

OR

With Continuity Correction

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y + \frac{1}{2}\left(\frac{1}{m}+\frac{1}{n}\right)}{\sqrt{p(1-p)\left(\frac{1}{m}+\frac{1}{n}\right)}} \\ &= \frac{\frac{825}{1200} - \frac{525}{11504} + \frac{1}{2}\left(\frac{1}{1200}+\frac{1}{1150}\right)}{\sqrt{\frac{1350}{2350}\left(1-\frac{1350}{2350}\right)\left(\frac{1}{1200}+\frac{1}{1150}\right)}} \\ &= 11.2791 \end{aligned}$$

Step 4: **Statistical Conclusion** — The $\wp$-value is $2 \times \mathbb{P}(Z \geq |z_{\text{obs}}|)$ and is approximately 0 for both cases. This is less than 0.05, so reject H_0.

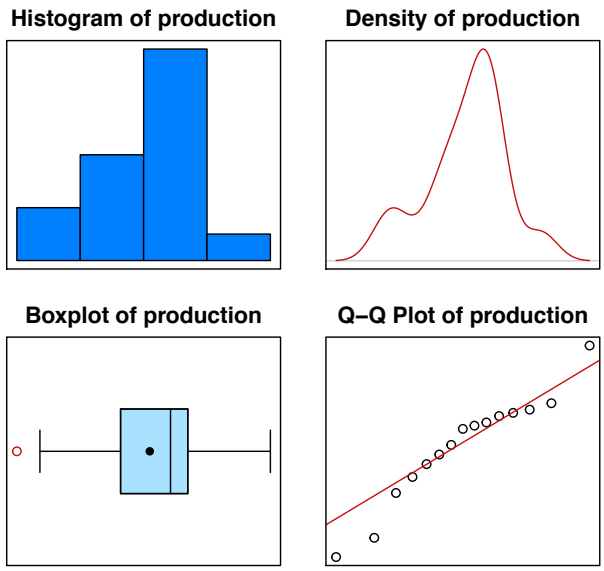
Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest a gender difference in the proportion of females and males that prefer low-fat yogurt at an α level of 0.01.

Solution for 11:

To see if `textileA` is profitable, start by verifying the normality assumption of the data using exploratory data analysis (`eda()`).

```
> woolA <- subset(x = WOOL, subset = location =="textileA")
> with(data = woolA,
+      eda(production)
+ )
```



The results from applying the function `eda()` to the production of wool suggest it is not unreasonable to assume production of wool for `textileA` follows a normal distribution. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — To test if wool production for `textileA` exceeds 1000 kilograms per day, the hypotheses are

$$H_0 : \mu = 1 \text{ versus } H_1 : \mu > 1$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{X}$ because $E\left[\overline{X}\right] = \mu$.

```
> xbar <- mean(woolA$production)
> xbar

[1] 1.226
```

The value of this test statistic is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.226$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{15-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_{14} and H_1 is an upper one-sided hypothesis, the rejection region is $t_{\text{obs}} > t_{0.95;\,14} = 1.7613$.

```
> RR <- qt(0.95, 14)
> RR

[1] 1.76131
```

```
> TR <- with(data = woolA,
+            t.test(production, mu = 1, alternative = "greater")
+ )
> TR


One Sample t-test

data:  production
t = 5.1942, df = 14, p-value = 6.801e-05
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 1.149365      Inf
sample estimates:
mean of x
    1.226
```

The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = 5.1942$.

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(t_{14} \geq 5.1942) = 1e-04$.

I. From the rejection region, reject H_0 because $t_{\text{obs}} = 5.1942$ is greater than 1.7613.

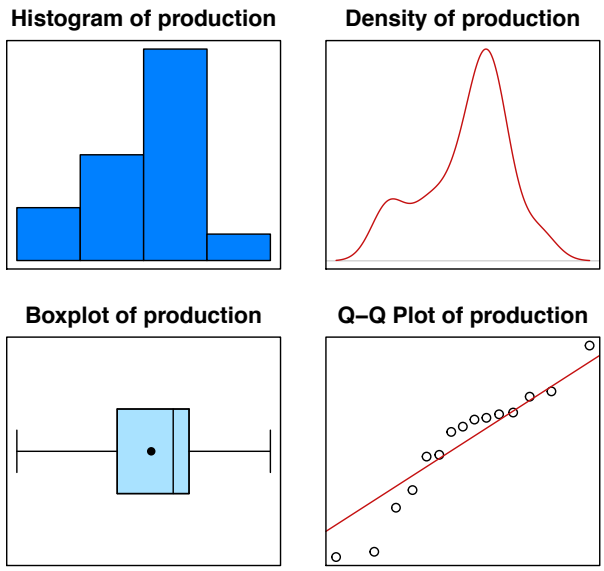
II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 1e-04$ is less than 0.05.

Reject H_0.

Step 5: **English Conclusion** — There is evidence to suggest `textileA` produces more than 1000 kilograms of refined wool per day.

To see if `textileB` is profitable, start by verifying the normality assumption of the data using exploratory data analysis (`eda()`).

```
> woolB <- subset(x = WOOL, subset = location =="textileB")
> with(data = woolB,
+      eda(production)
+ )
```



The results from applying the function `eda()` to the production of wool suggest it is not unreasonable to assume production of wool for `textileB` follows a normal distribution. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — To test if wool production for `textileB` exceeds 1000 kilograms per day, the hypotheses are

$$H_0 : \mu = 1 \text{ versus } H_1 : \mu > 1$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{X}$ because $E\left[\overline{X}\right] = \mu$.

```
> xbar <- mean(woolB$production)
> xbar

[1] 1.446
```

The value of this test statistic is $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 1.446$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{\overline{X}-\mu_0}{S/\sqrt{n}} \sim t_{15-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_{14} and H_1 is an upper one-sided hypothesis, the rejection region is $t_{\text{obs}} > t_{0.95;\,14} = 1.7613$.

```
> RR <- qt(0.95, 14)
> RR

[1] 1.76131
```

```
> TR <- with(data = woolB,
+            t.test(production, mu = 1, alternative = "greater")
+ )
> TR


One Sample t-test

data:  production
t = 5.1386, df = 14, p-value = 7.53e-05
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 1.293129      Inf
sample estimates:
mean of x
    1.446
```

The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = 5.1386$.

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(t_{14} \geq 5.1386) = 1e-04$.

I. From the rejection region, reject H_0 because $t_{\text{obs}} = 5.1386$ is greater than 1.7613.

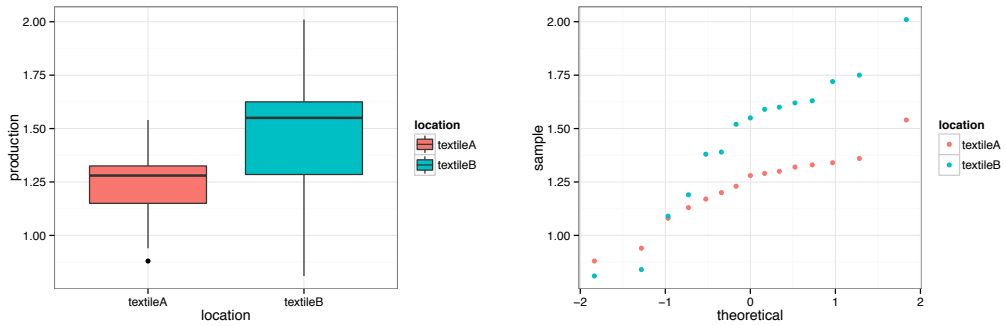
II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 1e-04$ is less than 0.05.

Reject H_0.

Step 5: **English Conclusion** — There is evidence to suggest `textileB` produces more than 1000 kilograms of refined wool per day.

To discover if one textile is superior to the other, start by verifying the reasonableness of the normality assumption.

```
> ggplot(data = WOOL, aes(x = location, y = production, fill = location)) +
+   geom_boxplot() +
+   theme_bw()
> ggplot(data = WOOL, aes(sample = production, color = location)) +
+   stat_qq() +
+   theme_bw()
```



The side-by-side boxplots and normal quantile-quantile plots suggest it may be reasonable to assume the wool production for both textile plants follow normal distributions; however, it is clear from the boxplot that the variances are different. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — Since the problem wants to test to see if the mean wool production for the textile plants is different and the problem does not suggest one textile plant is superior to the other, use a two-sided alternative hypothesis.

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y \neq 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{X} - \overline{Y}$ because $E\left[\overline{X} - \overline{Y}\right] = \mu_X - \mu_Y$.

```
> Means <- with(data = WOOL,
+       tapply(production, location , mean)
+       )
> Means

textileA textileB
   1.226    1.446
```

The value of this test statistic is $1.226 - 1.446 = -0.22$. The standardized test statistic under the assumption that H_0 is true and its approximate distribution are

$$\frac{\left[\left(\overline{X} - \overline{Y}\right) - \delta_0\right]}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \overset{\cdot}{\sim} t_\nu.$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately t_ν and H_1 is a two-sided hypothesis, the rejection region is $|t_{\text{obs}}| > t_{0.95;\,20.6186} = 1.7222$.

```
> TR <- t.test(production ~ location, data = WOOL)
> TR


Welch Two Sample t-test

data:  production by location
t = -2.266, df = 20.619, p-value = 0.03435
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.42213539 -0.01786461
sample estimates:
mean in group textileA mean in group textileB
                 1.226                  1.446

> RR <- qt(0.95, TR$parameter)
> RR

[1] 1.722211
```

The degrees of freedom are

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} = 20.6186,$$

and the value of the standardized test statistic is

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} = -2.266.$$

Step 4: **Statistical Conclusion** — The $\wp$-value is $2 \times \mathbb{P}(t_{20.6186} \geq 2.266) = 0.0344$.

I. From the rejection region, reject H_0 because $|t_{\text{obs}}| = 2.266$ is greater than 1.7222.

II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 0.0344$ is less than 0.05.

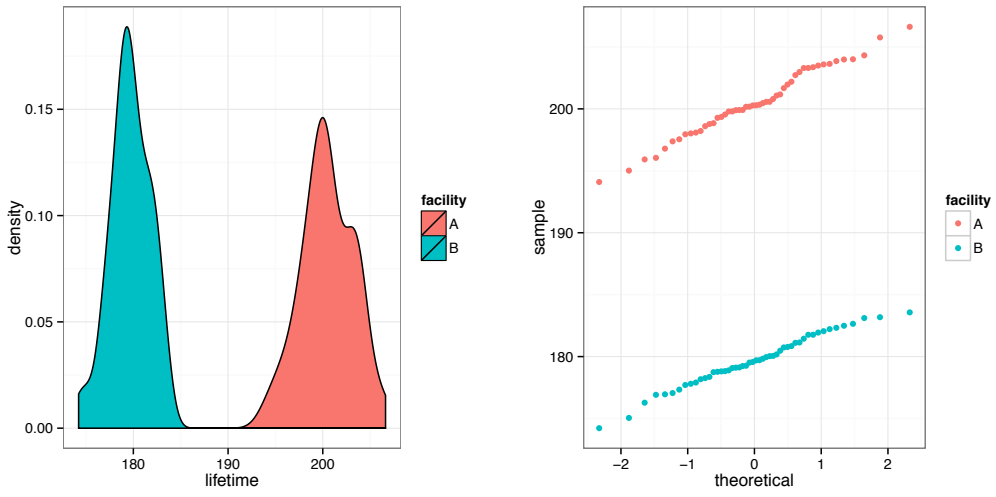
Reject H_0.

Step 5: **English Conclusion** — There is evidence to suggest different mean wool production for the two plants.

Solution for 13:

Prior to using a test that is very sensitive to departures in normality, density plots and quantile-quantile normal plots are created for both facilities.

```
> ggplot(data = BATTERY, aes(lifetime, fill = facility)) +
+   geom_density() +
+   theme_bw()
> ggplot(data = BATTERY, aes(sample = lifetime, color = facility)) +
+   stat_qq() +
+   theme_bw()
```



Based on the density plots and quantile-quantile normal plots, it seems reasonable to assume the battery life from both facilities follow normal distributions. Therefore, proceed with the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the variability in facility A's battery life (X) is greater than the variability in facility B's battery life (Y) are

$$H_0 : \sigma_X^2 = \sigma_Y^2 \text{ versus } H_1 : \sigma_X^2 > \sigma_Y^2.$$

Step 2: **Test Statistic** — The test statistics chosen are S_X^2 and S_Y^2 since $E\left[S_X^2\right] = \sigma_X^2$ and $E\left[S_Y^2\right] = \sigma_Y^2$.

```
> VAR <- tapply(BATTERY$lifetime, BATTERY$facility, var)
> VAR

       A        B
7.539291 4.347130
```

The values of these test statistics are $s_X^2 = 7.5393$ and $s_Y^2 = 4.3471$. The standardized test statistic under the assumption that H_0 is true and its distribution are $S_X^2/S_Y^2 \sim F_{50-1,50-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed $F_{49,49}$, and H_1 is an upper one-sided hypothesis, the rejection region is $f_{\text{obs}} > F_{0.95;\,49,49} = 1.6073$.

```
> RR <- qf(0.95, 49, 49)
> RR

[1] 1.607289

> TR <- var.test(lifetime ~ facility, data = BATTERY,
+                alternative = "greater")
> TR


F test to compare two variances

data:  lifetime by facility
F = 1.7343, num df = 49, denom df = 49, p-value = 0.02836
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 1.079031      Inf
sample estimates:
ratio of variances
          1.734314
```

The value of the standardized test statistic is $f_{\text{obs}} = (7.5393)/(4.3471) = 1.7343$.

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(F_{49,49} \geq 1.7343) = 0.0284$.

I. From the rejection region, reject H_0 because $f_{\text{obs}} = 1.7343$ is greater than 1.6073.

II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 0.0284$ is less than 0.05.

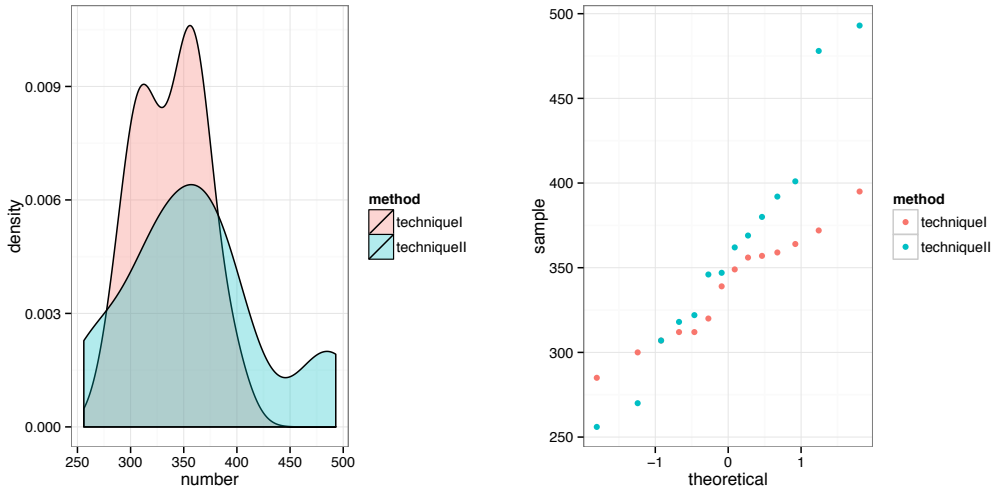
Reject H_0.

Step 5: **English Conclusion** — The evidence suggests the variability of battery life from facility A is greater than the variance for battery life from facility B.

Solution for 15:

To solve this problem, start by verifying the reasonableness of the normality assumption.

```
> ggplot(data = CHIPS, aes(number, fill = method)) +
+   geom_density(alpha = 0.3) +
+   theme_bw()
> ggplot(data = CHIPS, aes(sample = number, color = method)) +
+   stat_qq() +
+   theme_bw()
```



The density plots and normal quantile-quantile plots suggest it is reasonable to assume the number of usable chips from both techniques follow normal distributions; however, it is clear from the density plots that the variances are different. Now, proceed with the five-step procedure.

Step 1: **Hypotheses** — Since the problem wants to test if there are differences in the mean number of usable chips generated by the two techniques, use a two-sided alternative hypothesis.

$$H_0 : \mu_X - \mu_Y = 0 \text{ versus } H_1 : \mu_X - \mu_Y \neq 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{X} - \overline{Y}$ because $E\left[\overline{X} - \overline{Y}\right] = \mu_X - \mu_Y$.

```
> MEANS <- tapply(CHIPS$number, CHIPS$method, mean)
> MEANS

 techniqueI techniqueII
   337.6429    360.0714
```

The value of this test statistic is $337.6429 - 360.0714 = -22.4286$. The standardized test statistic under the assumption that H_0 is true and its approximate distribution are

$$\frac{\left[\left(\overline{X} - \overline{Y}\right) - \delta_0\right]}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \overset{\cdot}{\sim} t_\nu.$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed approximately t_ν, and H_1 is a two-sided hypothesis, the rejection region is $t_{\text{obs}} < t_{0.025;\,18.4541} = -2.0972$ or $t_{\text{obs}} > t_{0.975;\,18.4541} = 2.0972$.

```
> TR <- t.test(number ~ method, data = CHIPS)
> TR


        Welch Two Sample t-test

data:  number by method
t = -1.1175, df = 18.454, p-value = 0.2781
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -64.52203  19.66489
sample estimates:
 mean in group techniqueI mean in group techniqueII
                 337.6429                  360.0714

> RRupper <- qt(0.975, TR$parameter)
> RRlower <- qt(0.025, TR$parameter)
> c(RRlower, RRupper)

[1] -2.097223  2.097223
```

The degrees of freedom are

$$\nu = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{(s_X^2/n_X)^2}{n_X - 1} + \frac{(s_Y^2/n_Y)^2}{n_Y - 1}} = 18.4541,$$

and the value of the standardized test statistic is

$$t_{\text{obs}} = \frac{\bar{x} - \bar{y} - \delta_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} = -1.1175.$$

Step 4: **Statistical Conclusion** — The $\wp$-value is $2 \times \mathbb{P}(t_{18.4541} \geq |-1.1175|) = 0.2781$.

I. From the rejection region, fail to reject H_0 because $t_{\text{obs}} = -1.1175$ is greater than -2.0972.

II. From the $\wp$-value, fail to reject H_0 because the $\wp$-value $= 0.2781$ is greater than 0.05.

Fail to reject $\mathbf{H_0}$.

Step 5: **English Conclusion** — There is insufficient evidence to suggest the average number of usable chips from technique I is different from the average number of usable chips from technique II.

Solution for 17:
Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether immigration from Bolivia has increased are

$$H_0 : \pi = 0.004 \text{ versus } H_1 : \pi > 0.004.$$

Step 2: **Test Statistic** — The test statistic chosen is Y, where Y is the number of Bolivian immigrants. Provided H_0 is true, $Y \sim Bin(n, \pi_0)$. The value of the test statistic is $y_{\text{obs}} = 87$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

$$\begin{aligned}\wp\text{-value} = \mathbb{P}\left(Y \geq y_{\text{obs}} \,|\, H_0\right) &= \sum_{i=y_{\text{obs}}}^{n} \binom{n}{i} \pi_0^i \left(1-\pi_0\right)^{n-i} \\ &= \sum_{87}^{3740} \binom{3740}{i} 0.004^i \left(0.996\right)^{3740-i} \\ &= 0 \quad \text{Computed with R}\end{aligned}$$

```
> pvalue <- sum(dbinom(87:3740, 3740, 0.004))
> pvalue

[1] 1.505911e-37

> TR <- binom.test(x = 87, n = 3740, p = 0.004,
+                  alternative = "greater")
> TR


Exact binomial test

data:  87 and 3740
number of successes = 87, number of trials = 3740, p-value <
2.2e-16
alternative hypothesis: true probability of success is greater than 0.004
95 percent confidence interval:
 0.01935316 1.00000000
sample estimates:
probability of success
            0.02326203
```

Reject H_0.

Step 5: **English Conclusion** — There is evidence to suggest the proportion of Bolivian immigrants in Pamplona, Spain, has increased.

Solution for 19:

(a) Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to contradict the housing director's claim about the bathrooms are

$$H_0 : \pi = 0.50 \text{ versus } H_1 : \pi < 0.50.$$

Step 2: **Test Statistic** — The test statistic chosen is Y, where Y is the number of apartments that have more than one bathroom. Provided H_0 is true, $Y \sim Bin(n, \pi_0)$.

```
> FT <- xtabs(~toilets, data = VIT2005)
> FT

toilets
  1   2
116 102

> yobs <- FT[2]
> n <- sum(!is.na(VIT2005$toilets))
> c(yobs, n)

  2
102 218
```

The value of the test statistic is $y_{\text{obs}} = 102$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

$$\begin{aligned}\wp\text{-value} = \mathbb{P}\left(Y \leq y_{\text{obs}} \,|\, H_0\right) &= \sum_{i=0}^{y_{\text{obs}}} \binom{n}{i} \pi_0^i \left(1-\pi_0\right)^{n-i} \\ &= \sum_{0}^{102} \binom{218}{i} 0.5^i \left(0.5\right)^{218-i} \\ &= 0.1893 \quad \text{Computed with R}\end{aligned}$$

```
> pvalue <- sum(dbinom(0:yobs, n, 0.5))
> pvalue

[1] 0.1893231
```

```
> TR <- binom.test(x = yobs, n = n, p = 0.5, alternative = "less")
> TR


Exact binomial test

data:  yobs and n
number of successes = 102, number of trials = 218, p-value =
0.1893
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.000000 0.525842
sample estimates:
probability of success
             0.4678899
```

Thus, one fails to reject H_0 because 0.1893 is greater than 0.1.

Fail to reject H_0.

Step 5: **English Conclusion** — There is not sufficient evidence to contradict the claim that at least 50% of all apartments have more than one bathroom.

(b) Use the five-step procedure.

Exact method:

Step 1: **Hypotheses** — The null and alternative hypotheses to support the housing director's claim about elevators are

$$H_0 : \pi = 0.75 \text{ versus } H_1 : \pi > 0.75.$$

Step 2: **Test Statistic** — The test statistic chosen is Y, where Y is the number of apartments that have an elevator.

```
> FT <- xtabs(~ elevator, data = VIT2005)
> FT

elevator
  0   1
 44 174

> yobs <- FT[2]
> n <- sum(!is.na(VIT2005$elevator))
> c(yobs, n)

  1
174 218
```

Provided H_0 is true, $Y \sim Bin(n, \pi_0)$. The value of the test statistic is $y_{\text{obs}} = 174$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

$$\begin{aligned}\wp\text{-value} = \mathbb{P}\left(Y \geq y_{\text{obs}} \,|\, H_0\right) &= \sum_{i=y_{\text{obs}}}^{n} \binom{n}{i} \pi_0^i \left(1-\pi_0\right)^{n-i} \\ &= \sum_{174}^{218} \binom{218}{i} 0.75^i \left(0.25\right)^{218-i} \\ &= 0.0564 \quad \text{Computed with R}\end{aligned}$$

```
> pvalue <- sum(dbinom(yobs:n, n, 0.75))
> pvalue

[1] 0.05643458

> TR <- binom.test(x = yobs, n = n, p = 0.75,
+                  alternative = "greater")
> TR


Exact binomial test

data:  yobs and n
number of successes = 174, number of trials = 218, p-value =
0.05643
alternative hypothesis: true probability of success is greater than 0.75
95 percent confidence interval:
 0.748218 1.000000
sample estimates:
probability of success
             0.7981651
```

Thus, one rejects H_0 because 0.0564 is less than 0.1.

Reject H_0.

Step 5: **English Conclusion** — There is evidence to substantiate the claim of the housing director regarding elevators.

Approximate method: Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to substantiate the housing director's claim regarding elevators are

$$H_0 : \pi = 0.75 \text{ versus } H_1 : \pi > 0.75.$$

Step 2: **Test Statistic** — The test statistic chosen is P, where P is the proportion of apartments with elevators. Provided H_0 is true,

$$P \overset{\cdot}{\sim} N\left(\pi_0, \sqrt{\frac{\pi_0(1-\pi_0)}{n}}\right)$$

and the standardized test statistic is

$$Z = \frac{P - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \overset{\cdot}{\sim} N(0,1).$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0,1)$ distribution, and H_1 is an upper one-sided hypothesis, the rejection region is $z_{\text{obs}} > z_{0.9} = 1.2816$.

```
> RR <- qnorm(0.90)
> RR

[1] 1.281552

> TRnoCC <- prop.test(x = yobs, n = n, p = 0.75,
+                     alternative = "greater", correct = FALSE)
> TRwiCC <- prop.test(x = yobs, n = n, p = 0.75,
+                     alternative = "greater", correct = TRUE)
> TRnoCC


1-sample proportions test without continuity correction

data:  yobs out of n, null probability 0.75
X-squared = 2.6972, df = 1, p-value = 0.05026
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.7499209 1.0000000
sample estimates:
        p
0.7981651

> TRwiCC


1-sample proportions test with continuity correction

data:  yobs out of n, null probability 0.75
X-squared = 2.4465, df = 1, p-value = 0.05889
alternative hypothesis: true p is greater than 0.75
95 percent confidence interval:
 0.7474708 1.0000000
sample estimates:
        p
0.7981651
```

The value of the standardized test statistic is

Without Continuity Correction		With Continuity Correction
$z_{\text{obs}} = \dfrac{p-\pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$		$z_{\text{obs}} = \dfrac{p-\pi_0+\frac{1}{2n}}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$
$= \dfrac{\frac{174}{218}-0.75}{\sqrt{\frac{(0.75)(1-0.75)}{218}}}$	OR	$= \dfrac{\frac{174}{218}-0.75+\frac{1}{436}}{\sqrt{\frac{(0.75)(1-0.75)}{218}}}$
$= 1.6423$		$= 1.5641$

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(Z \geq 1.6423) = 0.0503$ or $\mathbb{P}(Z \geq 1.5641) = 0.0589$ for continuity corrections not used and used, respectively.

I. From the rejection region, reject H_0 because $z_{\text{obs}} = 1.6423$ (no continuity correction) is greater than 1.2816, and $z_{\text{obs}} = 1.5641$ (continuity correction) is greater than 1.2816.

II. From the $\wp$-value, reject H_0 because the $\wp$-value $= 0.0503$ (without continuity correction) or $\wp$-value $= 0.0589$ (with continuity correction) is less than 0.1.

Reject H_0.

Step 5: **English Conclusion** — There is evidence to support the housing director's claim about the percent of apartments with elevators.

(c) To solve this problem, use Fisher's exact test and the five-step procedure. Only Fisher's Exact Test will be completed to answer the question as the $n(1-\pi) > 10$ condition will not be satisfied for a large sample approximation.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of apartments built prior to 1980 with elevators that do not have garages is less than the proportion of apartments that do not have elevators or garages are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X > \pi_Y.$$

In this case, the random variable X will represent the number of apartments built prior to 1980 that do not have garages or elevators, and the random variable Y will represent the number of apartments built prior to 1980 that have an elevator but no garage.

Step 2: **Test Statistic** — The test statistic chosen is X, where X is the number of apartments built prior to 1980 that do not have garages or elevators.

```
> FT <- xtabs(~elevator + garage,
+             data = subset(VIT2005, subset = age > 25))
> FT

        garage
elevator  0  1
       0 19  0
       1 22  4
```

Table 9.1: Apartments built prior to 1980 classified by the presence of a garage and of an elevator

		Garage		
		NO	YES	
Elevators	NO	$19 = x$	0	$19 = m$
	YES	22	4	$26 = n$
		$41 = k$	$4 = N - k$	$45 = N$

The observed value of the test statistic is $x = 19$.

Provided H_0 is true, and conditioning on the fact that $X + Y = k$, $X \sim Hyper(m, n, k)$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion** — To compute the $\wp$-value, compute

$$\mathbb{P}(X \geq x \,|\, H_0) = \sum_{i=x}^{\min\{m,k\}} \frac{\binom{m}{i}\binom{n}{k-i}}{\binom{N}{k}} = \sum_{i=19}^{\min\{19,41\}} \frac{\binom{19}{i}\binom{26}{41-i}}{\binom{45}{41}} = 0.1003$$

```
> pvalue <- sum(dhyper(19:19, 19, 26, 41))
> pvalue

[1] 0.1003389

> TR <- fisher.test(FT, alternative = "greater")
> TR


Fisher's Exact Test for Count Data

data:  FT
p-value = 0.1003
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.6878007       Inf
sample estimates:
odds ratio
       Inf
```

Since the $\wp$-value is 0.1003, one fails to reject H_0 because 0.1003 is greater than 0.10.

Fail to reject H_0.

Step 5: **English Conclusion** — There is not sufficient evidence to suggest that the proportion of apartments built prior to 1980 with elevators and no garages is lower

than the proportion of apartments built prior to 1980 without elevators and no garages.

Solution for 21:

(a) To solve this problem, use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test if there is evidence to suggest that $\pi_{\text{male|urban}} < \pi_{\text{female|urban}}$ are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X < \pi_Y.$$

In this case, let the random variable X represent the number males given an urban area, and let the random variable Y represent the number of females given an urban area.

Step 2: **Test Statistic** — The test statistic chosen is $P_X - P_Y$ since $E[P_X - P_Y] = \pi_X - \pi_Y$. The standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{P_X - P_Y}{\sqrt{P(1-P)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0,1)$ distribution and H_1 is a lower one-sided hypothesis, the rejection region is $z_{\text{obs}} < z_{0.05} = -1.6449$.

```
> RR <- qnorm(0.05)
> RR

[1] -1.644854

> x <- 4734
> m <- 10895
> y <- 6161
> n <- 10895
> p <- (x + y)/(m + n)
> p

[1] 0.5

> TR <- prop.test(x = c(x, y), n = c(m, n),
+                 correct = FALSE, alternative = "less")
> TR


2-sample test for equality of proportions without continuity
correction

data:  c(x, y) out of c(m, n)
X-squared = 373.81, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
```

```
 -1.0000000 -0.1199306
sample estimates:
   prop 1    prop 2
0.4345112 0.5654888

> sts <- -1*sqrt(TR$statistic)
> names(sts) <- "z_obs"
> sts

    z_obs
-19.33416
```

The pooled estimate of π is $p = \frac{x+y}{m+n} = \frac{4734+6161}{10895+10895} = 0.5$. The value of the standardized test statistic is -19.3342.

Without a continuity correction,

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y}{\sqrt{p(1-p)\left(\frac{1}{m}+\frac{1}{n}\right)}} \\ &= \frac{\frac{4734}{10895} - \frac{6161}{10895}}{\sqrt{\frac{10895}{21790}\left(1-\frac{10895}{21790}\right)\left(\frac{1}{10895}+\frac{1}{10895}\right)}} \\ &= -19.3342 \end{aligned}$$

Notice that if z_{obs} is squared, it will be equal to the values of `X-squared` in the R output.

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(Z \leq z_{\text{obs}} = -19.3342) = 0$. This is less than 0.05, so reject H_0.

Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest the proportion of unemployed males given an urban area is less than the proportion of unemployed females given an urban area at an α level of 0.05.

(b) Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of unemployed females given an urban area is greater than 55% are

$$H_0 : \pi = 0.55 \text{ versus } H_1 : \pi > 0.55.$$

Step 2: **Test Statistic** — The test statistic chosen is Y, where Y is the number of unemployed females given an urban area. Provided H_0 is true, $Y \sim Bin(n, \pi_0)$. The value of the test statistic is $y_{\text{obs}} = 6161$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

```
> pvalue <- sum(dbinom(6161:10895, 10895, 0.55))
> pvalue

[1] 0.0005906443

> TR <- binom.test(x = 6161, n = 10895, p = 0.55,
+                  alternative = "greater")
> TR


Exact binomial test

data:  6161 and 10895
number of successes = 6161, number of trials = 10895, p-value =
0.0005906
alternative hypothesis: true probability of success is greater than 0.55
95 percent confidence interval:
 0.5576192 1.0000000
sample estimates:
probability of success
             0.5654888
```

Since the $\wp$-value is 6e-04, reject H_0.

Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest the proportion of unemployed females given an urban area is greater than 55%.

(c) Use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test whether the proportion of unemployed females given a rural area is greater than 50% are

$$H_0 : \pi = 0.50 \text{ versus } H_1 : \pi > 0.50.$$

Step 2: **Test Statistic** — The test statistic chosen is Y, where Y is the number of unemployed females given a rural area. Provided H_0 is true, $Y \sim Bin(n, \pi_0)$. The value of the test statistic is $y_{\text{obs}} = 4033$.

Step 3: **Rejection Region Calculations** — Rejection is based on the $\wp$-value, so none are required.

Step 4: **Statistical Conclusion — Likelihood Method:**

```
> pvalue <- sum(dbinom(4033:7292, 7292, 0.50))
> pvalue

[1] 6.48379e-20
```

```
> TR <- binom.test(x = 4033, n = 7292, p = 0.50,
+                  alternative = "greater")
> TR


Exact binomial test

data:  4033 and 7292
number of successes = 4033, number of trials = 7292, p-value <
2.2e-16
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.5434122 1.0000000
sample estimates:
probability of success
             0.5530719
```

Since the $\wp$-value is 0, reject H_0.

Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest the proportion of unemployed females given a rural area is greater than 50%.

(d) To solve this problem, use the five-step procedure.

Step 1: **Hypotheses** — The null and alternative hypotheses to test if there is evidence to suggest that $\pi_{\text{female|urban}} > \pi_{\text{female|rural}}$ are

$$H_0 : \pi_X = \pi_Y \text{ versus } H_1 : \pi_X > \pi_Y.$$

In this case, let the random variable X represent the number of females given an urban area, and let the random variable Y represent the number of females given a rural area.

Step 2: **Test Statistic** — The test statistic chosen is $P_X - P_Y$ since $E[P_X - P_Y] = \pi_X - \pi_Y$. The standardized test statistic under the assumption that H_0 is true is

$$Z = \frac{P_X - P_Y}{\sqrt{P(1-P)\left(\frac{1}{m} + \frac{1}{n}\right)}}$$

Step 3: **Rejection Region Calculations** — Because the standardized test statistic has an approximate $N(0,1)$ distribution and H_1 is an upper one-sided hypothesis, the rejection region is $z_{\text{obs}} > z_{0.95} = 1.6449$.

```
> x <- 6161
> y <- 4033
> m <- 10895
> n <- 7292
> p <- (x + y)/(m + n)
> p
```

```
[1] 0.5605103

> TR <- prop.test(x = c(x, y), n = c(m, n), correct = FALSE,
+                 alternative = "greater")
> TR


2-sample test for equality of proportions without continuity
correction

data:  c(x, y) out of c(m, n)
X-squared = 2.7341, df = 1, p-value = 0.04911
alternative hypothesis: greater
95 percent confidence interval:
 5.852303e-05 1.000000e+00
sample estimates:
   prop 1    prop 2
0.5654888 0.5530719

> sts <- sqrt(TR$statistic)
> names(sts) <- "z_obs"
> sts

   z_obs
1.653497
```

The pooled estimate of π is $p = \frac{x+y}{m+n} = \frac{6161+4033}{10895+7292} = 0.5605$. The value of the standardized test statistic is 1.6535.

Without a continuity correction,

$$\begin{aligned} z_{\text{obs}} &= \frac{p_X - p_Y}{\sqrt{p(1-p)\left(\frac{1}{m}+\frac{1}{n}\right)}} \\ &= \frac{\frac{6161}{10895} - \frac{4033}{7292}}{\sqrt{\frac{10194}{18187}\left(1-\frac{10194}{18187}\right)\left(\frac{1}{10895}+\frac{1}{7292}\right)}} \\ &= 1.6535 \end{aligned}$$

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(Z \geq z_{\text{obs}} = 1.6535) = 0.0491$. This is less than 0.05, so reject H_0.

Reject H_0.

Step 5: **English Conclusion** — There is sufficient evidence to suggest the proportion of unemployed females given an urban area is greater than the proportion of unemployed females given a rural area at an α level of 0.05.

Solution for 23:

The null and alternative hypotheses to test whether gamma radiation from iodine exceeds 0.3 μGy/h are

$$H_0 : \mu = 0.3 \text{ versus } H_1 : \mu > 0.3.$$

The power of a test is the probability that the null hypothesis is rejected when it is false. Here,

$$\begin{aligned}\text{Power}(\mu_1 = 0.3 \times 1.02) &= \mathbb{P}\left(\text{reject } H_0 \,|\, \mu_1 = 0.3 \times 1.02\right) \\ &= \mathbb{P}\left(\overline{X} > 95^{\text{th}} \text{ percentile of a } N\left(0.3,\, \tfrac{0.015}{\sqrt{n}}\right) \,|\, \mu_1 = 0.306\right)\end{aligned}$$

R is used in an iterative process to discover the value of $n = 99$.

$$\begin{aligned}&= \mathbb{P}\left(\overline{X} > 95^{\text{th}} \text{ percentile of a } N\left(0.3,\, \tfrac{0.015}{\sqrt{99}}\right) \,|\, \mu_1 = 0.306\right) \\ &= 0.9902\end{aligned}$$

```
> alpha <- 0.05
> mu0 <- 0.3
> mu1 <- 1.02*.3
> sigma <- 0.015
> BETA <- 0.01
> n <- 0
> Power <- 0
> while(Power < 1 - BETA){
+   n <- n + 1
+   cv <- qnorm(1 - alpha, mu0, sigma/sqrt(n))
+   Power <- (1 - pnorm(cv, mu1, sigma/sqrt(n)))
+ }
> n

[1] 99

> Power

[1] 0.9902308
```

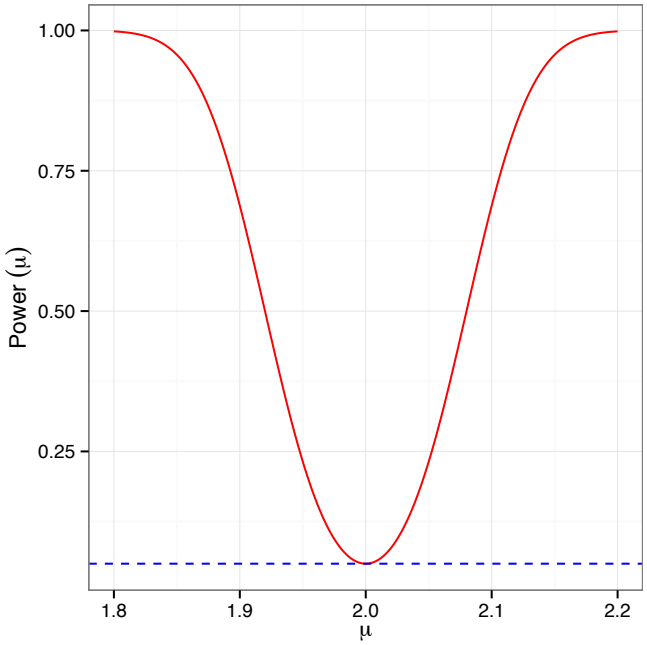
The required sample size (n) to test the hypothesis $H_0 : \mu = 0.3$ versus $H_1 : \mu > 0.3$ for the probability of a type I error (α) to be 0.05 and the probability of a type II error (β) to be no more than 0.01 is 99.

Solution for 25:

(a)

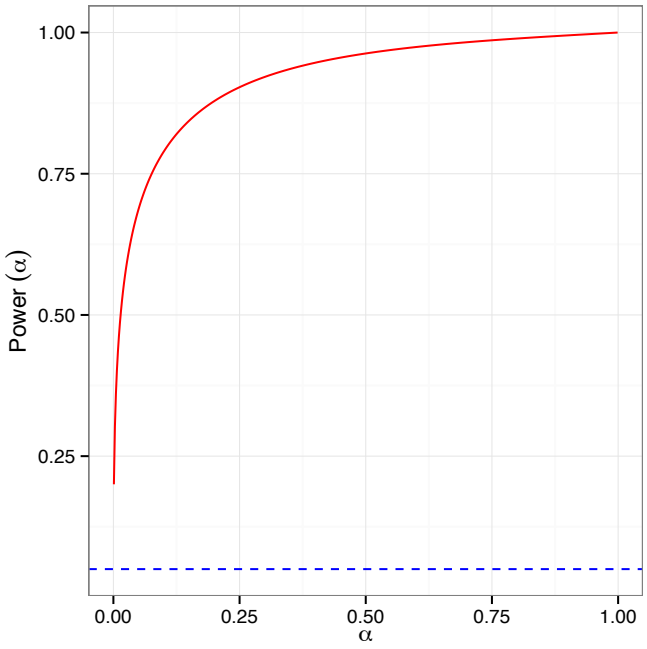
```
> mu <- seq(from = 1.8, to = 2.2, length = 500)
> n <- 150
> alpha <- 0.05
> sigma <- 0.5
> lcv <- qnorm(alpha/2, 2, sigma/sqrt(n))
> ucv <- qnorm(1 - alpha/2, 2, sigma/sqrt(n))
> Power <- pnorm(lcv, mu, sigma/sqrt(n)) +
+   pnorm(ucv, mu, sigma/sqrt(n), lower = FALSE)
> DF <- data.frame(mu, Power)
> ggplot(data = DF, aes(x = mu, y = Power)) +
```

```
+   geom_line(color = "red") +
+   theme_bw() +
+   labs(x = expression(mu), y = expression(Power~(mu))) +
+   geom_hline(yintercept = 0.05, color = "blue", lty = "dashed")
```



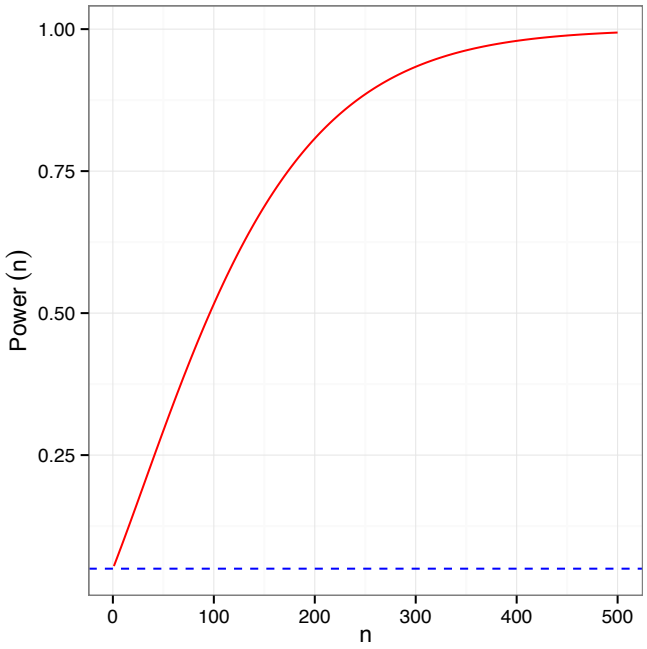
(b)

```
> alpha <- seq(from = 0.001, to = 0.999, length = 500)
> mu <- 2.1
> sigma <- 0.5
> n <- 150
> lcv <- qnorm(alpha/2, 2, sigma/sqrt(n))
> ucv <- qnorm(1 - alpha/2, 2, sigma/sqrt(n))
> Power <- pnorm(lcv, mu, sigma/sqrt(n)) +
+   pnorm(ucv, mu, sigma/sqrt(n), lower = FALSE)
> DF <- data.frame(mu, Power)
> ggplot(data = DF, aes(x = alpha, y = Power)) +
+   geom_line(color = "red") +
+   theme_bw() +
+   labs(x = expression(alpha), y = expression(Power~(alpha))) +
+   geom_hline(yintercept = 0.05, color = "blue", lty = "dashed")
```



(c)

```
> n <- 1:500
> mu <- 2.1
> sigma <- 0.5
> alpha <- 0.05
> lcv <- qnorm(alpha/2, 2, sigma/sqrt(n))
> ucv <- qnorm(1 - alpha/2, 2, sigma/sqrt(n))
> Power <- pnorm(lcv, mu, sigma/sqrt(n)) +
+   pnorm(ucv, mu, sigma/sqrt(n), lower = FALSE)
> DF <- data.frame(mu, Power)
> ggplot(data = DF, aes(x = n, y = Power)) +
+   geom_line(color = "red") +
+   theme_bw() +
+   labs(x = expression(n), y = expression(Power~(n))) +
+   geom_hline(yintercept = 0.05, color = "blue", lty = "dashed")
```



(d) For a fixed value of n and α, the power of a test increases as the true mean moves farther from the hypothesized mean. When the sample size is fixed, the power of a test increases as the level of α increases. When the α-level of the test is fixed, the power of the test increases as the sample size increases.

Solution for 27:

(a)

```
> set.seed(21)
> alpha <- 0.05
> n <- 49
> mu0 <- 100
> mu1 <- 108
> sigma <- 28
> m <- 10000
> xbar <- numeric(m)
> for(i in 1:m){
+   xbar[i] <- mean(rnorm(n, mu1, sigma))
+ }
> ucv <- qnorm(1 - alpha, mu0, sigma/sqrt(n))
> empiricalPOWERa <- mean(xbar >= ucv)
> empiricalPOWERa

[1] 0.6349
```

The empirical power is 0.6349.

(b)

```
> set.seed(21)
> alpha <- 0.20
> n <- 49
> mu0 <- 100
> mu1 <- 108
> sigma <- 28
> m <- 10000
> xbar <- numeric(m)
> for(i in 1:m){
+   xbar[i] <- mean(rnorm(n, mu1, sigma))
+ }
> ucv <- qnorm(1 - alpha, mu0, sigma/sqrt(n))
> empiricalPOWERb <- mean(xbar >= ucv)
> empiricalPOWERb

[1] 0.8795
```

The empirical power is 0.8795.

(c)

```
> cva <- qnorm(0.95, 100, 28/sqrt(49))
> theoPOWERa <- pnorm(cva, 108, 28/sqrt(49), lower = FALSE)
> theoPOWERa

[1] 0.63876

> cvb <- qnorm(0.80, 100, 28/sqrt(49))
> theoPOWERb <- pnorm(cvb, 108, 28/sqrt(49), lower = FALSE)
> theoPOWERb

[1] 0.8766452
```

The theoretical powers for parts (a) and (b) are 0.6388 and 0.8766, respectively. The simulated (empirical) power for (a) and (b) are 0.6349 and 0.8795, respectively. The empirical powers are reasonably close to the theoretical powers.

(d) As α increases, so does the power of the test.

Solution for 29:

(a)

```
> set.seed(42)
> m <- 10000
> n <- 25
> alpha <- 0.05
> mu <- 10
> sigma <- 2.5
> xbar <- numeric(m)
> for(i in 1:m){
+   xbar[i] <- mean(rnorm(n, mu, sigma))
+ }
> lcv <- qnorm(alpha, mu, sigma/sqrt(n))
```

```
> empiricalALPHA <- mean(xbar <= lcv)
> empiricalALPHA

[1] 0.0504
```

The empirical α (0.0504) is close to the theoretical α of 0.05.

(b)

```
> x <- sum(xbar <= lcv)
> n <- m
> c(x, n)

[1]   504 10000

> CI <- prop.test(x = x, n = n, correct = FALSE)$conf
> CI

[1] 0.04628220 0.05486309
attr(,"conf.level")
[1] 0.95
```

The 95% confidence interval for α is [0.0463 0.0549] based on the values from part (a).

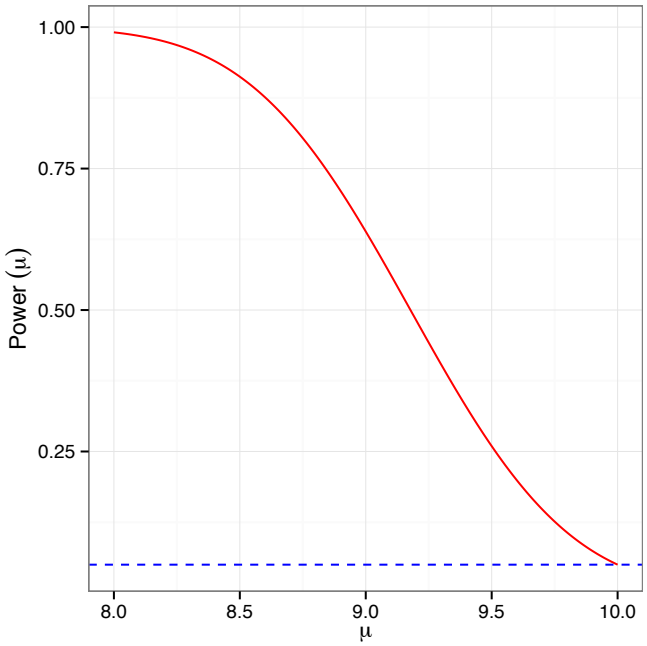
(c)

```
> lcv <- qnorm(0.05, 10, 2.5/sqrt(25))
> lcv

[1] 9.177573

> theoPOWER <- pnorm(lcv, 9.5, 2.5/sqrt(25))
> theoPOWER

[1] 0.259511
```

The theoretical power is 0.2595.

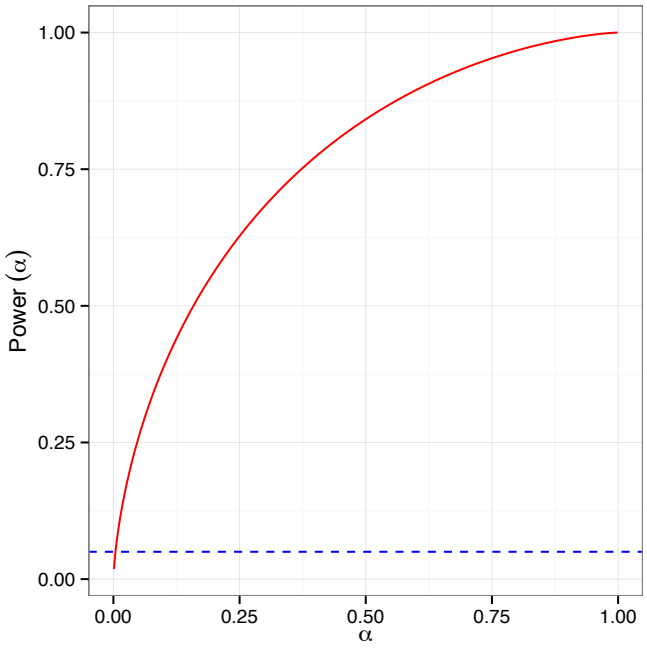
(d)

```
> mu <- seq(from = 8, to = 10, length = 500)
> n <- 25
> alpha = 0.05
> sigma <- 2.5
> lcv <- qnorm(alpha, 10, sigma/sqrt(n))
> Power <- pnorm(lcv, mu, sigma/sqrt(n))
> DF <- data.frame(mu, Power)
> ggplot(data = DF, aes(x = mu, y = Power)) +
+   geom_line(color = "red") +
+   theme_bw() +
+   labs(x = expression(mu), y = expression(Power~(mu))) +
+   geom_hline(yintercept = alpha, color = "blue", lty = "dashed")
```



(e)

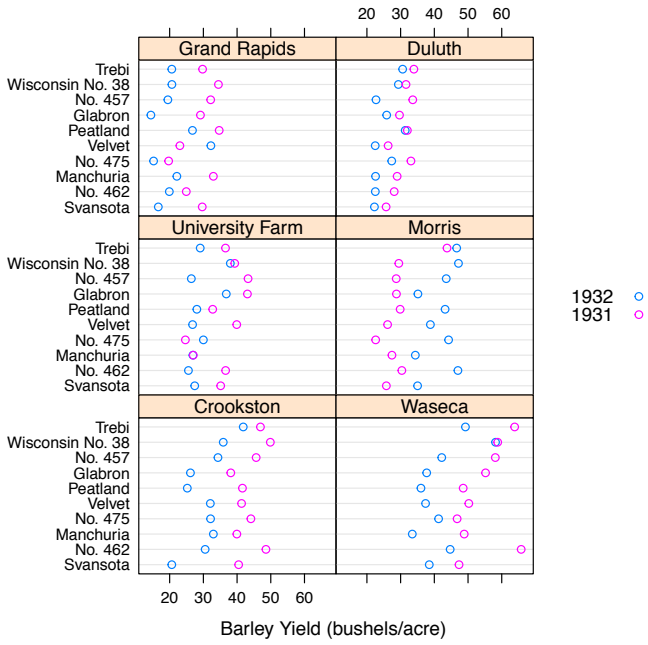
```
> alpha <- seq(from = 0.001, to = 0.999, length = 500)
> n <- 25
> mu1 <- 9.5
> sigma <- 2.5
> lcv <- qnorm(alpha, 10, sigma/sqrt(n))
> Power <- pnorm(lcv, mu1, sigma/sqrt(n))
> DF <- data.frame(mu, Power)
> ggplot(data = DF, aes(x = alpha, y = Power)) +
+   geom_line(color = "red") +
+   theme_bw() +
+   labs(x = expression(alpha), y = expression(Power~(alpha))) +
+   geom_hline(yintercept = 0.05, color = "blue", lty = "dashed")
```



Solution for 31:

(a)

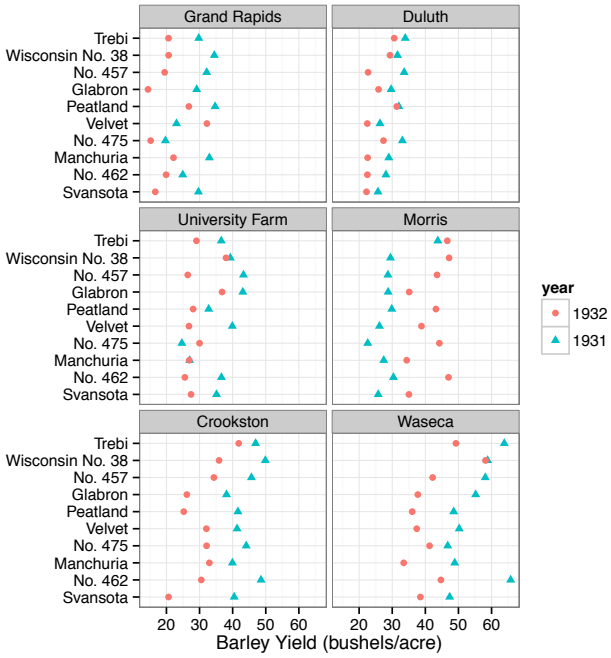
```
> dotplot(variety ~ yield | site, data = barley, groups = year,
+         key = simpleKey(levels(barley$year), space = "right"),
+         xlab = "Barley Yield (bushels/acre) ",
+         layout = c(2, 3), ylab = NULL, as.table = TRUE)
```



Note that `layout = c(1, 6)` shows the dots more clearly switching, for the Morris site; however, the fonts become illegible.

(b)

```
> ggplot(data = barley, aes(yield, variety, color = year, shape = year)) +
+   geom_point() +
+   facet_wrap(~site, ncol = 2) +
+   theme_bw() +
+   labs(x = "Barley Yield (bushels/acre)", y = "")
```



(c) Start my recoding the 1931 Morris barley yield.

```
> yieldMor32 <- with(data = barley,
+                    yield[year == "1931" & site == "Morris"]
+ )
> yieldCro32 <- with(data = barley,
+                    yield[year == "1932" & site == "Crookston"]
+ )
```

Step 1: **Hypotheses** — To test if the average 1932 (recorded as 1931's yield) barley yield from Morris is greater than the average 1932 barley yield from Crookston, the hypotheses are

$$H_0 : \mu_D = 0 \text{ versus } H_1 : \mu_D > 0$$

Step 2: **Test Statistic** — The test statistic chosen is $\overline{D}$ because $E\left[\overline{D}\right] = \mu_D$.

```
> Dif <- yieldMor32 - yieldCro32
> dbar <- mean(Dif)
> dbar

[1] -1.893329
```

The value of this test statistic is $\bar{d} = -1.8933$. The standardized test statistic under the assumption that H_0 is true and its distribution are $\frac{\overline{D}-\delta_0}{S_D/\sqrt{n_D}} \sim t_{10-1}$.

Step 3: **Rejection Region Calculations** — Because the standardized test statistic is distributed t_9, and H_1 is an upper one-sided hypothesis, the rejection region is $t_{\text{obs}} > t_{1-0.05;9} = t_{0.95;9} = 1.8331$.

```
> RR <- qt(0.95, 9)
> RR

[1] 1.833113

> TR <- t.test(Dif, alternative = "greater")
> TR


One Sample t-test

data:  Dif
t = -1.1307, df = 9, p-value = 0.8563
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 -4.962699       Inf
sample estimates:
mean of x
-1.893329
```

The value of the standardized test statistic is $t_{\text{obs}} = \frac{\bar{d}-\delta_0}{s_D/\sqrt{n_D}} = \frac{-1.8933-0}{5.2949/\sqrt{10}} = -1.1307$.

Step 4: **Statistical Conclusion** — The $\wp$-value is $\mathbb{P}(t_9 \geq -1.1307) = 0.8563$.

I. From the rejection region, fail to reject H_0 because $t_{\text{obs}} = -1.1307$ is less than 1.8331.

II. From the $\wp$-value, fail to reject H_0 because the $\wp$-value $= 0.8563$ is more than 0.05.

Reject H_0.

Step 5: **English Conclusion** — There is not sufficient evidence to suggest that the 1932 mean barley yield for Morris (as recorded under 1931) is greater than the 1932 mean barley yield for Crookston.

Chapter 10

Odd solutions

Solution for 1:

Pros: The assumptions with respect to the underlying population are not as strict as the assumptions required for parametric tests.

Cons: If the underlying population distribution is normal, the non-parametric tests are less powerful than their parametric counterparts.

Solution for 3:

The Wilcoxon signed-rank test test has more power than the sign test for testing the median difference of two dependent samples. The assumptions required for the Wilcoxon signed-rank test test are a continuous and symmetric distribution for the population of differences and a median that exists.

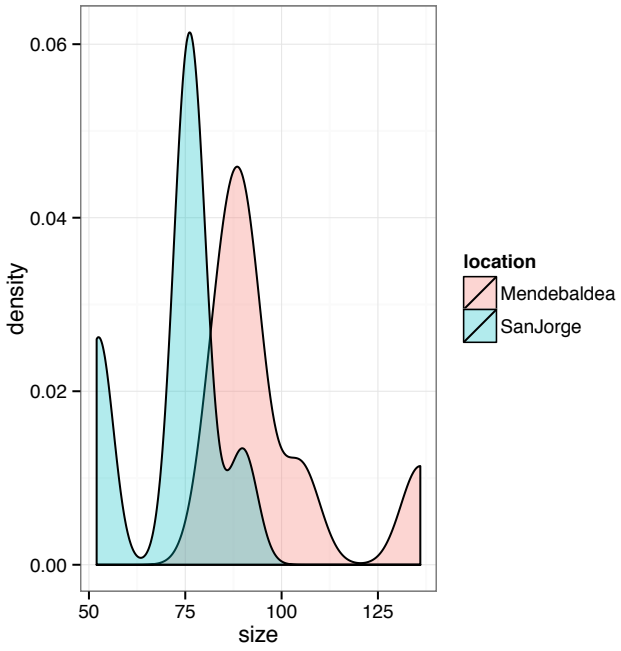
Solution for 5:

Many statistical procedures require knowledge of the population from which the sample is taken. Goodness-of-fit procedures are typically used when the form of the population is in question. The null hypothesis in a goodness-of-fit test is a statement about the form of the cumulative distribution.

Solution for 7:

(a)

```
> ggplot(data = APTSIZE) +
+   geom_density(aes(x = size, fill = location), alpha = 0.3) +
+   theme_bw()
```



Based on the densities, the distributional shapes and skews for the two apartments appear to be different; however, due to the small sample sizes (7 and 8), it is very hard to reject the null hypothesis that the two **cdf**s are the same. Consequently, one might assume the distributions are similar and proceed with a Wilcoxon rank-sum test procedure. An alternative approach might be to perform a permutation test.

(i) Exact procedure:

```
> library(coin)

Loading required package:  survival
   Attaching package:  'coin'
   The following object is masked _by_  '.GlobalEnv':
       alpha

> wilcox_test(size ~ location, data = APTSIZE, distribution = "exact",
+             alternative = "greater")


Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  size by location (Mendebaldea, SanJorge)
Z = 2.9167, p-value = 0.001088
alternative hypothesis: true mu is greater than 0

> SJ <- APTSIZE$size[APTSIZE$location == "SanJorge"]
> MB <- APTSIZE$size[APTSIZE$location == "Mendebaldea"]
> TR <- wilcoxe.test(MB, SJ, alternative = "greater")
> TR
```

```
Wilcoxon Rank Sum Test

data:  MB and SJ
w = 81, p-value = 0.001088
alternative hypothesis: true median is greater than 0
88.53147 percent confidence interval:
  15 Inf
sample estimates:
difference in location
                    16
```

Based on the small $\wp$-value (0.0011), reject the null hypothesis. The evidence supports the agent's claim that the size of Mendebaldea apartments is greater than the size of San Jorge apartments.

(ii) Approximate procedure:

```
> TR <- wilcox_test(size ~ location, data = APTSIZE,
+                   alternative = "greater")
> TR


Asymptotic Wilcoxon Mann-Whitney Rank Sum Test

data:  size by location (Mendebaldea, SanJorge)
Z = 2.9167, p-value = 0.001769
alternative hypothesis: true mu is greater than 0
```

Based on the small $\wp$-value (0.0018), reject the null hypothesis. The evidence supports the agent's claim that the size of Mendebaldea apartments is greater than the size of San Jorge apartments.

Approximate permutation approach:

```
> TM <- with(data = APTSIZE,
+            tapply(size, location, mean, trim = 0.10)
+ )
> TM

Mendebaldea    SanJorge
   97.28571    72.00000

> obsDiff <- TM[1] - TM[2]
> obsDiff

Mendebaldea
   25.28571

> Size <- APTSIZE$size
> N <- 10^4 - 1              # number of times to repeat the process
> Diff <- numeric(N)         # space to save the random differences
> set.seed(11)
> for (i in 1:N) {
```

```
+ # sample of size 8, from 1 to 15, without replacement
+ index <- sample(15, size = 8, replace = FALSE)
+ Diff[i] <- mean(Size[index], trim = 0.10) -
+             mean(Size[-index], trim = 0.10)
+ }
> pvalueTM <- (sum(Diff >= obsDiff) + 1)/(N + 1) # p-value
> pvalueTM # results will vary

[1] 7e-04
```

Based on the small $\wp$-value (7e-04), reject the null hypothesis. The evidence supports the agent's claim that the size of Mendebaldea apartments is greater than the size of San Jorge apartments.

(b)

```
> TR <- wilcox_test(size ~ location, data = APTSIZE, distribution = "exact",
+              conf.int = TRUE, conf.level = 0.90)
> TR


Exact Wilcoxon Mann-Whitney Rank Sum Test

data:  size by location (Mendebaldea, SanJorge)
Z = 2.9167, p-value = 0.001865
alternative hypothesis: true mu is not equal to 0
90 percent confidence interval:
 10 38
sample estimates:
difference in location
                    16
```

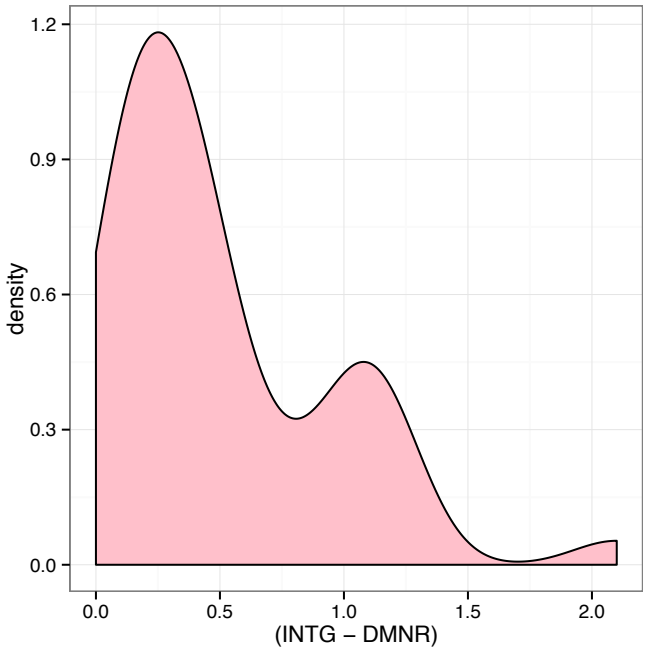
A 90% confidence interval using `wilcox_test()` for the median of Mendebaldea minus the median of San Jorge is [10, 38]. Note that the actual confidence level is higher than 90% when using `wilcox_test()`. To obtain the actual confidence level, one might use the function `wilcoxe.test()`, which reports the closest confidence level that can be achieved given the requested argument to `conf.level`.

```
> TR <- wilcoxe.test(MB, SJ, conf.level = 0.92)$conf
> TR

[1] 14 30
attr(,"conf.level")
[1] 0.9179487
```

A 91.7949% confidence interval using `wilcoxe.test()` for the median of Mendebaldea minus the median of San Jorge is [14, 30].

Solution for 9:

```
> ggplot(data = USJudgeRatings, aes(x = (INTG - DMNR))) +
+   geom_density(fill = "pink") +
+   theme_bw()
```



Since the density plot of `INTR − DMNR` is skewed to the right, use the sign test to test if lawyers are more likely to give a judge high integrity ratings rather than high demeanor ratings.

```
> Dif <- USJudgeRatings$INTG - USJudgeRatings$DMNR
> SIGN.test(Dif, md = 0, alternative = "greater")


One-sample Sign-Test

data:  Dif
s = 41, p-value = 4.552e-13
alternative hypothesis: true median is greater than 0
95 percent confidence interval:
 0.2563989       Inf
sample estimates:
median of x
        0.4
                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9369 0.3000    Inf
Interpolated CI       0.9500 0.2564    Inf
Upper Achieved CI     0.9670 0.2000    Inf
```

Based on the small $\wp$-value, reject the null hypothesis. The evidence suggests lawyers are more likely to give a judge high integrity ratings rather than high demeanor ratings.

(b)

```
> SIGN.test(Dif, md = 0, conf.level = 0.90)
```

```
One-sample Sign-Test

data:  Dif
s = 41, p-value = 9.104e-13
alternative hypothesis: true median is not equal to 0
90 percent confidence interval:
 0.2563989 0.4436011
sample estimates:
median of x
        0.4
                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.8737 0.3000 0.4000
Interpolated CI       0.9000 0.2564 0.4436
Upper Achieved CI     0.9340 0.2000 0.5000
```

An interpolated 90% confidence interval for the median differences (`INTR` − `DMNR`) is [0.2564, 0.4436].

Solution for 11:

(a)

```
> DATA <- c(0.98, 0.95, 0.91, 0.93, 0.94, 0.94, 0.89, 0.88, 0.90, 0.93)
> OBS <- mean(DATA[1:5]) - mean(DATA[6:10])
> OBS

[1] 0.034

> ANS <- t(combn(10, 5))
> head(ANS)

     [,1] [,2] [,3] [,4] [,5]
[1,]    1    2    3    4    5
[2,]    1    2    3    4    6
[3,]    1    2    3    4    7
[4,]    1    2    3    4    8
[5,]    1    2    3    4    9
[6,]    1    2    3    4   10

> nn <- dim(ANS)[1]
> nn

[1] 252

> means <- numeric(nn)
> for (i in 1:nn) {
+ means[i] <- mean(DATA[ANS[i, ]]) - mean(DATA[-ANS[i, ]])
+ }
> pvalue <- mean(means >= OBS)
> pvalue

[1] 0.04365079
```

The exact $\wp$-value is 0.0437.

(b)

```
> Company <- factor(c(rep("Japanese", 5), rep("American", 5)))
> DF <- data.frame(Time = DATA, Company)
> head(DF)

  Time  Company
1 0.98 Japanese
2 0.95 Japanese
3 0.91 Japanese
4 0.93 Japanese
5 0.94 Japanese
6 0.94 American

> library(coin)
> oneway_test(Time ~ Company, distribution = "exact",
+             alternative = "less", data = DF)


Exact 2-Sample Permutation Test

data:  Time by Company (American, Japanese)
Z = -1.7756, p-value = 0.04365
alternative hypothesis: true mu is less than 0
```

Note that the $\wp$-values for (a) and (b) agree.

(c)

```
> library(boot)


Attaching package:  'boot'
   The following object is masked from 'package:survival':
       aml
   The following object is masked from 'package:lattice':
       melanoma

> set.seed(12)
> meandiff <- function(data, i){
+   d <- data[i]
+   MD <- mean(d[1:5]) - mean(d[6:10])
+   MD
+ }
> ans <- boot(DATA, meandiff, sim = "permutation", R = 10^4 - 1)
> pvalue <- (sum(ans$t >= ans$t0) + 1)/(10^4 - 1 + 1)
> pvalue

[1] 0.0437
```

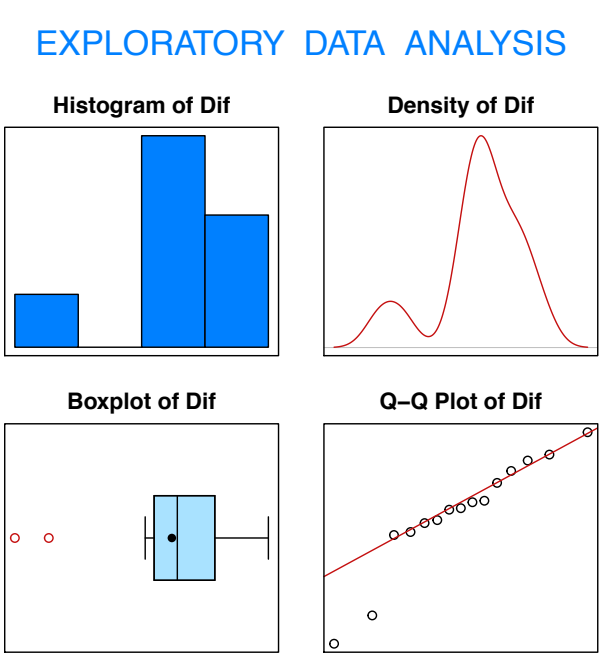
(d) All of the $\wp$-values are less than 0.05, suggesting the average time it takes the American company to transmit 1 terabyte is less than the average time it takes the Japanese company to transmit 1 terabyte.

Solution for 13:

(a) The samples are paired as each pair comes from a single pot.

(b)

```
> Dif <- FERTILIZE$height[FERTILIZE$fertilization == "cross"] -
+   FERTILIZE$height[FERTILIZE$fertilization == "self"]
> eda(Dif)

Size (n)  Missing  Minimum   1st Qu     Mean   Median   TrMean   3rd Qu
  15.000    0.000   -8.375    1.375    2.617    3.000    2.617    5.625
     Max    Stdev      Var  SE Mean   I.Q.R.    Range Kurtosis Skewness
   9.375    4.718   22.260    1.218    4.250   17.750    0.141   -0.895
SW p-val
   0.098
```



Normality is questionable.

```
> TR <- t.test(Dif)
> TR


One Sample t-test

data:  Dif
t = 2.148, df = 14, p-value = 0.0497
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.003899165 5.229434169
sample estimates:
mean of x
 2.616667
```

Based on a $\wp$-value of 0.0497, the evidence suggests the mean difference (cross fertilized plant height − self fertilized plant height) is not zero.

(c)

```
> TR <- wilcox.test(Dif)
> TR


Wilcoxon signed rank test

data:  Dif
V = 96, p-value = 0.04126
alternative hypothesis: true location is not equal to 0
```

Based on a $\wp$-value of 0.0413, the evidence suggests the mean difference (cross fertilized plant height − self fertilized plant height) is not zero.

(d)

```
> obsMdif <- mean(Dif)
> obsMdif

[1] 2.616667

> sims <- 10^4 - 1
> MDif <- numeric(sims)
> for (i in 1:sims) {
+ PM <- sample(c(-1, 1), size = length(Dif), replace = TRUE)
+ MDif[i] <- mean(Dif*PM)
+ }
> pvalue <- ((sum(MDif >= obsMdif) + 1)/(sims + 1))*2
> pvalue

[1] 0.0512
```

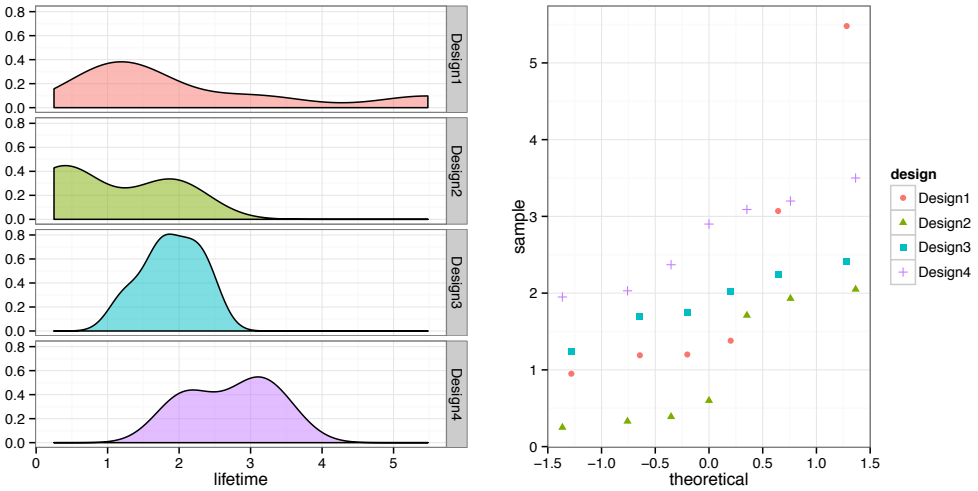
Based on a $\wp$-value of 0.0512, their is some evidence to suggest the mean difference (cross fertilized plant height − self fertilized plant height) is not zero.

(e) The $\wp$-values for the tests performed in (b), (c), and (d) are all close to 0.05.

Solution for 15:

(a)

```
> ggplot(data = CIRCUIT, aes(x = lifetime, fill = design)) +
+   geom_density(alpha = 0.5) +
+   facet_grid(design ~ .) +
+   guides(fill = FALSE) +
+   labs(y = "") +
+   theme_bw()
> ggplot(data = CIRCUIT, aes(sample = lifetime, shape = design,
+                            color = design)) +
+   stat_qq() +
+   theme_bw()
```



The density plots and quantile-quantile normal plots make normality questionable; however, ruling out normality with so few observations is difficult.

(b)

```
> TR <- kruskal.test(lifetime ~ design, data = CIRCUIT)
> TR


Kruskal-Wallis rank sum test

data:  lifetime by design
Kruskal-Wallis chi-squared = 10.245, df = 3, p-value = 0.0166
```

The $\wp$-value from the Kruskal-Wallis test of 0.0166 suggests differences exist among the mean lifetimes of different circuit designs.

(c)

```
> N <- 10^4 - 1
> set.seed(3)
> Fs <- numeric(N)
> Fobs <- summary(aov(lifetime ~ design, data = CIRCUIT))[[1]][1, 4]
> for (i in 1:N) {
+   Fs[i] <- summary(aov(lifetime ~ sample(design),
+                        data = CIRCUIT))[[1]][1, 4]
+ }
> pvalue <- (sum(Fs >= Fobs) + 1)/(N + 1)
> pvalue

[1] 0.0317
```

The permutation $\wp$-value of 0.0317 suggests differences exist among the mean lifetimes of different circuit designs.

Solution for 17:

(a)

```
> FT <- xtabs(breaks ~ wool + tension, data = warpbreaks)
> FT

    tension
wool   L   M   H
   A 401 216 221
   B 254 259 169
```

(b) Hypothesis: Wool and tension are independent.

```
> TR <- chisq.test(FT)
> TR


Pearson's Chi-squared test

data:  FT
X-squared = 28.102, df = 2, p-value = 7.901e-07
```

The $\wp$-value of 0 suggest there is an association between wool and tension.

Solution for 19:

(a)

```
> number <- c(47, 8, 54, 43,
+             31, 5, 21, 17,
+             64, 4, 43, 43,
+             134, 11, 104, 66)
> bank <- factor(rep(c("BBVA", "CM", "LC", "BS"), times = 4))
> region <- factor(rep(c("Navarra", "Alava", "Guipuzcoa", "Vizcaya"),
+                      each = 4))
> DF <- data.frame(number, bank, region)
> rm(number, bank, region)  # clean up
> FT <- xtabs(number ~ region + bank, data = DF)
> FT

           bank
region      BBVA  BS  CM  LC
  Alava       31  17   5  21
  Guipuzcoa   64  43   4  43
  Navarra     47  43   8  54
  Vizcaya    134  66  11 104
```

(b) There are only four observations in the Guizpuzcoa region of the CM bank. Banks CM and BS are combined to form a single category.

```
> fix.table <- cbind(FT[, c(1, 4)], apply(FT[, 2:3], 1, sum))
> dimnames(fix.table)[[2]] <- c("BBVA", "LC", "CM & BS")
> fix.table

          BBVA  LC CM & BS
Alava       31  21      22
Guipuzcoa   64  43      47
```

```
Navarra     47  54      51
Vizcaya    134 104      77

> TR <- chisq.test(fix.table)
> TR


Pearson's Chi-squared test

data:  fix.table
X-squared = 9.0633, df = 6, p-value = 0.1701
```

There is little evidence to suggest any association exists between `region` and `bank` based on the $\wp$-value of 0.1701.

(c)

```
> DFFT <- as.data.frame(FT)
> DF2 <- vcdExtra::expand.dft(DFFT)
> N <- 10^4 - 1
> Chi <- numeric(N)
> FRT <- xtabs(~ region + bank, data = DF2)
> ChiObs <- chisq.test(FRT)$statistic
> ChiObs

X-squared
 11.89246

> for(i in 1:N){
+   SFT <- xtabs(~ sample(region) + bank, data = DF2)
+   Chi[i] <- chisq.test(SFT)$statistic
+ }
> pvalue <- (sum(Chi >= ChiObs) + 1)/(N + 1)
> pvalue

[1] 0.2194
```

Based on the permutation $\wp$-value of 0.2194, there is little evidence to suggest any association exists between `region` and `bank`.

Solution for 21:

```
> TR <- ks.test(x = PHONE$call.time, y = "pexp", rate = 1/3.7)
> TR


One-sample Kolmogorov-Smirnov test

data:  PHONE$call.time
D = 0.15339, p-value = 0.6514
alternative hypothesis: two-sided
```

Yes, it is reasonable to assume `call.time` follows an $Exp(\lambda = 3.7)$ distribution. Based on a $\wp$-value of 0.6514, there is little evidence to suggest the distribution is not an $Exp(\lambda = 3.7)$.

Solution for 23:

(a)

```
> connection.time <- c(0.03, 0.48, 0.49, 0.52, 0.66, 0.69, 0.70,
+                      0.76, 0.82, 1.20, 1.22, 1.39, 1.62, 1.85,
+                      1.97, 2.25, 2.84, 3.44, 3.48, 4.02)
> TR <- ks.test(x = connection.time, y = "pexp", rate = 1/1.5)
> TR


One-sample Kolmogorov-Smirnov test

data:  connection.time
D = 0.22385, p-value = 0.2315
alternative hypothesis: two-sided
```

Based on a $\wp$-value of 0.2315, there is not sufficient evidence to suggest the distribution is not an exponential with a mean of 1.5.

(b)

```
> Categories <- cut(connection.time, breaks = c(0, 1, 2, Inf))
> Obs <- xtabs(~ Categories)
> Obs

Categories
  (0,1]   (1,2] (2,Inf]
      9       6       5

> Rate <- 1 / 1.5
> Prob <- c(pexp(1, rate = Rate), pexp(2, rate = Rate) -
+           pexp(1, rate = Rate), pexp(2, rate = Rate, lower = FALSE))
> Prob

[1] 0.4865829 0.2498200 0.2635971

> TR <- chisq.test(x = Obs, p = Prob)

Warning in chisq.test(x = Obs, p = Prob):  Chi-squared approximation may be
incorrect

> TR


Chi-squared test for given probabilities

data:  Obs
X-squared = 0.27062, df = 2, p-value = 0.8734
```

There is not sufficient evidence ($\wp$-value = 0.8734) to suggest the distribution is not an exponential with a mean of 1.5.

(c)

```
> SIGN.test(connection.time, md = 1, alternative = "greater")


One-sample Sign-Test

data:  connection.time
s = 11, p-value = 0.4119
alternative hypothesis: true median is greater than 1
95 percent confidence interval:
 0.697928      Inf
sample estimates:
median of x
       1.21
                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.9423 0.7000    Inf
Interpolated CI       0.9500 0.6979    Inf
Upper Achieved CI     0.9793 0.6900    Inf
```

There is not sufficient evidence to suggest the median connection time is greater than 1 second.

Solution for 25:

(a)

```
> MALES <- HairEyeColor[, , 1]
> MALES

       Eye
Hair    Brown Blue Hazel Green
  Black    32   11    10     3
  Brown    53   50    25    15
  Red      10   10     7     7
  Blond     3   30     5     8

> TR <- chisq.test(MALES)
> TR


Pearson's Chi-squared test

data:  MALES
X-squared = 41.28, df = 9, p-value = 4.447e-06

> DFT <- as.data.frame(as.table(MALES))
> DF <- vcdExtra::expand.dft(DFT)
> sims <- 10^4 - 1
> Chi <- numeric(sims)
> FT <- xtabs(~ Hair + Eye, data = DF)
> ChiObs <- chisq.test(FT)$statistic
> ChiObs

X-squared
 41.28029
```

```
> for(i in 1:sims){
+   SFT <- xtabs(~ sample(Hair) + Eye, data = DF)
+   Chi[i] <- chisq.test(SFT)$statistic
+ }
> pvalue <- (sum(Chi >= ChiObs) + 1)/(sims + 1)
> pvalue

[1] 1e-04
```

The $\wp$-value from testing independence using `chisq.test()` is 0, while the $\wp$-value from the randomization test of independence is 1e-04. Both $\wp$-values suggest a strong association exists for males between hair and eye color.

(b)

```
> FEMALES <- HairEyeColor[, , 2]
> FEMALES

       Eye
Hair    Brown Blue Hazel Green
  Black    36    9     5     2
  Brown    66   34    29    14
  Red      16    7     7     7
  Blond     4   64     5     8

> TR <- chisq.test(FEMALES)
> TR


Pearson's Chi-squared test

data:  FEMALES
X-squared = 106.66, df = 9, p-value < 2.2e-16

> DFT <- as.data.frame(as.table(FEMALES))
> DF <- vcdExtra::expand.dft(DFT)
> sims <- 10^4 - 1
> Chi <- numeric(sims)
> FT <- xtabs(~ Hair + Eye, data = DF)
> ChiObs <- chisq.test(FT)$statistic
> ChiObs

X-squared
 106.6637

> for(i in 1:sims){
+   SFT <- xtabs(~ sample(Hair) + Eye, data = DF)
+   Chi[i] <- chisq.test(SFT)$statistic
+ }
> pvalue <- (sum(Chi >= ChiObs) + 1)/(sims + 1)
> pvalue

[1] 1e-04
```

The $\wp$-value from testing independence using `chisq.test()` is 0, while the $\wp$-value from the randomization test of independence is 1e-04. Both $\wp$-values suggest a strong association exists for females between hair and eye color.

Solution for 27:

Due to the small sample sizes and the discrete nature of the scores, a permutation test is used to see if the new test yields higher scores for patients who are known to suffer from delusions than patients that do not suffer from delusions.

```
> present <- c(5, 5, 4, 5, 4, 5, 5)
> absent <- c(1, 0, 5, 0, 4, 4, 0)
> Number <- c(present, absent)
> Delusions <- factor(c(rep("present", 7), rep("absent", 7)))
> DF <- data.frame(Number, Delusions)
> rm(present, absent, Number, Delusions)
> library(coin)
> oneway_test(Number ~ Delusions, data = DF, distribution = "exact",
+             alternative = "less")


Exact 2-Sample Permutation Test

data:  Number by Delusions (absent, present)
Z = -2.4205, p-value = 0.01166
alternative hypothesis: true mu is less than 0
```

Evidence suggests ($\wp$-value = 0.0117) the new test yields higher scores for patients who are known to suffer from delusions than patients who do not suffer from delusions.

Solution for 29:

```
> B <- 10000
> set.seed(10)
> n <- 10
> xbar <-  numeric(B)
> xs <- rnorm(n, 0, 1)
> for(i in 1:B){
+   xbar[i] <- mean(sample(xs, n, replace = TRUE))
+ }
> sd(xbar)

[1] 0.2105907

> PE10 <- (abs(sd(xbar) - 1/sqrt(n))/(1/sqrt(n)))*100
> PE10

[1] 33.40538

> library(boot)
> MEAN <- function(data, i){
+   d <- data[i]
+   M <- mean(d)
+   M
```

```
+ }
> set.seed(10)
> boot10 <- boot(xs, MEAN, R = B)
> boot10


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = xs, statistic = MEAN, R = B)


Bootstrap Statistics :
      original       bias    std. error
t1* -0.4906568 0.0006681112   0.2088944

> sd(boot10$t)

[1] 0.2088944

> PEboot10 <- (abs(sd(boot10$t) - 1/sqrt(n))/(1/sqrt(n)))*100
> PEboot10

[1] 33.94179
```

```
> B <- 10000
> set.seed(10)
> n <- 100
> xbar <-  numeric(B)
> xs <- rnorm(n, 0, 1)
> for(i in 1:B){
+   xbar[i] <- mean(sample(xs, n, replace = TRUE))
+ }
> sd(xbar)

[1] 0.0928912

> PE100 <- (abs(sd(xbar) - 1/sqrt(n))/(1/sqrt(n)))*100
> PE100

[1] 7.108795

> set.seed(10)
> boot100 <- boot(xs, MEAN, R = B)
> boot100


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
```

```
boot(data = xs, statistic = MEAN, R = B)


Bootstrap Statistics :
      original        bias    std. error
t1* -0.1365489 -0.0005420164  0.09386073

> sd(boot100$t)

[1] 0.09386073

> PEboot100 <- (abs(sd(boot100$t) - 1/sqrt(n))/(1/sqrt(n)))*100
> PEboot100

[1] 6.139268


> B <- 10000
> set.seed(10)
> n <- 1000
> xbar <-  numeric(B)
> xs <- rnorm(n, 0, 1)
> for(i in 1:B){
+   xbar[i] <- mean(sample(xs, n, replace = TRUE))
+ }
> sd(xbar)

[1] 0.03128147

> PE1000 <- (abs(sd(xbar) - 1/sqrt(n))/(1/sqrt(n)))*100
> PE1000

[1] 1.079309

> set.seed(10)
> boot1000 <- boot(xs, MEAN, R = B)
> boot1000


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = xs, statistic = MEAN, R = B)


Bootstrap Statistics :
      original        bias    std. error
t1* 0.01137474 -0.0003534521  0.03133319

> sd(boot1000$t)

[1] 0.03133319
```

```
> PEboot1000 <- (abs(sd(boot1000$t) - 1/sqrt(n))/(1/sqrt(n)))*100
> PEboot1000

[1] 0.9157565
```

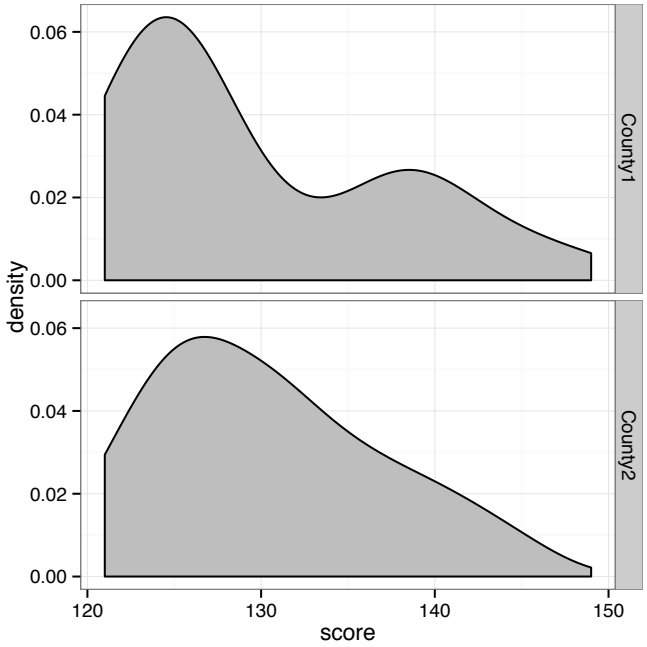
The percent difference from what the standard error should be decreases as the sample size increases.

Solution for 31:

(a) The standard deviation for the Stanford-Binet IQ test is for the general population not gifted and talented students. Consequently, the value 16 should not be used.

(b) Note that both counties have distributions that are skewed to the right. The normality assumptions required to use a confidence interval based on the t-distribution are most likely not satisfied. In this scenario, it would be appropriate to construct a confidence interval based on the Wilcoxon Rank Sum test.

```
> ggplot(data = SBIQ, aes(x = score)) +
+   geom_density(fill = "gray") +
+   facet_grid(county ~.) +
+   theme_bw()
> CI <- wilcox.test(score ~ county, data = SBIQ, conf.int = TRUE,
+                   conf.level = 0.96)$conf
> CI

[1] -4.000024  1.999988
attr(,"conf.level")
[1] 0.96
```



A 96% confidence interval for the true average IQ difference between gifted and talented students in `County1` and in `County2` is $[-4, 2]$ based on the Wilcoxon Rank Sum test.

(c)

```
> set.seed(12)
> sims <- 10^4 - 1
> DM <- numeric(sims)
> ScoreCounty1 <- subset(SBIQ, county == "County1",
+                        select = score, drop = TRUE)
> ScoreCounty2 <- subset(SBIQ, county == "County2",
+                        select = score, drop = TRUE)
> for(i in 1:sims){
+   SC1 <- sample(ScoreCounty1, size = 40, replace = TRUE)
+   SC2 <- sample(ScoreCounty2, size = 40, replace = TRUE)
+   DM[i] <- mean(SC1) - mean(SC2)
+ }
> CIV <- sort(DM)
> BCIp <- c(CIV[(sims + 1)*0.02], CIV[(sims + 1)*0.98])
> BCIp

[1] -3.850  2.775

> library(boot)
> MDS <- function(data, i){
+   d <- data[i, ]
+   m <- tapply(d$score, d$county, mean)
+   md <- m[1] - m[2]
+   md
+ }
> set.seed(1)
> boot.obj <- boot(data = SBIQ, statistic = MDS, R = sims)
> BBCI <- boot.ci(boot.obj, type = "perc", conf = 0.96)
> BBCI

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 9999 bootstrap replicates

CALL :
boot.ci(boot.out = boot.obj, conf = 0.96, type = "perc")

Intervals :
Level     Percentile
96%   (-3.9214,  2.7500 )
Calculations and Intervals on Original Scale
```

Two bootstrap percentile confidence intervals are constructed. The first uses the function `sample()` to do the resampling while the second bootstrap percentile confidence interval uses the function `boot()` inside the package `boot` to do the resampling. The first 96% bootstrap percentile confidence interval for the true average IQ difference between gifted and talented students in `County1` and in `County2` is $[-3.85, 2.775]$. The second 96% bootstrap percentile confidence interval for the true average IQ difference between gifted and talented students in `County1` and in `County2` is $[-3.9214, 2.75]$. Using different seed values will result in slightly different confidence intervals.

Chapter 11

Odd solutions

Solution for 1:

```
> set.seed(1)
> population <- rep(LETTERS[1:3], 5)
> treatment  <- sample(population)
> DF       <- data.frame(run = 1:15, treatment)
> head(DF)

  run treatment
1   1         A
2   2         C
3   3         B
4   4         B
5   5         C
6   6         C
```

The full data frame `DF` is the randomization scheme.

Solution for 3:

```
> set.seed(1)
> factor1 <- factor(rep(LETTERS[1:4], times = 9))
> factor2 <- factor(rep(c("I", "II", "III"), each = 12))
> expt_units <- rep(1:3, 12)
> ftable(xtabs(sample(1:36) ~ factor1 + factor2 + expt_units))

                expt_units  1  2  3
factor1 factor2
A       I                  10  7 18
        II                 17 15 33
        III                16 31 13
B       I                   2 14 28
        II                  4  9 19
        III                22 25  3
C       I                  29  6 20
        II                 32 36 24
        III                 8 23  1
D       I                  30 34  5
        II                 11 35 27
        III                12 26 21
```

Solution for 5:

(a)

```
> dof <- c(4, 3, round(.01703/.00142, 0))
> SS <- c(0.0073, 3*0.35431, 0.01703)
> MS <- c(0.00073/4, 0.35431, 0.00142)
> Fobs <- c(MS[1]/MS[3], MS[2]/MS[3], NA)
> PrF <- c(pf(Fobs[1], dof[1], dof[3], lower = FALSE),
+          pf(Fobs[2], dof[2], dof[3], lower = FALSE), NA)
> TABLE <- cbind(dof, SS, MS, Fobs, PrF)
> rownames(TABLE) <- c("block", "factor", "Residuals")
> TABLE

          dof      SS        MS        Fobs          PrF
block       4 0.00730 0.0001825   0.1285211 9.691238e-01
factor      3 1.06293 0.3543100 249.5140845 4.494132e-11
Residuals  12 0.01703 0.0014200          NA           NA
```

(b) There are four levels for `factor`.

(c) There are five blocks in the design.

(d) The model's parameters are the positive and negative increments to the global mean.

(e) Answers will vary.

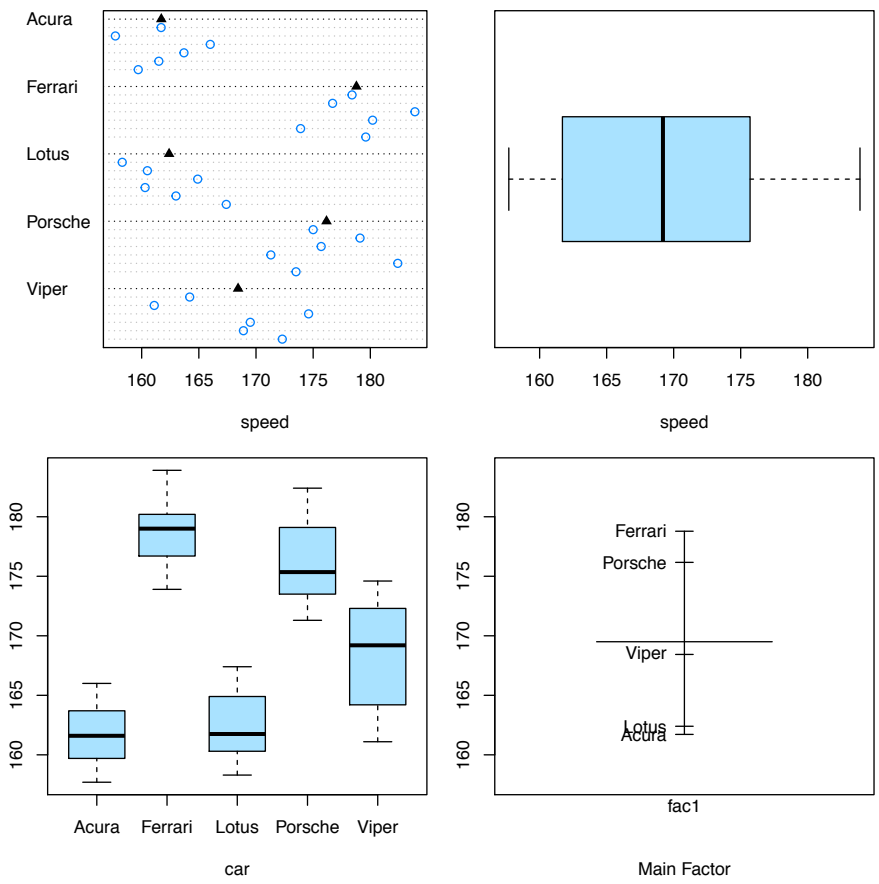
Solution for 7:

(a) A complete randomized design such as

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, 2, 3, 4, 5, \quad j = 1, \ldots, 6, \quad \varepsilon_{ij} \sim N(0, \sigma)$$

should be used to analyze the experiment. Before proceeding with formal inferential procedures, the data are examined with the function `oneway.plots()`.

```
> speed <- c(159.7, 161.5, 163.7, 166.0, 157.7, 161.7,
+            179.6, 173.9, 180.2, 183.9, 176.7, 178.4,
+            167.4, 163.0, 160.3, 164.9, 160.5, 158.3,
+            173.5, 182.4, 171.3, 175.7, 179.1, 175.0,
+            172.3, 168.9, 169.5, 174.6, 161.1, 164.2)
> car <- factor(c(rep("Acura", times = 6), rep("Ferrari", times = 6),
+                 rep("Lotus", times = 6), rep("Porsche", times = 6),
+                 rep("Viper", times = 6)))
> oneway.plots(Y = speed, fac1 = car)
```



Based on the output from `oneway.plots()`, one can see the fastest speeds have been recorded by Ferrari and Porsche, while the slowest speeds have been recorded by Acura and Lotus.

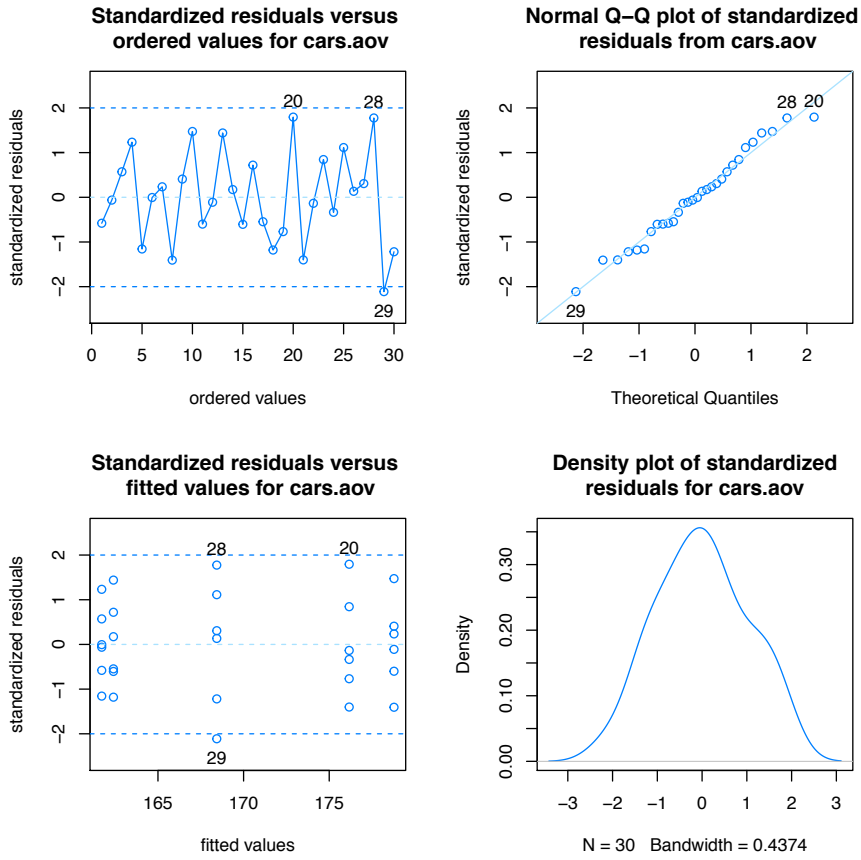
(b)

```
> DF <- data.frame(speed, car)
> rm(speed, car)  # clean up
> cars.aov <- aov(speed ~ car, data = DF)
> TR <- summary(cars.aov)
> TR

            Df Sum Sq Mean Sq F value  Pr(>F)
car          4   1456   364.1   25.14 1.9e-08 ***
Residuals   25    362    14.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The $\wp$-value $= 0$ from the ANOVA output suggests differences exist among the mean maximum speeds for the five vehicles.

(c)

```
> checking.plots(cars.aov)
```



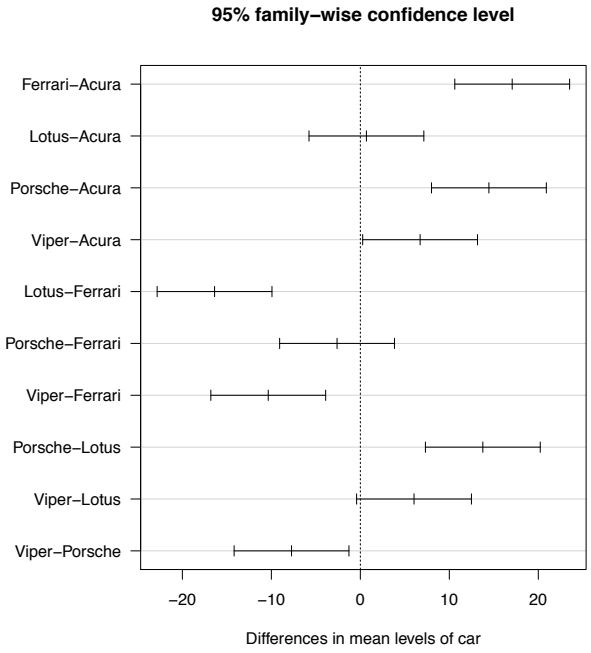
The assumptions are satisfied for the model in part (a). Specifically, no discernible pattern is seen in the top left graph that would threaten the assumption of independence. The top and bottom right graphs suggest the assumption of normality for the errors is a reasonable assumption. The bottom left graph makes the assumption of constant variance appear reasonable.

(d) The mean squared error value for the model in part (a) is 14.4809.

(e)

```
> car.mc <- TukeyHSD(cars.aov)
> car.mc

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = speed ~ car, data = DF)

$car
                      diff         lwr       upr     p adj
Ferrari-Acura    17.0666667  10.6142478 23.519086 0.0000004
Lotus-Acura       0.6833333  -5.7690855  7.135752 0.9978264
```

```
Porsche-Acura    14.4500000   7.9975811 20.902419 0.0000064
Viper-Acura       6.7166667   0.2642478 13.169086 0.0384079
Lotus-Ferrari   -16.3833333 -22.8357522 -9.930914 0.0000008
Porsche-Ferrari  -2.6166667  -9.0690855  3.835752 0.7563408
Viper-Ferrari   -10.3500000 -16.8024189 -3.897581 0.0006917
Porsche-Lotus    13.7666667   7.3142478 20.219086 0.0000137
Viper-Lotus       6.0333333  -0.4190855 12.485752 0.0749333
Viper-Porsche    -7.7333333 -14.1857522 -1.280914 0.0132379

> opar <- par(no.readonly = TRUE)
> par(mar = c(5.1, 10.1, 4.1, 2.1))
> plot(car.mc, las = 1)
> par(opar)
```



All of the car differences are significant with the exception of Lotus−Acura, Porsche−Ferrari, and Viper−Lotus.

Solution for 9:

(a) The design structure is a completely randomized design (CRD).

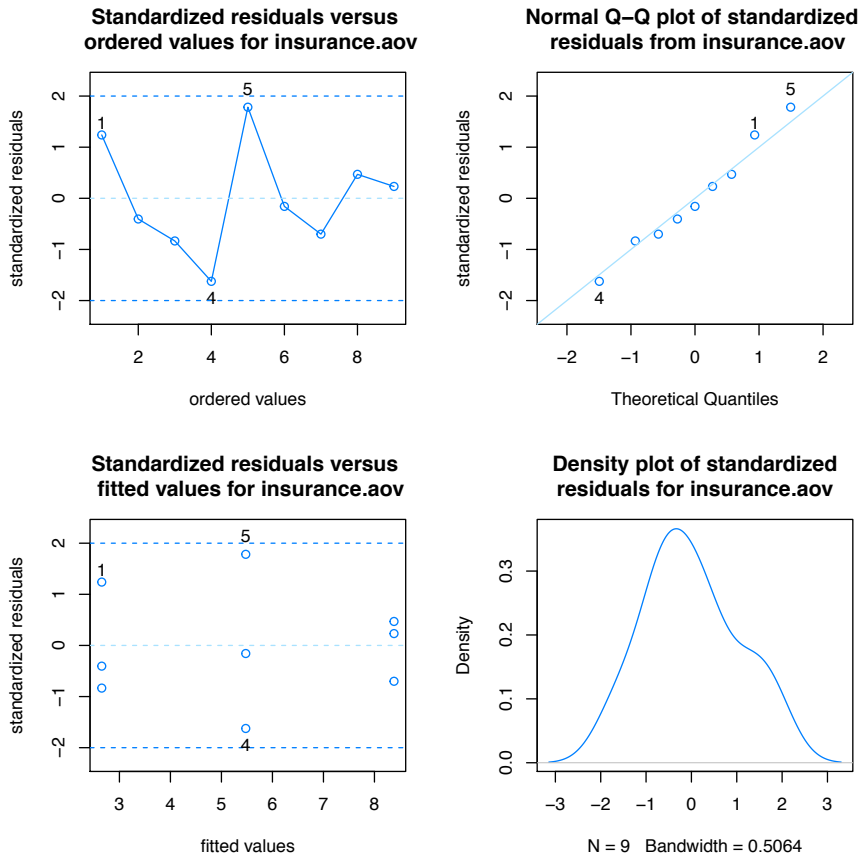
(b)

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, 2, 3 \quad j = 1, 2, 3 \quad \varepsilon_{ij} \sim N(0, \sigma)$$

(c) The three basic assumptions concerning the errors: independence, normal distribution, and constant variance are assessed with the `checking.plots()` function.

```
> time <- c(3.49, 2.38, 2.09, 4.38, 6.68, 5.37, 7.91, 8.70, 8.54)
> treatment <- factor(rep(c("telephone", "internet", "in person"),
+                         each = 3))
> DF <- data.frame(time, treatment)
```

```
> rm(time, treatment)  # clean up
> insurance.aov <- aov(time ~ treatment, data = DF)
> checking.plots(insurance.aov)
```



The model appears adequate.

(d)

```
> TR <- summary(insurance.aov)
> TR

            Df Sum Sq Mean Sq F value   Pr(>F)
treatment    2  49.25  24.626   36.01 0.000455 ***
Residuals    6   4.10   0.684
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the $\wp$-value $= 5e-04$, there is strong evidence to suggest differences exist among the methods used to issue insurance policies.

(e)

```
> MT <- model.tables(insurance.aov, type = "means")
> MT

Tables of means
Grand mean

5.504444

 treatment
treatment
in person  internet telephone
    8.383     5.477     2.653
```

The estimate for μ is 5.5044. The estimates for τ_i are 8.3833, 5.4767, and 2.6533.

(f)

```
> MSE <- summary(insurance.aov)[[1]][2, 3]
> MSE

[1] 0.6838333

> sde <- sqrt(MSE)
> sde

[1] 0.8269422
```

An estimate of the standard deviation of the errors is the square root of the mean squared error (0.8269).

(g)

```
> EFF <- proj(insurance.aov)[,]
> EFF

  (Intercept)   treatment  Residuals
1    5.504444 -2.85111111  0.8366667
2    5.504444 -2.85111111 -0.2733333
3    5.504444 -2.85111111 -0.5633333
4    5.504444 -0.02777778 -1.0966667
5    5.504444 -0.02777778  1.2033333
6    5.504444 -0.02777778 -0.1066667
7    5.504444  2.87888889 -0.4733333
8    5.504444  2.87888889  0.3166667
9    5.504444  2.87888889  0.1566667

> MeanMat <- matrix(EFF[, 1], byrow = TRUE, nrow = 3)
> MeanMat

         [,1]     [,2]     [,3]
[1,] 5.504444 5.504444 5.504444
[2,] 5.504444 5.504444 5.504444
[3,] 5.504444 5.504444 5.504444
```

```
> TreatMat <- matrix(EFF[, 2], byrow = TRUE, nrow = 3)
> TreatMat

            [,1]        [,2]        [,3]
[1,] -2.85111111 -2.85111111 -2.85111111
[2,] -0.02777778 -0.02777778 -0.02777778
[3,]  2.87888889  2.87888889  2.87888889

> ResidMat <- matrix(EFF[, 3], byrow = TRUE, nrow = 3)
> ResidMat

           [,1]       [,2]       [,3]
[1,]  0.8366667 -0.2733333 -0.5633333
[2,] -1.0966667  1.2033333 -0.1066667
[3,] -0.4733333  0.3166667  0.1566667

> Values <- MeanMat + TreatMat + ResidMat
> Values

     [,1] [,2] [,3]
[1,] 3.49 2.38 2.09
[2,] 4.38 6.68 5.37
[3,] 7.91 8.70 8.54
```

(h) The residuals sum to zero.

```
> sum(resid(insurance.aov))

[1] -1.387779e-16

> # Or
> sum(ResidMat)

[1] -1.387779e-16
```

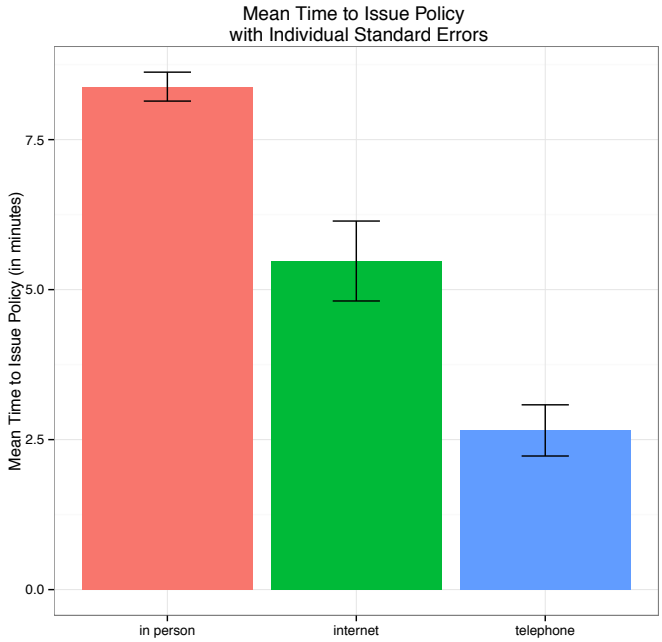
(i)

```
> TukeyHSD(insurance.aov)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = time ~ treatment, data = DF)

$treatment
                          diff       lwr        upr     p adj
internet-in person  -2.906667 -4.978352 -0.8349817 0.0119941
telephone-in person -5.730000 -7.801685 -3.6583151 0.0003595
telephone-internet  -2.823333 -4.895018 -0.7516484 0.0137052
```

The three methods of issuing insurance are all significantly different.

(j)

```
> library(plyr)
> mdf <- ddply(DF, "treatment", summarize, MeanTreat = mean(time),
+              SE = sd(time)/sqrt(3))
> mdf

  treatment MeanTreat        SE
1 in person  8.383333 0.2411316
2  internet  5.476667 0.6660914
3 telephone  2.653333 0.4266276

> ggplot(data = mdf, aes(x = treatment, y = MeanTreat, fill = treatment)) +
+   geom_bar(stat = "identity") +
+   geom_errorbar(aes(ymin = MeanTreat - SE, ymax = MeanTreat + SE),
+                 width = 0.25) +
+   guides(fill = FALSE) +
+   labs(x = "", y = "Mean Time to Issue Policy (in minutes)",
+   title = "Mean Time to Issue Policy \n with Individual Standard Errors") +
+   theme_bw()
```

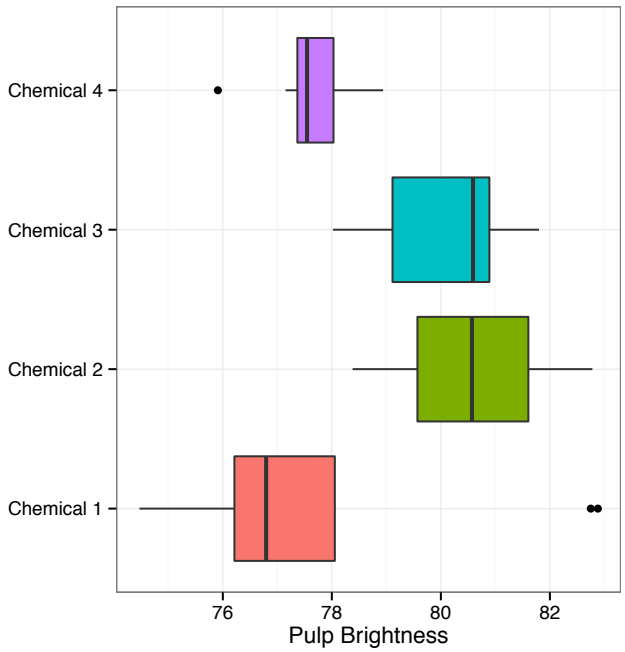


Solution for 11:

(a)

```
> pulpbright <- c(77.20, 74.47, 82.75, 76.21, 82.88,
+                 76.22, 78.06, 76.39, 76.16, 78.04,
+                 80.52, 79.31, 81.91, 80.35, 78.38,
+                 81.84, 82.78, 80.90, 79.18, 80.62,
+                 79.42, 78.02, 81.60, 80.80, 80.63,
+                 79.01, 80.55, 78.48, 81.80, 80.92,
+                 78.00, 78.36, 77.54, 77.36, 77.55,
```

```
+                 75.91, 78.04, 78.94, 77.15, 77.39)
> chemical <- factor(rep(c("Chemical 1", "Chemical 2", "Chemical 3",
+                          "Chemical 4"), each = 10))
> DF <- data.frame(pulpbright, chemical)
> rm(pulpbright, chemical)  # clean up
> ggplot(data = DF, aes(x = chemical, y = pulpbright, fill = chemical)) +
+   geom_boxplot() +
+   coord_flip() +
+   guides(fill = FALSE) +
+   labs(x= "", y = "Pulp Brightness") +
+   theme_bw()
```



Chemical 1 has less pulp brightness and a larger range than the other three chemicals. Chemical 2 and Chemical 3 appear to have higher pulp brightness values compared to Chemical 1 and Chemical 4.

(b) Since the chemicals used in the experiment are a sample from a much larger group of chemicals, the chemicals are treated as random effects and the model used is the random effects model.

```
> chem.aov <- aov(pulpbright ~ chemical, data = DF)
> TR <- summary(chem.aov)  # Test results
> TR

            Df Sum Sq Mean Sq F value   Pr(>F)
chemical     3  69.91  23.304   7.654 0.000441 ***
Residuals   36 109.61   3.045
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the $\wp$-value $= 4e{-}04$ from the ANOVA table, evidence suggests the variance of the random effects is greater than zero. In essence, there are significant differences between chemicals.

(c)

```
> MST <- TR[[1]][1, 3]  # MS chemicals
> MSE <- TR[[1]][2, 3]  # MS error
> sig2tau <- (MST - MSE)/10
> sig2tau

[1] 2.02595
```

$$\begin{aligned}\hat{\sigma}^2_\tau &= \frac{SS_{\texttt{chemical}} - SS_{\text{Error}}}{n} \\ &= \frac{23.3042 - 3.0447}{10} = 2.026\end{aligned}$$

The estimated component of variance for the chemicals ($\hat{\sigma}^2_\tau$) is 2.026.

(d)

```
> totalVar <- sig2tau + MSE
> totalVar

[1] 5.070669
```

The total variability of the data is estimated as $\hat{\sigma}^2_\tau + MSE$ which equals 5.0707.

(e) Given that

$$\frac{MST/(n\sigma^2_\tau + \sigma^2)}{MSE/\sigma^2} \sim F_{a-1,\, a\cdot N-a},$$

$$\begin{aligned}
\mathbb{P}\left(f_{\alpha/2;\, a-1, N-a} \leq \frac{\sigma^2 MST}{(n\sigma^2_\tau + \sigma^2)MSE} \leq f_{1-\alpha/2;\, a-1, N-a}\right) &= 1-\alpha \\
\mathbb{P}\left(f_{\alpha/2;\, a-1, N-a} \cdot \frac{MSE}{MST} \leq \frac{\sigma^2}{n\sigma^2_\tau + \sigma^2} \leq f_{1-\alpha/2;\, a-1, N-a} \cdot \frac{MSE}{MST}\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{MST}{f_{1-\alpha/2;\, a-1, N-a} MSE} \leq \frac{n\sigma^2_\tau + \sigma^2}{\sigma^2} \leq \frac{MST}{f_{\alpha/2;\, a-1, N-a} MSE}\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{MST}{f_{1-\alpha/2;\, a-1, N-a} MSE} - 1 \leq \frac{n\sigma^2_\tau}{\sigma^2} \leq \frac{MST}{f_{\alpha/2;\, a-1, N-a} MSE} - 1\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{1}{n} \cdot \left[\frac{MST}{f_{1-\alpha/2;\, a-1, N-a} MSE} - 1\right] \leq \frac{\sigma^2_\tau}{\sigma^2} \leq \frac{1}{n} \cdot \left[\frac{MST}{f_{\alpha/2;\, a-1, N-a} MSE} - 1\right]\right) &= 1-\alpha
\end{aligned}$$

Let $L = \frac{1}{n} \cdot \left[\frac{MST}{f_{1-\alpha/2;\, a-1, N-a} MSE} - 1\right]$ and $U = \frac{1}{n} \cdot \left[\frac{MST}{f_{\alpha/2;\, a-1, N-a} MSE} - 1\right]$.

Then

$$\begin{aligned}
\mathbb{P}\left(L \le \frac{\sigma_\tau^2}{\sigma^2} \le U\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{1}{U} \le \frac{\sigma^2}{\sigma_\tau^2} \le \frac{1}{L}\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{1}{U}+1 \le \frac{\sigma^2}{\sigma_\tau^2}+1 \le \frac{1}{L}+1\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{1+U}{U} \le \frac{\sigma^2+\sigma_\tau^2}{\sigma_\tau^2} \le \frac{1+L}{L}\right) &= 1-\alpha \\
\mathbb{P}\left(\frac{L}{1+L} \le \frac{\sigma_\tau^2}{\sigma^2+\sigma_\tau^2} \le \frac{U}{1+U}\right) &= 1-\alpha \\
\Longrightarrow CI_{1-\alpha}\left(\frac{\sigma_\tau^2}{\sigma^2+\sigma_\tau^2}\right) &= \left[\frac{L}{1+L}, \frac{U}{1+U}\right]
\end{aligned}$$

where $L = \frac{1}{n} \cdot \left[\frac{MST}{f_{1-\alpha/2;\, a-1, N-a} MSE} - 1\right]$ and $U = \frac{1}{n} \cdot \left[\frac{MST}{f_{\alpha/2;\, a-1, N-a} MSE} - 1\right]$.

```
> dfn <- summary(chem.aov)[[1]][1, 1] # a - 1 = 3
> dfd <- summary(chem.aov)[[1]][2, 1] # N - a = 36
> f.lo <- qf(0.025, dfn, dfd)
> f.up <- qf(0.975, dfn, dfd)
> n <- 10
> L <- 1/n*(MST/(MSE*f.up) - 1)
> U <- 1/n*(MST/(MSE*f.lo) - 1)
> CI <- c(L/(L + 1), U/(U + 1))
> CI  # CI for ratio

[1] 0.1058598 0.9141983
```

The $CI_{0.95}\left(\frac{\sigma_\tau^2}{\sigma_\tau^2+\sigma^2}\right)$ is [0.1059, 0.9142].

Solution for 13:

(a) The design structure used by the EPA is a completely randomized design (CRD).

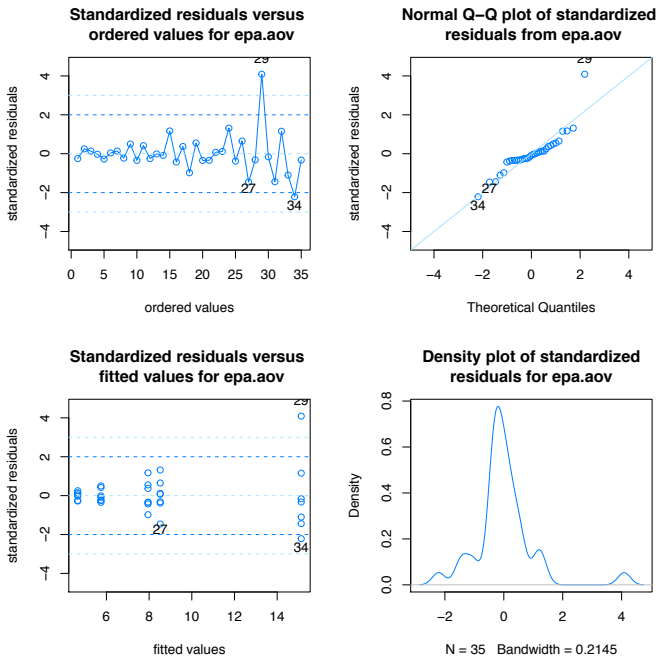
(b) The model to use is

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, 2, 3, 4, 5 \quad j = 1, \ldots, 7, \quad \varepsilon_{ij} \sim N(0, \sigma).$$

(c) The three basic assumptions concerning the errors: independence, normal distribution, and constant variance are assessed with the `checking.plots()` function.

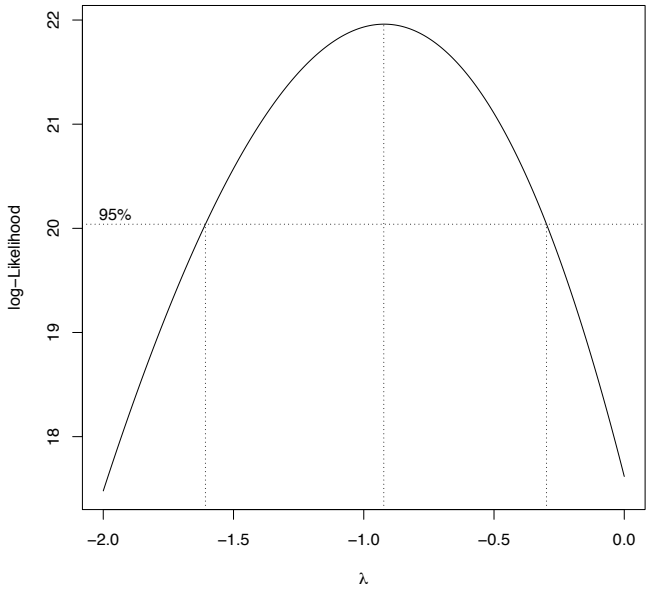
```
> fuel <- c(4.35, 4.96, 4.82, 4.62, 4.32, 4.70, 4.82,
+           5.47, 6.35, 5.33, 6.25, 5.44, 5.73, 5.64,
+           9.37, 7.43, 8.40, 6.76, 8.62, 7.53, 7.54,
+           8.61, 8.66, 10.12, 8.06, 9.31, 6.75, 8.14,
+           20.09, 14.93, 13.38, 16.53, 13.79, 12.44, 14.73)
> vehicle <- factor(rep(c("Compact", "Station Wagon", "Minivan",
+                         "Van", "Pickup Truck"), each = 7))
> DF <- data.frame(fuel, vehicle)
```

```
> rm(fuel, vehicle)  # clean up
> epa.aov <- aov(fuel ~ vehicle, data = DF)
> checking.plots(epa.aov)
```



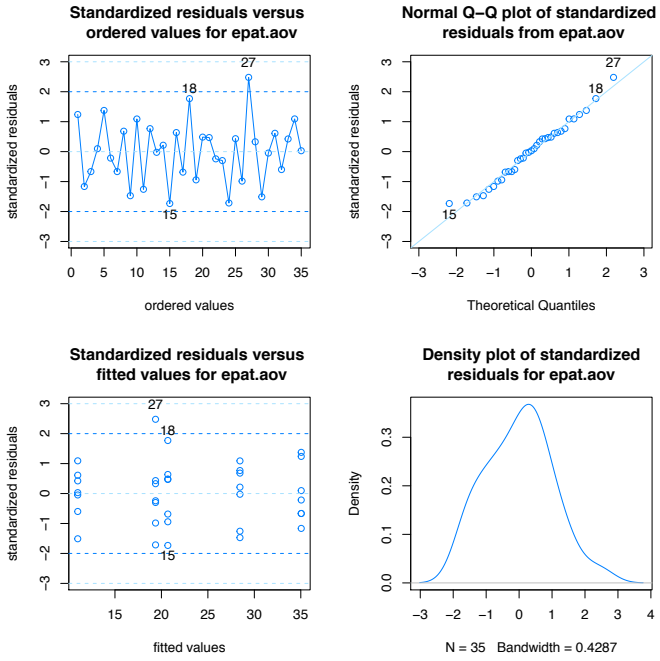
The assumptions of normality and constant variance are not satisfied. Consequently, a transformation for the response variable is investigated using the `boxcox()` function from the `MASS` package.

```
> library(MASS)
> boxcox(epa.aov, lambda = seq(-2, 0, length = 300))
```



Based on the graph, a reciprocal transformation is suggested for the variable `fuel`. A new model is fit with the response (162.78/fuel), which corresponds to miles/gallon.

```
> epat.aov <- aov(162.78/fuel ~ vehicle, data = DF)
> checking.plots(epat.aov)
```



The errors for the new model stored in `epat.aov` appear to be independent, normal, and have constant variance.

(d)

```
> TR <- summary(epat.aov)
> TR

            Df Sum Sq Mean Sq F value Pr(>F)
vehicle      4 2362.5   590.6   138.1 <2e-16 ***
Residuals   30  128.3     4.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The $\wp$-value $= 0$ suggests there are significant differences in average miles per gallon for the different vehicle types.

(e) The mean squared error for the model `epat.aov` is 4.2779.

(f) There are significant differences in average miles per gallon for the different vehicle types.

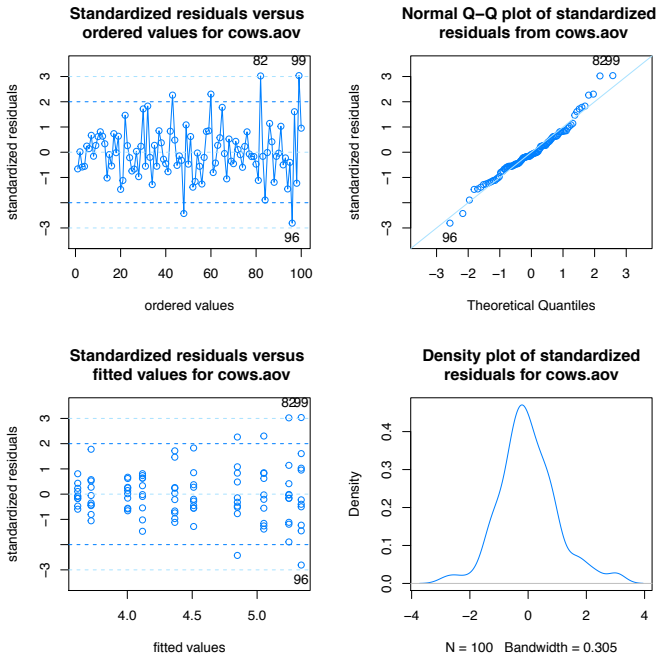
Solution for 15:

(a)

```
> cows.aov <- aov(butterfat ~ age*breed, data = COWS)
> summary(cows.aov)

            Df Sum Sq Mean Sq F value Pr(>F)
age          1   0.21   0.207   1.172  0.282
breed        4  34.32   8.580  48.595 <2e-16 ***
age:breed    4   0.27   0.067   0.381  0.821
Residuals   90  15.89   0.177
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b)

```
> checking.plots(cows.aov)
```



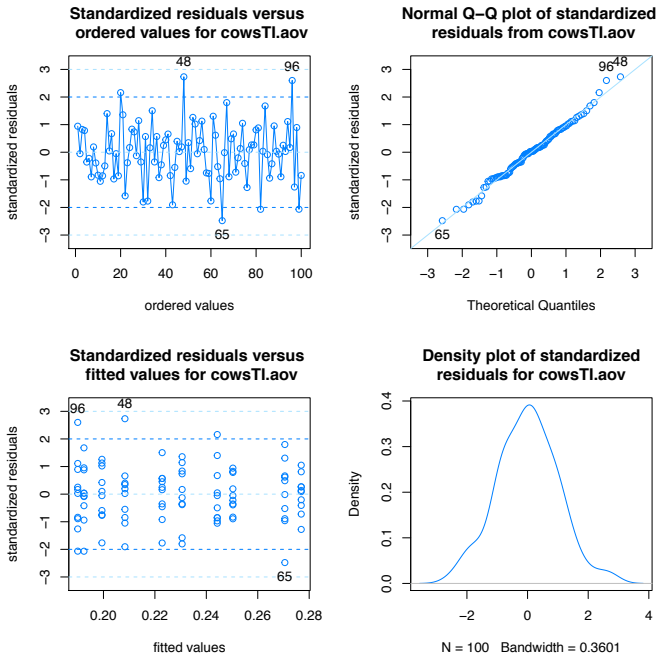
There appears to be an increasing variance with larger butterfat values. The model does not satisfy the homogeneity of variance assumption.

(c) The `boxCox()` function from the `car` package is used on the object `cows.aov`.

```
> boxCox(cows.aov, lambda = seq(-3, 0, length = 300))

Error in eval(expr, envir, enclos):  could not find function "boxCox"
```

The 95% confidence interval for λ extends from roughly -2.4 to -0.5. Since a transformation using $\lambda = -1$ is inside the 95% confidence interval as well as being a monotonic transformation, the decision to use $\lambda = -1$ is made.

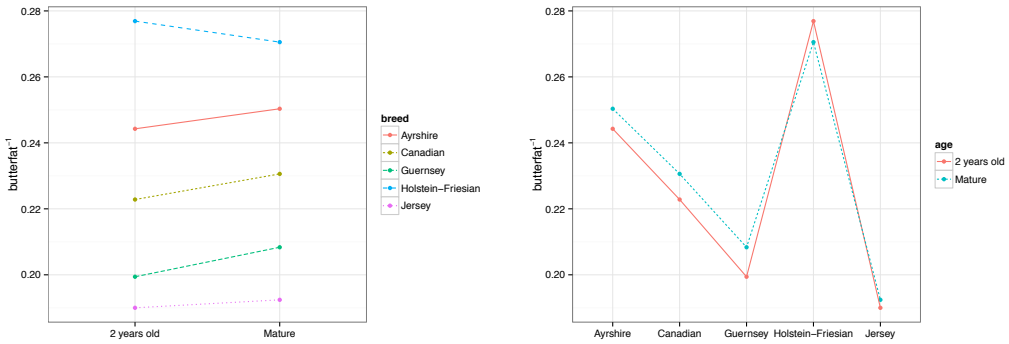
```
> cowsTI.aov <- aov(I(butterfat^-1) ~ age*breed, data = COWS)
> TR <- summary(cowsTI.aov)
> TR

            Df  Sum Sq  Mean Sq F value Pr(>F)
age          1 0.00035 0.000355   0.977  0.326
breed        4 0.08797 0.021993  60.599 <2e-16 ***
age:breed    4 0.00076 0.000191   0.526  0.717
Residuals   90 0.03266 0.000363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> checking.plots(cowsTI.aov)
```



After the butterfat values have been transformed, increasing variance no longer appears problematic. Although there are five observations whose standardized residuals are greater in absolute value than two, this is to be expected with one hundred observations.

(d)

```
> ggplot(data = COWS, aes(x = age, y = butterfat^-1, colour = breed,
+                         group = breed, linetype = breed)) +
+   stat_summary(fun.y = mean, geom = "point") +
+   stat_summary(fun.y = mean, geom = "line") +
+   theme_bw() +
+   labs(x = "", y = expression(butterfat^{-1}))
> ggplot(data = COWS, aes(x = breed, y = butterfat^-1, colour = age,
+                         group = age, linetype = age)) +
+   stat_summary(fun.y = mean, geom = "point") +
+   stat_summary(fun.y = mean, geom = "line") +
+   theme_bw() +
+   labs(x = "", y = expression(butterfat^{-1}))
```



The interaction plots show relatively parallel lines suggesting `age` and `breed` do not interact which is corroborated with the interaction $\wp$-value = 0.7169 computed in (d).

(e)

```
> model.tables(cowsTI.aov, type = "means")

Tables of means
Grand mean

0.2285625

 age
age
2 years old      Mature
    0.22668     0.23045

 breed
breed
         Ayrshire          Canadian          Guernsey Holstein-Friesian
          0.24730           0.22671           0.20388           0.27372
           Jersey
          0.19121

 age:breed
             breed
age           Ayrshire Canadian Guernsey Holstein-Friesian Jersey
  2 years old 0.24426  0.22282  0.19941  0.27691           0.18999
  Mature      0.25034  0.23059  0.20835  0.27054           0.19242

> model.tables(cowsTI.aov, type = "effects")

Tables of effects

 age
age
2 years old      Mature
 -0.0018833   0.0018833

 breed
breed
         Ayrshire          Canadian          Guernsey Holstein-Friesian
          0.01874          -0.00186          -0.02468           0.04516
           Jersey
         -0.03736

 age:breed
             breed
age           Ayrshire  Canadian  Guernsey  Holstein-Friesian Jersey
  2 years old -0.001156 -0.002000 -0.002586  0.005069          0.000673
  Mature       0.001156  0.002000  0.002586 -0.005069         -0.000673
```

Table 11.1: Group means and parameter estimates for the object `cowsTI.aov`

		Breed					$\overline{Y}_{i\bullet\bullet}$	$\hat{\alpha}_i = \overline{Y}_{i\bullet\bullet} - \overline{Y}_{\bullet\bullet\bullet}$
		Ayrshire	Canadian	Guernsey	H-F	Jersey		
Age	2 year old	$\overline{Y}_{11\bullet} = 0.2443$ $\widehat{\alpha\beta}_{11} = -0.0011$	$\overline{Y}_{12\bullet} = 0.2228$ $\widehat{\alpha\beta}_{12} = -0.0020$	$\overline{Y}_{13\bullet} = 0.1994$ $\widehat{\alpha\beta}_{13} = -0.0025$	$\overline{Y}_{14\bullet} = 0.2769$ $\widehat{\alpha\beta}_{14} = 0.0051$	$\overline{Y}_{15\bullet} = 0.1900$ $\widehat{\alpha\beta}_{15} = 0.0007$	$\overline{Y}_{1\bullet\bullet} =$ 0.2267	$\hat{\alpha}_1 =$ -0.001883
	mature	$\overline{Y}_{21\bullet} = 0.2503$ $\widehat{\alpha\beta}_{21} = 0.0011$	$\overline{Y}_{22\bullet} = 0.2306$ $\widehat{\alpha\beta}_{22} = 0.0020$	$\overline{Y}_{23\bullet} = 0.2084$ $\widehat{\alpha\beta}_{23} = 0.0026$	$\overline{Y}_{24\bullet} = 0.2705$ $\widehat{\alpha\beta}_{24} = 0.0051$	$\overline{Y}_{25\bullet} = 0.1924$ $\widehat{\alpha\beta}_{25} = -.0007$	$\overline{Y}_{2\bullet\bullet} =$ 0.2304	$\hat{\alpha}_2 =$ 0.001883
	$\overline{Y}_{\bullet j\bullet}$	$\overline{Y}_{\bullet 1\bullet} = 0.2473$	$\overline{Y}_{\bullet 2\bullet} = 0.2267$	$\overline{Y}_{\bullet 3\bullet} = 0.2039$	$\overline{Y}_{\bullet 4\bullet} = 0.2737$	$\overline{Y}_{\bullet 5\bullet} = 0.1912$		
	$\hat{\beta}_j =$ $\overline{Y}_{\bullet j\bullet} - \overline{Y}_{\bullet\bullet\bullet}$	$\hat{\beta}_1 = 0.1874$	$\hat{\beta}_2 = -0.00186$	$\hat{\beta}_3 = -0.02468$	$\hat{\beta}_4 = 0.04516$	$\hat{\beta}_5 = -0.03736$		$\overline{Y}_{\bullet\bullet\bullet} =$ 0.2286

Using the results from the function `model.tables()`, Table 11.1 is created.

(f) To determine which breeds have higher butterfat production, Tukey's HSD pairwise confidence intervals are created using the function `TukeyHSD()`.

```
> CIt <- TukeyHSD(cowsTI.aov, which = "breed")
> CIt

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = I(butterfat^-1) ~ age * breed, data = COWS)

$breed
                                  diff          lwr          upr     p adj
Canadian-Ayrshire          -0.02059248 -0.037363459 -0.003821510 0.0082292
Guernsey-Ayrshire          -0.04341678 -0.060187757 -0.026645808 0.0000000
Holstein-Friesian-Ayrshire  0.02642525  0.009654271  0.043196220 0.0002965
Jersey-Ayrshire            -0.05609240 -0.072863376 -0.039321427 0.0000000
Guernsey-Canadian          -0.02282430 -0.039595272 -0.006053323 0.0024803
Holstein-Friesian-Canadian  0.04701773  0.030246756  0.063788705 0.0000000
Jersey-Canadian            -0.03549992 -0.052270892 -0.018728942 0.0000006
Holstein-Friesian-Guernsey  0.06984203  0.053071053  0.086613003 0.0000000
Jersey-Guernsey            -0.01267562 -0.029446594  0.004095355 0.2274154
Jersey-Holstein-Friesian   -0.08251765 -0.099288622 -0.065746673 0.0000000
```

At the $\alpha_e = 0.05$ all breeds are significantly different from one another with the exception of Jersey and Guernsey.

Solution for 17:

(a) The contests of `SUNFLOWER` are explored using the functions `xtabs()` and `ftable()`.

(i)

```
> xtabs(yield ~ stage + defoli, data = SUNFLOWER)

        defoli
stage    control treat1 treat2 treat3
  stage1    7865   6673   3576   1999
  stage2    7213   8945   7089    704
```

```
  stage3    6561   8971   6527   1943
  stage4    5619   9361   6082   5673
  stage5    5380   2418   5664   5227
```

(ii)

```
> ftable(xtabs(numseed ~ defoli + location + stage, data = SUNFLOWER))

                stage stage1 stage2 stage3 stage4 stage5
defoli  location
control A                877    981    738   1073      0
        B               1054    685   1154    954    949
        C                948    695      0    646    612
        D                  0    758   1049    799   1263
treat1  A                706    954   1052    729      0
        B               1206   1066   1337   1378    901
        C                745    695      0   1001    691
        D                  0    839   1007    845   1552
treat2  A                519    476    629    545      0
        B                591    639    629   1159    117
        C                837    735      0    747    730
        D                  0    788    853    761   1497
treat3  A                277    204    298   1258      0
        B                317    253    209   1209   1568
        C                563    174      0    408    631
        D                  0    142    409    267   1742
```

(b)

```
> xtabs(~ stage + defoli, data = SUNFLOWER)

        defoli
stage    control treat1 treat2 treat3
  stage1       3      3      3      3
  stage2       4      4      4      4
  stage3       3      3      3      3
  stage4       4      4      4      4
  stage5       4      4      4      4
```

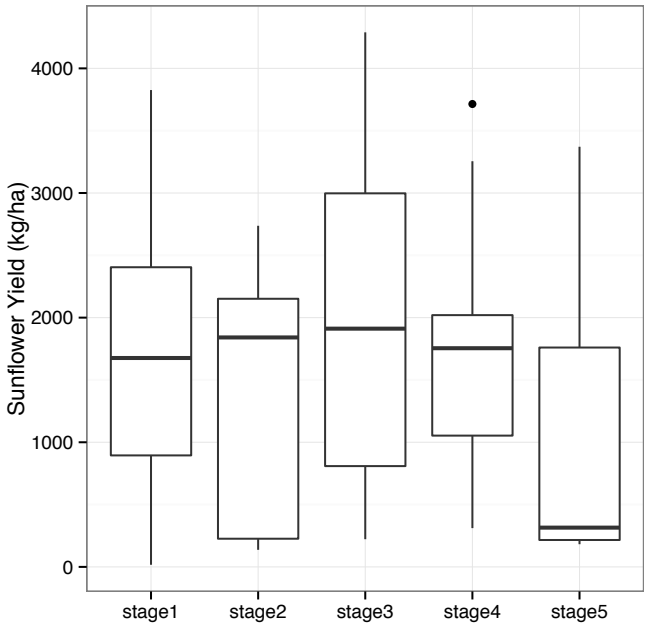
There are three observations for every combination of `defoli` with stages 1 and 3. All other combinations of `defoli` and `stage` have four observations.

(c) The design is complete.

(d) The design is unbalanced as the number of observations for each treatment combination are not the same.

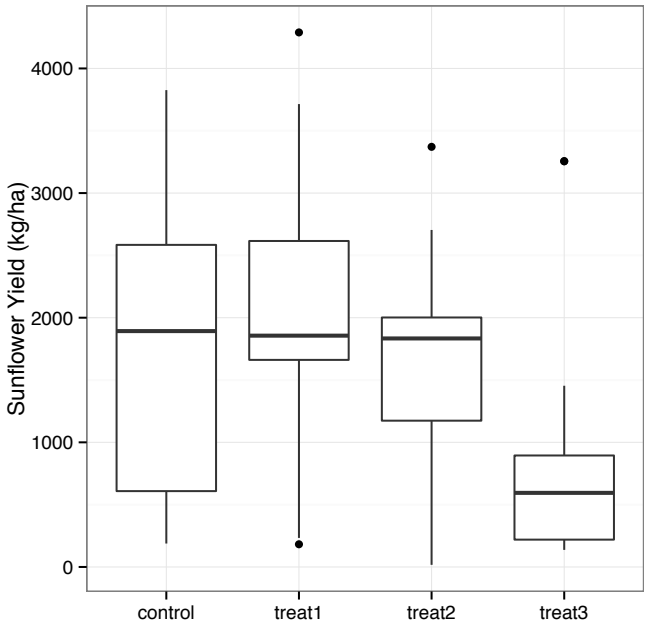
(e)

```
> ggplot(data = SUNFLOWER, aes(x = stage, y = yield)) +
+   geom_boxplot() +
+   theme_bw() +
+   labs(x = "", y = "Sunflower Yield (kg/ha)")
```



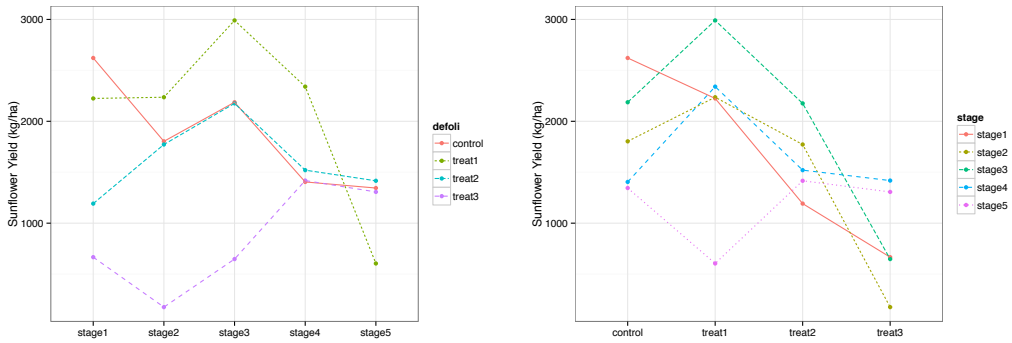
(f)

```
> ggplot(data = SUNFLOWER, aes(x = defoli, y = yield)) +
+   geom_boxplot() +
+   theme_bw() +
+   labs(x = "", y = "Sunflower Yield (kg/ha)")
```



(g)

```
> ggplot(data = SUNFLOWER, aes(x = stage, y = yield,
+                              colour = defoli, group = defoli,
+                              linetype = defoli)) +
+   stat_summary(fun.y = mean, geom = "point") +
+   stat_summary(fun.y = mean, geom = "line") +
+   theme_bw() +
+   labs(x = "", y = "Sunflower Yield (kg/ha)")
> ggplot(data = SUNFLOWER, aes(x = defoli, y = yield,
+                              colour = stage, group = stage,
+                              linetype = stage)) +
+   stat_summary(fun.y = mean, geom = "point") +
+   stat_summary(fun.y = mean, geom = "line") +
+   theme_bw() +
+   labs(x = "", y = "Sunflower Yield (kg/ha)")
```



The lack of parallel lines in the plots suggest interaction between the factors `stage` and `defoli`.

Model (A)

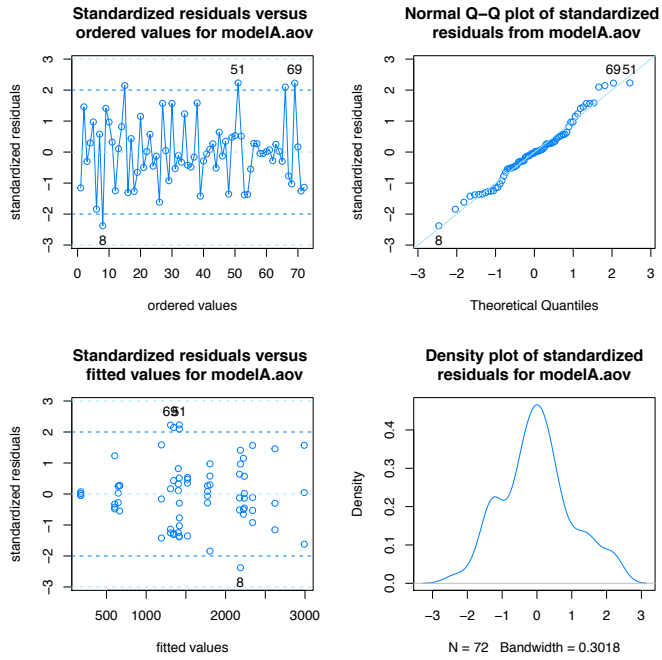
(i)

```
> modelA.aov <- aov(yield ~ stage*defoli, data = SUNFLOWER)
> TR <- summary(modelA.aov)
> TR

             Df   Sum Sq Mean Sq F value  Pr(>F)
stage         4  5186036 1296509   1.267 0.29495
defoli        3 13720078 4573359   4.468 0.00726 **
stage:defoli 12 16236084 1353007   1.322 0.23500
Residuals    52 53224683 1023552
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The interaction between `stage` and `defoli` is not significant as the $\wp$-value $= 0.235$.

(ii) The three basic assumptions concerning the errors: independence, normal distribution, and constant variance are assessed with the `checking.plots()` function.

```
> checking.plots(modelA.aov)
```



The assumption of constant variance is slightly questionable. The equality of variance assumption is tested with the function `leveneTest()` from the `car` package.

```
> library(car)
> TR1 <- leveneTest(yield ~ stage, data = SUNFLOWER)
> TR2 <- leveneTest(yield ~ defoli, data = SUNFLOWER)
> TR1

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  0.4987 0.7367
      67

> TR2

Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  3  0.8974 0.4471
      68
```

Levene's test does not reject the null hypothesis of constant variance for either `stage` ($\wp$-value = 0.7367) or `defoliation` ($\wp$-value = 0.4471). The assumptions for Model (A) appear to be satisfied.

Model (B)

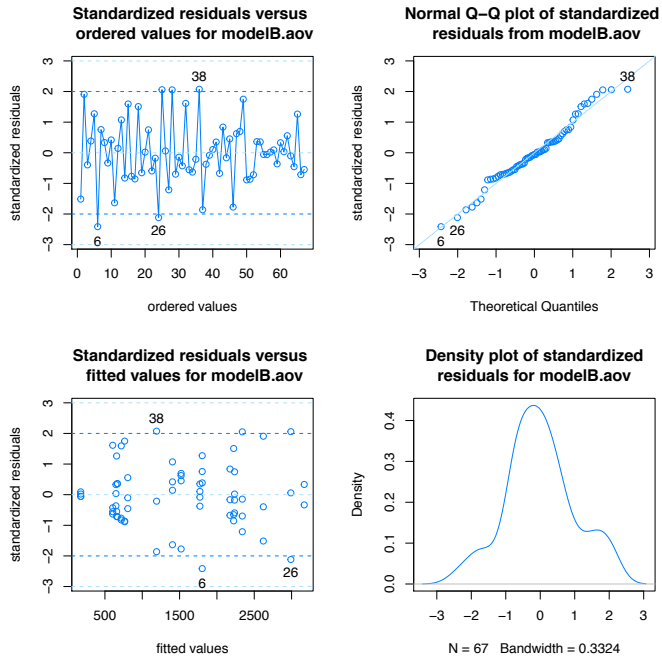
```
> TooBig <- which(abs(rstandard(modelA.aov)) > 2)
> TooBig

 8 15 51 66 69
```

```
 8 15 51 66 69

> modelB.aov <- aov(yield ~ stage*defoli, data = SUNFLOWER[-TooBig, ])
> TR <- summary(modelB.aov)
> TR

             Df   Sum Sq Mean Sq F value   Pr(>F)
stage         4 13997052 3499263   5.860 0.000655 ***
defoli        3 21878479 7292826  12.213 4.95e-06 ***
stage:defoli 12 10573606  881134   1.476 0.167430
Residuals    47 28064990  597127
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) The interaction term `stage:defoli` is not significant ($\wp$-value = 0.1674).

(ii) The three basic assumptions concerning the errors: independence, normal distribution, and constant variance are assessed with the `checking.plots()` function.

```
> checking.plots(modelB.aov)
```



The model appears adequate. The equality of variance assumption is tested with the function `leveneTest()` from the `car` package.

```
> library(car)
> TR1 <- leveneTest(yield ~ stage, data = SUNFLOWER[-TooBig, ])
> TR2 <- leveneTest(yield ~ defoli, data = SUNFLOWER[-TooBig, ])
> TR1

Levene's Test for Homogeneity of Variance (center = median)
```

```
      Df F value Pr(>F)
group  4   1.214  0.314
      62

> TR2

Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  3  2.5623 0.06266 .
      63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

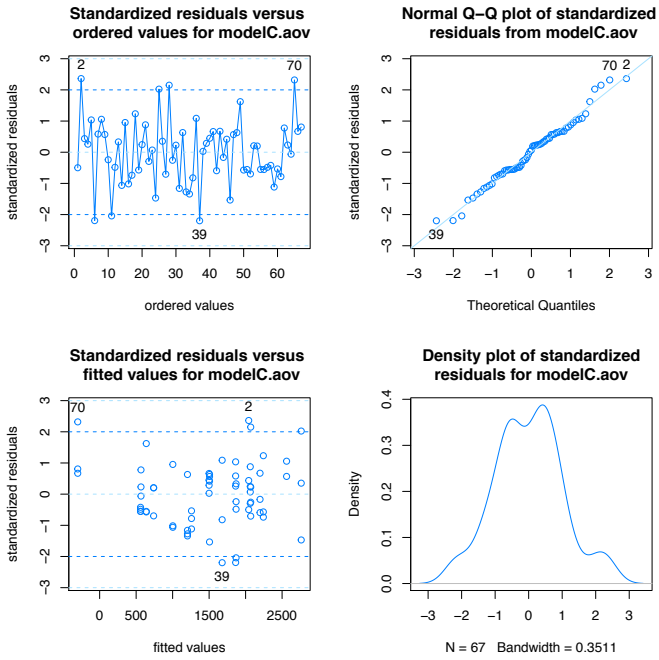
Levene's test does not reject the null hypothesis of constant variance at the $\alpha = 0.05$ level for either `stage` ($\wp$-value $= 0.314$) or `defoliation` ($\wp$-value $= 0.0627$). The assumptions for Model (B) appear to be satisfied.

Model (C)

```
> modelC.aov <- aov(yield ~ stage + defoli, data = SUNFLOWER[-TooBig, ])
> TR <- summary(modelC.aov)
> TR

            Df   Sum Sq Mean Sq F value   Pr(>F)
stage        4 13997052 3499263   5.343 0.000968 ***
defoli       3 21878479 7292826  11.136 6.85e-06 ***
Residuals   59 38638596  654891
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) The three basic assumptions concerning the errors: independence, normal distribution, and constant variance are assessed with the `checking.plots()` function.

```
> checking.plots(modelC.aov)
```



The model appears adequate.

(ii) The model's effects are estimated with the function `model.tables()`. The values for the decomposition of the $Y_{i,j}$s are obtained using the function `proj()` and stored in the object `EFF`.

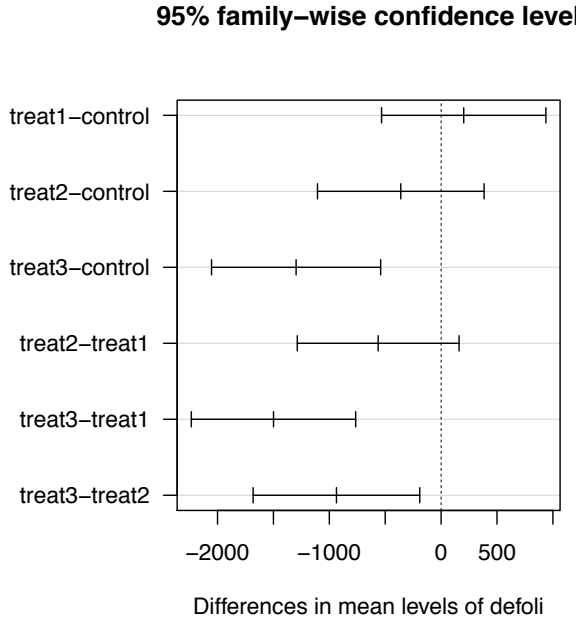
```
> model.tables(modelC.aov, type = "effects")

Tables of effects

 stage
    stage1 stage2 stage3 stage4 stage5
     181.1  1.982  666.9  70.38 -815.1
rep   12.0 16.000   11.0  15.00   13.0

 defoli
    control treat1 treat2 treat3
        347  549.2 -13.84 -950.1
rep      16   18.0  17.00   16.0

> EFF <- proj(modelC.aov)[,]
> head(EFF)

  (Intercept)      stage   defoli  Residuals
1    1494.955 181.128109 366.5917  -375.6750
2    1494.955 181.128109 366.5917  1783.3250
3    1494.955 181.128109 366.5917   329.3250
4    1494.955   1.982276 366.5917   197.4708
5    1494.955   1.982276 366.5917   791.4708
6    1494.955   1.982276 366.5917 -1675.5292
```

```
> VAL <- apply(EFF, 1, sum)
> VAL[1:10]

   1    2    3    4    5    6    7    9   10   11
1667 3826 2372 2061 2655  188 2309 3352 2987 1685

> SUNFLOWER[-TooBig, ]$yield[1:10]

 [1] 1667 3826 2372 2061 2655  188 2309 3352 2987 1685
```

Note that the values stored in the object `VAL` are the same as the original yield values used to construct Model (C).

(iii)

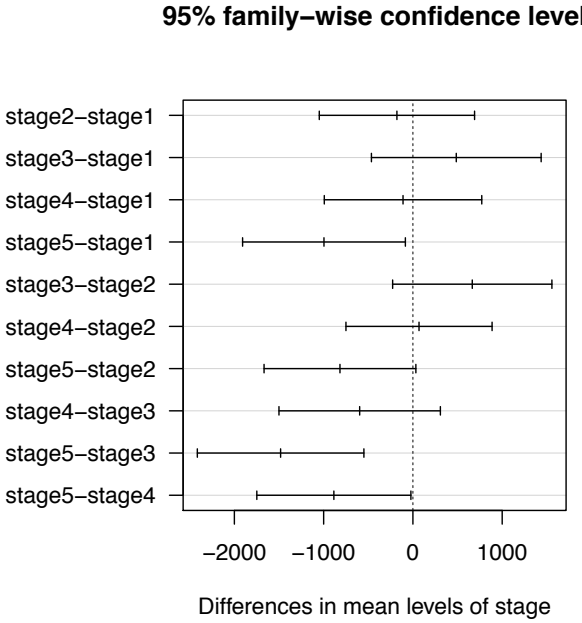
```
> CI <- TukeyHSD(modelC.aov, which = "defoli")
> CI

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ stage + defoli, data = SUNFLOWER[-TooBig, ])

$defoli
                     diff        lwr       upr     p adj
treat1-control   202.2444  -532.8707  937.3595 0.8857552
treat2-control  -360.8110 -1106.0312  384.4093 0.5791269
treat3-control -1297.0928 -2053.5200 -540.6656 0.0001663
treat2-treat1   -563.0554 -1286.6335  160.5228 0.1792500
treat3-treat1  -1499.3372 -2234.4523 -764.2221 0.0000075
treat3-treat2   -936.2818 -1681.5021 -191.0616 0.0081636

> opar <- par(no.readonly = TRUE)
> par(mar = c(5, 8, 5, 3))
> plot(CI, las = 1)
> par(opar)
```



The mean of `treat3` is significantly less than the mean of `treat2`, the mean of `treat1`, and the mean of `control`.

(iv)

```
> CI <- TukeyHSD(modelC.aov, which = "stage")
> CI

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = yield ~ stage + defoli, data = SUNFLOWER[-TooBig, ])

$stage
                    diff        lwr        upr     p adj
stage2-stage1  -179.14583 -1048.7577  690.46602 0.9775671
stage3-stage1   485.73485  -464.8130 1436.28274 0.6060833
stage4-stage1  -110.75000  -992.6974  771.19739 0.9965671
stage5-stage1  -996.23718 -1907.8381  -84.63626 0.0254664
stage3-stage2   664.88068  -227.0325 1556.79390 0.2348146
stage4-stage2    68.39583  -750.0167  886.80838 0.9993030
stage5-stage2  -817.09135 -1667.3761   33.19339 0.0653308
stage4-stage3  -596.48485 -1500.4293  307.45962 0.3517722
stage5-stage3 -1481.97203 -2414.8711 -549.07297 0.0003382
stage5-stage4  -885.48718 -1748.3838  -22.59057 0.0415604

> opar <- par(no.readonly = TRUE)
> par(mar = c(5, 8, 5, 3))
> plot(CI, las = 1)
> par(opar)
```



The mean of `stage5` is significantly less than the mean of `stage3` and the mean of `stage1`.

(h) The the function `contr.helmert(4)` will create three orthogonal contrasts where the values in the third contrast (-1, -1, -1, 3) are such that this contrast can be used to investigate whether there are statistical differences between 100% defoliation and the remaining three levels of defoliation.

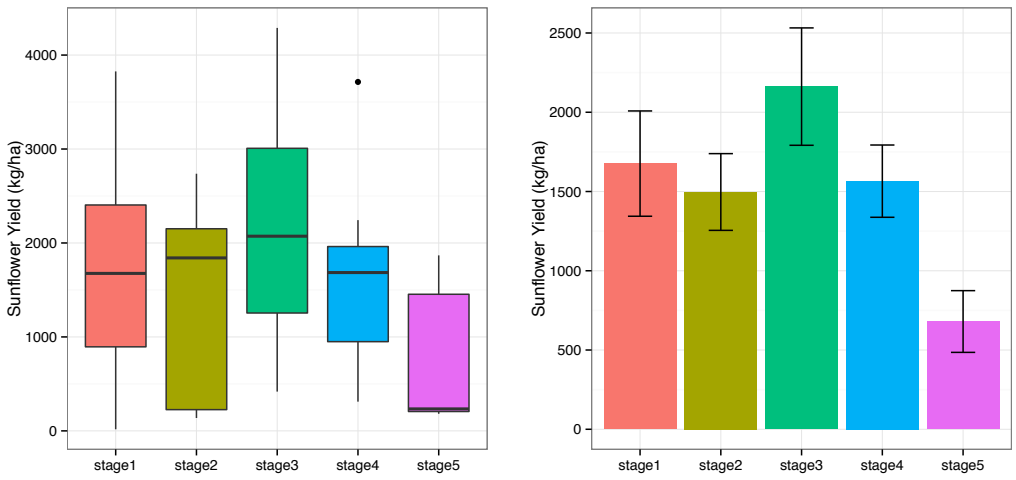
```
> ND <- SUNFLOWER[-TooBig, ]  # shorter data name
> contrasts(ND$defoli) <- contr.helmert(levels(ND$defoli))
> CO <- contrasts(ND$defoli)
> CO

        [,1] [,2] [,3]
control   -1   -1   -1
treat1     1   -1   -1
treat2     0    2   -1
treat3     0    0    3

> partH.aov <- aov(yield ~ stage + C(defoli, CO, 1) + C(defoli, CO, 2) +
+                    C(defoli, CO, 3), data = ND)
> TR <- summary(partH.aov)
> TR

                Df   Sum Sq  Mean Sq F value   Pr(>F)
stage            4 13997052  3499263   5.343 0.000968 ***
C(defoli, CO, 1)  1   555284   555284   0.848 0.360896
C(defoli, CO, 2)  1  2407979  2407979   3.677 0.060016 .
C(defoli, CO, 3)  1 18915215 18915215  28.883 1.37e-06 ***
Residuals        59 38638596   654891
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the $\wp$-value of 0 from the ANOVA, one can conclude that there is a statistical difference between the mean for 100% defoliation and the mean for the remaining three levels of defoliation.

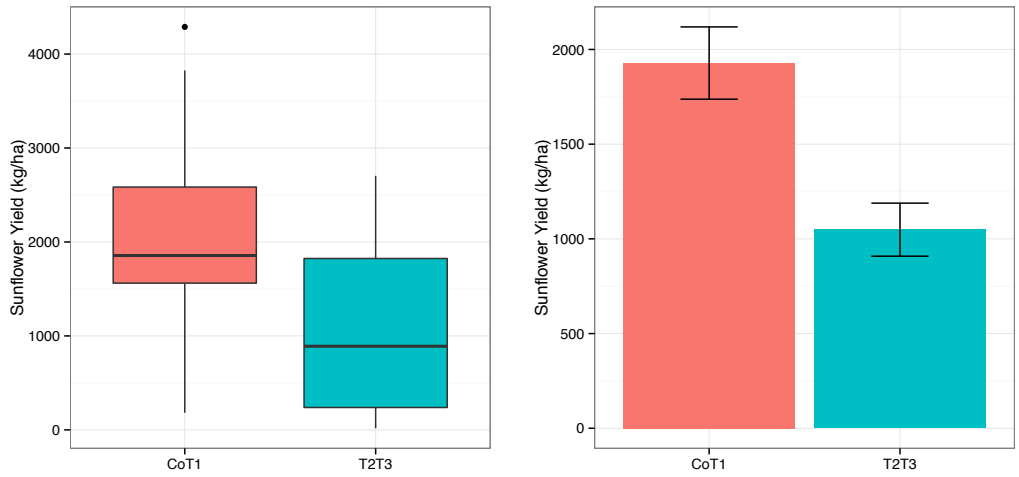
(i)

```
> ggplot(data = ND, aes(x = stage, y = yield, fill = stage)) +
+   geom_boxplot() +
+   theme_bw() +
+   labs(y = "Sunflower Yield (kg/ha)", x = "") +
+   guides(fill = FALSE)
> library(plyr)
> mdf <- ddply(ND, "stage", summarize, MeanStage = mean(yield),
+              SE = sd(yield)/sqrt(length(yield)))
> mdf

   stage MeanStage       SE
1 stage1 1676.0833 332.0844
2 stage2 1496.9375 242.0135
3 stage3 2161.8182 370.2027
4 stage4 1565.3333 227.9503
5 stage5  679.8462 194.7349

> ggplot(data = mdf, aes(x = stage, y = MeanStage, fill = stage)) +
+   geom_bar(stat = "identity") +
+   geom_errorbar(aes(ymin = MeanStage - SE, ymax = MeanStage + SE),
+                 width = 0.30) +
+   guides(fill = FALSE) +
+   theme_bw() +
+   labs(y = "Sunflower Yield (kg/ha)", x = "")
```



(j)

```
> levels(ND$defoli) <- list(CoT1 = c("control", "treat1"),
+                           T2T3 = c("treat2", "treat3"))
> levels(ND$defoli)

[1] "CoT1" "T2T3"

> ggplot(data = ND, aes(x = defoli, y = yield, fill = defoli)) +
+   geom_boxplot() +
+   theme_bw() +
+   labs(y = "Sunflower Yield (kg/ha)", x = "") +
+   guides(fill = FALSE)
> library(plyr)
> mdf <- ddply(ND, "defoli", summarize, MeanStage = mean(yield),
+              SE = sd(yield)/sqrt(length(yield)))
> mdf

  defoli MeanStage       SE
1   CoT1  1928.235 190.7991
2   T2T3  1048.545 140.0398

> ggplot(data = mdf, aes(x = defoli, y = MeanStage, fill = defoli)) +
+   geom_bar(stat = "identity") +
+   geom_errorbar(aes(ymin = MeanStage - SE, ymax = MeanStage + SE),
+                 width = 0.30) +
+   guides(fill = FALSE) +
+   theme_bw() +
+   labs(y = "Sunflower Yield (kg/ha)", x = "")
```



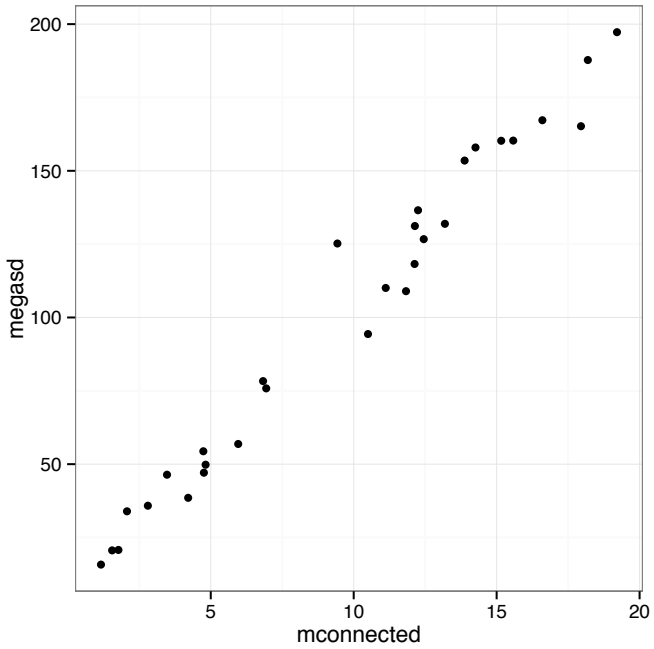
Chapter 12

Odd solutions

Solution for 1:

(a)

```
> ggplot(data = URLADDRESS, aes(x = mconnected, y = megasd)) +
+   geom_point() +
+   theme_bw()
```

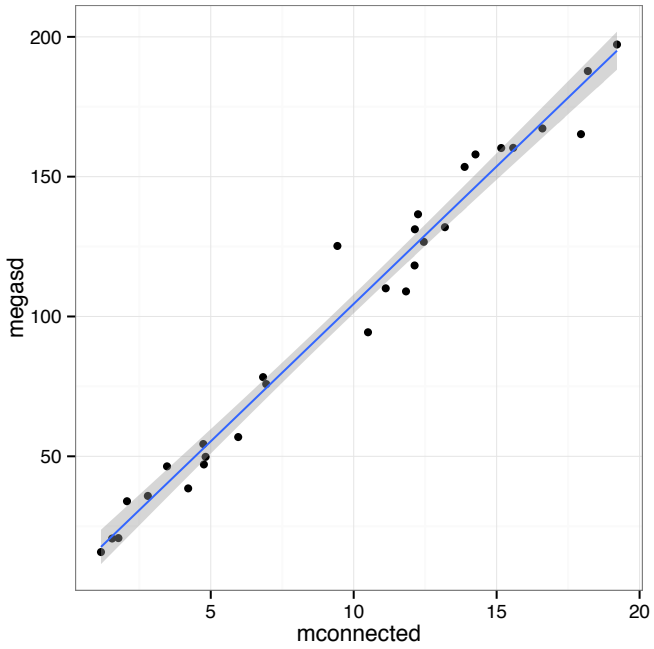


Based on the graph, the relationship between `megasd` and `mconnected` is positive, linear, and strong.

(b)

```
> mod1a <- lm(megasd ~ mconnected, data = URLADDRESS)
> mod1a


Call:
lm(formula = megasd ~ mconnected, data = URLADDRESS)

Coefficients:
(Intercept)   mconnected
      6.189        9.831
```

```
> ggplot(data = URLADDRESS, aes(x = mconnected, y = megasd)) +
+   geom_point() +
+   theme_bw() +
+   geom_smooth(method = "lm")
```



The least squares regression line is $\widehat{Y} = 6.189 + 9.8313x$.

(c) The variance matrix of the $\hat{\beta}$s is computed with the function `vcov()`.

```
> vcov(mod1a)

            (Intercept)  mconnected
(Intercept)  10.8661271 -0.85425034
mconnected   -0.8542503  0.08931002
```

(d)

```
> seb1hat <- sqrt(vcov(mod1a)[2, 2])
> seb1hat
[1] 0.2988478
> # or
> TR <- coef(summary(mod1a))
> TR
            Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 6.188972  3.2963809  1.877505 7.090274e-02
mconnected  9.831263  0.2988478 32.897221 6.369479e-24
> seb1hat <- TR[2, 2]
> seb1hat
[1] 0.2988478
```

The standard error of $\hat{\beta}_1$, $\hat{\sigma}_{\hat{\beta}_1} = s_{\hat{\beta}_1} = 0.2988$.

(e)

```
> vcov(mod1a)

            (Intercept)  mconnected
(Intercept)  10.8661271 -0.85425034
mconnected   -0.8542503  0.08931002

> covb0b1 <- vcov(mod1a)[1, 2]
> covb0b1

[1] -0.8542503
```

The covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ i s-0.8543.

(f)

```
> CI <- confint(mod1a, level = 0.95)
> CI

                2.5 %   97.5 %
(Intercept) -0.5633578 12.94130
mconnected   9.2191007 10.44342
```

The 95% confidence interval for the slope of the regression line from part (b) is $CI_{0.95}(\beta_1) = [9.2191, 10.4434]$.

(g)

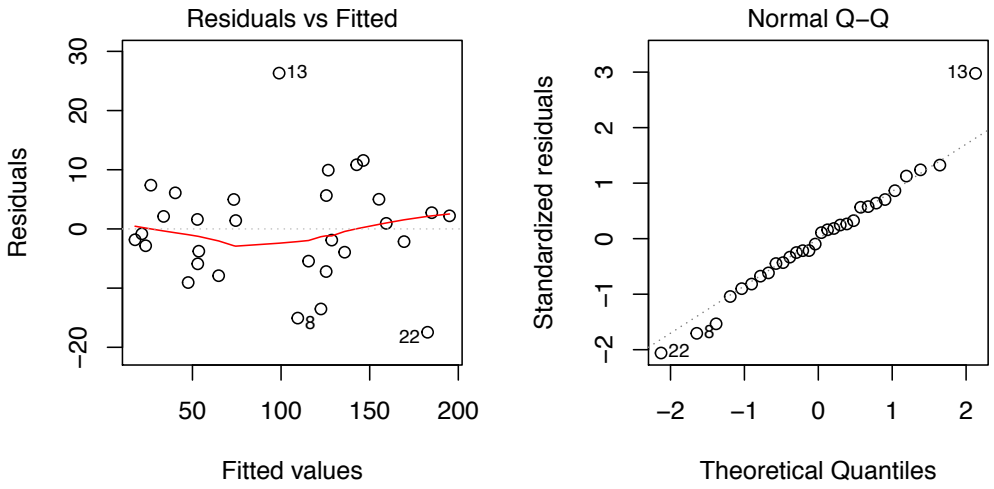
```
> summary(mod1a)


Call:
lm(formula = megasd ~ mconnected, data = URLADDRESS)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4601  -5.0653   0.0563   5.0002  26.3222

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.1890     3.2964   1.878   0.0709 .
mconnected    9.8313     0.2988  32.897   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.992 on 28 degrees of freedom
Multiple R-squared:  0.9748,Adjusted R-squared:  0.9739
F-statistic:  1082 on 1 and 28 DF,  p-value: < 2.2e-16

> summary(mod1a)$r.squared

[1] 0.9747799
```

```
> summary(mod1a)$adj.r.squared

[1] 0.9738792

> summary(mod1a)$sigma

[1] 8.992034

> summary(mod1a)$sigma^2

[1] 80.85668
```

The R^2, R_a^2, and residual variance are 0.9748, 0.9739, and 80.8567, respectively.

(h) The errors are assumed to be independent, follow a normal distribution with mean zero, and have constant variance. Since the errors are unobservable, the residuals are analyzed.

```
> par(mfrow = c(1, 2))
> plot(mod1a, which = 1:2)
> par(mfrow = c(1, 1))
```



The normality assumptions for the errors appears reasonable based on the graphs of the residuals.

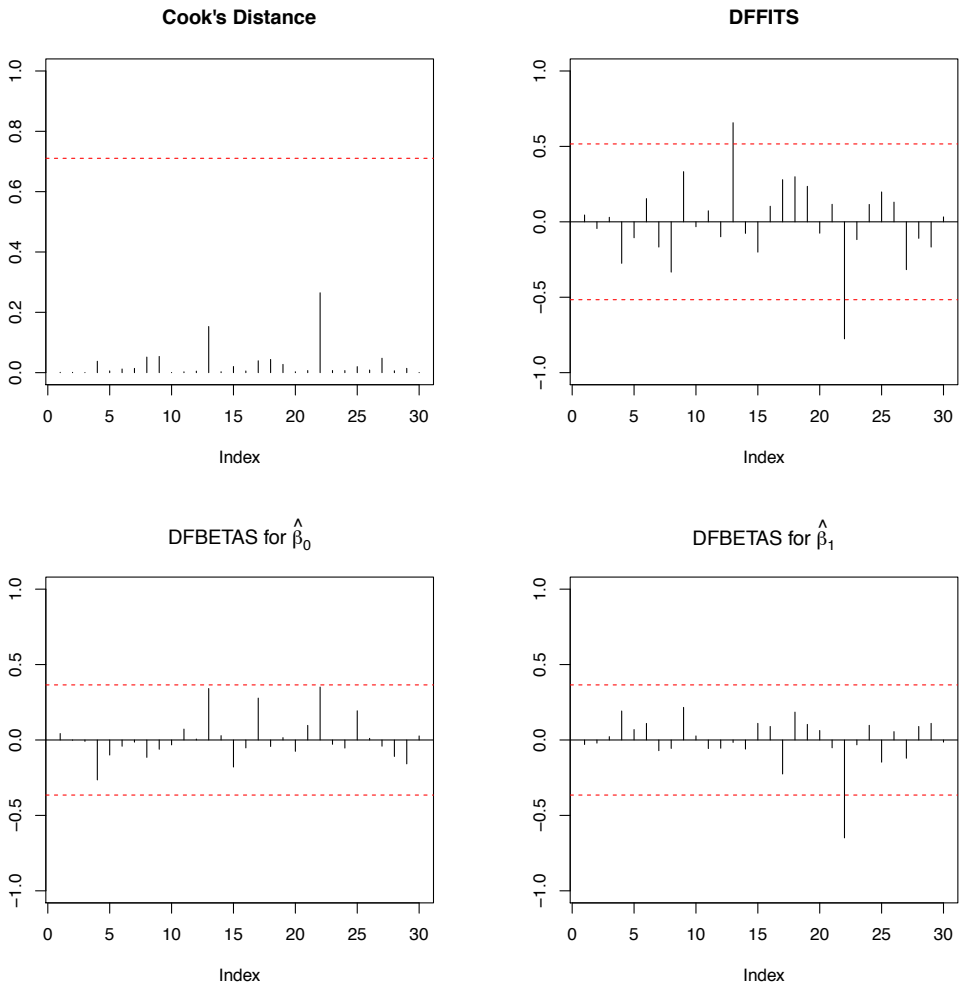
(i)

```
> TR <- outlierTest(mod1a)
> TR

   rstudent unadjusted p-value Bonferonni p
13 3.536669          0.0014863      0.04459
```

Using the Bonferroni approach to test for outliers with the function `outlierTest()` from the `car` package, observation 13 is considered an outlier ($\wp$-value = 0.0446) at the $\alpha = 0.05$ level.

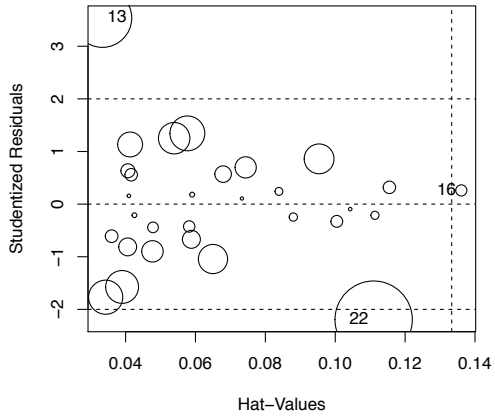
(j)

```
> n <- length(URLADDRESS$mconnected)
> n

[1] 30

> p <- 2
> plot(cooks.distance(mod1a), type = "h", ylim = c(0, 1),
+      ylab = "", main = "Cook's Distance")
> CF <- qf(0.50, p, n - p)
> abline(h = CF, lty = "dashed", col = "red")
> #
> plot(dffits(mod1a), type = "h", ylim = c(-1, 1), ylab = "",
+      main = "DFFITS")
> CV <- 2*sqrt(p/n)
> abline(h = c(CV, -CV), lty = "dashed", col = "red")
> abline(h = 0)
> which(abs(dffits(mod1a)) > CV)

13 22
13 22

> #
> plot(dfbetas(mod1a)[,1], type = "h", ylim = c(-1, 1), ylab = "",
+      main = substitute(paste("DFBETAS for ", hat(beta)[0])))
> CV <- 2/sqrt(n)
> abline(h = c(CV, -CV), lty = "dashed", col = "red")
> abline(h = 0)
> which(abs(dfbetas(mod1a)[,1]) > CV)

named integer(0)

> #
> plot(dfbetas(mod1a)[,2], type = "h", ylim = c(-1, 1), ylab = "",
+      main = substitute(paste("DFBETAS for ", hat(beta)[1])))
> CV <- 2/sqrt(n)
> abline(h = c(CV, -CV), lty = "dashed", col = "red")
> abline(h = 0)
> which(abs(dfbetas(mod1a)[,2]) > CV)

22
22
```



Based on the graphs, DFFITs flags observations 13 and 22 while DFBETAS for $\hat{\beta}_1$ flags observation 22 for further scrutiny.

```
> influencePlot(mod1a)

      StudRes        Hat      CookD
13  3.5366686 0.03335346 0.39106806
16  0.2602743 0.13608497 0.07429144
22 -2.1953572 0.11099209 0.51453521
```



The bubble plot also flags observations 13 and 22 for further study. Observations 13 and 22 are potentially influential observations; however, without further knowledge of the data, there is little one can do other than omit the values and see if the regression line changes significantly.

```
> mod1b <- lm(megasd ~ mconnected, data = URLADDRESS[-c(13, 22),])
> coef(summary(mod1a))

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 6.188972  3.2963809  1.877505 7.090274e-02
mconnected  9.831263  0.2988478 32.897221 6.369479e-24

> coef(summary(mod1b))

              Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  4.231873  2.5714045  1.645744 1.118536e-01
mconnected  10.008235  0.2390842 41.860720 2.176567e-25
```

Observations 13 and 22 are marginally influential as the estimates for the intercept and slope as well as the standard errors for the intercept and slope are marginally different when observations 13 and 22 are removed.

(k)

```
> CI <- predict(mod1a, newdata = data.frame(mconnected = c(5, 10, 15)),
+               interval = "conf", level = 0.90)
> CI

        fit       lwr       upr
1  55.34529  51.71411  58.97646
2 104.50160 101.70009 107.30311
3 153.65791 149.72931 157.58652
```

The estimated mean value of megabytes downloaded by clients spending 5, 10, and 15 minutes on line is 55.3453, 104.5016, and 153.6579 megabytes, respectively. The individual

90% confidence intervals for clients spending 5, 10, and 15 minutes on line, respectively, are

$$CI_{0.90}\left[E(Y_h)\right] = [51.7141, 58.9765],$$
$$CI_{0.90}\left[E(Y_h)\right] = [101.7001, 107.3031], \text{ and}$$
$$CI_{0.90}\left[E(Y_h)\right] = [149.7293, 157.5865].$$

(l)

```
> PI <- predict(mod1a, newdata = data.frame(mconnected = 30),
+               interval = "pred", level = 0.90)
> PI

       fit      lwr      upr
1 301.1269 282.4263 319.8274
```

The predicted megabytes downloaded by a client spending 30 minutes on line is 301.1269. The 90% prediction interval for megabytes downloaded for a client spending 30 minutes on line is

$$PI_{0.90}\left[Y_{h(new)}\right] = [282.4263,\ 319.8274].$$

Solution for 3:

(a)

```
> DF <- c(1, 10)
> SST <- 267 - 2*53/12*53 + 12*(53/12)^2
> SSE <- 22.08
> SSR <- SST - SSE
> SS <- c(SSR, SSE)
> MS <- SS/DF
> Fobs <- c((MS[1])/(MS[2]), NA)
> PrF <- c(pf(Fobs[1], DF[1], DF[2], lower = FALSE), NA)
> ANOVA <- cbind(DF, SS, MS, Fobs, PrF)
> side <- c("Regression","Error")
> top <- c("Df", "Sum Sq", "Mean Sq", "F value", "Pr(>F)")
> dimnames(ANOVA) <- list(side, top)
> ANOVA

           Df   Sum Sq  Mean Sq  F value    Pr(>F)
Regression  1 10.83667 10.83667 4.907911 0.0510929
Error      10 22.08000  2.20800       NA        NA
```

(b) If the user is testing the regression at the $\alpha = 0.05$ level, then, strictly speaking, the regression is not significant since $\wp$-value $= 0.0511$.

(c)

```
> R2 <- SSR/SST
> R2

[1] 0.3292152
```

According to the linear model, 32.9215% of the variability in the response is accounted for by variation in the predictor.

(d) The model's residual variance is $MSE = 2.208$.

(e)

```
> b1 <- (2630 - (53*581)/12) / (28507 - 581^2/12)     # 12.18
> b0 <- 53/12 - b1*581/12                             # 12.14
> c(b0, b1)

[1] -3.7937210  0.1695777

> XTXI22 <- 1/(28507 - 2*581/12*581 + 12*(581/12)^2) # 12.26 entry 2, 2
> XTXI22

[1] 0.002653106

> MSE <- ANOVA[2, 3]
> sb1 <- sqrt(MSE*XTXI22)
> sb1

[1] 0.07653796

> CI <- c(b1 + c(-1, 1)*qt(.975, 12 - 2)*sb1)
> CI

[1] -0.000959481  0.340114909
```

The 95% confidence interval for β_1 is $CI_{0.95}(\beta_1) = [-0.001, 0.3401]$.

Solution for 5:

(a) It must be shown that $E\left[\frac{\sum_i \hat{\varepsilon}_i^2}{n-2}\right] = \sigma^2$. In a simple linear regression model, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Summing over all i and dividing by n yields

$$\overline{Y} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon}_i.$$

Since $\hat{\varepsilon}_i = Y_i - \widehat{Y}_i$ and $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$,

$$\hat{\varepsilon}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

The point $(\bar{x}, \overline{Y})$ is always on the simple linear regression line, so

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \bar{x}.$$

Substituting into the expression of $\hat{\varepsilon}_i$ for $\hat{\beta}_0$ gives

$$\hat{\varepsilon}_i = (Y_i - \overline{Y}) + \hat{\beta}_1(\bar{x} - x_i).$$

Since $Y_i - \overline{Y} = \beta_1(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$,

$$\hat{\varepsilon}_i = (\beta_1 - \hat{\beta}_1)(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}). \tag{12.1}$$

Squaring and summing both sides of (12.1) yields

$$\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \sum_{i=1}^{n}\left[(\beta_1-\hat{\beta}_1)^2(x_i-\bar{x})^2 + 2(\beta_1-\hat{\beta}_1)(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon}) + (\varepsilon_i-\bar{\varepsilon})^2\right]$$
$$= \underbrace{(\beta_1-\hat{\beta}_1)^2\sum_{i=1}^{n}(x_i-\bar{x})^2}_{\text{First}} + \underbrace{2(\beta_1-\hat{\beta}_1)\sum_{i=1}^{n}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})}_{\text{Middle}} + \underbrace{\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2}_{\text{Last}}$$

The expectations of the First, Middle, and Last expressions will be taken to ascertain $E\left[\sum_{i=1}^{n}\hat{\varepsilon}_i^2\right]$. Recall that

$$Var\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \tag{12.2}$$

The expected value of First is

$$\begin{aligned}
E\left[(\beta_1-\hat{\beta}_1)^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right] &= \sum_{i=1}^{n}(x_i-\bar{x})^2\cdot E\left[(\beta_1-\hat{\beta}_1)^2\right] \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2\cdot Var\left[\hat{\beta}_1\right] \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2\cdot\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \\
&= \sigma^2
\end{aligned}$$

The expected value of Middle is

$$\begin{aligned}
&E\left[2(\beta_1-\hat{\beta}_1)\sum_{i=1}^{n}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})\right] \\
&\quad= -2E\left[(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}(x_i-\bar{x})(\varepsilon_i-\bar{\varepsilon})\right] \\
&\quad= -2E\left[(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}(x_i-\bar{x})\underbrace{\left(\hat{\varepsilon}_i-(\beta_1-\hat{\beta}_1)(x_i-\bar{x})\right)}_{\text{From (12.1)}}\right] \\
&\quad= -2E\left[(\hat{\beta}_1-\beta_1)\left\{\sum_{i=1}^{n}\hat{\varepsilon}_i(x_i-\bar{x}) + (\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}\right]
\end{aligned}$$

By property 3 of the fitted regression line, $\sum_{i=1}^{n}x_i\hat{\varepsilon}_i = 0$, and property 1 of the fitted regression line, $\sum_{i=1}^{n}\hat{\varepsilon}_i = 0$:

$$\begin{aligned}
&= -2E\left[(\hat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}(x_i-\bar{x})^2\right] \\
&= -2\sum_{i=1}^{n}(x_i-\bar{x})^2\cdot Var\left[\hat{\beta}_1\right] \\
&= -2\sigma^2 \quad \text{By (12.2)}
\end{aligned}$$

Recall that $\varepsilon_i \sim N(0, \sigma^2)$. This means $E[\varepsilon_i^2] = E[(\varepsilon_i - 0)^2] = Var[\varepsilon_i] = \sigma^2$. For simple linear regression, it is also true that the covariance of any two error terms is zero. This implies that $E[\varepsilon_i \varepsilon_j] = 0$ if $i \neq j$.

The expected value of LAST is

$$\begin{aligned} E\left[\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2\right] &= E\left[\sum_{i=1}^{n}\left(\varepsilon_i^2 - 2\varepsilon_i\bar{\varepsilon} + \bar{\varepsilon}^2\right)\right] \\ &= \sum_{i=1}^{n} E\left[\varepsilon_i^2\right] - 2\sum_{i=1}^{n} E\left[\varepsilon_i\bar{\varepsilon}\right] + nE\left[\bar{\varepsilon}^2\right] \\ &= n\sigma^2 - 2\sum_{i=1}^{n} E\left[\varepsilon_i\left(\frac{\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_n}{n}\right)\right] + n\,Var[\bar{\varepsilon}] \\ &= n\sigma^2 - 2\sigma^2 + n \cdot \frac{\sigma^2}{n} \\ &= n\sigma^2 - \sigma^2 = \sigma^2(n-1) \end{aligned}$$

$$\begin{aligned} E\left[\sum_i \hat{\varepsilon}_i^2\right] &= E[\text{FIRST} + \text{MIDDLE} + \text{LAST}] \\ &= \sigma^2 - 2\sigma^2 + \sigma^2(n-1) = (n-2)\sigma^2 \\ \implies E\left[\frac{\sum_i \hat{\varepsilon}_i^2}{n-2}\right] &= E\left[\hat{\sigma}^2\right] = \sigma^2 \quad \text{Check} \end{aligned}$$

(b)

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\begin{bmatrix} \sum_{i-1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}$$

$$\begin{aligned} h_{ii} &= \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \\ &= [1 \quad x_i]\frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}\begin{bmatrix} 1 \\ x_i \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}[1 \quad x_i]\begin{bmatrix} \sum_{i=1}^{n} x_i^2 - x_i\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i + nx_i \end{bmatrix} \\ &= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\left[\sum_{i=1}^{n} x_i^2 - x_i\sum_{i=1}^{n} x_i + x_i\left(-\sum_{i=1}^{n} x_i + nx_i\right)\right] \\ &= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\left[\sum_{i=1}^{n} x_i^2 - 2x_i\sum_{i=1}^{n} x_i + nx_i^2\right] \end{aligned}$$

Recall that $\frac{\sum_{i=1}^n x_i^2}{n} - \bar{x} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$ and $\sum_{i=1}^n x_i = n\bar{x}$.

$$\begin{aligned}
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\frac{\sum_{i=1}^n x_i^2}{n} \underbrace{-\bar{x}^2 + \bar{x}^2}_{\text{Add for form.}} - 2x_i\bar{x} + x_i^2 \right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \left[\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} + (x_i - \bar{x})^2 \right] \\
&= \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{Check}
\end{aligned}$$

Solution for 7:

Recall that $r_i^* = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$.

Also note that

$$\widehat{Y}_i - \widehat{Y}_{i(i)} = \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}$$

The expression in (12.5) as well as that for h_{ii} allow simplification to

$$\begin{aligned}
&= \mathbf{x}_i' \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{\varepsilon}_i}{1 - h_{ii}} \\
\widehat{Y}_i - \widehat{Y}_{i(i)} &= \frac{h_{ii}\hat{\varepsilon}_i}{1 - h_{ii}}
\end{aligned}$$

The original expression becomes

$$\begin{aligned}
\text{DFFITS}_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} &= \frac{h_{ii}\hat{\varepsilon}_i}{1 - h_{ii}} \cdot \frac{1}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}} \\
&= \frac{\hat{\varepsilon}_i \cdot \sqrt{h_{ii}}}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}} \cdot \sqrt{1 - h_{ii}}} \\
&= r_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \quad \text{Check}
\end{aligned}$$

Solution for 9:

It is known that $SSR = SST - SSE$, so

$$\begin{aligned}
SSR = \sum_{i=1}^n (\widehat{Y}_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \bar{Y})^2 - \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2 \\
&= \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} - \left[\underbrace{\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}}_{} \right] \\
&= \mathbf{Y}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} - \left[\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} \right] \\
&= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \quad \text{Check}
\end{aligned}$$

Solution for 11:

R Code 12.1

```
> mod1 <- lm(hwfat ~ abs + triceps, data = HSWRESTLER)
> anova(mod1)

Analysis of Variance Table

Response: hwfat
          Df Sum Sq Mean Sq F value    Pr(>F)
abs        1 5072.8  5072.8 541.365 < 2.2e-16 ***
triceps    1  242.2   242.2  25.844 2.639e-06 ***
Residuals 75  702.8     9.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> SSR <- sum(anova(mod1)[1:2, 2])
> SSR

[1] 5315.008
```

The value obtained previously with the `lm()` and `anova()` functions for $SSR = 5315.0081$. Recall that

$$\begin{aligned} SSR &= \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - \tfrac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y} \\ &= \mathbf{Y}'(\mathbf{H} - \tfrac{1}{n}\mathbf{J})\mathbf{Y} \end{aligned}$$

```
> n <- dim(HSWRESTLER)[1]
> Y <- matrix(HSWRESTLER$hwfat, nrow = n)
> dim(Y)

[1] 78  1

> X <- model.matrix(mod1)
> dim(X)

[1] 78  3

> H <- X%*%(solve(t(X)%*%X)%*%t(X))
> dim(H)

[1] 78 78

> J <- matrix(rep(1, n*n), nrow = n)
> dim(J)

[1] 78 78

> SSR <- t(Y)%*%(H - 1/n*J)%*%Y
> SSR

         [,1]
[1,] 5315.008
```

Using quadratic forms, the value for $SSR = 5315.0081$, the same value computed from using R Code 12.1.

Solution for 13:

(a)

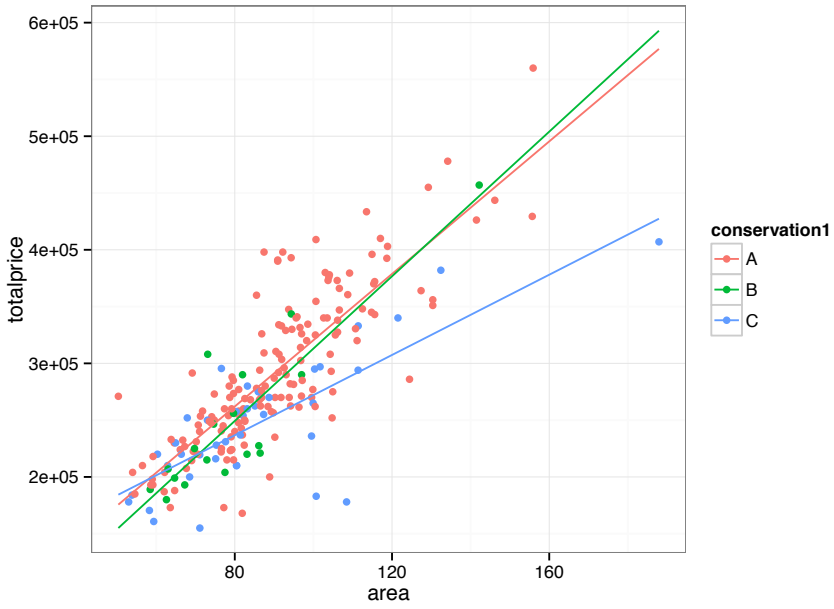
```
> levels(VIT2005$conservation)

[1] "1A" "2A" "2B" "3A"

> xtabs(~conservation, data = VIT2005)

conservation
 1A  2A  2B  3A
161  18  36   3

> VIT2005$conservation1 <- VIT2005$conservation
> levels(VIT2005$conservation1) <- list(A = "1A", B = "2A",
+                                       C = c("2B", "3A"))
> xtabs(~conservation1, data = VIT2005)

conservation1
  A   B   C
161  18  39
```

(b)

```
> line1 <- lm(totalprice ~ area, data = VIT2005,
+             subset = VIT2005$conservation1 == "A")
> coef(summary(line1))

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 28634.285 14411.9012  1.98685 4.865701e-02
area         2917.545   155.6748 18.74128 4.204588e-42

> line2 <- lm(totalprice ~ area, data = VIT2005,
+             subset = VIT2005$conservation1 == "B")
> coef(summary(line2))

             Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -5373.946 35455.1129 -0.1515704 8.814204e-01
area         3183.757   433.2997  7.3477006 1.641688e-06

> line3 <- lm(totalprice ~ area, data = VIT2005,
+             subset = VIT2005$conservation1 == "C")
> coef(summary(line3))

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 95394.348 20555.5010 4.640818 4.254153e-05
area         1766.622   231.2392 7.639803 4.051498e-09
```

(c)

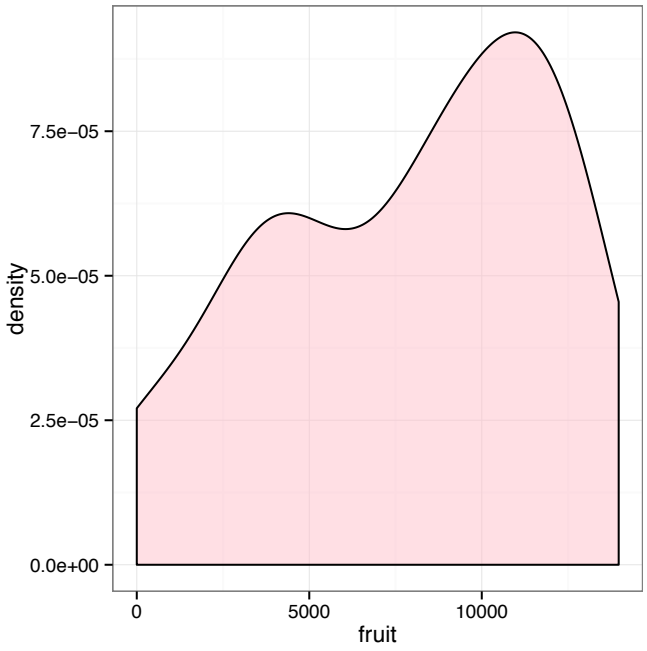
```
> ggplot(data = VIT2005, aes(x = area, y = totalprice,
+                            color = conservation1)) +
+   geom_point() +
+   geom_smooth(method = "lm", se = FALSE, fullrange = TRUE) +
+   theme_bw()
> rm(VIT2005)  # Clean up
```



Solution for 15:

(a)

```
> ggplot(data = SATFRUIT, aes(x = fruit)) +
+   geom_density(fill = "pink", alpha = 0.5) +
+   theme_bw()
> median(SATFRUIT$fruit)

[1] 8536.259

> IQR(SATFRUIT$fruit)

[1] 7115.129
```



The distribution of `fruit` appears to be bimodal with modes occurring around 4000 m^2 and 11000 m^2. The median is 8536.2585 m^2 and the IQR is 7115.1291 m^2.

(b)

```
> mfa <- max(SATFRUIT$fruit)
> mfa

[1] 13968.61
```

The maximum number of m^2 classified fruits in a segment is 13968.608 m^2.

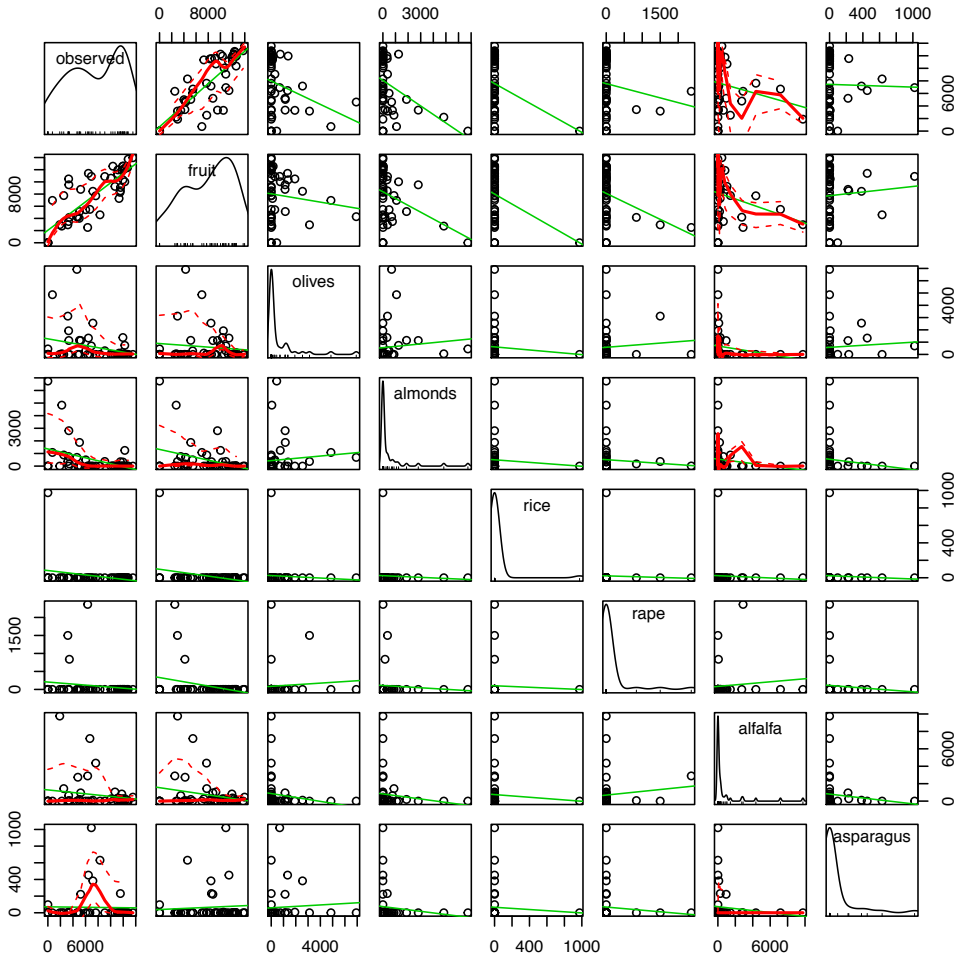
(c)

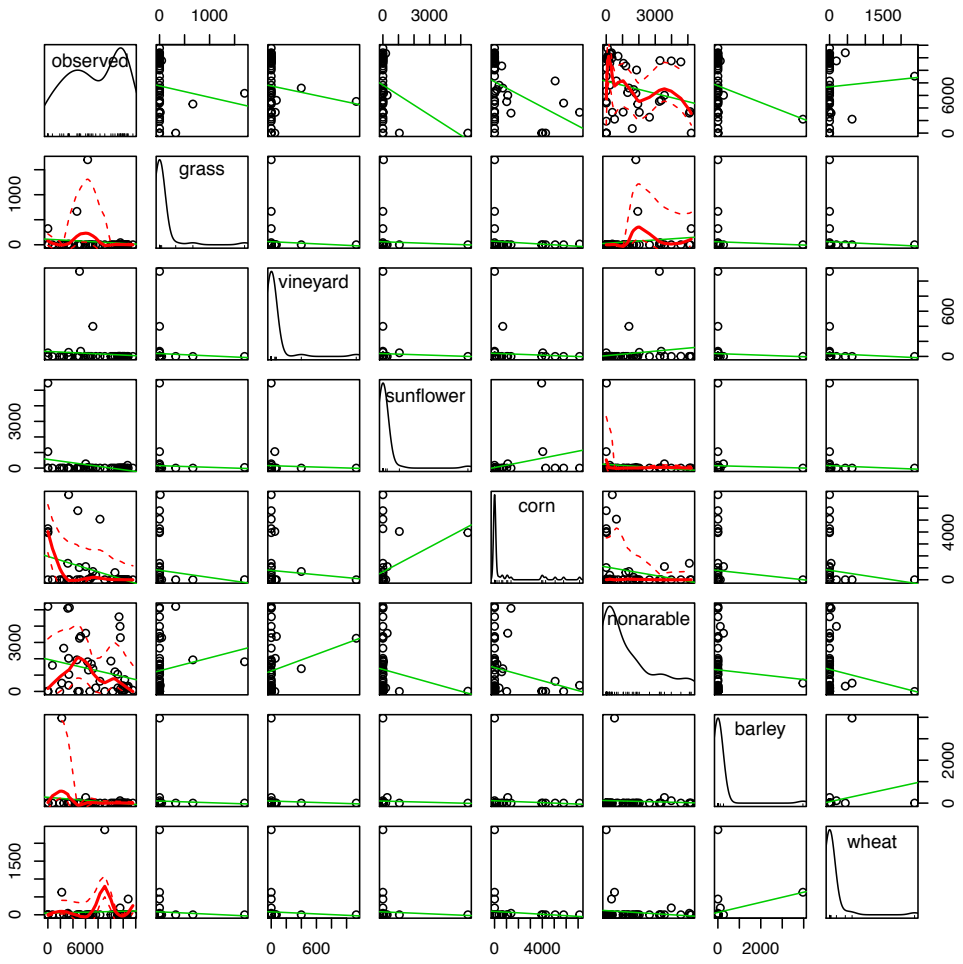
```
> T1 <- xtabs(~smallarea, data = SATFRUIT)
> T1

smallarea
R63 R67 R68
  3  32  12
```

There are 3 observations in small area `R63`, 32 observations in small area `R67`, and 12 observations in small area `R68`.

(d) The function `scatterplotMatrix()` is used twice, each time plotting `observed` versus seven different numerical variables from `SATFRUIT`.

```
> library(car)
> scatterplotMatrix(SATFRUIT[, c(17, 16, 15, 14, 13, 12, 11, 10)])
> scatterplotMatrix(SATFRUIT[, c(17, 9, 8, 7, 6, 5, 4, 3)])
```

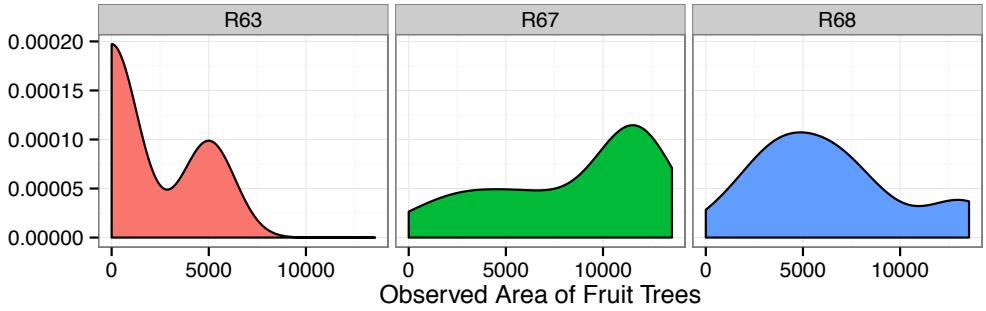

observed
fruit
olives
almonds
rice
rape
alfalfa
asparagus
0 8000
0 3000
0 1500
0 400 1000
0 6000
0 4000
0 400 1000
0 6000



There is a positive linear association between `observed` and `fruit`.

(e)

```
> ggplot(data = SATFRUIT, aes(x = observed, fill = smallarea)) +
+   geom_density() +
+   facet_grid(.~smallarea) +
+   guides(fill = FALSE) +
+   labs(x = "Observed Area of Fruit Trees", y = "") +
+   theme_bw()
```

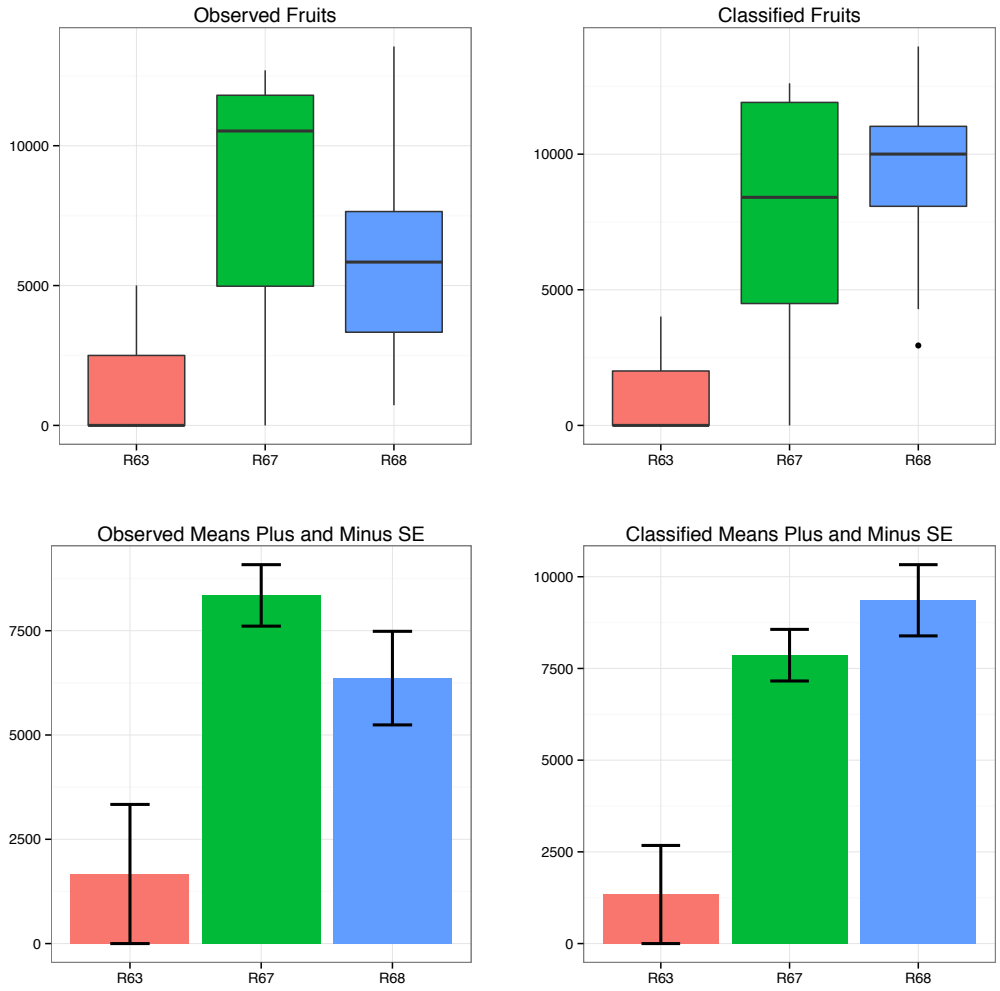


The density plot of `observed` for small area `R63` is bimodal with a positive skew. The density plot of `observed` for small area `R67` is unimodal with a negative skew. The density plot of `observed` for small area `R68` is unimodal with a slight positive skew.

(f)

```
> library(plyr)
> mdf <- ddply(SATFRUIT, "smallarea", summarize,
+              mobserved = mean(observed),
+              mfruit = mean(fruit),
+              seobserved = sd(observed)/sqrt(length(observed)),
+              sefruit = sd(fruit)/sqrt(length(fruit)))
> mdf

  smallarea mobserved   mfruit seobserved   sefruit
1       R63  1667.786 1337.629  1667.7857 1337.5750
2       R67  8346.330 7861.341   736.7979  703.4567
3       R68  6363.758 9358.520  1121.6687  970.7940

> ggplot(data = SATFRUIT, aes(x = smallarea, y = observed,
+                             fill = smallarea)) +
+   geom_boxplot() +
+   guides(fill = FALSE) +
+   theme_bw() +
+   labs(x = "", y = "", title = "Observed Fruits")
> ggplot(data = SATFRUIT, aes(x = smallarea, y = fruit,
+                             fill = smallarea)) +
+   geom_boxplot() +
+   guides(fill = FALSE) +
+   theme_bw() +
+   labs(x = "", y = "", title = "Classified Fruits")
> ggplot(data = mdf, aes(x = smallarea, y = mobserved, fill = smallarea)) +
+   geom_bar(stat = "identity") +
+   geom_errorbar(aes(ymin = mobserved - seobserved,
+                     ymax = mobserved + seobserved),
+                 width = 0.3, size = 1) +
+   guides(fill = FALSE) +
+   labs(x = "", y = "", title = "Observed Means Plus and Minus SE") +
+   theme_bw()
> ggplot(data = mdf, aes(x = smallarea, y = mfruit, fill = smallarea)) +
+   geom_bar(stat = "identity") +
```

```
+   geom_errorbar(aes(ymin = mfruit - sefruit,
+                     ymax = mfruit + sefruit), width = 0.3, size = 1) +
+   guides(fill = FALSE) +
+   labs(x = "", y = "", title = "Classified Means Plus and Minus SE") +
+   theme_bw()
```



The boxplots and barplots of observed fruits (`observed`) and classified fruits (`fruit`) by small areas allow one to see the similarities in measurements.

(g)

```
> NUM <- c("wheat", "barley", "nonarable", "corn", "sunflower",
+          "vineyard", "grass", "asparagus", "alfalfa", "rape",
+          "rice", "almonds", "olives", "fruit")
> COR <- cor(SATFRUIT[ ,"observed"], SATFRUIT[, NUM])
> COR

          wheat     barley  nonarable       corn  sunflower   vineyard
[1,] 0.05068929 -0.1781316 -0.2402526 -0.4022165 -0.2973749 -0.1012791
```

```
          grass   asparagus    alfalfa      rape       rice    almonds
[1,] -0.1122948 -0.01936316 -0.1713412 -0.146768 -0.2553356 -0.3988465
        olives     fruit
[1,] -0.289251 0.8186904
```

The highest three correlations with `observed` occur with `fruit` (0.8187), `corn` (-0.4022), and `almonds` (-0.3988).

Model (A)

```
> model.all <- lm(observed ~ ., data = SATFRUIT[, -c(1, 2)])
> drop1(model.all, test = "F")

Single term deletions

Model:
observed ~ wheat + barley + nonarable + corn + sunflower + vineyard +
    grass + asparagus + alfalfa + rape + rice + almonds + olives +
    fruit
           Df Sum of Sq       RSS    AIC F value   Pr(>F)
<none>                 187855866 744.45
wheat       1    749537 188605403 742.64  0.1277 0.723197
barley      1   1241554 189097420 742.76  0.2115 0.648711
nonarable   1    449872 188305738 742.56  0.0766 0.783694
corn        1    220048 188075914 742.50  0.0375 0.847707
sunflower   1   9457413 197313279 744.76  1.6110 0.213504
vineyard    1     77109 187932975 742.47  0.0131 0.909472
grass       1  19149030 207004896 747.01  3.2619 0.080321 .
asparagus   1   8582514 196438380 744.55  1.4620 0.235474
alfalfa     1   9188703 197044569 744.69  1.5652 0.219970
rape        1   7781620 195637486 744.36  1.3255 0.258129
rice        1   7943324 195799190 744.40  1.3531 0.253339
almonds     1  28339632 216195498 749.05  4.8275 0.035367 *
olives      1  51082691 238938557 753.75  8.7016 0.005902 **
fruit       1  48852251 236708117 753.31  8.3217 0.006955 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.be <- update(model.all, .~. -vineyard)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ wheat + barley + nonarable + corn + sunflower + grass +
    asparagus + alfalfa + rape + rice + almonds + olives + fruit
           Df Sum of Sq       RSS    AIC F value   Pr(>F)
<none>                 187932975 742.47
wheat       1    762090 188695066 740.66  0.1338 0.716840
barley      1   1188596 189121571 740.76  0.2087 0.650773
nonarable   1    449333 188382308 740.58  0.0789 0.780547
corn        1    258617 188191592 740.53  0.0454 0.832560
```

```
sunflower  1   9649791 197582766 742.82  1.6945 0.202029
grass      1  19072158 207005133 745.01  3.3490 0.076293 .
asparagus  1   8527358 196460333 742.55  1.4974 0.229744
alfalfa    1   9439172 197372148 742.77  1.6575 0.206902
rape       1   7880704 195813680 742.40  1.3838 0.247870
rice       1   8092214 196025189 742.45  1.4209 0.241749
almonds    1  28294466 216227442 747.06  4.9684 0.032740 *
olives     1  51005719 238938694 751.75  8.9563 0.005204 **
fruit      1  49360377 237293352 751.43  8.6674 0.005894 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.be <- update(model.be, .~. -corn)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ wheat + barley + nonarable + sunflower + grass + asparagus +
    alfalfa + rape + rice + almonds + olives + fruit
           Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                 188191592 740.53
wheat      1    695993 188887586 738.71  0.1257  0.725077
barley     1   1688126 189879718 738.95  0.3050  0.584384
nonarable  1   1312294 189503886 738.86  0.2371  0.629442
sunflower  1   9465632 197657224 740.84  1.7101  0.199743
grass      1  21506668 209698260 743.62  3.8855  0.056887 .
asparagus  1   9384987 197576579 740.82  1.6956  0.201623
alfalfa    1  10829975 199021568 741.16  1.9566  0.170932
rape       1   7891052 196082645 740.46  1.4257  0.240739
rice       1   7834334 196025926 740.45  1.4154  0.242403
almonds    1  28985865 217177457 745.27  5.2368  0.028453 *
olives     1  53442129 241633721 750.28  9.6552  0.003799 **
fruit      1 132702772 320894364 763.61 23.9750 2.339e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.be <- update(model.be, .~. -wheat)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ barley + nonarable + sunflower + grass + asparagus +
    alfalfa + rape + rice + almonds + olives + fruit
           Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                 188887586 738.71
barley     1   1402403 190289989 737.05  0.2599  0.613417
nonarable  1   1567279 190454865 737.09  0.2904  0.593370
sunflower  1   9331527 198219113 738.97  1.7291  0.197078
grass      1  21850524 210738110 741.85  4.0488  0.051949 .
asparagus  1   9104030 197991616 738.92  1.6869  0.202498
```

```
alfalfa    1  10478270 199365856 739.24  1.9416  0.172284
rape       1   8174004 197061590 738.70  1.5146  0.226645
rice       1   7727390 196614975 738.59  1.4318  0.239506
almonds    1  29526603 218414189 743.53  5.4711  0.025169 *
olives     1  53249957 242137543 748.38  9.8670  0.003414 **
fruit      1 132045934 320933520 761.62 24.4675 1.886e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -nonarable)
> drop1(model.be, test = "F")
```

```
Single term deletions

Model:
observed ~ barley + sunflower + grass + asparagus + alfalfa +
    rape + rice + almonds + olives + fruit
          Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                 190454865 737.09
barley     1    654522 191109387 735.26  0.1237  0.727087
sunflower  1  12467556 202922421 738.07  2.3566  0.133494
grass      1  20352171 210807036 739.87  3.8470  0.057605 .
asparagus  1   8548831 199003697 737.16  1.6159  0.211815
alfalfa    1  14110835 204565700 738.45  2.6672  0.111150
rape       1   6618738 197073603 736.70  1.2511  0.270754
rice       1  10266657 200721522 737.56  1.9406  0.172145
almonds    1  28172638 218627503 741.58  5.3252  0.026878 *
olives     1  51778032 242232897 746.40  9.7871  0.003473 **
fruit      1 136557997 327012862 760.50 25.8124 1.174e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -barley)
> drop1(model.be, test = "F")
```

```
Single term deletions

Model:
observed ~ sunflower + grass + asparagus + alfalfa + rape + rice +
    almonds + olives + fruit
          Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                 191109387 735.26
sunflower  1  12581665 203691052 736.25  2.4359  0.127100
grass      1  19816183 210925570 737.89  3.8365  0.057717 .
asparagus  1   8478560 199587947 735.30  1.6415  0.208091
alfalfa    1  14216422 205325809 736.63  2.7524  0.105564
rape       1   6382540 197491927 734.80  1.2357  0.273472
rice       1  10368281 201477668 735.74  2.0074  0.164902
almonds    1  30140317 221249704 740.14  5.8354  0.020762 *
olives     1  53865187 244974574 744.93 10.4286  0.002603 **
fruit      1 136019618 327129005 758.52 26.3343 9.384e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -rape)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ sunflower + grass + asparagus + alfalfa + rice + almonds +
    olives + fruit
           Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                  197491927 734.80
sunflower  1   9247243 206739170 734.95  1.7793  0.190177
grass      1  14256487 211748414 736.08  2.7431  0.105909
asparagus  1   7010826 204502752 734.44  1.3490  0.252701
alfalfa    1  11009029 208500956 735.35  2.1183  0.153762
rice       1   7581297 205073224 734.57  1.4587  0.234594
almonds    1  24163064 221654991 738.22  4.6493  0.037460 *
olives     1  49218231 246710158 743.26  9.4702  0.003863 **
fruit      1 179931492 377423419 763.24 34.6211 8.213e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.be <- update(model.be, .~. -asparagus)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ sunflower + grass + alfalfa + rice + almonds + olives +
    fruit
           Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                  204502752 734.44
sunflower  1   7708043 212210795 734.18  1.4700  0.232644
grass      1  15967531 220470283 735.97  3.0451  0.088857 .
alfalfa    1   8717497 213220250 734.40  1.6625  0.204863
rice       1   6239692 210742445 733.85  1.1899  0.282034
almonds    1  20447995 224950748 736.92  3.8996  0.055409 .
olives     1  50571125 255073878 742.83  9.6442  0.003532 **
fruit      1 188035615 392538367 763.09 35.8596 5.377e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> model.be <- update(model.be, .~. -rice)
> drop1(model.be, test = "F")

Single term deletions

Model:
observed ~ sunflower + grass + alfalfa + almonds + olives + fruit
           Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>                  210742445 733.85
sunflower  1   4806006 215548450 732.91  0.9122   0.34527
grass      1  18831533 229573978 735.87  3.5743   0.06594 .
```

```
alfalfa    1   5762467 216504911 733.12  1.0937   0.30192
almonds    1  15909834 226652278 735.27  3.0198   0.08995 .
olives     1  46346966 257089411 741.19  8.7969   0.00507 **
fruit      1 277244647 487987092 771.32 52.6225 8.323e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -sunflower)
> drop1(model.be, test = "F")
```

```
Single term deletions

Model:
observed ~ grass + alfalfa + almonds + olives + fruit
       Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>               215548450 732.91
grass   1  21162588 236711038 735.31  4.0254  0.051447 .
alfalfa 1   3748504 219296954 731.72  0.7130  0.403346
almonds 1  12553969 228102419 733.57  2.3879  0.129961
olives  1  43040568 258589019 739.47  8.1869  0.006613 **
fruit   1 392515189 608063639 779.66 74.6613 8.898e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -alfalfa)
> drop1(model.be, test = "F")
```

```
Single term deletions

Model:
observed ~ grass + almonds + olives + fruit
       Df Sum of Sq       RSS    AIC F value    Pr(>F)
<none>               219296954 731.72
grass   1  19596107 238893061 733.74  3.7531  0.059455 .
almonds 1  10196514 229493468 731.86  1.9528  0.169619
olives  1  39717334 259014288 737.55  7.6067  0.008573 **
fruit   1 449327366 668624320 782.12 86.0557 1.005e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> model.be <- update(model.be, .~. -almonds)
> drop1(model.be, test = "F")
```

```
Single term deletions

Model:
observed ~ grass + olives + fruit
       Df Sum of Sq       RSS    AIC  F value    Pr(>F)
<none>               229493468 731.86
grass   1  20200760 249694228 733.82   3.7850  0.058270 .
olives  1  42521560 272015027 737.85   7.9672  0.007186 **
fruit   1 556838390 786331858 787.74 104.3343 4.516e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> modelA <- lm(observed ~ grass + olives + fruit, data = SATFRUIT)
```

The independent variables included in Model (A) are `grass`, `olives`, and `fruit`.

(i)

```
> modelA <- lm(observed ~ grass + olives + fruit, data = SATFRUIT)
> modelAg <- glm(observed ~ grass + olives + fruit, data = SATFRUIT)
> library(boot)
> cv.errorN <- cv.glm(SATFRUIT, modelAg)
> CVNa <- cv.errorN$delta[1]
> CVNa

[1] 9131968

> set.seed(5)
> cv.error5 <- cv.glm(SATFRUIT, modelAg, K=5)
> CV5a <- cv.error5$delta[1]
> CV5a

[1] 10000800
```

The $CV_n = 9131967.681$ for model (A), and $CV_5 = 10000799.7019$ for model (A).

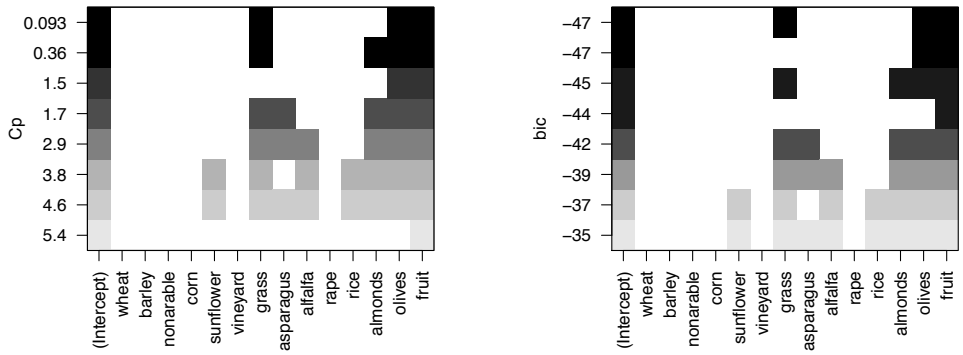
(ii) Since this problem and a few more will request many goodness of fit statistics, a function called `mgof()` is written to compute the requested values.

```
> mgof <- function(model = model, data = DF, ...){
+   R2a <- summary(model)$adj.r.squared
+   R2 <- summary(model)$r.squared
+   aic <- AIC(model)
+   bic <- AIC(model, k = log(nrow(data)))
+   se <- summary(model)$sigma
+   form <- formula(model)
+   ANS <- c(R2 = R2, R2.adj = R2a, AIC = aic, BIC = bic, SE = se)
+   ANS
+ }
> MGOF <- mgof(model = modelA, data = SATFRUIT)
> MGOF

         R2       R2.adj          AIC          BIC           SE
  0.7335810    0.7149936  867.2383856  876.4891237 2310.2072172
```

The R^2, R_a^2, AIC, BIC, and standard error for `modelA` are 0.7336, 0.715, 867.2384, 876.4891, and 2310.2072, respectively. The total proportion of variability explained by `modelA` is 0.7336.

Model (B)

```
> library(leaps)
> models.exh <- regsubsets(observed ~., data = SATFRUIT[, -c(1, 2)])
> plot(models.exh, scale = "Cp")    # returns grass, olives, and fruit
> plot(models.exh, scale = "bic")   # returns grass, olives, and fruit
> modelB <- lm(observed ~ grass + olives + fruit, data = SATFRUIT)
```



Based on the plots, the model suggested with the C_p criteria is the same as the model suggested with the BIC criteria. Both the BIC and the C_p criteria suggest a model with the variables `grass`, `olives`, and `fruit`, which are the same as the variables in model (A) (`modelA`).

Model (C)

```
> predict.regsubsets <- function(object, newdata, id, ...){
+   form <- as.formula(object$call[[2]])
+   mat <- model.matrix(form, newdata)
+   coefi = coef(object, id = id)
+   xvars <- names(coefi)
+   mat[, xvars]%*%coefi
+ }
> n <- nrow(SATFRUIT)
> k <- n              # set the number of folds equal to n
> folds <- sample(x = 1:k, size = nrow(SATFRUIT), replace = FALSE)
> cv.errors <- matrix(NA, k, 8, dimnames = list(NULL, paste(1:8)))
> for(j in 1:k){
+   best.fit <- regsubsets(observed ~.,
+                          data = SATFRUIT[folds != j, -c(1, 2)])
+   for(i in 1:8){
+     pred <- predict(best.fit,
+                     newdata = SATFRUIT[folds == j, -c(1, 2)], id = i)
+     cv.errors[j, i] <- mean((SATFRUIT$observed[folds == j] - pred)^2)
+   }
+ }

Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in
= force.in, :  1  linear dependencies found

Reordering variables and trying again:

> mean.cv.errors <- apply(cv.errors, 2, mean)
> mean.cv.errors
```

```
      1       2       3       4       5       6       7       8
 7622559  7015165  9846285 14783252 20398603 16846213 20298590 19530744

> which.min(mean.cv.errors)

2
2

> coef(models.exh, which.min(mean.cv.errors))

 (Intercept)       olives        fruit
1253.8210813   -0.6455802    0.8366379

> ### k-fold (5)
> n <- nrow(SATFRUIT)
> k <- 5               # set the number of folds equal to n
> set.seed(5)          # set for reproducible results
> folds <- sample(x = 1:k, size = nrow(SATFRUIT), replace = TRUE)
> cv.errors <- matrix(NA, k, 8, dimnames = list(NULL, paste(1:8)))
> for(j in 1:k){
+   best.fit <- regsubsets(observed ~.,
+                          data = SATFRUIT[folds != j, -c(1, 2)])
+   for(i in 1:8){
+     pred <- predict(best.fit,
+                     newdata = SATFRUIT[folds == j, -c(1, 2)], id = i)
+     cv.errors[j, i] <- mean((SATFRUIT$observed[folds == j] - pred)^2)
+   }
+ }

Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in
= force.in, :  1  linear dependencies found

Reordering variables and trying again:

> mean.cv.errors <- apply(cv.errors, 2, mean)
> mean.cv.errors

      1       2       3       4       5       6       7       8
10000416  6761528 10962898 19364798 21858719 25082236 65280034 26785702

> which.min(mean.cv.errors)

2
2

> coef(models.exh, which.min(mean.cv.errors))

 (Intercept)       olives        fruit
1253.8210813   -0.6455802    0.8366379

> modelD <- lm(observed ~ olives + fruit, data = SATFRUIT)
```

Using leave-one-out cross validation as well as $k = 5$ fold cross validation suggests using a model with two variables (`olives`, and `fruit`) produces a model with the smallest possible

MSPE.

(i)

```
> modelC <- lm(observed ~ olives + fruit, data = SATFRUIT)
> modelCg <- glm(observed ~ olives + fruit, data = SATFRUIT)
> library(boot)
> cv.errorN <- cv.glm(SATFRUIT, modelCg)
> CVNc <- cv.errorN$delta[1]
> CVNc

[1] 7045110

> set.seed(5)
> cv.error5 <- cv.glm(SATFRUIT, modelCg, K=5)
> CV5c <- cv.error5$delta[1]
> CV5c

[1] 8505432
```

The $CV_n = 7045109.8173$ for model (C), and $CV_5 = 8505431.931$ for model (C).

(ii)

```
> MGOF <- mgof(model = modelC, data = SATFRUIT)
> MGOF

         R2       R2.adj          AIC          BIC           SE
  0.7101299    0.6969540  869.2034239  876.6040143 2382.1983162
```

The R^2, R_a^2, AIC, BIC, and standard error for `modelC` are 0.7101, 0.697, 869.2034, 876.604, and 2382.1983, respectively. The total proportion of variability explained by `modelC` is 0.7101.

Model (D) Since model (C) had the smaller cross validation error between models (A) and (C), model (D) is created by adding the variable `smallarea` to model (C).

```
> modelD <- lm(observed ~ olives + fruit + smallarea, data = SATFRUIT)
```

(i)

```
> drop1(modelD, test = "F")

Single term deletions

Model:
observed ~ olives + fruit + smallarea
          Df Sum of Sq       RSS    AIC  F value    Pr(>F)
<none>                 187854627 724.45
olives     1    302009 188156636 722.52   0.0675  0.796249
fruit      1 472362181 660216808 781.52 105.6094 4.937e-13 ***
smallarea  2  61839601 249694228 733.82   6.9130  0.002539 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> library(car)
> Anova(modelD)

Anova Table (Type II tests)

Response: observed
             Sum Sq Df  F value    Pr(>F)
olives       302009  1   0.0675  0.796249
fruit     472362181  1 105.6094 4.937e-13 ***
smallarea  61839601  2   6.9130  0.002539 **
Residuals 187854627 42
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> modelD <- update(modelD, .~. - olives)
> Anova(modelD)

Anova Table (Type II tests)

Response: observed
             Sum Sq Df F value    Pr(>F)
fruit     533135095  1 121.839 3.961e-14 ***
smallarea  95886674  2  10.957 0.0001427 ***
Residuals 188156636 43
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The variable `olives` is not significant and is removed from the model.

(ii)

```
> modelD <- lm(observed ~ fruit + smallarea, data = SATFRUIT)
> modelDg <- glm(observed ~ fruit + smallarea, data = SATFRUIT)
> library(boot)
> cv.errorN <- cv.glm(SATFRUIT, modelDg)
> CVNd <- cv.errorN$delta[1]
> CVNd

[1] 4663010

> set.seed(5)
> cv.error5 <- cv.glm(SATFRUIT, modelDg, K=5)
> CV5d <- cv.error5$delta[1]
> CV5d

[1] 4947248
```

The $CV_n = 4663009.5389$ for model (D), and $CV_5 = 4947248.4954$ for model (D).

(iii)

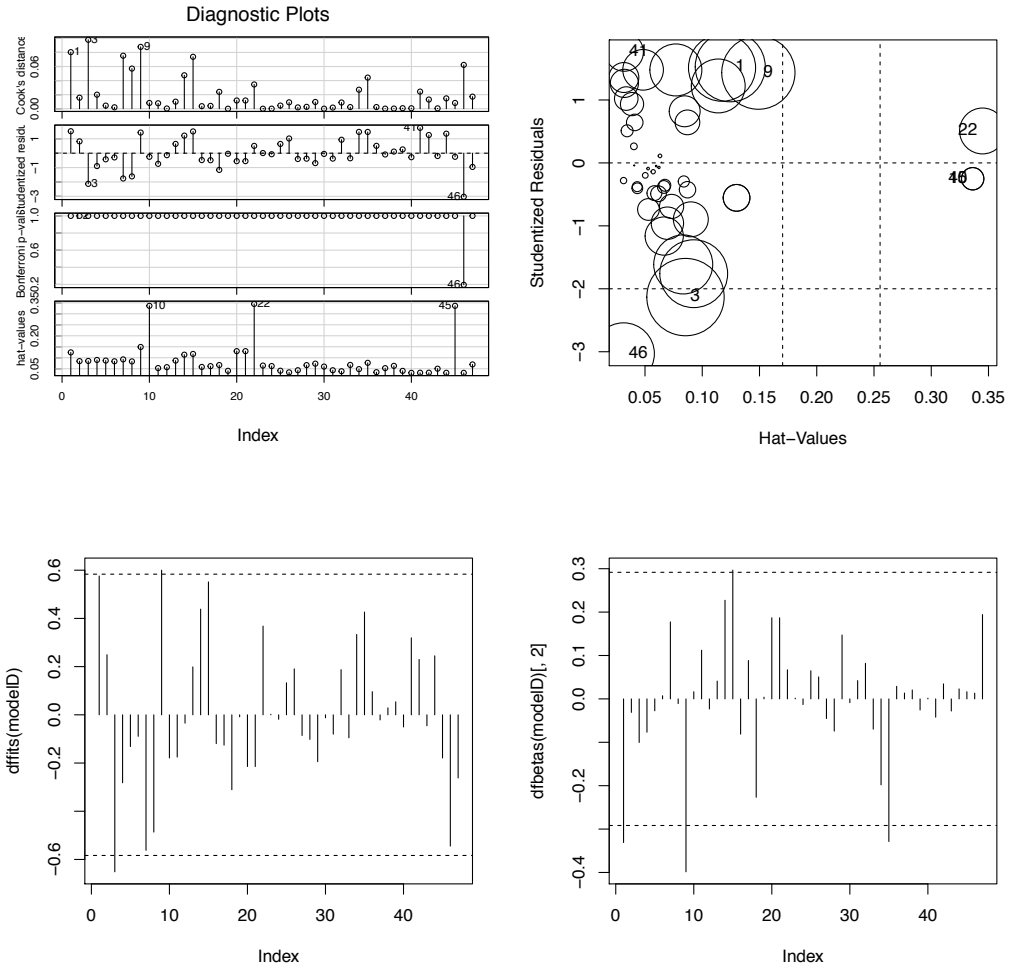
```
> MGOF <- mgof(model = modelD, data = SATFRUIT)
> MGOF

         R2       R2.adj          AIC          BIC           SE
  0.7815689    0.7663295  857.9041960  867.1549340 2091.8259301
```

The R^2, R_a^2, AIC, BIC, and standard error for `modelC` are 0.7816, 0.7663, 857.9042, 867.1549, and 2091.8259, respectively. The total proportion of variability explained by `modelD` is 0.7816.

(iv) Yes, the cross validation error for model (D) is smaller than the cross validation error for both models (A) and (C). The CV_n = 9131967.681 for model (A), and CV_5 = 10000799.7019 for model (A). The CV_n = 7045109.8173 for model (C), and CV_5 = 8505431.931 for model (C). The CV_n = 4663009.5389 for model (D), and CV_5 = 4947248.4954 for model (D).

(v) The first three requested graphs are computed with `influenceIndexPlot()` from the `car` package. The `influencePlot()` function is used to see which points might be influential. In this case, none of the observations are overly influential.

```
> influenceIndexPlot(modelD, id.n = 3)
> influencePlot(modelD, id.n = 3)

      StudRes        Hat      CookD
1   1.5275478 0.12444800 0.28358712
3  -2.1319340 0.08536298 0.31300494
9   1.4342951 0.14901774 0.29647838
10 -0.2515599 0.33619097 0.09050352
22  0.5075109 0.34476479 0.18567791
41  1.7624163 0.03180587 0.15594357
45 -0.2515024 0.33619143 0.09048297
46 -3.0306678 0.03126857 0.24953058

> CVdffits <- 2*sqrt(4/47) # 2*sqrt(p/n)
> plot(dffits(modelD), type = "h")
> abline(h = c(CVdffits, -CVdffits), lty = "dashed")
> CVdfbetas <- 2/sqrt(47)  # 2/sqrt(n)
> plot(dfbetas(modelD)[, 2], type = "h")
> abline(h = c(CVdfbetas, -CVdfbetas), lty = "dashed")
```



(vi)

```
> hcv <- 2*4/47  # 2*p/n
> hcv            # hii CV

[1] 0.1702128

> which(hatvalues(modelD) > hcv)

10 22 45
10 22 45

> outlierTest(modelD)


No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
46 -3.030668          0.0041662      0.19581
```

There are three observations (10, 22, and 45) with a leverage value that exceeds 0.1702.

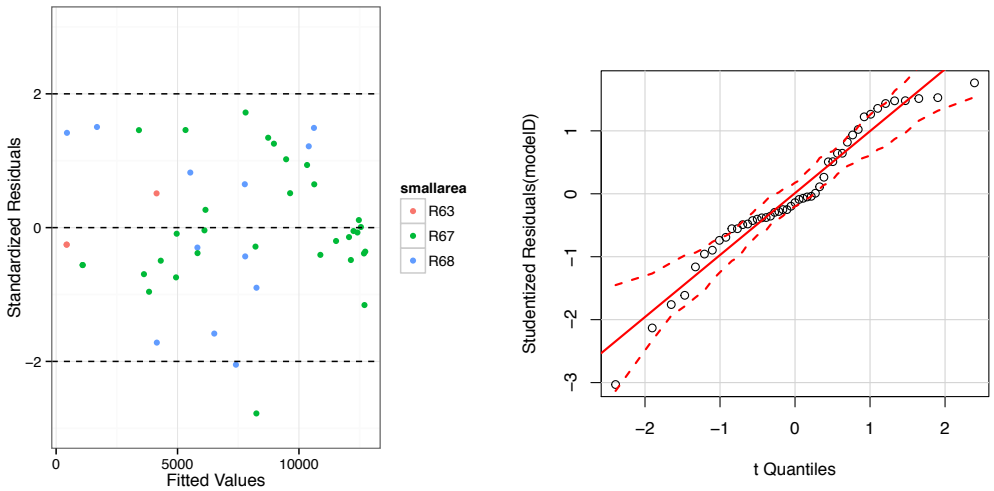
(vii)

```
> outlierTest(modelD)


No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
    rstudent unadjusted p-value Bonferonni p
46 -3.030668          0.0041662      0.19581
```

There are no outliers according to a Bonferroni test.

(viii)

```
> DF <- fortify(modelD)
> ggplot(data = DF, aes(x = .fitted, y = .stdresid, color = smallarea)) +
+   geom_point() +
+   ylim(-3, 3) +
+   geom_hline(yintercept = c(0, 2, -2), lty = "dashed") +
+   labs(x = "Fitted Values", y = "Standardized Residuals") +
+   theme_bw()
> set.seed(7)
> qqPlot(modelD)
> shapiro.test(resid(modelD))


Shapiro-Wilk normality test

data:  resid(modelD)
W = 0.95681, p-value = 0.08035
```



The standardized residuals appear to be equally scattered between -2 and 2 when plotted against the fitted values. The q-q plot of the residuals for `modelD` against the quantiles for a t distribution does not reveal any serious problems, and the Shapiro Wilk normality test has a $\wp$-value above 0.05.

(ix)

```
> CI <- confint(modelD)
> CI

                     2.5 %      97.5 %
(Intercept)  -2012.6283481 2879.394461
fruit            0.7542251    1.091433
smallareaR67 -2116.2680000 3432.813298
smallareaR68 -5746.3330265  334.455286
```

The 95% confidence interval for the `fruit` coefficient is [0.7542, 1.0914].

(h)

```
> coef(summary(modelD))

                  Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)    433.3830564 1.212883e+03  0.3573165 7.226027e-01
fruit            0.9228291 8.360421e-02 11.0380687 3.960570e-14
smallareaR67   658.2726489 1.375788e+03  0.4784696 6.347399e-01
smallareaR68 -2705.9388703 1.507614e+03 -1.7948481 7.970919e-02

> IOF <- coef(summary(modelD))[2, 1]*10000
> IOF

[1] 9228.291
```

Holding all other quantities in the model constant, an increase in `fruit` of $10,000$ m^2 would increase the expected observed fruits by 9228.2905 m^2.

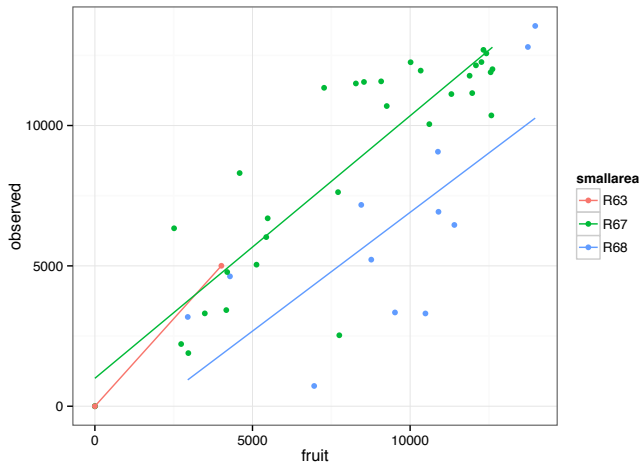
(i)

```
> newdata <- data.frame(smallarea = c("R63", "R67", "R68"),
+                       fruit = c(97044.28, 4878603.43, 2883488.24))
> PredictEstimate <- predict(modelD, newdata = newdata)
> PredictEstimate

       1          2          3
89988.66 4503208.64 2658694.17
```

The predicted area of fruit trees for small areas `R63`, `R67`, and `R68` are 89988.6641 m^2, 4503208.6404 m^2, and 2658694.1669 m^2.

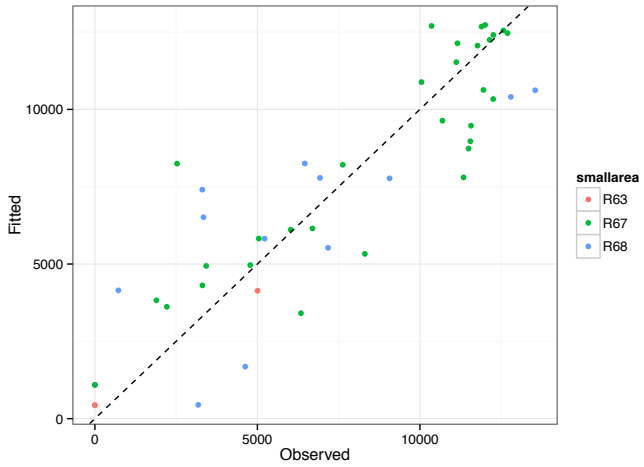
(j) The function `ggplot()` is used with the aesthetic `color = smallarea` to distinguish the small areas in the plot. The regression lines are nearly parallel suggesting a model with an identical slope but different intercepts for each small area may be reasonable.

```
> ggplot(data = SATFRUIT, aes(x = fruit, y = observed, color = smallarea)) +
+   geom_point() +
+   stat_smooth(method = "lm", se = FALSE) +
+   theme_bw()
```



(k)

```
> ggplot(data = DF, aes(x = observed, y = .fitted, color = smallarea)) +
+   geom_point()  +
+   theme_bw() +
+   geom_abline(intercept = 0, slope = 1, lty = "dashed") +
+   labs(x = "Observed", y = "Fitted")
```

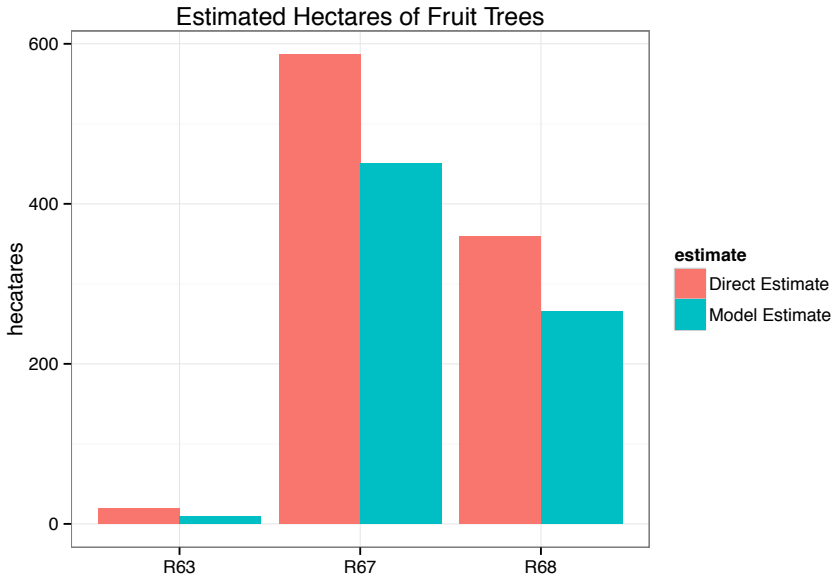


A straight line appears to model the relationship between fitted and observed values.

(l) Recall that the direct technique estimates the total surface area by multiplying the mean of the observed surface area in the sampled segments by the total number of segments in every small area. The direct and model estimates initially in m^2 are converted to hectares by dividing each estimate by 10,000.

```
> DirectEstimate <- tapply(SATFRUIT$observed,
+                          SATFRUIT$smallarea, mean) * c(119, 703, 564)
> DirectEstimate

      R63       R67       R68
198466.5 5867470.0 3589159.5
```

```
> newdata <- data.frame(smallarea = c("R63", "R67", "R68"),
+                       fruit = c(97044.28, 4878603.43, 2883488.24))
> PredictEstimate <- predict(modelD, newdata = newdata)
> DPestimates <- rbind(DirectEstimate, PredictEstimate)
> DPE <- data.frame(DPestimates/10000)
> DPE

                     R63      R67      R68
DirectEstimate  19.846650 586.7470 358.9159
PredictEstimate  8.998866 450.3209 265.8694

> DF <- stack(DPE)
> DF$estimate <- rep(c("Direct Estimate", "Model Estimate"),
+                    times = 3)
> DF

      values ind        estimate
1  19.846650 R63 Direct Estimate
2   8.998866 R63  Model Estimate
3 586.746996 R67 Direct Estimate
4 450.320864 R67  Model Estimate
5 358.915945 R68 Direct Estimate
6 265.869417 R68  Model Estimate

> ggplot(data = DF, aes(x = ind, y = values, fill = estimate)) +
+   geom_bar(stat="identity", position = "dodge") +
+   labs(x = "", y = "hecatares",
+        title = "Estimated Hectares of Fruit Trees") +
+   theme_bw()
```



The predictions from the small area model for all small areas are less than the direct estimate predictions.