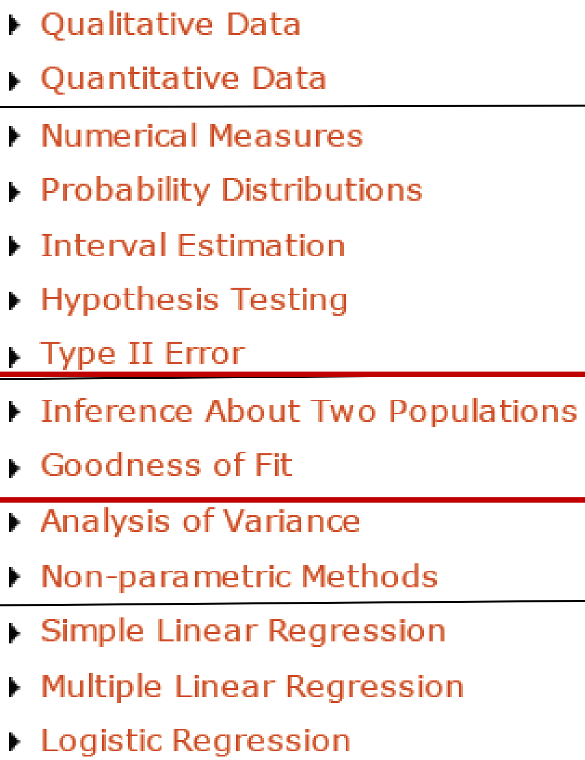
Primjena računara u biologiji

Vladimir Filipović

vladaf@matf.bg.ac.rs

Inference About Two Populations and Goodness of Fit

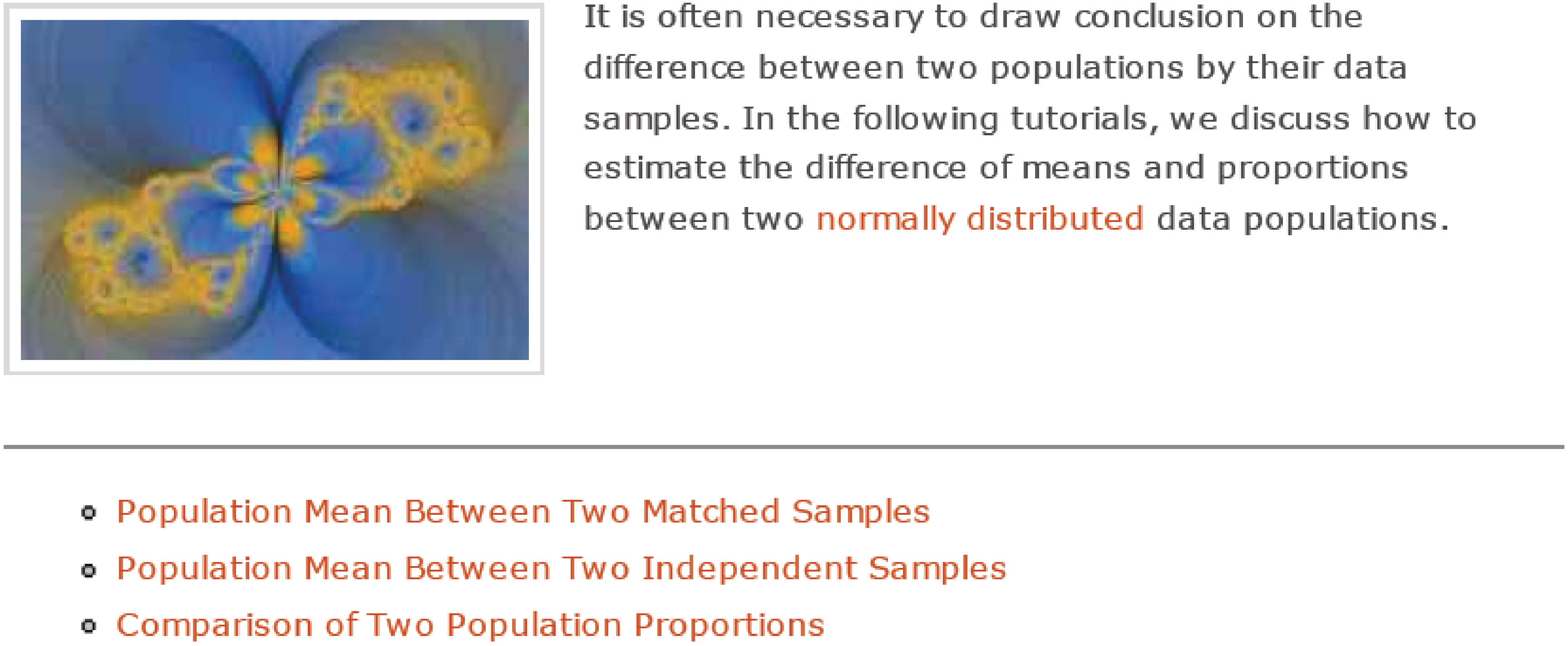
Elementary Statistics with R



- Qualitative Data
- Quantitative Data
- Numerical Measures
- Probability Distributions
- Interval Estimation
- Hypothesis Testing
- Type II Error
- Inference About Two Populations
- Goodness of Fit
- Analysis of Variance
- Non-parametric Methods
- Simple Linear Regression
- Multiple Linear Regression
- Logistic Regression

Inference About Two Populations

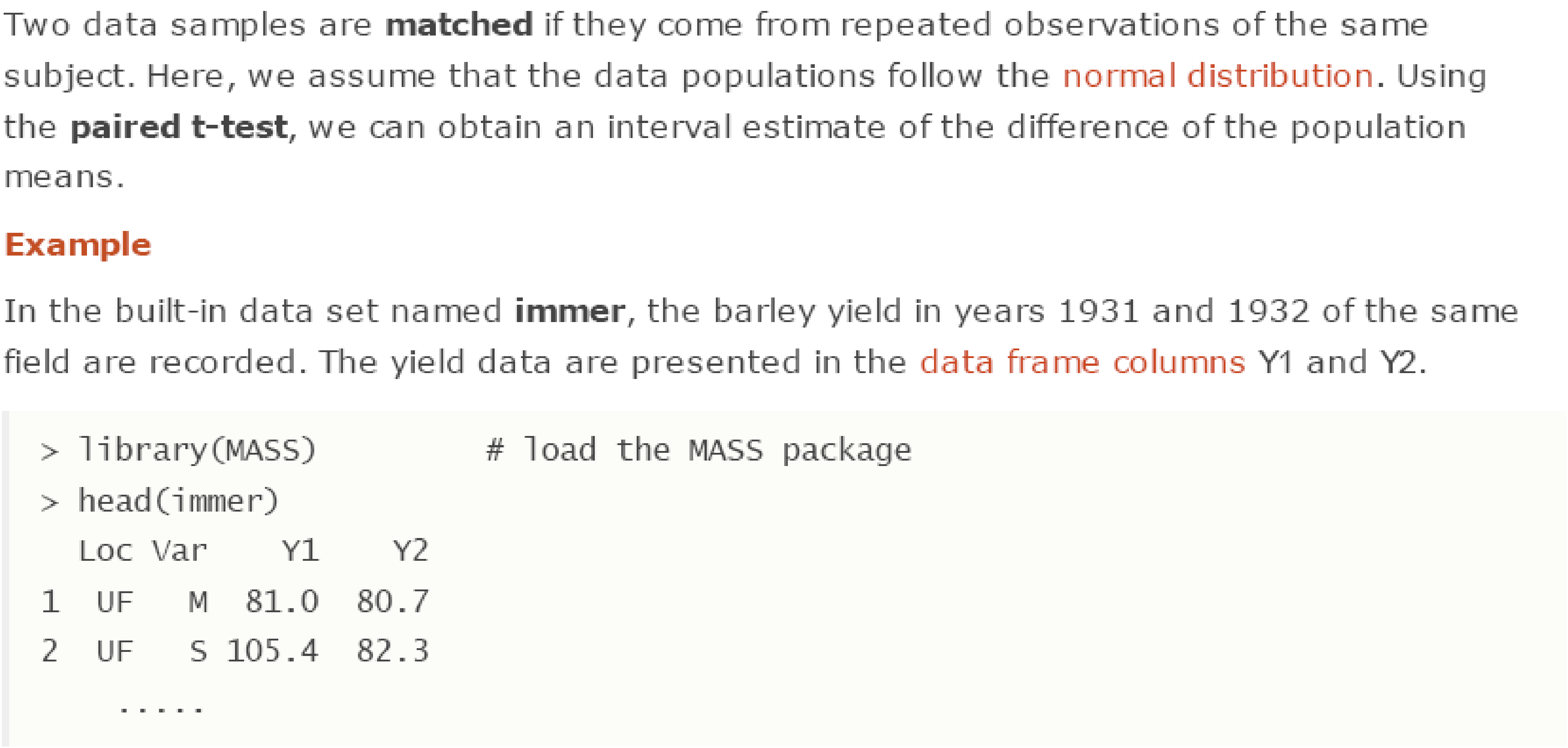
Inference About Two Populations



It is often necessary to draw conclusion on the difference between two populations by their data samples. In the following tutorials, we discuss how to estimate the difference of means and proportions between two normally distributed data populations.

- Population Mean Between Two Matched Samples
- Population Mean Between Two Independent Samples
- Comparison of Two Population Proportions

Population Mean Between Two Matched Samples



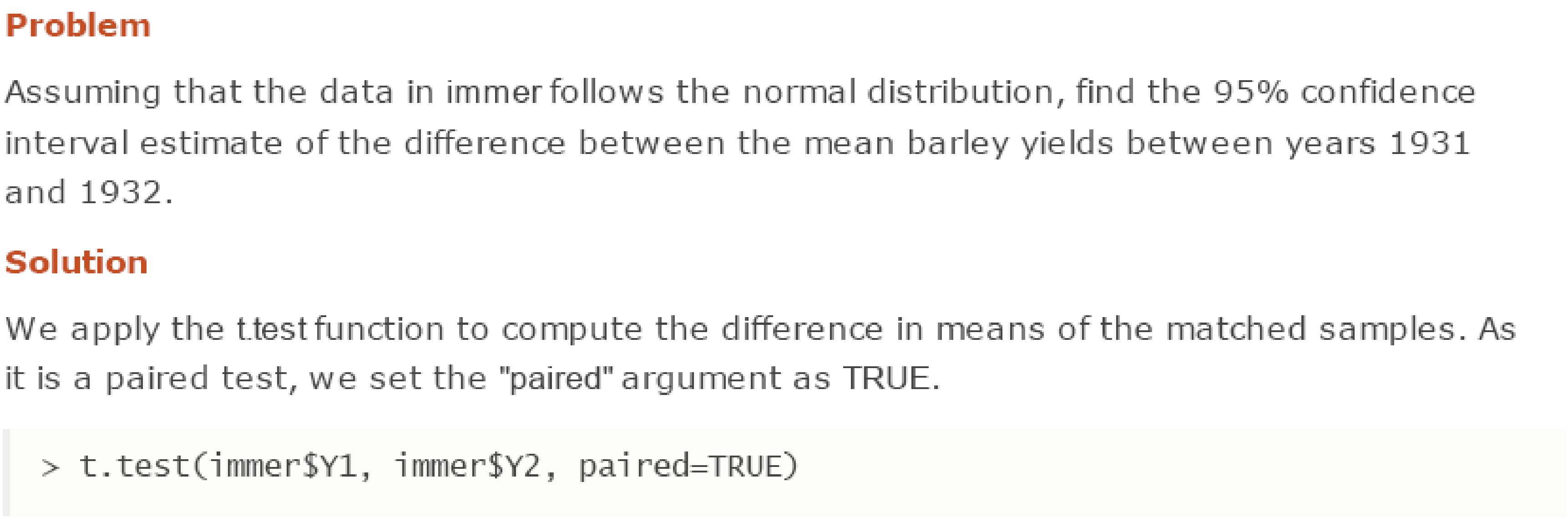
Two data samples are **matched** if they come from repeated observations of the same subject. Here, we assume that the data populations follow the normal distribution. Using the **paired t-test**, we can obtain an interval estimate of the difference of the population means.

Example

In the built-in data set named **immer**, the barley yield in years 1931 and 1932 of the same field are recorded. The yield data are presented in the data frame columns Y1 and Y2.

```
> library(MASS)         # load the MASS package
> head(immer)
  Loc Var    Y1   Y2
1  UF   M  81.0  80.7
2  UF   S 105.4  82.3
     .....
```

Population Mean Between Two Matched Samples (2)



Problem

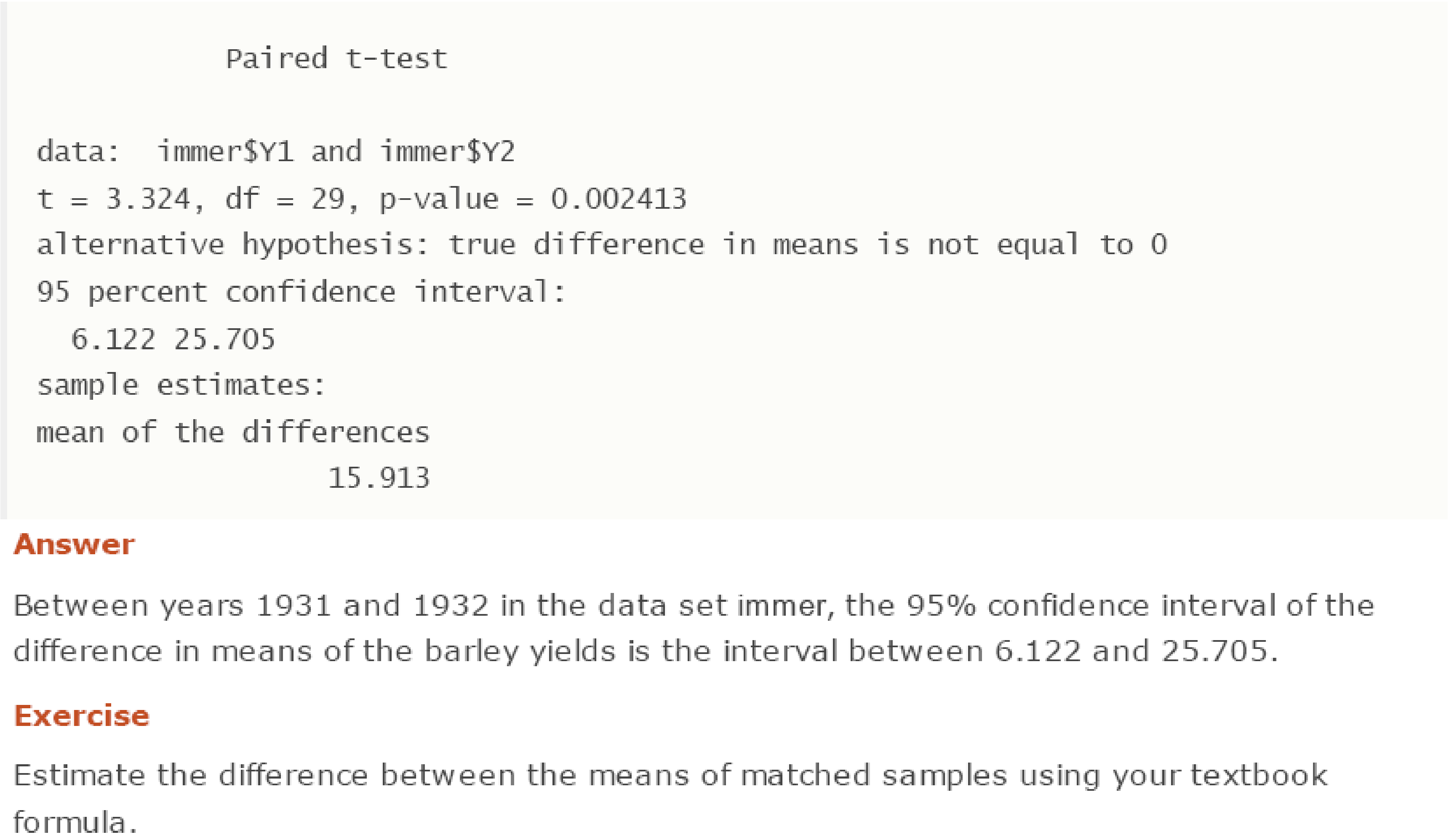
Assuming that the data in immer follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean barley yields between years 1931 and 1932.

Solution

We apply the t.test function to compute the difference in means of the matched samples. As it is a paired test, we set the "paired" argument as TRUE.

```
> t.test(immer$Y1, immer$Y2, paired=TRUE)
```

Population Mean Between Two Matched Samples (3)



```

	Paired t-test

data:  immer$Y1 and immer$Y2
t = 3.324, df = 29, p-value = 0.002413
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.122 25.705
sample estimates:
mean of the differences
                 15.913
```

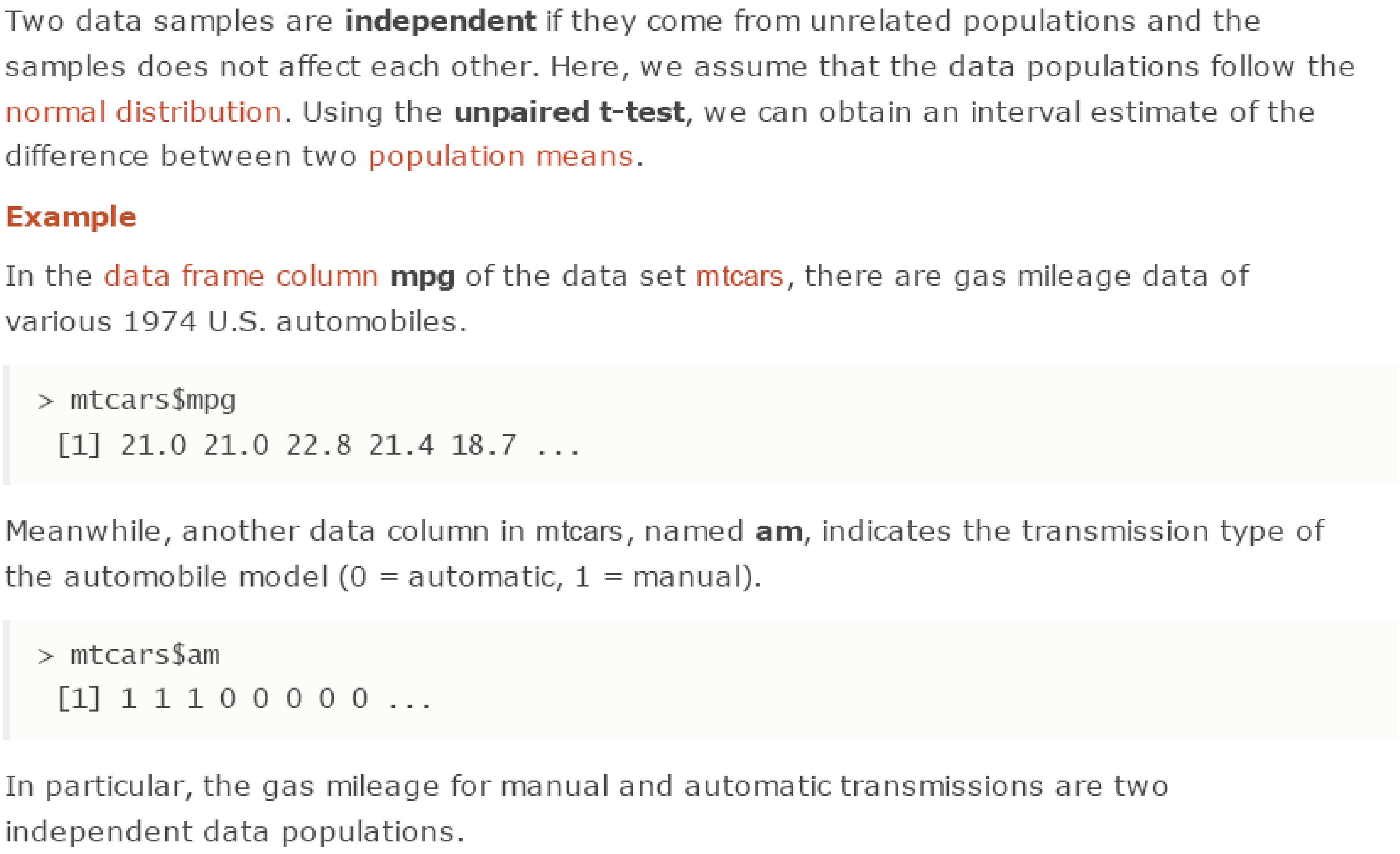
Answer

Between years 1931 and 1932 in the data set immer, the 95% confidence interval of the difference in means of the barley yields is the interval between 6.122 and 25.705.

Exercise

Estimate the difference between the means of matched samples using your textbook formula.

Population Mean Between Two Independent Samples



Two data samples are **independent** if they come from unrelated populations and the samples does not affect each other. Here, we assume that the data populations follow the normal distribution. Using the **unpaired t-test**, we can obtain an interval estimate of the difference between two population means.

Example

In the data frame column **mpg** of the data set mtcars, there are gas mileage data of various 1974 U.S. automobiles.

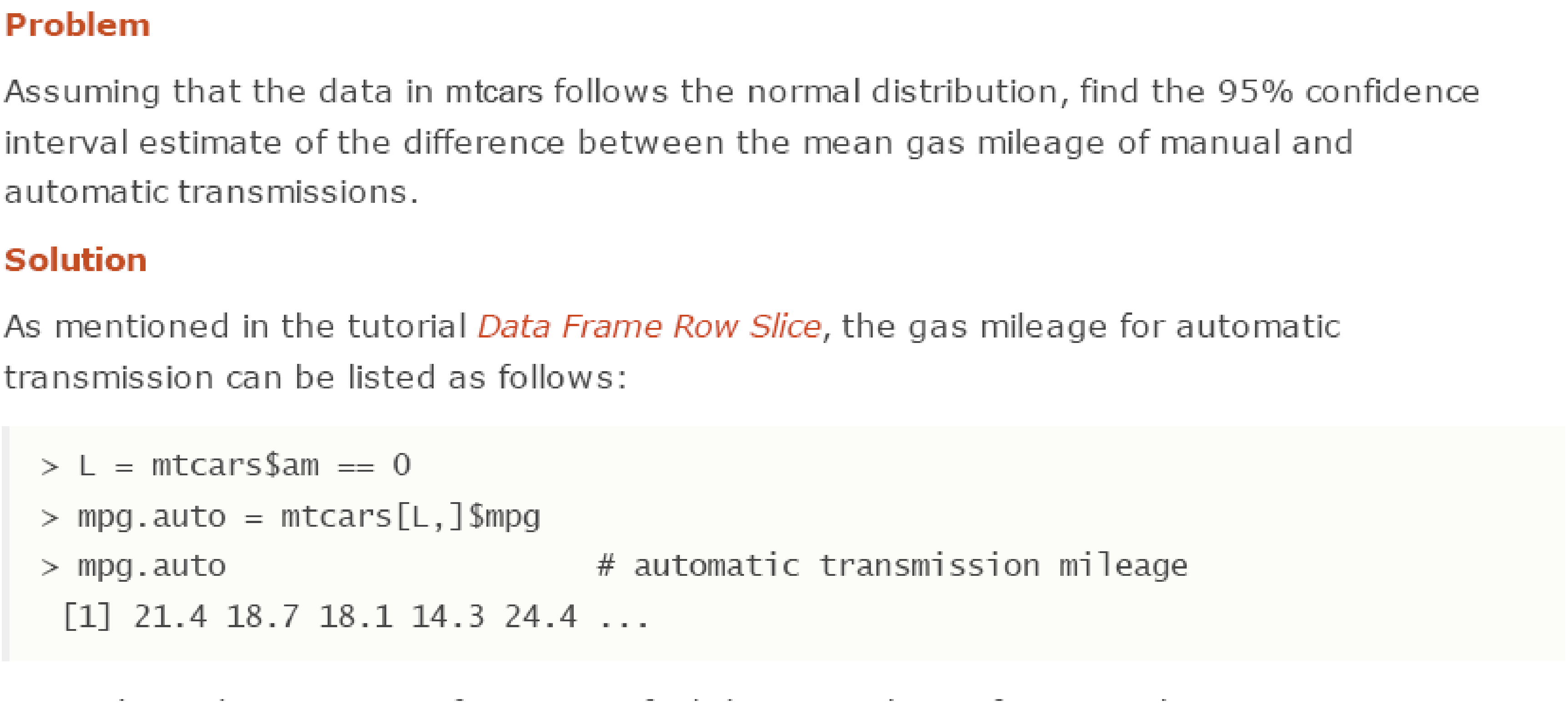
```
> mtcars$mpg
 [1] 21.0 21.0 22.8 21.4 18.7 ...
```

Meanwhile, another data column in mtcars, named **am**, indicates the transmission type of the automobile model (0 = automatic, 1 = manual).

```
> mtcars$am
 [1] 1 1 1 0 0 0 0 0 ...
```

In particular, the gas mileage for manual and automatic transmissions are two independent data populations.

Population Mean Between Two Independent Samples (2)



Problem

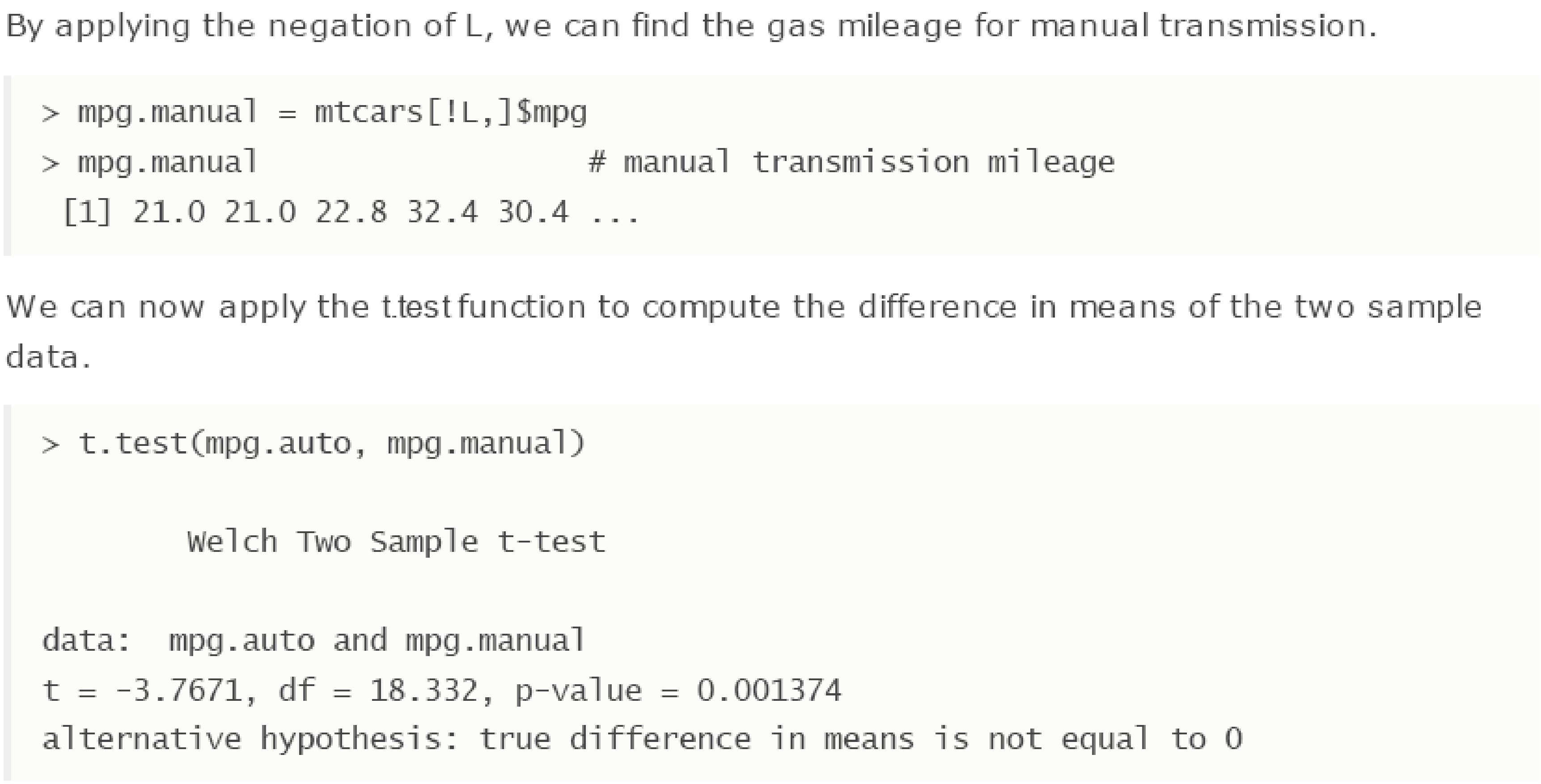
Assuming that the data in mtcars follows the normal distribution, find the 95% confidence interval estimate of the difference between the mean gas mileage of manual and automatic transmissions.

Solution

As mentioned in the tutorial *Data Frame Row Slice*, the gas mileage for automatic transmission can be listed as follows:

```
> L = mtcars$am == 0
> mpg.auto = mtcars[L,]$mpg
> mpg.auto                    # automatic transmission mileage
 [1] 21.4 18.7 18.1 14.3 24.4 ...
```

Population Mean Between Two Independent Samples (3)



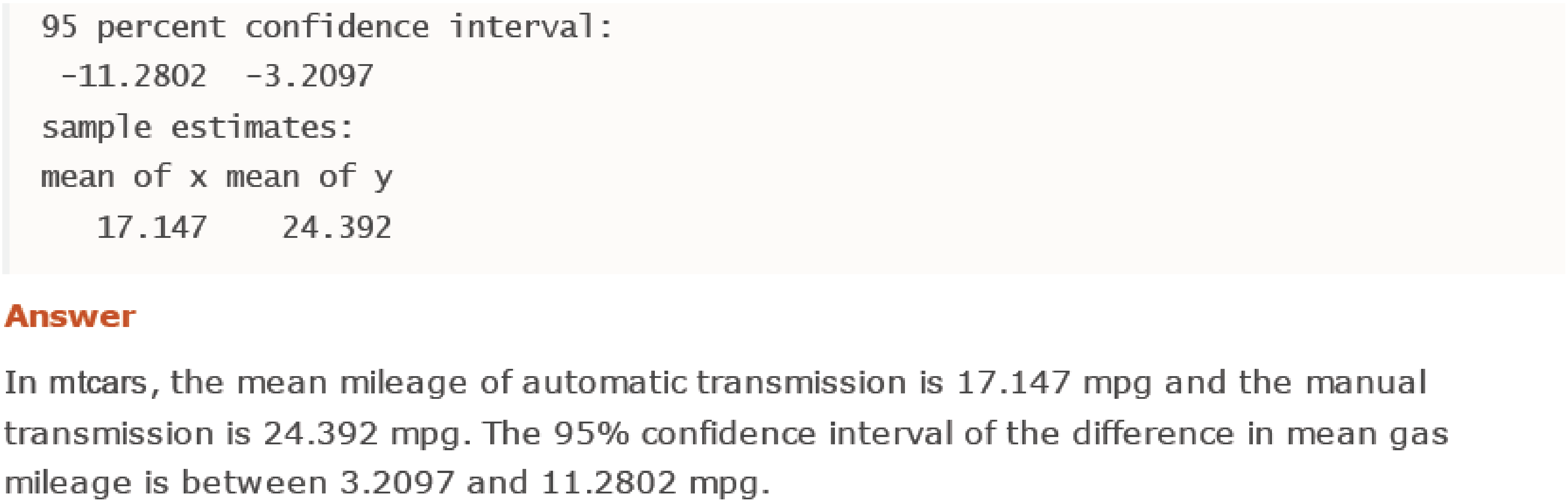
By applying the negation of L, we can find the gas mileage for manual transmission.

```
> mpg.manual = mtcars[!L,]$mpg
> mpg.manual                        # manual transmission mileage
 [1] 21.0 21.0 22.8 32.4 30.4 ...
```

We can now apply the t.test function to compute the difference in means of the two sample data.

```
> t.test(mpg.auto, mpg.manual)

        Welch Two Sample t-test

data:  mpg.auto and mpg.manual
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
```

Population Mean Between Two Independent Samples (4)

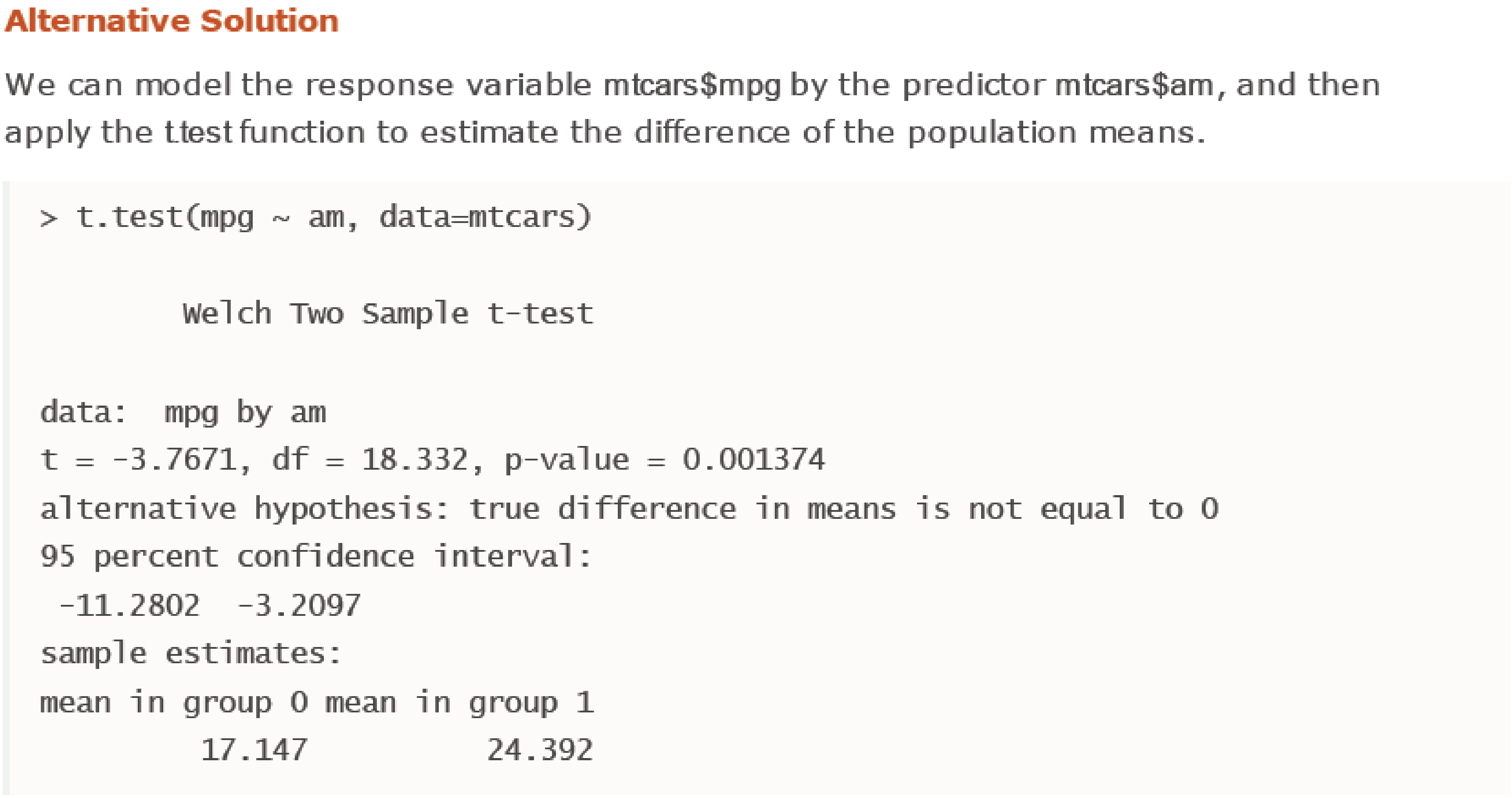


```
95 percent confidence interval:
 -11.2802  -3.2097
sample estimates:
mean of x mean of y
   17.147    24.392
```

Answer

In mtcars, the mean mileage of automatic transmission is 17.147 mpg and the manual transmission is 24.392 mpg. The 95% confidence interval of the difference in mean gas mileage is between 3.2097 and 11.2802 mpg.

Population Mean Between Two Independent Samples (5)



Alternative Solution

We can model the response variable mtcars$mpg by the predictor mtcars$am, and then apply the ttest function to estimate the difference of the population means.

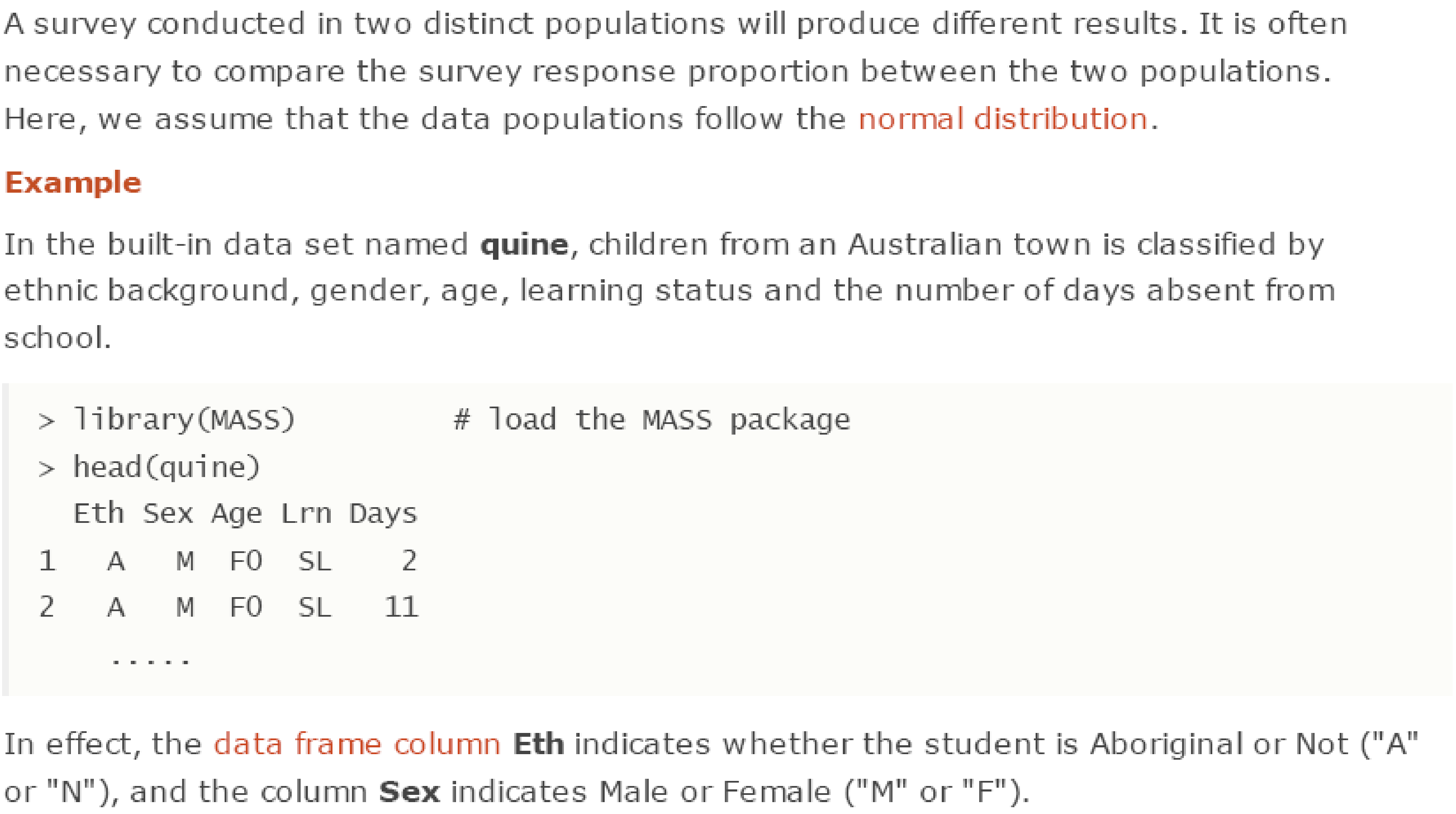
```
> t.test(mpg ~ am, data=mtcars)

        Welch Two Sample t-test

data:  mpg by am
t = -3.7671, df = 18.332, p-value = 0.001374
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.2802  -3.2097
sample estimates:
mean in group 0 mean in group 1
         17.147          24.392
```

Population Mean Between Two Independent Samples (6)

Note

Some textbooks truncate down the degree of freedom to an integer, and the result would differ from the t test.

Comparison of Two Population Proportions



A survey conducted in two distinct populations will produce different results. It is often necessary to compare the survey response proportion between the two populations. Here, we assume that the data populations follow the normal distribution.

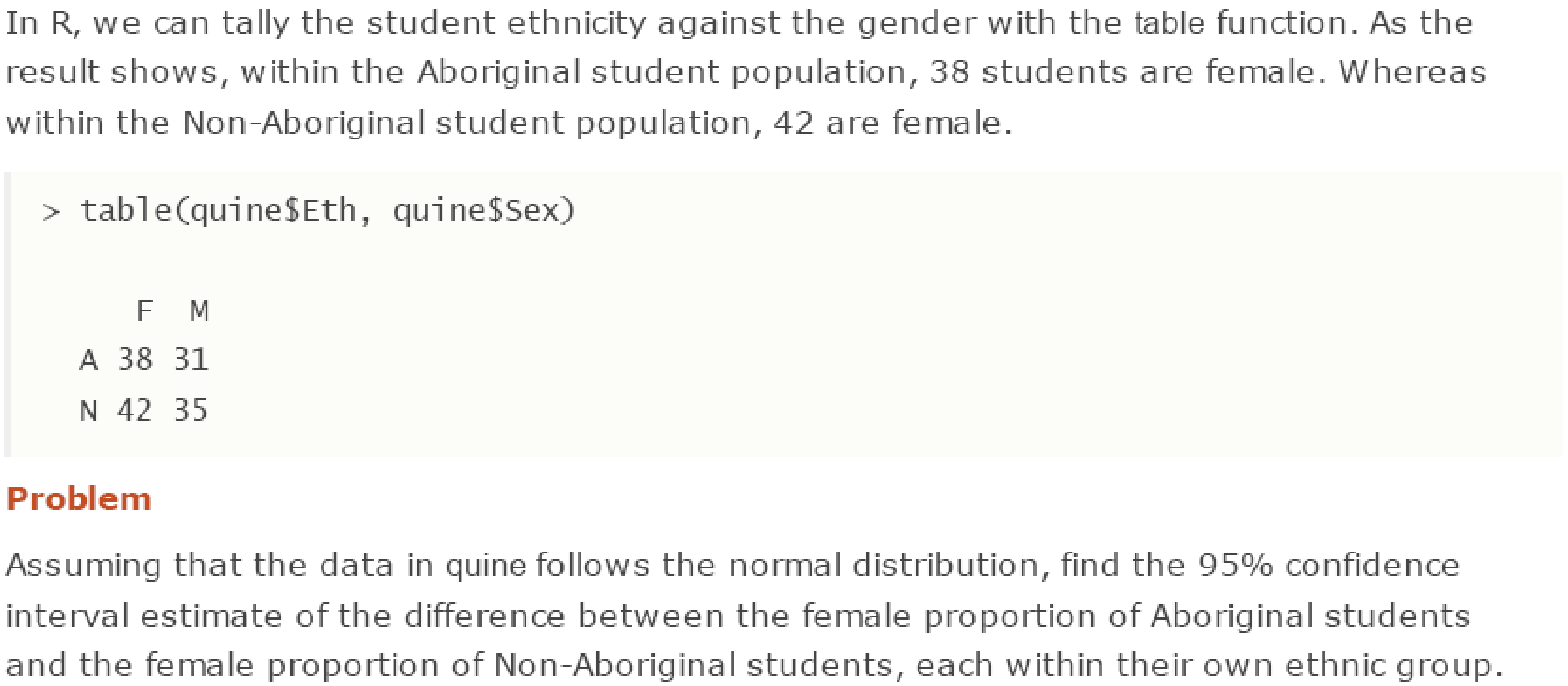
Example

In the built-in data set named **quine**, children from an Australian town is classified by ethnic background, gender, age, learning status and the number of days absent from school.

```
> library(MASS)          # load the MASS package
> head(quine)
  Eth Sex Age Lrn Days
1   A   M  F0  SL    2
2   A   M  F0  SL   11
     .....
```

In effect, the data frame column **Eth** indicates whether the student is Aboriginal or Not ("A" or "N"), and the column **Sex** indicates Male or Female ("M" or "F").

Comparison of Two Population Proportions (2)



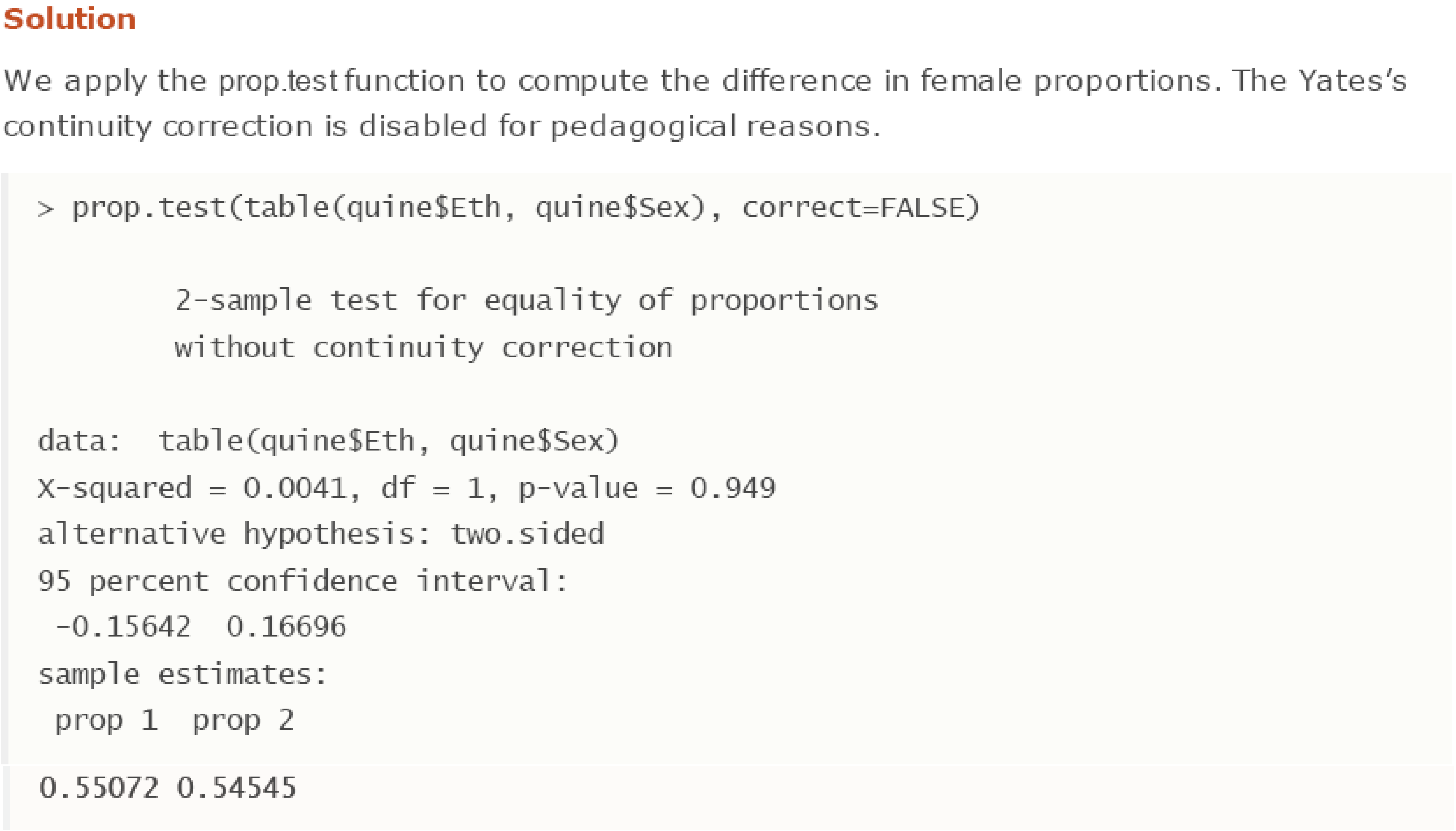
In R, we can tally the student ethnicity against the gender with the table function. As the result shows, within the Aboriginal student population, 38 students are female. Whereas within the Non-Aboriginal student population, 42 are female.

```
> table(quine$Eth, quine$Sex)

     F  M
  A 38 31
  N 42 35
```

Problem

Assuming that the data in quine follows the normal distribution, find the 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students, each within their own ethnic group.

Comparison of Two Population Proportions (3)

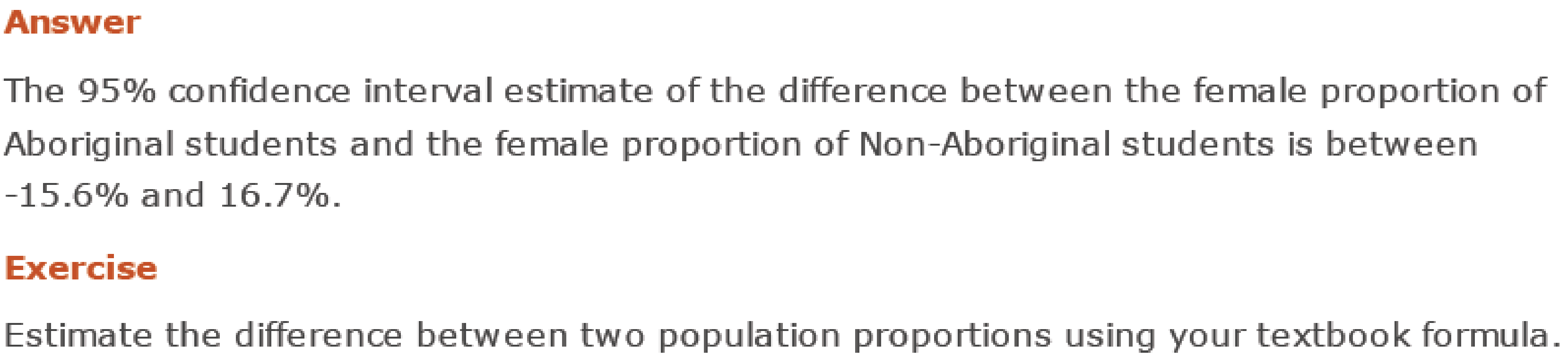


Solution

We apply the prop.test function to compute the difference in female proportions. The Yates's continuity correction is disabled for pedagogical reasons.

```
> prop.test(table(quine$Eth, quine$Sex), correct=FALSE)

        2-sample test for equality of proportions
        without continuity correction

data:  table(quine$Eth, quine$Sex)
X-squared = 0.0041, df = 1, p-value = 0.949
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.15642  0.16696
sample estimates:
 prop 1  prop 2
0.55072 0.54545
```

Comparison of Two Population Proportions (4)



Answer

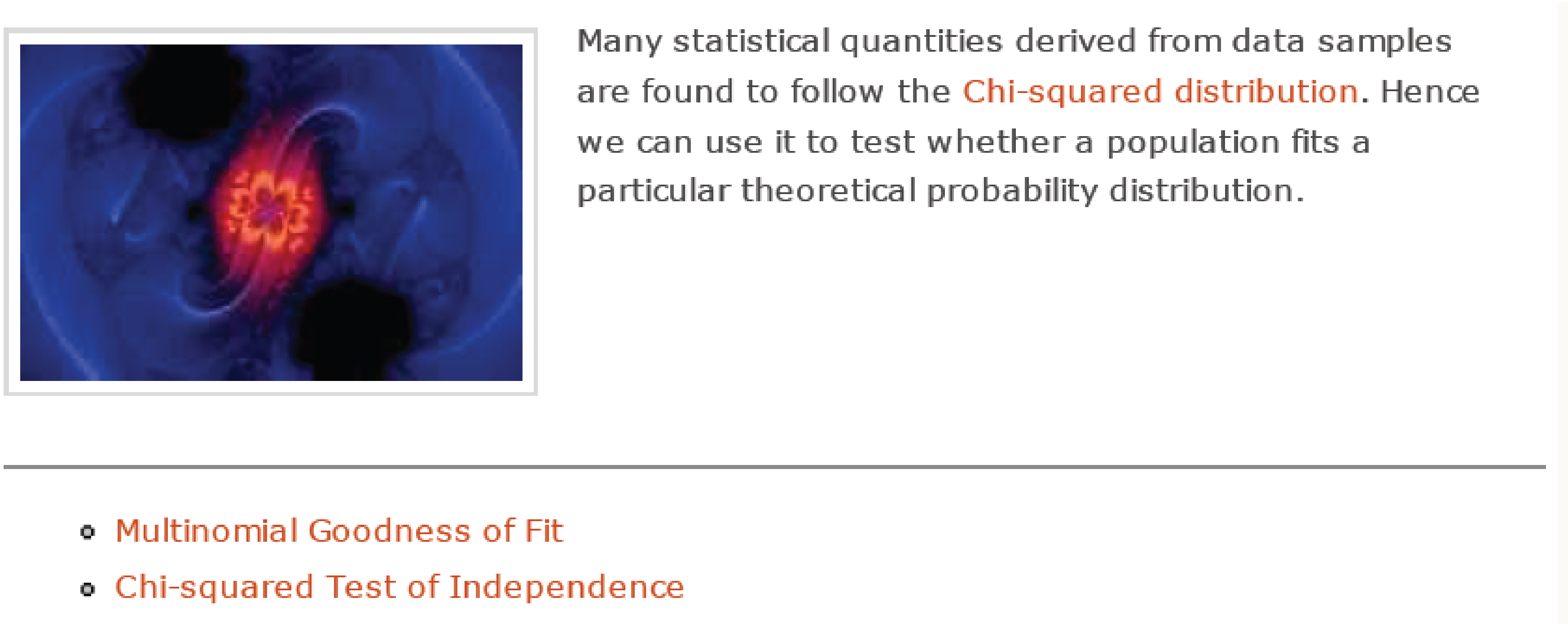
The 95% confidence interval estimate of the difference between the female proportion of Aboriginal students and the female proportion of Non-Aboriginal students is between -15.6% and 16.7%.

Exercise

Estimate the difference between two population proportions using your textbook formula.

Goodness of Fit

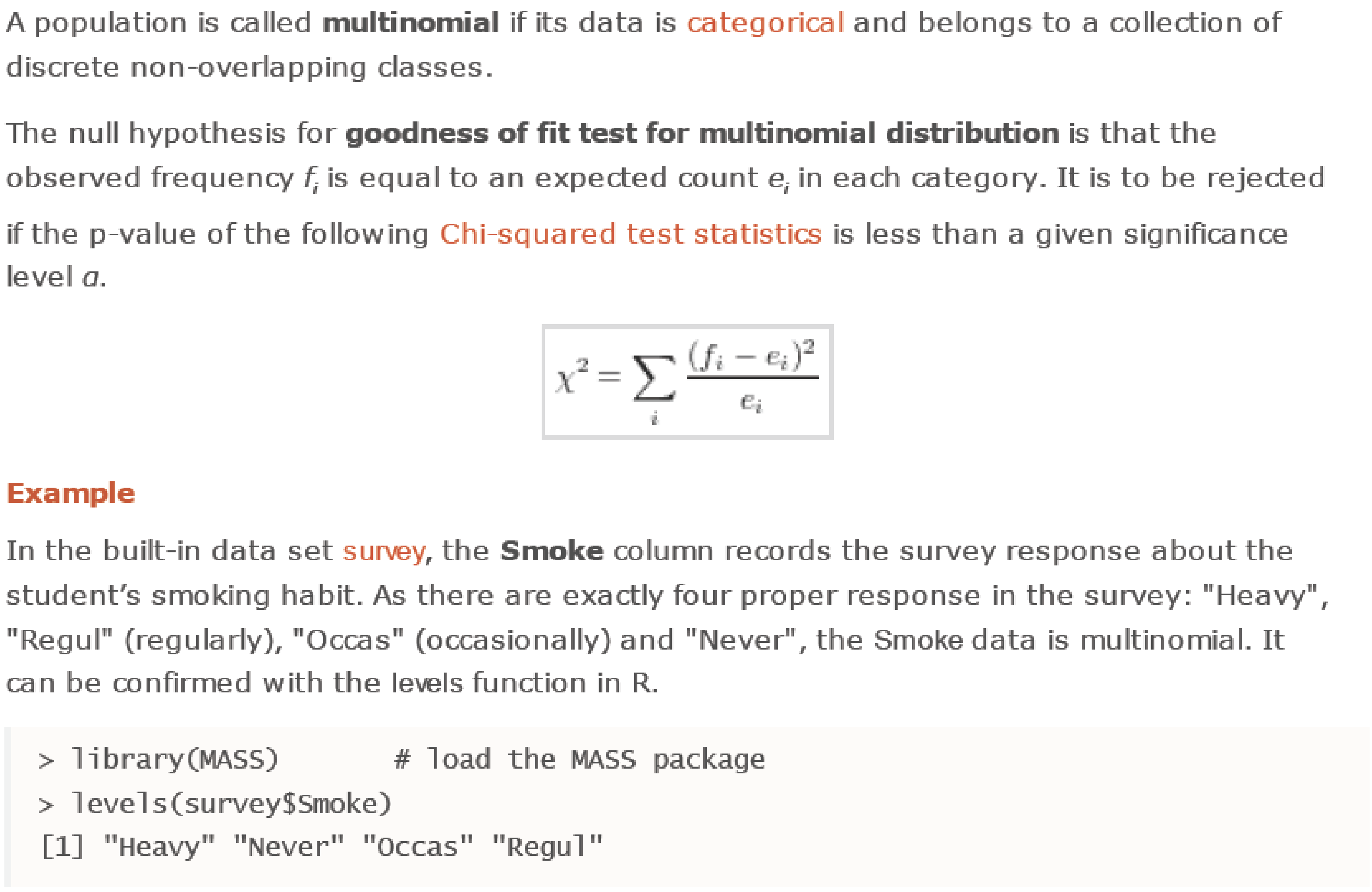
Goodness of Fit



Many statistical quantities derived from data samples are found to follow the Chi-squared distribution. Hence we can use it to test whether a population fits a particular theoretical probability distribution.

- Multinomial Goodness of Fit
- Chi-squared Test of Independence

Multinomial Goodness of Fit



A population is called **multinomial** if its data is categorical and belongs to a collection of discrete non-overlapping classes.

The null hypothesis for **goodness of fit test for multinomial distribution** is that the observed frequency f_i is equal to an expected count e_i in each category. It is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level *α*.

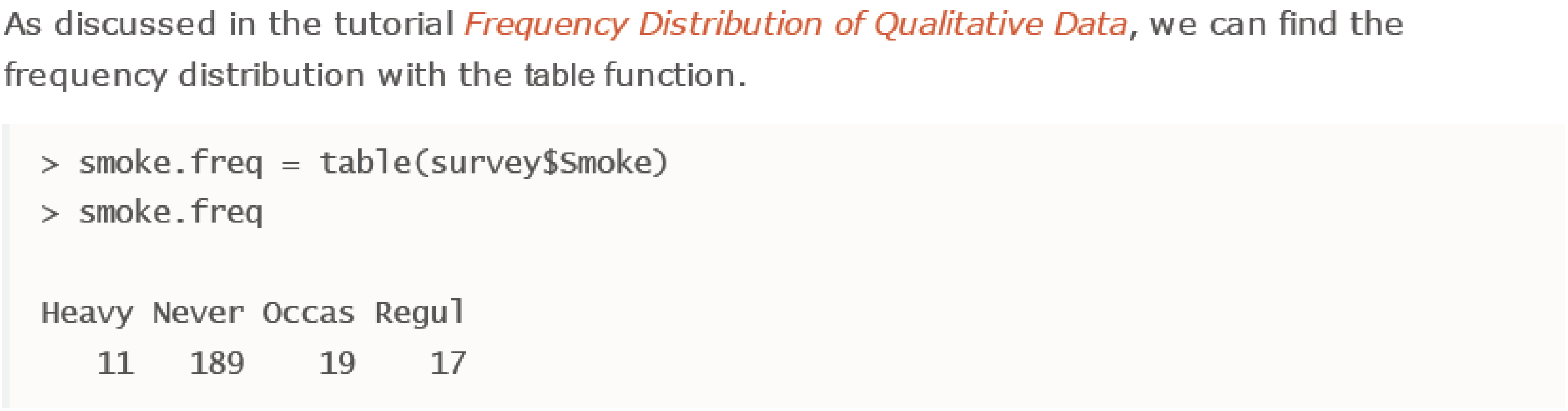
$$\chi^2 = \sum_i \frac{(f_i - e_i)^2}{e_i}$$

Example

In the built-in data set survey, the **Smoke** column records the survey response about the student's smoking habit. As there are exactly four proper response in the survey: "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never", the Smoke data is multinomial. It can be confirmed with the levels function in R.

```
> library(MASS)       # load the MASS package
> levels(survey$Smoke)
[1] "Heavy" "Never" "Occas" "Regul"
```

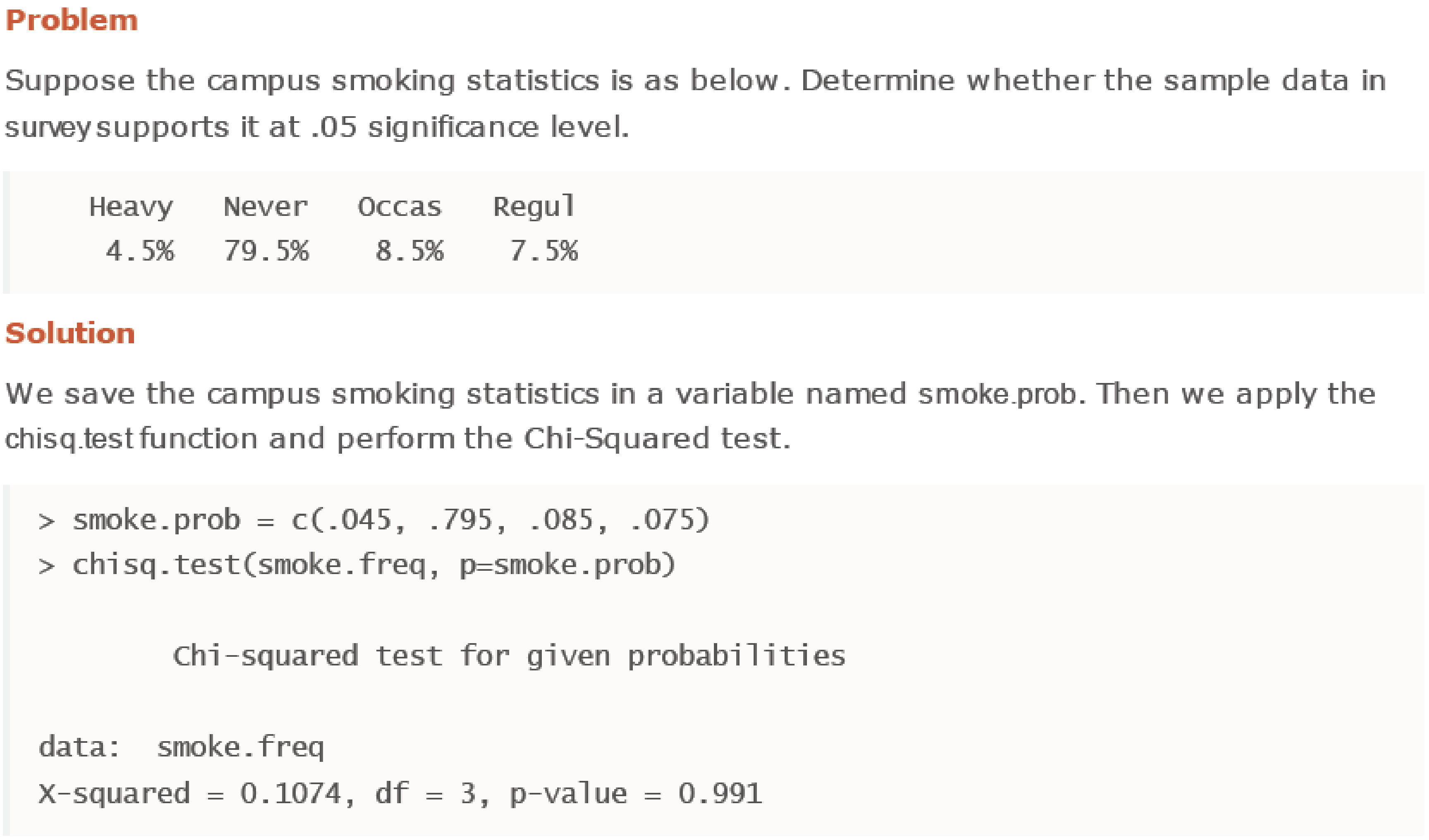
Multinomial Goodness of Fit (2)



As discussed in the tutorial *Frequency Distribution of Qualitative Data*, we can find the frequency distribution with the table function.

```
> smoke.freq = table(survey$Smoke)
> smoke.freq

Heavy Never Occas Regul
   11   189    19    17
```

Multinomial Goodness of Fit (3)



Problem

Suppose the campus smoking statistics is as below. Determine whether the sample data in survey supports it at .05 significance level.

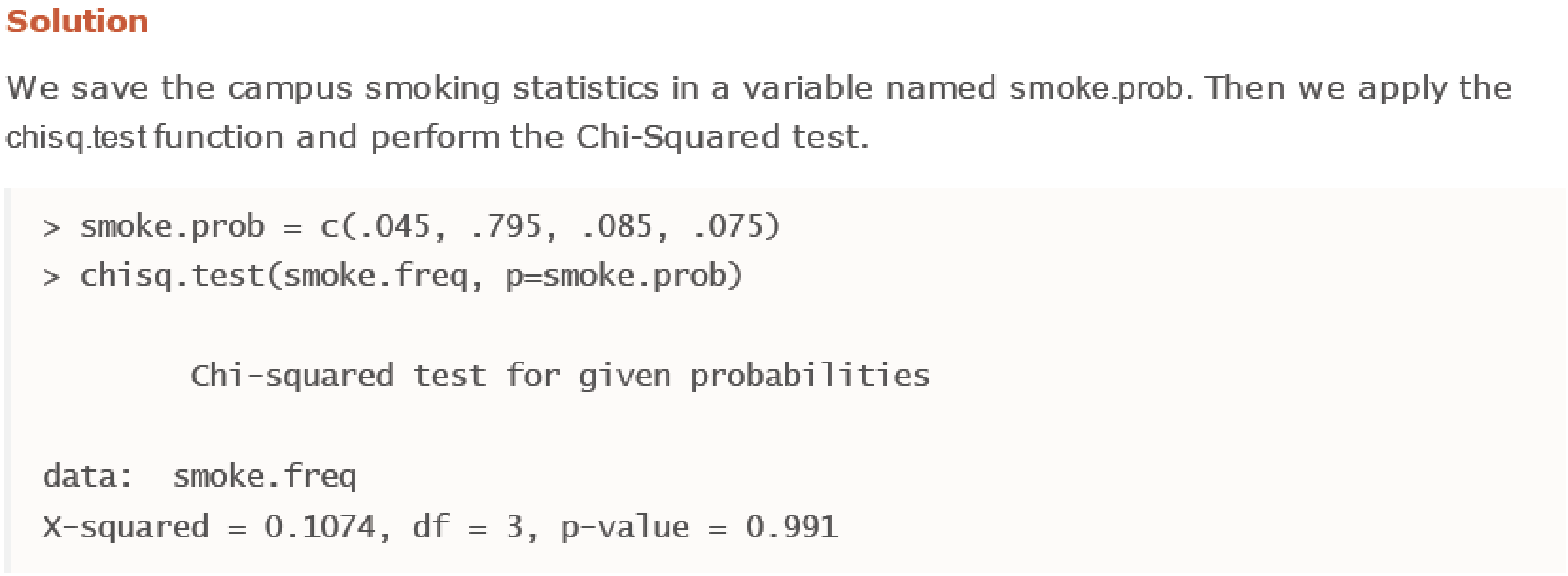
```
     Heavy    Never    Occas    Regul
      4.5%    79.5%     8.5%     7.5%
```

Solution

We save the campus smoking statistics in a variable named smoke.prob. Then we apply the chisq.test function and perform the Chi-Squared test.

```
> smoke.prob = c(.045, .795, .085, .075)
> chisq.test(smoke.freq, p=smoke.prob)

        Chi-squared test for given probabilities

data:  smoke.freq
X-squared = 0.1074, df = 3, p-value = 0.991
```

Multinomial Goodness of Fit (4)

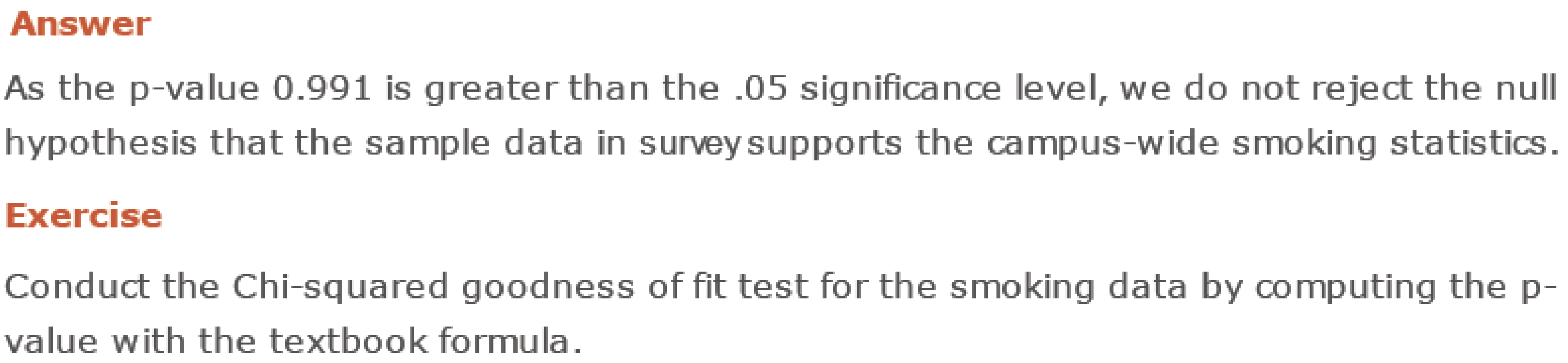


Solution

We save the campus smoking statistics in a variable named smoke.prob. Then we apply the chisq.test function and perform the Chi-Squared test.

```
> smoke.prob = c(.045, .795, .085, .075)
> chisq.test(smoke.freq, p=smoke.prob)

        Chi-squared test for given probabilities

data:  smoke.freq
X-squared = 0.1074, df = 3, p-value = 0.991
```

Multinomial Goodness of Fit (5)



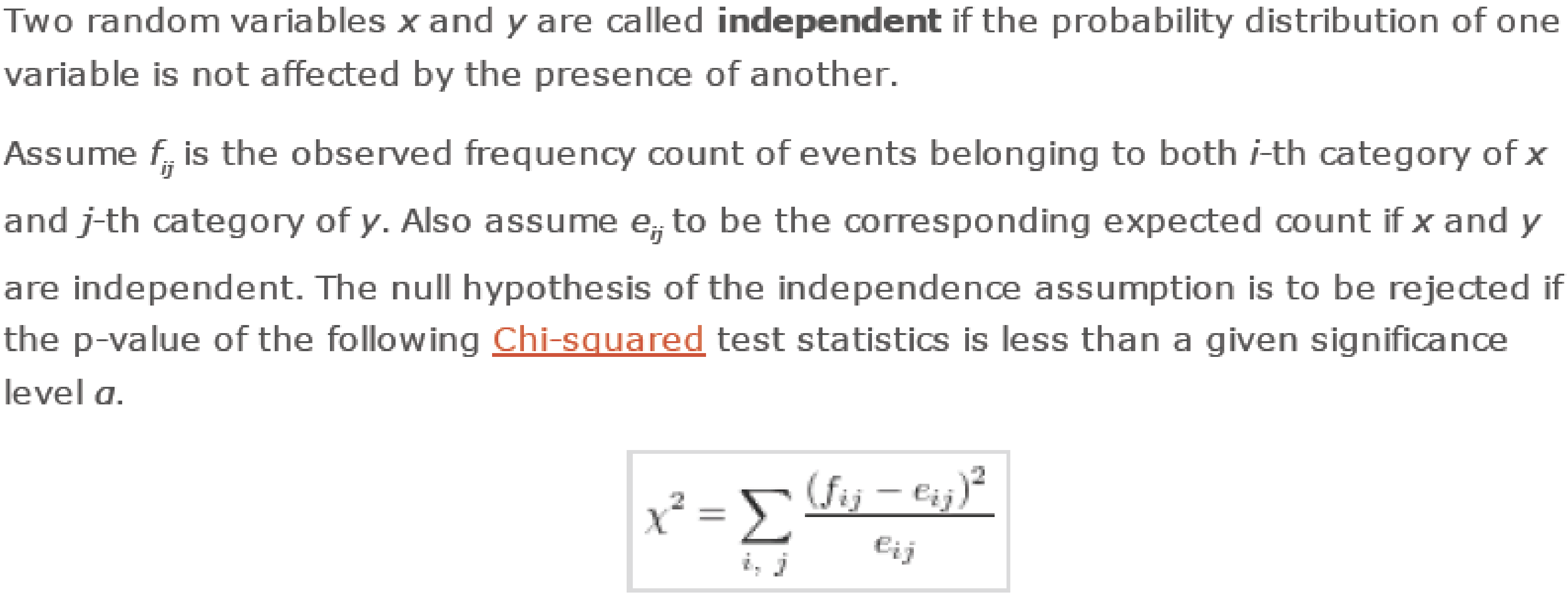
Answer

As the p-value 0.991 is greater than the .05 significance level, we do not reject the null hypothesis that the sample data in survey supports the campus-wide smoking statistics.

Exercise

Conduct the Chi-squared goodness of fit test for the smoking data by computing the p-value with the textbook formula.

Chi-squared Test of Independence

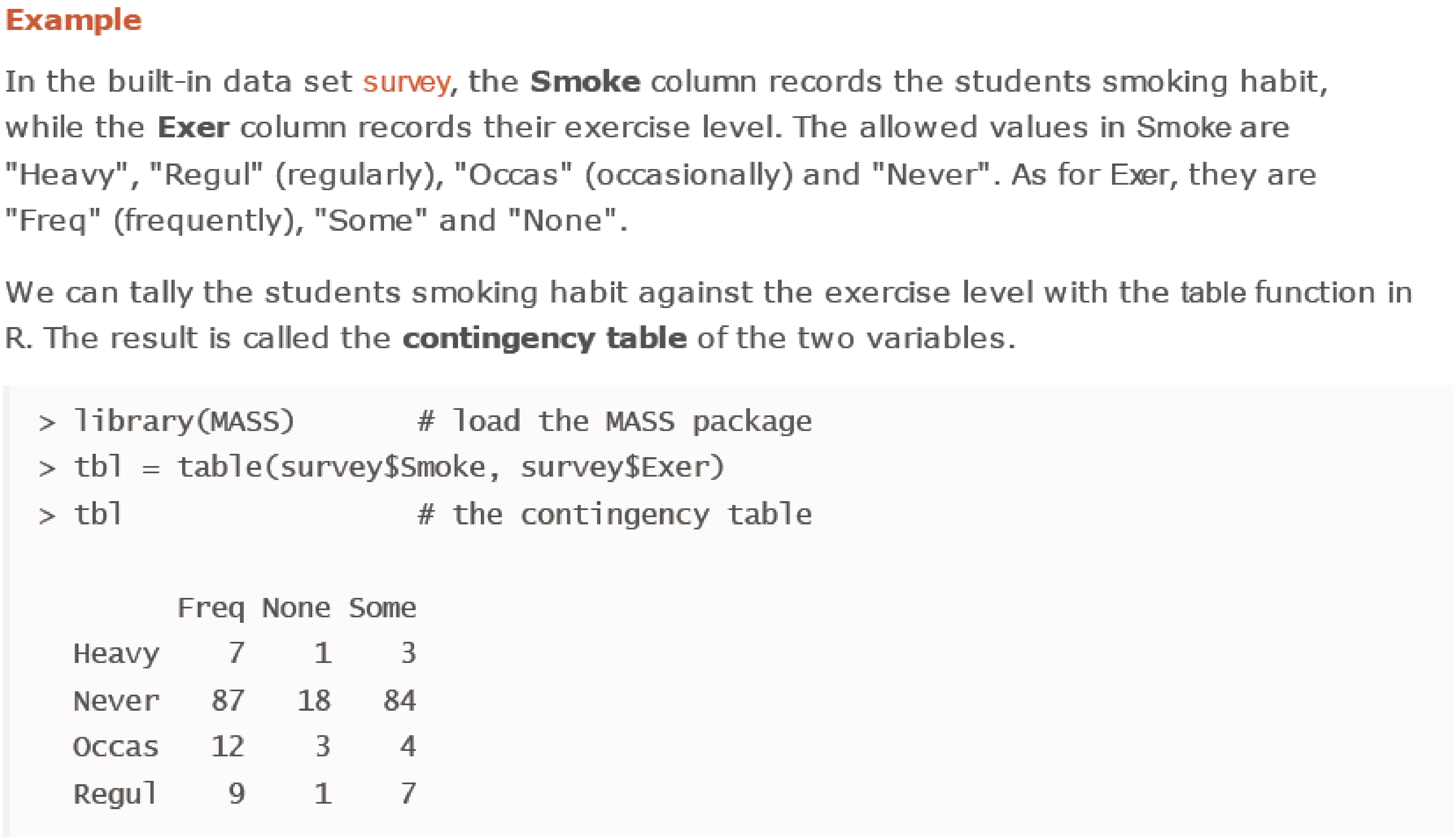


Two random variables *x* and *y* are called **independent** if the probability distribution of one variable is not affected by the presence of another.

Assume f_{ij} is the observed frequency count of events belonging to both *i*-th category of *x* and *j*-th category of *y*. Also assume e_{ij} to be the corresponding expected count if *x* and *y* are independent. The null hypothesis of the independence assumption is to be rejected if the p-value of the following Chi-squared test statistics is less than a given significance level *α*.

$$\chi^2 = \sum_{i,\ j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

Chi-squared Test of Independence (2)



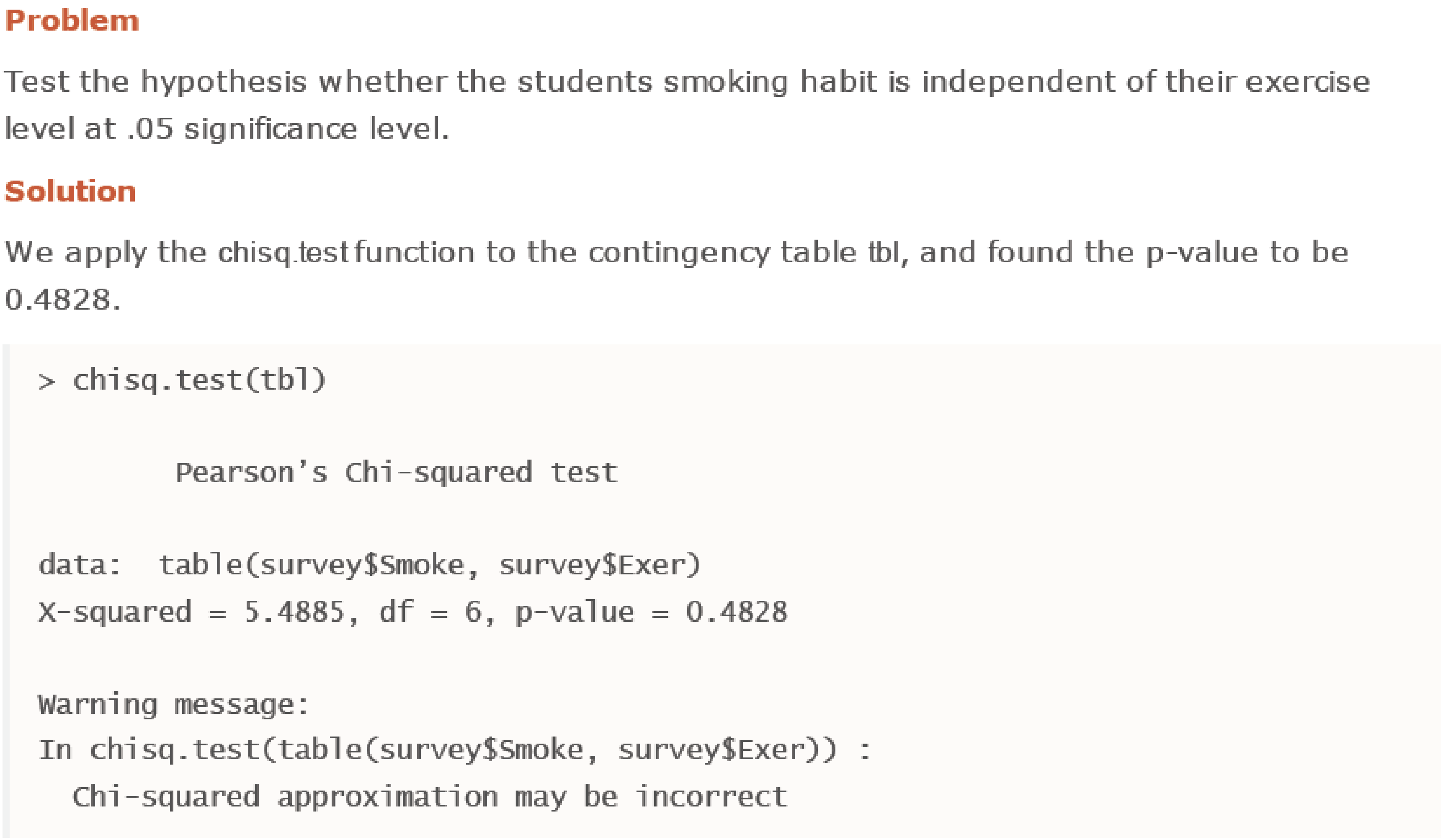
Example

In the built-in data set survey, the **Smoke** column records the students smoking habit, while the **Exer** column records their exercise level. The allowed values in Smoke are "Heavy", "Regul" (regularly), "Occas" (occasionally) and "Never". As for Exer, they are "Freq" (frequently), "Some" and "None".

We can tally the students smoking habit against the exercise level with the table function in R. The result is called the **contingency table** of the two variables.

```
> library(MASS)       # load the MASS package
> tbl = table(survey$Smoke, survey$Exer)
> tbl                 # the contingency table


        Freq None Some
  Heavy    7    1    3
  Never   87   18   84
  Occas   12    3    4
  Regul    9    1    7
```

Chi-squared Test of Independence (3)



Problem

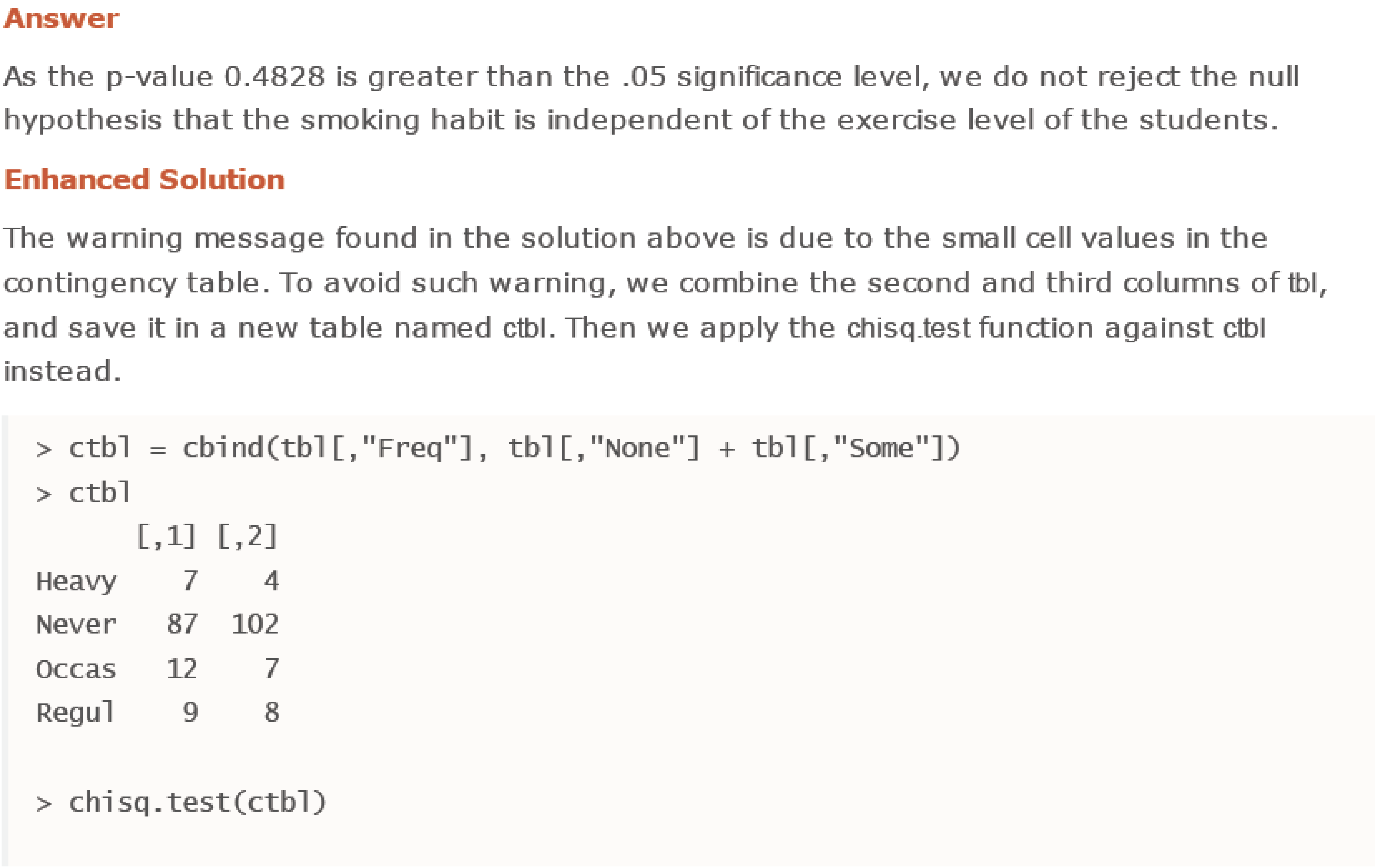
Test the hypothesis whether the students smoking habit is independent of their exercise level at .05 significance level.

Solution

We apply the chisq.test function to the contingency table tbl, and found the p-value to be 0.4828.

```
> chisq.test(tbl)

        Pearson’s Chi-squared test

data:  table(survey$Smoke, survey$Exer)
X-squared = 5.4885, df = 6, p-value = 0.4828

Warning message:
In chisq.test(table(survey$Smoke, survey$Exer)) :
  Chi-squared approximation may be incorrect
```

Chi-squared Test of Independence (4)



Answer

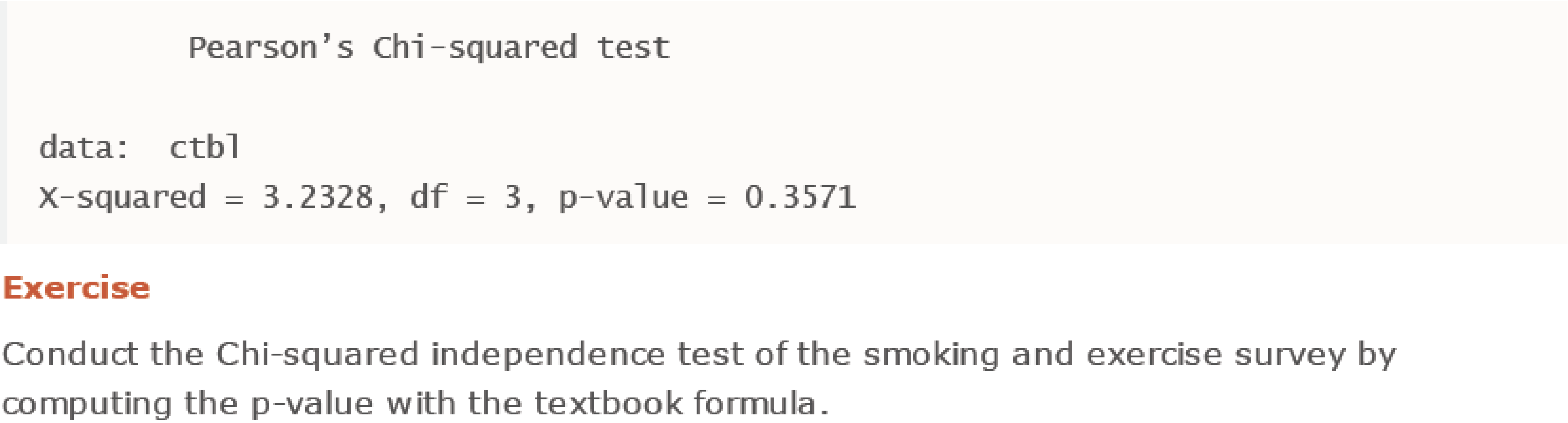
As the p-value 0.4828 is greater than the .05 significance level, we do not reject the null hypothesis that the smoking habit is independent of the exercise level of the students.

Enhanced Solution

The warning message found in the solution above is due to the small cell values in the contingency table. To avoid such warning, we combine the second and third columns of tbl, and save it in a new table named ctbl. Then we apply the chisq.test function against ctbl instead.

```
> ctbl = cbind(tbl[,"Freq"], tbl[,"None"] + tbl[,"Some"])
> ctbl
      [,1] [,2]
Heavy    7    4
Never   87  102
Occas   12    7
Regul    9    8

> chisq.test(ctbl)
```

Chi-squared Test of Independence (5)



```
        Pearson's Chi-squared test

data:  ctbl
X-squared = 3.2328, df = 3, p-value = 0.3571
```

Exercise

Conduct the Chi-squared independence test of the smoking and exercise survey by computing the p-value with the textbook formula.

Acknowlegments

Some parts of the material in this presentation is taken from http://www.r-tutor.com/

Deo materijala je preuzet sa sajta http://www.e-statistika.rs