# 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-square 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	4.87	2.90			
Summer	3.04	4.50			
Fall	1.65	4.94			
Winter	2.88	3.28			
Total	2.31	4.73			

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

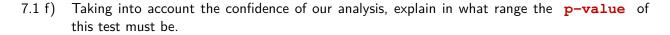
Answer 7.1b:		

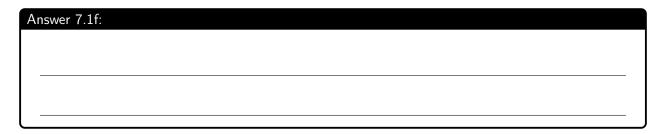
7.1 c)	Are you	allowed to do	chi-square	test	using these	data?	Explain	why.
--------	---------	---------------	------------	------	-------------	-------	---------	------



Hint 7.1: Take into account the assumptions of a chi-square test.

Answer 7.1 c:
7.1 d) Fill in the rest of the table to calculate the <b>chi-square value</b> $(X^2)$ for this sample.
Answer 7.1d:
$X^2$ :
The critical chi-square value using a confidence of 95% and 3 degrees of freedom is $X_{0.05}^2[df=3]=7.815.$
7.1 e) What is your conclusion on the basis of these results? Include the following elements:
☐ Show how the calculated Chi-square value relates to the critical Chi-square value.
$\square$ Discuss whether $H_0$ is rejected or not.
☐ Describe what this tells us about the historical <b>distribution</b> and the 2019 <b>distribution</b> .
$\Box$ Describe what type of error is relevant (type-I or type-II).
Answer 7.1e:





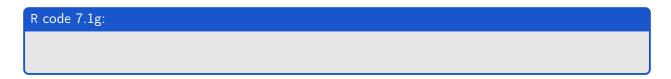
The bed and breakfast owner does not believe in your calculations and wants you to recalculate it in R.



Run the following code in R to import the data:

```
Observed <- c(37, 27, 12, 8)
Historical <- c(0.3, 0.4, 0.15, 0.15)
Expected <- c(25.2, 33.6, 12.6, 12.6)
```

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



```
Answer 7.1g: X^2:
```

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

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

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

```
Answer 7.1h:

YES / NO
```

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

```
# chi-square 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 7.1 i:			

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

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

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

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 7.1 j:			

An	swer 7.1j:				
_					
_					

#### Assignment 7.2: Chi-square 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-square**  $(X^2)$  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:

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

	test on these data?
Answe	er 7.2b:
_	
7.2 c)	Formulate the <b>null hypothesis</b> $H_0$ and the <b>alternative hypothesis</b> $H_1$ for this test in words.
Answe	er 7.2c:
$H_0$ :	:
$H_1$ :	:
7.2 d)	Perform a <b>Chi-square test</b> in R using these data and draw the conclusion for the hypotheses. Include the following elements:
	$\square$ Show how the calculated <b>Chi-square value</b> relates to the <b>critical value</b> . $\square$ Discuss whether $H_0$ is rejected or not.
	<ul> <li>□ Describe what this tells us about the historical and the new distribution.</li> <li>□ Describe what type of error is relevant (type-I or type-II).</li> </ul>
R code	e 7.2d:
Answe	er 7.2d:

7.2 b) Looking at the graph and considering the values in each month, can you perform a Chi-square

### 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  $(n_1 \text{ and } n_2)$ , one in the afternoon shift and one in the evening shift, and records the number of calls that lead to a subscription  $(k_1 \text{ and } k_2)$ :

Afternoon:  $n_1 = 71, k_1 = 8$ 

Evening:  $n_1 = 111, k_1 = 16$ 

7.3 a) What is the best estimate for the **population proportions**  $\pi_1$  and  $\pi_2$ ?

Answer	7.3a:		
$\pi_1$ :		$\pi_2$ :	

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



Hint 7.3: You can find the formula for a **confidence interval** for a **proportion** in the formula sheet. Use z=1.960 (see also Table 4 on page 98)

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 s	uccess	probability	for both	samples together.
---------------------------------	--------	-------------	----------	-------------------



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

Answer 7.3d:
Combined success probability:
7.3 e) Calculate the <b>combined standard error</b> for the two proportion z-test.
7.3 e) Calculate the Companied Standard Error for the two proportion 2-test.
Answer 7.3e:
Combined standard error:
7.3 f) Calculate the <b>z-score</b> for the two proportion z-test.
, constant in a constant in property of
Answer 7.3f:
z-score:
7.3 g) Drawn the conclusion based on the <b>z-score</b> using a <b>critical z-value</b> of 1.645 for
one-sided proportion test with 95% confidence. Include the following four elements:
$\square$ Show how the calculated <b>z-value</b> relates to the <b>critical z-value</b> .
$\square$ Discuss whether $H_0$ is rejected or not.
$\square$ Describe what this tells us about $\pi_1$ and $\pi_2$ .
$\Box$ 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)			
<pre>prop.test(x = k, n =</pre>	= n)		

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

Ar	iswer 7.3h:			