Chapter 11: Statistics with R - 2nd Edition

Robert Stinerock

Student Exercises

The csv data sets used in these exercises can be found on the website:

1. `holidays.csv`
2. `temps.csv`

1. Two independent random samples are drawn from two populations. For the first sample: $n_1 = 36$, $\overline{x}_1 = 26$, and $\sigma_1 = 3.25$; for the second sample: $n_2 = 31$, $\overline{x}_2 = 23$, and $\sigma_2 = 2.75$. For parts b, c, and d, use the form: $(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$.

   (a) What is the point estimate of the difference between the two population means?

   Answer:

   $$(\overline{x}_1 - \overline{x}_2) = 26 - 23 = 3$$

   (b) What is the 99% confidence interval estimate of the difference between the two population means?

   $$3 \pm (2.576)\sqrt{(3.25)^2/36 + (2.75)^2/31}$$

```
qnorm(0.995)
```

```
## [1] 2.575829
```

```
# The margin of error: moe

moe <- qnorm(0.995) * sqrt((3.25) ^ 2 / 36 + (2.75) ^ 2 / 31)


# What is the  margin of error?
```

```
moe
```

```
## [1] 1.888198
```

```
# The upper bound of the confidence interval.
```

```
(26 - 23) + moe
```

```
## [1] 4.888198
```

```
# The lower bound of the confidence interval.
```

```
(26 - 23) - moe
```

```
## [1] 1.111802
```

Answer:

$$3 \pm 1.89$$

$$[1.11, 4.89]$$

(c) What is the 95% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$3 \pm (1.96)\sqrt{(3.25)^2/36 + (2.75)^2/31}$$

```
qnorm(0.975)
```

```
## [1] 1.959964
```

```
# The margin of error.

moe <- qnorm(0.975) * sqrt((3.25) ^ 2 / 36 + (2.75) ^ 2 / 31)


# What is the  margin of error?

moe
```

```
## [1] 1.436741

# The upper bound of the confidence interval.

(26 - 23) + moe

## [1] 4.436741

# The lower bound of the confidence interval.

(26 - 23) - moe

## [1] 1.563259
```

Answer:

$$3 \pm 1.44$$

$$[1.56, 4.44]$$

(d) What is the 90% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$3 \pm (1.645)\sqrt{(3.25)^2/36 + (2.75)^2/31}$$

```
qnorm(0.95)

## [1] 1.644854
```

```
# The margin of error.

moe <- qnorm(0.95) * sqrt((3.25) ^ 2 / 36 + (2.75) ^ 2 / 31)


# What is the  margin of error?

moe
```

```
## [1] 1.205751

# The upper bound of the confidence interval.

(26 - 23) + moe

## [1] 4.205751

# The lower bound of the confidence interval.

(26 - 23) - moe

## [1] 1.794249
```

Answer:

$$3 \pm 1.21$$

$$[1.79, 4.21]$$

2. Two independent random samples are drawn from two populations. For the first sample: $n_1 = 54$, $\overline{x}_1 = -3$, and $\sigma_1 = 5.70$; for the second sample: $n_2 = 47$, $\overline{x}_2 = -7$, and $\sigma_2 = 5.10$. For parts b, c, and d, use the form: $(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$.

   (a) What is the point estimate of the difference between the two population means?

   Answer:

$$(\overline{x}_1 - \overline{x}_2) = -3 - (-7) = -3 + 7 = 4$$

   (b) What is the 99% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$4 \pm (2.576)\sqrt{(5.70)^2/54 + (5.10)^2/47}$$

```r
qnorm(0.995)
```

```
## [1] 2.575829
```

```r
# The margin of error.

moe <- qnorm(0.995) * sqrt((5.70) ^ 2 / 54 + (5.10) ^ 2 / 47)


# What is the  margin of error?

moe
```

```
## [1] 2.768353
```

```r
# The upper bound of the confidence interval.

(-3 - (- 7)) + moe
```

```
## [1] 6.768353
```

```r
# The lower bound of the confidence interval.

(-3 - (-7)) - moe
```

```
## [1] 1.231647
```

Answer:

$$4 \pm 2.77$$

$$[1.23, 6.77]$$

(c) What is the 95% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$4 \pm (1.96)\sqrt{(5.70)^2/54 + (5.10)^2/47}$$

```r
qnorm(0.975)
```

```
## [1] 1.959964
```

```r
# The margin of error.

moe <- qnorm(0.975) * sqrt((5.70) ^ 2 / 54 + (5.10) ^ 2 / 47)


# What is the  margin of error?

moe
```

```
## [1] 2.106456
```

```r
# The upper bound of the confidence interval.

(-3 - (- 7)) + moe
```

```
## [1] 6.106456
```

```r
# The lower bound of the confidence interval.

(-3 - (-7)) - moe
```

```
## [1] 1.893544
```

Answer:

$$4 \pm 2.11$$

$$[1.89, 6.11]$$

(d) What is the 90% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$4 \pm (1.645)\sqrt{(5.70)^2/54 + (5.10)^2/47}$$

```r
qnorm(0.95)
```

```
## [1] 1.644854
```

```r
# The margin of error.

moe <- qnorm(0.95) * sqrt((5.70) ^ 2 / 54 + (5.10) ^ 2 / 47)


# What is the  margin of error?

moe
```

```
## [1] 1.767794
```

```r
# The upper bound of the confidence interval.

(-3 - (- 7)) + moe
```

```
## [1] 5.767794
```

```r
# The lower bound of the confidence interval.

(-3 - (-7)) - moe
```

```
## [1] 2.232206
```

Answer:

$$4 \pm 1.77$$

$$[2.23, 5.77]$$

3. Two independent random samples are drawn from two populations. For the first sample: $n_1 = 107$, $\overline{x}_1 = 3$, and $\sigma_1 = 8.90$; for the second sample: $n_2 = 121$, $\overline{x}_2 = -2$, and $\sigma_2 = 9.80$.

   (a) Please provide the point estimate of the difference between the two population means.

   Answer:

$$(\overline{x}_1 - \overline{x}_2) = 3 - (-2) = 3 + 2 = 5$$

(b) What is the 99% confidence interval estimate of the difference between the two population means?

$$(\overline{x}_1 - \overline{x}_2) \pm Z_{\alpha/2}\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}$$

$$5 \pm (2.576)\sqrt{(8.90)^2/107 + (9.80)^2/121}$$

```
qnorm(0.995)
```

```
## [1] 2.575829
```

```
# The margin of error.

moe <- qnorm(0.995) * sqrt((8.90) ^ 2 / 107 + (9.80) ^ 2 / 121)


# What is the  margin of error?

moe
```

```
## [1] 3.190286
```

```
# The upper bound of the confidence interval.

(3 - (- 2)) + moe
```

```
## [1] 8.190286
```

```
# The lower bound of the confidence interval.

(3 - (-2)) - moe
```

```
## [1] 1.809714
```

Answer:

$$5 \pm 3.19$$

$$[1.81, 8.19]$$

(c) What is the 95% confidence interval estimate of the difference between the two population means?

```
qnorm(0.975)
```

```
## [1] 1.959964
```

```
# The margin of error.

moe <- qnorm(0.975) * sqrt((8.90) ^ 2 / 107 + (9.80) ^ 2 / 121)


# What is the  margin of error?

moe
```

```
## [1] 2.427508
```

```
# The upper bound of the confidence interval.

(3 - (- 2)) + moe
```

```
## [1] 7.427508
```

```
# The lower bound of the confidence interval.

(3 - (-2)) - moe
```

```
## [1] 2.572492
```

Answer:

$$5 \pm 2.43$$

$$[2.57, 7.43]$$

(d) What is the 90% confidence interval estimate of the difference between the two population means?

```r
qnorm(0.95)
```

```
## [1] 1.644854
```

```r
# The margin of error.

moe <- qnorm(0.95) * sqrt((8.90) ^ 2 / 107 + (9.80) ^ 2 / 121)


# What is the  margin of error?

moe
```

```
## [1] 2.037229
```

```r
# The upper bound of the confidence interval.

(3 - (- 2)) + moe
```

```
## [1] 7.037229
```

```r
# The lower bound of the confidence interval.

(3 - (-2)) - moe
```

```
## [1] 2.962771
```

Answer:

$$5 \pm 2.04$$

$$[2.96, 7.04]$$

4. Two independent random samples have been selected from two populations for which the population standard deviations are unknown. For the first sample, $n_1 = 24$, $\overline{x}_1 = 717$, and $s_1 = 81$; for the second sample, $n_2 = 31$, $\overline{x}_2 = 658$, and $s_2 = 77$. Find the 90% confidence interval estimate of the difference between two population means, and remember to use the $n_1 + n_2 - 2$ expression for the degrees of freedom.

```r
qt(0.95, 24 + 31 - 2)
```

```
## [1] 1.674116
```

```r
# The margin of error.

moe <- qt(0.95, 53) * sqrt((81) ^ 2 / 24 + (77) ^ 2 / 31)


# What is the  margin of error?

moe

## [1] 36.08616

# The upper bound of the confidence interval.

(717 - 658) + moe

## [1] 95.08616

# The lower bound of the confidence interval.

(717 - 658) - moe

## [1] 22.91384
```

Answer:

$$59 \pm 36.09$$

$$[22.91, 95.09]$$

5. Two independent random samples have been selected from two populations for which the population standard deviations are unknown. For the first sample, $n_1 = 34$, $\overline{x}_1 = -124$, and $s_1 = 14$; for the second sample, $n_2 = 28$, $\overline{x}_2 = -132$, and $s_2 = 12$. Find the 95% confidence interval estimate of the difference between two population means, and use the $n_1 + n_2 - 2$ expression for the degrees of freedom.

```r
qt(0.975, 34 + 28 - 2)

## [1] 2.000298
```

```r
# The margin of error.
```

```r
moe <- qt(0.975, 60) * sqrt((14) ^ 2 / 34 + (12) ^ 2 / 28)


# What is the  margin of error?

moe

## [1] 6.606304

# The upper bound of the confidence interval.

(-124 - (-132)) + moe

## [1] 14.6063

# The lower bound of the confidence interval.

(-124 - (-132)) - moe

## [1] 1.393696
```

Answer:

$$8 \pm 6.61$$

$$[1.39, 14.61]$$

6. Two independent random samples have been selected from two populations for which the population standard deviations are unknown. For the first sample, $n_1 = 27$, $\overline{x}_1 = 2$, and $s_1 = 3.70$; for the second sample, $n_2 = 22$, $\overline{x}_2 = -3$, and $s_2 = 2.45$. Find the 99% confidence interval estimate of the difference between two population means, and use the $n_1 + n_2 - 2$ expression for the degrees of freedom.

```r
qt(0.995, 27 + 22 - 2)

## [1] 2.684556
```

```r
# The margin of error.

moe <- qt(0.995, 47) * sqrt((3.70) ^ 2 / 27 + (2.45) ^ 2 / 22)
```

```r
# What is the  margin of error?

moe

## [1] 2.37075

# The upper bound of the confidence interval.

(2 - (-3)) + moe

## [1] 7.37075

# The lower bound of the confidence interval.

(2 - (-3)) - moe

## [1] 2.62925
```

Answer:

$$5 \pm 2.371$$

$$[2.63, 7.37]$$

7. A recent study of the cost of living in various U.S. cities has found regional differences in home prices. Two of the cities considered are Dallas, Texas and Minneapolis, Minnesota. The following data have been collected in the studies for those two cities: a sample of 47 homes in Dallas are for sale for an average of $\$151,800$, with a standard deviation of $\$17,457$; a sample of 34 comparable homes are for sale in Minneapolis for an average of $\$207,100$, with a standard deviation of $\$26,510$.

   (a) Are these data paired or independent? Why?

   Answer: Since the Dallas data are collected independently of the Minneapolis data, they are independent, not paired.

   (b) Find the 95% confidence interval estimate of the difference between the mean price for a home in Minneapolis and a comparable home in Dallas.

```r
qt(0.975, 47 + 34 - 2)

## [1] 1.99045
```

```r
# The margin of error.

moe <- qt(0.975, 79) * sqrt((26510) ^ 2 / 34 + (17457) ^ 2 / 47)


# What is the  margin of error?

moe

## [1] 10372.13

# The upper bound of the confidence interval.

(207100 - 151800) + moe

## [1] 65672.13

# The lower bound of the confidence interval.

(207100 - 151800) - moe

## [1] 44927.87
```

Answer: There is a 0.95 probability that the difference between the mean price of a home in Minneapolis and the mean price of a home in Dallas falls in the interval from \$44,928 to \$65,672. Home prices in Minneapolis are higher than those in Dallas.

$$\$55,300 \pm \$10,372$$

$$[\$44,928,\ \$65,672]$$

8. Two independent simple random samples are drawn from two populations. For the first sample: $n_1 = 63$, $\overline{x}_1 = 73$, and $\sigma_1 = 11.1$; for the second sample: $n_2 = 68$, $\overline{x}_2 = 76$, and $\sigma_2 = 4.4$. Test the hypothesis of no difference between the population means at the $\alpha = 0.05$ level of significance. Use the six-step hypothesis-testing framework.

   (a) **Develop the null hypothesis H_0 in statistical terms.**

$$H_0 : \mu_1 - \mu_2 = 0$$

(b) **Develop the alternative hypothesis H_a in statistical terms.**

$$H_a : \mu_1 - \mu_2 \neq 0$$

(c) **Set the level of significance α, and decide on the sample sizes n_1 and n_2.**

$$n_1 = 63 \text{ and } n_2 = 68$$

$$\alpha = 0.05$$

(d) **Use α to specify the rejection region RR.**

We specify the rejection region in terms of z:

Reject H_0 if $z \geq z_{\alpha/2} = z_{0.025} = 1.96$ or $z \leq -z_{\alpha/2} = -z_{0.025} = -1.96$

```
qnorm(0.975)

## [1] 1.959964

qnorm(0.025)

## [1] -1.959964
```

that is, the rejection region is

$$RR : z \geq 1.96 \text{ and } z \leq -1.96$$

where

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

where δ_0 is the hypothesized difference between μ_1 and μ_2.

(e) **Collect the data and calculate the test statistic.**

Since $n_1 = 63$, $\overline{x}_1 = 73$, $\sigma_1 = 11.1$, $n_2 = 68$, $\overline{x}_2 = 76$, $\sigma_2 = 4.4$, $\delta_0 = 0$, then

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

$$z = \frac{(73 - 76) - 0}{\sqrt{(11.1)^2/63 + (4.4)^2/68}} = -2.00$$

(f) **Use the value of the test statistic and the rejection region RR to decide whether to reject H_0. If the test statistic falls in the rejection region, reject H_0; if the test statistic falls outside the rejection region, do not reject H_0.**

Since $z = -2.00 < -1.96$, we reject H_0.

9. In the case of the previous exercise, what is the p-value? What is the conclusion based on the p-value?

The p-value $= 2(p(z \leq -2.00)) = 2(0.0228) = 0.0455$.

```
2 * pnorm(-2)

## [1] 0.04550026
```

Answer: Since $\alpha = 0.05$ and the p-value $= 0.0455 < 0.05$, we reject H_0.

10. Two independent simple random samples are drawn from two populations. For the first sample: $n_1 = 81$, $\overline{x}_1 = 17.1$, and $\sigma_1 = 2.3$; for the second sample: $n_2 = 64$, $\overline{x}_2 = 17.6$, and $\sigma_2 = 2.1$. Test $H_0 : \mu_1 - \mu_2 \geq 0$ at the $\alpha = 0.05$ level of significance. Use the six-step hypothesis-testing framework.

(a) **Develop the null hypothesis H_0 in statistical terms.**

$$H_0 : \mu_1 - \mu_2 \geq 0$$

(b) **Develop the alternative hypothesis H_a in statistical terms.**

$$H_a : \mu_1 - \mu_2 < 0$$

(c) **Set the level of significance α, and decide on the sample sizes n_1 and n_2.**

$$n_1 = 81 \text{ and } n_2 = 64$$

$$\alpha = 0.05$$

(d) **Use α to specify the rejection region RR.**

We specify the rejection region in terms of z:

Reject H_0 if $z < -z_\alpha = -z_{0.05} = -1.645$

```
qnorm(0.05)
```

```
## [1] -1.644854
```

that is, the rejection region is

$$RR : z < -1.645$$

where

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

and where δ_0 is the hypothesized difference between μ_1 and μ_2.

(e) **Collect the data and calculate the test statistic.**

Since $n_1 = 81$, $\overline{x}_1 = 17.1$, $\sigma_1 = 2.3$, $n_2 = 64$, $\overline{x}_2 = 17.6$, $\sigma_2 = 2.1$, $\delta_0 = 0$, then

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

$$z = \frac{(17.1 - 17.6) - 0}{\sqrt{(2.3)^2/81 + (2.1)^2/64}} = -1.36$$

(f) **Use the value of the test statistic and the rejection region RR to decide whether to reject H_0. If the test statistic falls in the rejection region, reject H_0; if the test statistic falls outside the rejection region, do not reject H_0.**

Since $z = -1.36 > -1.645$, we do not reject H_0.

11. In the case of the preceding exercise, what is the p-value? What is the conclusion based on the p-value?

The p-value $= p(z \leq -1.36) = 0.0869$.

```
pnorm(-1.36)

## [1] 0.08691496
```

Answer: Since $\alpha = 0.05$ and the p-value $= 0.0869 > 0.05$, we do not reject H_0.

12. Two independent simple random samples are drawn from two populations. For the first sample: $n_1 = 100$, $\overline{x}_1 = 120$, and $\sigma_1 = 20$; for the second sample: $n_2 = 100$, $\overline{x}_2 = 112$, and $\sigma_2 = 20$. Test $H_0 : \mu_1 - \mu_2 \leq 0$ at the $\alpha = 0.01$ level of significance. Use the six-step hypothesis-testing framework.

(a) **Develop the null hypothesis H_0 in statistical terms.**

$$H_0 : \mu_1 - \mu_2 \leq 0$$

(b) **Develop the alternative hypothesis H_a in statistical terms.**

$$H_a : \mu_1 - \mu_2 > 0$$

(c) **Set the level of significance α, and decide on the sample sizes n_1 and n_2.**

$$n_1 = 100 \text{ and } n_2 = 100$$

$$\alpha = 0.01$$

(d) **Use α to specify the rejection region RR.**

We specify the rejection region in terms of z:

Reject H_0 if $z \geq z_\alpha = z_{0.01} = 2.33$

```
qnorm(0.01, lower.tail = FALSE)
```

```
## [1] 2.326348
```

that is, the rejection region is

$$RR : z \geq 2.33$$

where

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

and where δ_0 is the hypothesized difference between μ_1 and μ_2.

(e) **Collect the data and calculate the test statistic.**

Since $n_1 = 100$, $\overline{x}_1 = 120$, $\sigma_1 = 20$, $n_2 = 100$, $\overline{x}_2 = 112$, $\sigma_2 = 20$, $\delta_0 = 0$, then

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{{\sigma_1}^2/n_1 + {\sigma_2}^2/n_2}}$$

$$z = \frac{(120 - 112) - 0}{\sqrt{(20)^2/100 + (20)^2/100}} = 2.83$$

(f) **Use the value of the test statistic and the rejection region RR to decide whether to reject H_0. If the test statistic falls in the rejection region, reject H_0; if the test statistic falls outside the rejection region, do not reject H_0.**

Since $z = 2.83 > 2.33$, we reject H_0.

13. In the case of the previous exercise, what is the p-value? What is the conclusion based on the p-value?

The p-value $= p(z \geq 2.83) = 0.0023$.

```
pnorm(2.83, lower.tail = FALSE)

## [1] 0.0023274
```

Answer: Since $\alpha = 0.01$ and the p-value $= 0.0023 < 0.01$, we reject H_0.

14. A statistics instructor is concerned that after her students perform well on the first of two major examinations in the introductory-level class, their performance appears to drop off on the second. Since this pattern appears to repeat itself across many sections of the same statistics class at her university, she wants to confirm that the downward trend in performance on the two 100-point examinations is real. To this end, she collects the examination results from a random sample of $n = 12$ students from the previous academic year. The scores on Exam 1 are: 79, 92, 81, 80, 79, 80, 78, 88, 86, 88, 77, and 93. On Exam 2, they are: 80, 75, 67, 82, 76, 71, 78, 78, 80, 77, 78,and 75. Create a data frame that organizes this data into two variables and twelve observations and use R to answer questions (b) and (c).

(a) Are these data independent or paired? Why?

Answer: Since there are two measurements on each of the twelve students, these data are paired.

(b) What is the point estimate of the difference between the two population means, $\mu_1 - \mu_2$?

Answer: 7

```
ex1 <- c(79, 92, 81, 80, 79, 80, 78, 88, 86, 88, 77, 93)
ex2 <- c(80, 75, 67, 82, 76, 71, 78, 78, 80, 77, 78, 75)
scores <- data.frame(Exam1 = ex1, Exam2 = ex2)
scores
```

```
##    Exam1 Exam2
## 1     79    80
## 2     92    75
## 3     81    67
## 4     80    82
## 5     79    76
## 6     80    71
## 7     78    78
## 8     88    78
## 9     86    80
## 10    88    77
## 11    77    78
## 12    93    75
```

```
mean(scores$Exam1)
```

```
## [1] 83.41667
```

```
mean(scores$Exam2)
```

```
## [1] 76.41667
```

$$\overline{x}_1 - \overline{x}_2 = 83.42 - 76.42 = 7$$

```
mean(scores$Exam1) - mean(scores$Exam2)
```

```
## [1] 7
```

(c) What is the 99% confidence interval estimate of the difference between the two population means, $\mu_1 - \mu_2$?

Answer: There is a 0.99 probability that the difference between the mean performance on the two examinations falls in the interval from 0.5235 to 13.4765.

That is, the 99% confidence interval estimate of the difference in means is $[0.5235, 13.4765]$. The mean score on the first exam is higher than the mean score on the second exam.

```
t.test(scores$Exam1, scores$Exam2, conf.level = 0.99,
       paired = TRUE)

##
##  Paired t-test
##
## data:  scores$Exam1 and scores$Exam2
## t = 3.3568, df = 11, p-value = 0.0064
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##   0.5234552 13.4765448
## sample estimates:
## mean of the differences
##                       7
```

15. Using the test score data from the previous question, test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.10$ level of significance.

```
t.test(scores$Exam1, scores$Exam2, paired = TRUE)


##
##  Paired t-test
##
## data:  scores$Exam1 and scores$Exam2
## t = 3.3568, df = 11, p-value = 0.0064
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.410281 11.589719
## sample estimates:
## mean of the differences
##                       7
```

Answer: Since p-value=0.0064$<\alpha = 0.10$, we reject H_0 that there is no difference between student performance on the two examinations.

Note: for a hypothesis test, unlike the case of the confidence interval estimate, we do not need to include the `conf.level=` argument in the `t.test()` function.

16. Using the `holidays.csv` data (see the website), answer the following questions. For a description of this data set, see Section 11.1 in the textbook.

(a) Are these data paired or independent? Why?

Answer: Since these data are collected from two independent populations, they are independent, not paired.

(b) Find the 90% confidence interval estimate of the difference between the mean monthly apartment rent in Cascais and in the Algarve.

Answer: There is a 0.90 probability that the difference between the mean apartment rents in the Algarve and Cascais falls in the interval from €39.32 to €302.75. That is, the 90% confidence interval estimate of the difference in means is $[39.32, 302.75]$. The mean monthly rent in the Algarve is higher than the mean monthly rent in Cascais.

```
holidays <- read.csv('holidays.csv') # Import the data set.
```

```
names(holidays) # Identify the variable names.

## [1] "algarve" "cascais"
```

```
t.test(holidays$algarve, holidays$cascais, conf.level = 0.90)

##
##  Welch Two Sample t-test
##
## data:  holidays$algarve and holidays$cascais
## t = 2.1751, df = 51.413, p-value = 0.03425
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##   39.3218 302.7545
## sample estimates:
## mean of x mean of y
##  2030.853  1859.815
```

Note: since these data are collected from two independent populations, we do not include the `paired = TRUE` argument.

17. This exercise uses the `Cars93` data from the `MASS` package. For starters, read the `Cars93` data into `E11_1`.

(a) What are the variable names? How many observations are included? Find (1) the minimum and maximum values, (2) the median and mean, (3) the first and third quartiles, and (4) the standard deviation of the two variables, `MPG.city` and `MPG.highway`. Also, find the distribution of vehicles from non-USA countries and from the USA. Comment on the initial findings.

```
library(MASS) # Load MASS package; it includes the Cars93 data.
```

```r
# Read the Cars93 data into E11_1.

E11_1 <- Cars93


# Use nrow() function to find number of observations.

nrow(E11_1) # Has 93 observations.
```

```
## [1] 93
```

```r
# Use the names() function to list variable names.

names(E11_1) # Has 27 variables.
```

```
##  [1] "Manufacturer"       "Model"              "Type"
##  [4] "Min.Price"          "Price"              "Max.Price"
##  [7] "MPG.city"           "MPG.highway"        "AirBags"
## [10] "DriveTrain"         "Cylinders"          "EngineSize"
## [13] "Horsepower"         "RPM"                "Rev.per.mile"
## [16] "Man.trans.avail"    "Fuel.tank.capacity" "Passengers"
## [19] "Length"             "Wheelbase"          "Width"
## [22] "Turn.circle"        "Rear.seat.room"     "Luggage.room"
## [25] "Weight"             "Origin"             "Make"
```

```r
# Use summary() function to explore MPG.city and MPG.highway.

summary(E11_1$MPG.city)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   18.00   21.00   22.37   25.00   46.00
```

```r
summary(E11_1$MPG.highway)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   26.00   28.00   29.09   31.00   50.00
```

```r
# Use sd() function for the standard deviation of each variable.

sd(E11_1$MPG.city)
```

```
## [1] 5.619812
```

```r
sd(E11_1$MPG.highway)
```

```
## [1] 5.331726
```

```r
# Use the table() function to find the distribution of
# vehicles from non-USA countries and from the USA.

table(E11_1$Origin)
```

```
##
##     USA non-USA
##      48      45
```

Answer: The `Cars93` data include 93 observations on 27 variables. The descriptive statistics for `MPG.city` and `MPG.highway` are provided above. Of the 93 vehicles in `Cars93`, 48 are from the USA, 45 are from non-USA countries.

(b) For the two variables `MPG.city` and `MPG.highway`, do you think the data are paired or independent? Why?

Answer: The data appear to be paired, not independent, since there are two measurements on each of the 93 vehicles.

(c) For practice structuring the data before analyzing it, we now subset `E11_1` by `Origin`. For this part of the exercise, create a new object from `E11_1` that includes only vehicles of USA origin (thus excluding all vehicles of non-USA origin) and name it `E11_2`. Check the contents of `E11_2` to make sure that it includes only vehicles of USA origin (there should be only 48 if the subsetting procedure worked correctly).

```
# Subset the data so that it includes only vehicles of
# USA origin.  Name the new object E11_2.

E11_2 <- E11_1[which(E11_1$Origin == 'USA'), ]


# Use the table() function to make sure E11_2 includes
# only USA vehicles and no non-USA vehicles.

table(E11_2$Origin)

##
##     USA non-USA
##      48       0
```

Answer: `E11_2` has 48 vehicles of USA origin; none is from a non-USA country.

(d) Among vehicles of USA-origin, what is the 90% confidence interval estimate of the difference between the mean miles per gallon (mpg) for city versus highway driving? Use the `E11_2` data.

Answer: For vehicles from the USA, there is a 0.90 probability that the difference between the mean mpg in the city and on the highway falls in the interval from 6.69 mpg to 7.68 mpg. That is, the 90% confidence interval estimate of the difference in means is $[6.69, 7.68]$. The mean mpg is higher for highway driving than for city driving.

```
t.test(E11_2$MPG.highway, E11_2$MPG.city, conf.level = 0.90,
       paired = TRUE)

##
```

```
##  Paired t-test
##
## data:  E11_2$MPG.highway and E11_2$MPG.city
## t = 24.428, df = 47, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  6.693794 7.681206
## sample estimates:
## mean of the differences
##                  7.1875
```

Note: since the data are paired, we must include the `paired = TRUE` argument in the `t.test()` function; we must also introduce the `conf.level = 0.90` argument if we want a 90% interval estimate of the difference of the means.

18. Using the `E11_2` data from the previous question, test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.05$ level of significance.

```
options(scipen = 999) # Suppress scientifc notation in output.


t.test(E11_2$MPG.highway, E11_2$MPG.city, paired = TRUE)


##
##  Paired t-test
##
## data:  E11_2$MPG.highway and E11_2$MPG.city
## t = 24.428, df = 47, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  6.595574 7.779426
## sample estimates:
## mean of the differences
##                  7.1875
```

Answer: Since p-value=0.00000000000000022$< \alpha = 0.05$, we reject the H_0 that there is no difference between the mean mpg for city versus highway driving.

19. For additional practice structuring the data before analyzing it, subset `E11_1` by `Origin` (once again) in a slightly different way. See part (a).

    (a) Create a new object from `E11_1` that includes only vehicles of non-USA origin (thus excluding all vehicles of USA origin) and name it `E11_3`. Check to make sure that `E11_3` includes only non-USA vehicles (there should be only 45).

```r
E11_3 <- E11_1[which(E11_1$Origin == 'non-USA'), ]

table(E11_3$Origin)

##
##     USA non-USA
##       0      45
```

Answer: There are 45 vehicles of non-USA origin; none is from the USA.

(b) What is the 90% confidence interval estimate of the difference between the mean miles per gallon (mpg) for city versus highway driving? Be sure to use the `E11_3` data?

Answer: For vehicles of non-USA origin, there is a 0.90 probability that the difference between the mean mpg in the city and on the highway falls in the interval from 5.85 mpg to 6.60 mpg. That is, the 90% confidence interval estimate of the difference in means is $[5.85, 6.60]$. The mean mpg is higher for highway driving than for city driving.

```r
t.test(E11_3$MPG.highway, E11_3$MPG.city, conf.level = 0.90,
       paired = TRUE)

##
##  Paired t-test
##
## data:  E11_3$MPG.highway and E11_3$MPG.city
## t = 27.718, df = 44, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  5.845038 6.599406
## sample estimates:
## mean of the differences
##                6.222222
```

20. Using the `E11_3` data from the previous question, test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.01$ level of significance.

```r
options(scipen = 999) # Suppress scientific notation.


t.test(E11_3$MPG.highway, E11_3$MPG.city, paired = TRUE)

##
##  Paired t-test
##
## data:  E11_3$MPG.highway and E11_3$MPG.city
```

```
## t = 27.718, df = 44, p-value < 0.00000000000000022
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.769806 6.674639
## sample estimates:
## mean of the differences
##                6.222222
```

Answer: Since p-value=0.00000000000000022$< \alpha = 0.01$, we reject the H_0 that there is no difference between the mean mpg for city versus highway driving.

21. Using the `temps.csv` data (on the website), answer the following questions. The data include the high and low temperatures (Celsius) for ten European cities.

```
temps <- read.csv('temps.csv') # Import temps.csv.
```

```
names(temps) # Identify variable names.

## [1] "City"      "Daytemp"   "Nighttemp"
```

```
temps # Print out full data set.

##          City Daytemp Nighttemp
## 1      Athens      21        12
## 2   Barcelona      12         9
## 3      Dublin       6         1
## 4      Lisbon      15         9
## 5  Luxembourg       3        -2
## 6      Moscow       2         1
## 7      Munich       4        -2
## 8      Naples      14        11
## 9       Paris       7        -1
## 10  Stockholm       2        -4
```

(a) Are these data paired or independent? Why?

Since there are two measurements on the same city, these data are paired.

(b) What is the 90% confidence interval estimate of $\mu_1 - \mu_2$?

Answer: There is a 0.90 probability that the difference between the mean daytime and nighttime temperatures (for these selected cities) falls in the interval from 3.81 to 6.59 degrees. That is, the 90% confidence interval estimate of the difference in daytime and nighttime temperatures is $[3.81, 6.59]$.

```
t.test(temps$Daytemp, temps$Nighttemp, conf.level = 0.90,
       paired = TRUE)

##
##  Paired t-test
##
## data:  temps$Daytemp and temps$Nighttemp
## t = 6.8675, df = 9, p-value = 0.00007328
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  3.811989 6.588011
## sample estimates:
## mean of the differences
##                     5.2
```

Note that for a 90% confidence interval estimate of the difference between the means of paired data, we must include both the `conf.level=0.90` and `paired = TRUE` arguments in the `t.test()` function.

22. Using the `temps` data, test $H_0 : \mu_1 - \mu_2 = 0$ against $H_a : \mu_1 - \mu_2 \neq 0$ at the $\alpha = 0.01$ level of significance.

```
t.test(temps$Daytemp, temps$Nighttemp, paired = TRUE)


##
##  Paired t-test
##
## data:  temps$Daytemp and temps$Nighttemp
## t = 6.8675, df = 9, p-value = 0.00007328
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  3.487122 6.912878
## sample estimates:
## mean of the differences
##                     5.2
```

Answer: Since p-value=0.00007328$< \alpha = 0.01$, we reject the H_0 that there is no difference between the daytime and nighttime temperatures. Unsurprisingly, the daytime temperatures are higher than the nighttime temperatures.

23. During two recent tax years, the Inland Revenue and Customs department (UK) conducted an in-house investigation of the accuracy of tax filling-advice given by agents to individuals who call with questions about how to handle various tax issues. During the first phase, conducted in 2020, calls were placed to a random sample of 900 agents for tax advice. After reviewing the accuracy of the advice provided, the investigation found that on 82 occasions the advice was incorrect. In a follow-up investigation in 2021, calls were placed to an independent (i.e., different) random

sample of 800 agents. On 28 occasions the advice was incorrect. Does it appear that the departmental effort to improve accuracy of the advice provided by their agents has been successful? Find the 95% confidence interval estimate of $p_1 - p_2$. (Note: $\overline{p}_1 = 82/900 = 0.09111$ and $\overline{p}_2 = 28/800 = 0.0350$)

Answer: There is a 0.95 probability that the difference between the proportion of agents providing incorrect tax advice two years ago and the proportion providing incorrect advice a year ago falls in the interval from 0.03340 to 0.07882. That is, the 95% confidence interval estimate of the difference in proportions (from two years ago to one year ago) is $[0.03340, 0.07882]$, and the data suggest the departmental effort to improve accuracy of the advice provided by their agents has been successful.

$$(\overline{p}_1 - \overline{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\overline{p}_1(1-\overline{p}_1)}{n_1} + \frac{\overline{p}_2(1-\overline{p}_2)}{n_2}}$$

$$(0.09111 - 0.0350) \pm 1.96\sqrt{\frac{0.09111(0.9089)}{900} + \frac{0.0350(0.9650)}{800}}$$

$$0.05611 \pm 0.0227$$

$$[0.03340, 0.07882]$$

```
bad <- c(82, 28)
total <- c(900, 800)
prop.test(bad, total, conf.level = 0.95, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  bad out of total
## X-squared = 22.034, df = 1, p-value = 0.000002679
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03340345 0.07881877
## sample estimates:
##     prop 1     prop 2
## 0.09111111 0.03500000
```

24. For the previous exercise, test $H_0 : p_1 - p_2 = 0$ against $H_a : p_1 - p_2 \neq 0$ at the $\alpha = 0.01$ level of significance.

```r
options(scipen = 999)

bad <- c(82, 28)
total <- c(900, 800)
prop.test(bad, total, correct = FALSE)

##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  bad out of total
## X-squared = 22.034, df = 1, p-value = 0.000002679
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03340345 0.07881877
## sample estimates:
##     prop 1     prop 2
## 0.09111111 0.03500000
```

Answer: Since p-value=0.000002679$<$ $\alpha = 0.01$, we reject the H_0 that there is no difference (from the first year to the second) between the proportion of agents providing inaccurate tax advice. There is a difference, and in fact the proportion providing inaccurate advice appears to have fallen from 2020 to 2021.

25. In a recent consumer confidence survey of 400 Portuguese adults, 54 of 200 men and 36 of 200 women expressed agreement with the statement, "I would have trouble paying an unexpected bill of €1,000 without borrowing from someone or selling something." Do men and women differ on their answer to this question? Use the six-step framework to test $H_0 : p_1 - p_2 = 0$ against $H_a : p_1 - p_2 \neq 0$ at the $\alpha = 0.05$ level of significance. What is the p-value?

(a) **Develop the null hypothesis in statistical terms.**

$$H_0 : p_1 - p_2 = 0$$

(b) **Develop the alternative hypothesis in statistical terms.**

$$H_a : p_1 - p_2 \neq 0$$

(c) **Set the level of significance α and decide on the sample size n.**

$$\alpha = 0.05$$

$$n_1 = 200 \text{ and } n_2 = 200$$

(d) **Use α to specify the rejection region RR.**

$$RR : z \geq 1.96 \text{ and } z \leq -1.96$$

```
qnorm(0.025)

## [1] -1.959964

qnorm(0.975)

## [1] 1.959964
```

where

$$z = \frac{(\overline{p}_1 - \overline{p}_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

and where

$$p = \frac{n_1\overline{p}_1 + n_2\overline{p}_2}{n_1 + n_2} = \frac{(200)(54/200) + (200)(36/200)}{200 + 200} = 0.2250$$

(e) **Collect the data and calculate the test statistic.**

$$z = \frac{(\overline{p}_1 - \overline{p}_2)}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.27 - 0.18}{\sqrt{0.2250(0.7750)\left(\frac{1}{200} + \frac{1}{200}\right)}} = 2.1553$$

and

p-value=$p(z>2.1553) + p(z< -2.1553) = 0.03114$

```
pnorm(2.1553, lower.tail = FALSE) + pnorm(-2.1553)

## [1] 0.03113837
```

(f) **Use the test statistic and RR to decide whether to reject H_0.**

Recall that the rejection region is $RR : z \geq 1.96$ and $z \leq -1.96$. Since $z = 2.1553 > 1.96$, we reject $H_0 : p_1 - p_2 = 0$. Moreover, since p-value$= 0.03114 < \alpha = 0.05$, we reject H_0.

26. For the previous exercise, use the `prop.test()` function to test $H_0 : p_1 - p_2 = 0$ against $H_a : p_1 - p_2 \neq 0$ at the $\alpha = 0.05$ level of significance. Is the p-value the same as it is in the previous exercise?

```
illiquid <- c(54, 36)
total <- c(200, 200)
prop.test(illiquid, total, correct = FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  illiquid out of total
## X-squared = 4.6452, df = 1, p-value = 0.03114
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.008631982 0.171368018
## sample estimates:
## prop 1 prop 2
##   0.27   0.18
```

Answer: The p-value of 0.03114 is the same using the `prop.test()` function as it was using the six-step framework in the previous exercise.