Chapter 7

Comparing proportions and distributions

Please read Chapter 12 from *Learning Statistics with R* before starting these assignments.

Content *Page*

Chapter 7: Comparing proportions and distributions

Learning objectives of this chapter:

- Chi-square testing by hand
- Performing a Chi-square test in R
- Proportion estimation and testing by hand and in R

Assignment 7.1: Chi-squared testing by hand

A small bed and breakfast wants to know if the **distribution** of the number of guests is changing. They compare their historical **distribution** with the number of guests in 2019 and want to show, with 95% confidence, that the shape of the 2019 **distribution** has changed with respect to the past years.

7.1 a) Formulate the **null hypothesis** H_0 and the **alternative hypothesis** H_1 for this test in words.

Answer 7.1a:

H_0: _____

H_1: _____

The table below shows the historical **distribution** of the bed and breakfast guests alongside the number of guests for 2019.

	Historical	Observed (O)	Expected (E)	$O-E$	$\frac{(O-E)^2}{E}$
Spring	48.7	29.0			
Summer	30.4	45.0			
Fall	16.5	49.4			
Winter	28.8	32.8			
Total	124.4	156.2			

7.1 b) Fill in the expected (E) column in the table and show how you calculated it below.

Answer 7.1b:

_____

_____

7.1 c) Are you allowed to do a **Chi-squared test** using these data? Explain why.

Hint 7.1: Take into account the assumptions of a ***Chi-squared test*** .

Answer 7.1c:

7.1 d) Fill in the rest of the table to calculate the **Chi-squared value** (X^2) for this sample.

Answer 7.1d:

X^2: _____

The critical **Chi-squared value** using a confidence of 95% and 3 **degrees of freedom** is $X^2_{0.05}[df = 3] = 7.815$.

7.1 e) What is your conclusion on the basis of these results? Include the following elements:

- ☐ Show how the calculated **Chi-squared value** relates to the **critical Chi-squared value** .
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about the historical **distribution** and the 2019 **distribution** .
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 7.1e:

7.1 f) Taking into account the confidence of our analysis, explain in what range the **p-value** of this test must be.

Answer 7.1f:

The bed and breakfast owner does not believe in your calculations and wants you to recalculate it in R.

Run the following code in R to import the data:

```
Observed <- c(29, 45, 49.4, 32.8)
Historical <- c(0.39, 0.25, 0.13, 0.23)
Expected <- c(60.918, 39.05, 20.306, 35.926)
```

7.1 g) Calculate the **Chi-squared value** for these data using R.

R code 7.1g:

Answer 7.1g:

X^2: _____

Run the following code in R to find out the **critical Chi-squared value** :

```
qchisq(p = 0.95, df = 3)
```

7.1 h) Does your conclusion hold up when you recalculate the **Chi-squared value** in R?

Answer 7.1h:

YES / NO

Use the following code in R to perform a **Chi-squared test** :

```
# Chi-squared test: x = observations p = model distribution
# rescale makes sure the model distribution adds up to 100%
chisq.test(x = Observed, p = Historical, rescale.p = TRUE)
```

7.1 i) Is the result what you expected?

Answer 7.1i:

As it turns out the function **chisq.test()** will always calculate the **expected values** by itself. You therefore do not need to make sure your total amounts sum to one. Therefore, in the code after **p =** you can supply a distribution that adds up to one (and set **rescale.p = FALSE**), or expected amounts that R will recalculate into a distribution anyway. If you do not supply the **p** argument, R will test against the **uniform distribution** .

Run the following code in R to store the results of the **Chi-squared test** in an object called **chisq** :

```
chisq <- chisq.test(x = Observed, p = Historical)
```

7.1 j) Find out how to extract the **expected values** from the **chisq** object. Do they match the **expected values** you wrote down in the table above?

R code 7.1j:

Answer 7.1j:

Assignment 7.2: Chi-squared testing in R

A gift company in the Netherlands wants to apply statistics to gain insight into their sales activity. This gift company normally generates most of its revenue in the period before summer and on Christmas holidays. Recently they opened a new store in a different country. The gift company wants to check, using a **Chi-squared** (X^2) test, whether the new store has a different seasonal pattern than the stores in the Netherlands.

Run the following code in R to create the data set and store it in an object named **sales** .

```
# These are the values for the sales data set
sales <- data.frame(month = seq(from = 1, to = 12, by = 1),
                    historical = c(5.1, 5.1, 6.7, 10, 11.4, 10,
                                   6.7, 5.1, 6.7, 10, 11.7, 11.7),
                    newstore = c(5.6, 6.2, 9.4, 8.6, 6.8, 4.8,
                                 5.6, 4.8, 8.8, 12.6, 13.1, 13.7))
```

7.2 a) Explore the **sales** object and describe what it contains.

Hint 7.2: You can use the **summary()** *function to find out some quick information.*

Answer 7.2a:

Run the following code in R to create a graphical representation of the data:

```
# Create a barplot
barplot(t(matrix(c(sales$historical, sales$newstore), ncol = 2)),
        beside = TRUE, names.arg = sales$month, las = 1, xlab = 'Month',
        ylab = 'Percentage of yearly sales',
        main = 'Seasonal sales',
        col = c('aquamarine3','coral'))

# Add a legend
legend('topleft', bty = 'n',
       legend = c('Historical', 'New store'),
       fill = c('aquamarine3','coral'))
```

The company wants to perform a **Chi-squared test** , with 90% confidence, on these data using the historical sales as the baseline values and the new store sales as the observed values.

7.2 b) Looking at the graph and considering the values in each month, can you perform a Chi-squared test on these data?

Answer 7.2b:

7.2 c) Formulate the **null hypothesis** H_0 and the **alternative hypothesis** H_1 for this test in words.

Answer 7.2c:

H_0 :

H_1 :

7.2 d) Perform a **Chi-squared test** in R using these data and draw the conclusion for the hypotheses. Include the following elements:

- ☐ Discuss the **p-value** of this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about the historical and the new **distribution** .
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

R code 7.2d:

Answer 7.2d:

Assignment 7.3: Proportion testing by hand and in R

The P.H.O.N.E. company sells subscriptions to magazines by phone. The commercial director wants to know what **proportion** of calls actually lead to a subscription and whether that depends on the time of day. He therefore takes two samples (n_1 and n_2), one in the afternoon shift and one in the evening shift, and records the number of calls that lead to a subscription (k_1 and k_2):

Afternoon: $n_1 = 71, k_1 = 8$ Evening: $n_2 = 111, k_2 = 16$

7.3 a) What is the best estimate for the **population proportions** π_1 and π_2?

Answer 7.3a:

π_1: _____ π_2: _____

7.3 b) Calculate the two-sided 95% **confidence interval** for the population proportion π for both samples.

*Hint 7.3: You can find the formula for a **confidence interval** for a **proportion** in the formula sheet on page 140. Use $z = 1.960$ (see also Table 4 on page 145).*

Answer 7.3b:

`Confidence interval sample 1:` _____

`Confidence interval sample 2:` _____

The success rate appears to be higher in the evening than in the afternoon. Unfortunately you cannot deduce from these estimated intervals for the population proportions whether this difference is significant. However, you can do this using a two-sample **z-test** for a **proportion** .

7.3 c) Write down the **null hypothesis** H_0 and **alternative hypothesis** H_1 for a test to find out of the evening success rate is higher than the afternoon success rate.

Answer 7.3c:

H_0: _____ H_1: _____

7.3 d) Calculate the **combined success probability** for both samples together.

Hint 7.4: You can find the formulas for the following assignments in the formula sheet on page 140.

Answer 7.3d:

Combined success probability: _____

7.3 e) Calculate the **combined standard error** for the two proportion z-test.

Answer 7.3e:

Combined standard error: _____

7.3 f) Calculate the **z-score** for the two proportion z-test.

Answer 7.3f:

z-score: _____

7.3 g) Drawn the conclusion based on the **z-score** using a **critical z-value** of 1.645 for a one-sided proportion test with 95% confidence. Include the following four elements:

- ☐ Show how the calculated **z-value** relates to the **critical z-value** .
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about π_1 and π_2.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 7.3g:

_____

_____

_____

_____

_____

_____

The commercial director does not believe your result and wants you to recalculate it in R.

Run the following code in R:

```
n <- c(71, 111)
k <- c(8, 16)
prop.test(x = k, n = n)
```

7.3 h) Interpret the results and compare the outcome with your answer for assignment 7.3g. Do your results match?

Answer 7.3h: