Chapter 3: Confidence intervals and hypothesis testing

Learning objectives of this chapter:

- Confidence interval and hypothesis testing by hand
- Testing the homogeneity of variance by hand
- Calculating confidence intervals for the mean and testing for normality
- Inspect and test the normality of a sample in R
- Using R to inspect and test the homogeneity of variance

Assignment 3.1: Confidence interval and hypothesis testing by hand

A call center with 38 employees handles thousands of calls a day. The company wants to get more insights into the duration of calls and the workload of the employees. On a day with a total of 4513 calls, the company decides to randomly pick 100 calls and measure how long they take. The **mean** call duration for this sample is 145 seconds with a **standard deviation** of 25 seconds. The call center wants to use this information to estimate the **mean** call duration for that day.

3.1 a) Write down the relevant information from this case. Use the symbols N, n, $\bar{x}$, s, σ, and μ.

Hint 3.1: Not all symbols are known and some should be left empty.

Answer 3.1a:

N: _____ s: _____

n: _____ σ: _____

$\bar{x}$: _____ μ: _____

3.1 b) What is the best estimate of the **mean** call duration (μ) based on this sample?

Answer 3.1b:

$\mu =$ _____

Explanation:

_____

3.1 c) Calculate the **standard error** of the mean (SE_μ) based on this sample.

Hint 3.2: Check the formula sheet on page 111 to find out how to calculate SE_μ.

Answer 3.1c:

$SE_\mu =$ _____

Calculation:

_____

3.1 d) Calculate the 99% **confidence interval** for the mean call duration.

*Hint 3.3: You can find the formula for the **lower bound**, **upper bound** and **z-value** in the formula sheet on page 111.*

Answer 3.1d:

z-value: _____

Calculation: _____

Lower bound: _____

Calculation: _____

Upper bound: _____

Calculation: _____

The manager of the call center does not really care about the **lower bound** and he does not require such high confidence. He just wants to make sure the average workload per employee is not too high, because otherwise he is forced by employment laws to hire more people. He calculated that if the **mean** call duration is below 150 seconds the workload is acceptable, and wants to use this sample to show with 95% confidence that the **mean** call duration is below 150 seconds.

3.1 e) Write down the hypotheses H_0 and H_1 for the manager's test. What is the value of μ_0? Also describe μ_0 for this case.

Hint 3.4: This is a one-sided test. Consider this when you formulate the hypotheses.

Answer 3.1e:

μ_0: _____

H_0: _____ H_1: _____

3.1 f) Do you need the **lower bound** or the **upper bound** of the **confidence interval** for this test? Calculate this bound.

Answer 3.1f:

`........` `bound:` _____

`Calculation:` _____

3.1 g) Draw the conclusion for the manager. Include the following four elements:

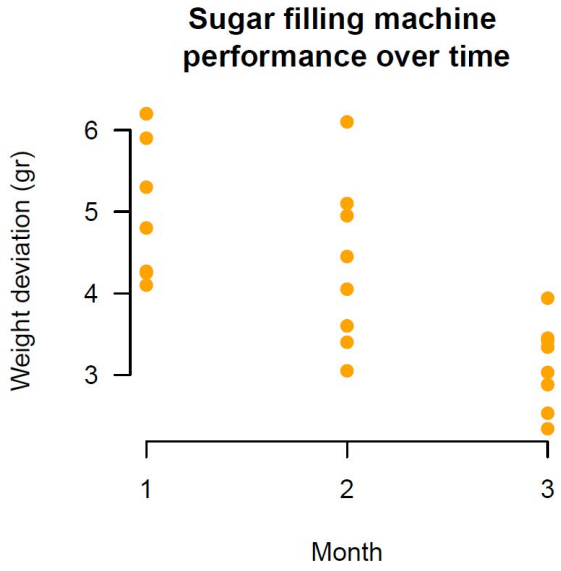
- ☐ Show how the **confidence interval** relates to μ_0.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about μ and μ_0.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 3.1g:

Assignment 3.2: Testing for homogeneity of variance by hand

A manufacturer of machines for food processing has created a machine that fills bags with sugar. A machine like this is never perfect and when a 1-kilo bag of sugar is filled there are always small deviations. The actual content of a bag is often a few grams more or less than a kilo. The manufacturer knows these deviations are bigger for new machines but become smaller as the machine is longer used due to better calibration and the smoothing out of the moving parts. To be able to show this to his clients they take a random sample of eight bags of sugar on the first day of the three months the machine is in use and plots the (absolute) weight deviations.

The manufacturer wants to use the chart below for a model that shows how much the machine improves over time, but this requires **homogeneity of variances**. The **variance** in month 1 is 0.801 gr, the **variance** in month 2 is 1.113 gr. The **variance** for month 3 has not been calculated yet.



i	x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
1	3.03		
2	3.45		
3	3.94		
4	2.34		
5	3.34		
6	2.53		
7	2.88		
8	3.42		

$\sum x_i =$ ..... $\sum (x_i - \bar{x})^2 =$ .....

$\bar{x} =$ ..... $s^2 =$ .....

3.2 a) Use the table to the right of the figure to calculate the **variance** s^2 for month 3 by filling in the calculations beneath the table.

Hint 3.5: you can find the formula for the **variance** *in the formula sheet on page 111.*

Answer 3.2a:

Variance: _____

3.2 b) Calculate **Hartley's F** (the variance ratio) for the sugar machine data set.

Hint 3.6: you can find the formula for **Hartley's F** *in the formula sheet on page 111.*

Answer 3.2b:

Hartley's F: _____

3.2 c) Formulate the **null hypothesis** H_0 and **alternative hypothesis** H_1 to test homogeneity of variance for this case.

Answer 3.2c:

H_0: _____ H_1: _____

3.2 d) Determine the **critical value** for Hartley's F ($F_{\max}$) for the sugar machine data set.

Hint 3.7: Check Table 1 on page 113 for the critical values of ***Hartley's F*** .

Answer 3.2d:

Hartley's $F_{\max}$: _____

3.2 e) Draw the conclusion on the homogeneity of variance for the sugar machine data set. Include the following elements:

- ☐ Show how the calculated **Hartley's F** relates to the critical value F_{max}.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about the homogeneity of the three variances.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 3.2e:

Assignment 3.3: Calculating a confidence interval and testing for normality

For this assignment, you will need the **`populations.csv`** data file, which contains four different **`populations`** of 10,000 observations called **`P1 - P4`** .

3.3 a) Use the **`read.csv()`** function (and **`setwd()`** function if you prefer) to import the data set into an object named **`dataset5`** .

R code 3.3a:

3.3 b) Run the following code in R and explain what it does. Why do you have to use a seed?

```
set.seed(54321) # You can replace 54321 with your own seed number

sample1 <- sample(dataset5$P1, size = 90)
sample2 <- sample(dataset5$P2, size = 90)
sample3 <- sample(dataset5$P3, size = 90)
sample4 <- sample(dataset5$P4, size = 90)
```

Hint 3.8: Use R's help function (the **`?`** *symbol) to find out more information about the* **`sample()`** *and* **`set.seed()`** *functions.*

Answer 3.3b:

3.3 c) Calculate the **`mean`** and **`standard deviation`** for each **`sample`** and save them into variables **`x1`** - **`x4`** and **`s1`** - **`s4`** . Use these results to also calculate the **`standard errors`** and same them into variables **`se1`** - **`se4`** .

R code 3.3c:

Answer 3.3c:

```
Standard error sample 1:  _____________

Standard error sample 2:  _____________

Standard error sample 3:  _____________

Standard error sample 4:  _____________
```

The R function **qnorm(p, mean, sd)** returns the **z-value** for **quantile p** from a **normal distribution** with a certain **mean** μ (**mean**) and **standard deviation** σ (**sd**). If you do not specify a mean or standard deviation for the function it will assume the standard normal distribution $N(\mu = 0, \sigma = 1)$.

3.3 d) Run the following code in R and explain the value that you see.

```
qnorm(p = 0.95)
```

Answer 3.3d:

The **qnorm()** function cannot return the **z-value** for the two-sided **confidence interval** , but you can work around that by realizing that in a two-sided interval you have to split the risk over both sides of the standard normal distribution.

3.3 e) Run the following code in R and explain why you can use the value in **z** for a two-sided 95% **confidence interval** .

```
z <- qnorm(p = 0.975)
```

Answer 3.3e:

3.3 f) Calculate the 95% **confidence interval** for each sample. Store the lower bounds **lb1** - **lb4** and store the upper bounds in **ub1** - **ub4** .

R code 3.3f:

This is a rare case in which you actually have the full **populations**, so you can check if the estimates based on your **samples** are actually close to the real value in the populations.

3.3 g) Calculate the actual **population means**, call them **mu1** - **mu4**. Next, fill the schema below with all known values for **ub**, **x**, **lb**, and **mu**.

R code 3.3g:

Answer 3.3g:

	sample1	sample2	sample3	sample4
ub				
x				
lb				

	P1	P2	P3	P4
mu				
mu in interval?	YES / NO	YES / NO	YES / NO	YES / NO

3.3 h) Are all **population means** inside the interval you calculated?

Answer 3.3h:

YES / NO

3.3 i) How often do you expect this to happen given the confidence level (95%) used?

Answer 3.3i:

Assignment 3.4: Inspecting and testing the normality of a sample in R

In this assignment you are going to work with an R package called **car** . Packages are extensions for R created by its community. They contain functions that are not available in the basic R version and can be really useful. There are many packages on the internet, so only use the one you really need and always make sure you use packages from a reliable source (like the official **CRAN** servers). For clarity, it will be indicated what functions come from what packages in the chapters. If there is no package mentioned, functions come from base R.

A package has to be installed once. Run the following code in R to install the car package:

```
install.packages('car')
```

A package needs to be loaded in **every script** that uses functions from this package.

```
library(car)
```

This assignment assumes you have done assignment 3.3 and therefore have four $n = 90$ samples called **sample1** - **sample4** from populations **P1** - **P4** from **dataset5** and calculated the **confidence intervals** for the population means.

3.4 a) Create a histogram for each sample in **sample1** - **sample4** .

R code 3.4a:

```

```

3.4 b) Which samples look like they could have been taken from a **normal distribution** ?

Answer 3.4b:

3.4 c) Use this R code (from the **car** package) to create four **qq-plots** for the four **samples**.

```
qqPlot(sample1, distribution = 'norm') # qqPlot for sample 1
```

R code 3.4c:

3.4 d) Evaluate each **qq-plot** and explain the deviations from the diagonal by referencing features of the histograms. Do the **qq-plots** confirm your answer for question 3.4b?

Answer 3.4d:

3.4 e) Formulate the **null hypothesis** H_0 and **alternative hypothesis** H_1 for a test of normality in these samples.

Answer 3.4e:

H_0: _____ H_1: _____

3.4 f) Use the following R code to perform a **Shapiro-Wilk** normality test on the four samples. Use a confidence level of 99% and write down the conclusion for each sample.

```
shapiro.test(sample1) # Normality test for sample 1
```

Hint 3.9: First think about what ***p-values*** *lead to rejecting the* ***null hypothesis*** H_0.

R code 3.4f:

Answer 3.4f:

In assignment 3.3 you used these **samples** and the **normal distribution** to estimate the **confidence interval** for the **population mean** . You also tested whether the actual population means were inside these intervals, which they almost always were.

Now you found out that most of these samples are actually not normally distributed at all.

3.4 g) If a sample is not normally distributed, does that mean you cannot use it to estimate the population mean? Explain your answer.

Answer 3.4g:

When the sample is not normally distributed, the sample **can** / **cannot** be used to estimate the population mean.

Explanation:

Assignment 3.5: Using R to inspect and test the homogeneity of variance

In this assignment, you will use the **iris** data set that is built into R. It is assumed that you have also installed the **car** package (from assignment 4.4).

Run the following code in R to load the **car** library and load the **iris** data set into the environment.

```
library(car)
data(iris)
```

3.5 a) Use the following code to create a **box plot** for the sepal length (**Sepal.Length**) per flower species (**Species**). Then rewrite the code to do the same for the sepal width (**Sepal.Width**) per flower species.

```
plot(x = iris$Species, y = iris$Sepal.Length,
     col = 'grey', main = 'Sepal Length')
```

R code 3.5a:

3.5 b) Visually inspect the graphs for **homogeneity of variance** . What can you say about the spread of the sepal length within the different iris species?

Answer 3.5b:

3.5 c) Formulate the **null hypothesis** H_0 and **alternative hypothesis** H_1 to test for **homogeneity of variance** in these samples.

Answer 3.5c:

H_0 : _____ H_1 : _____

3.5 d) Use the (**car** package) function **leveneTest()** to test the **homogeneity of variance** for the sepal length and width of the three species. Evaluate the hypotheses with a 90% confidence. Include the following elements:

- ☐ Discuss what the **p-value** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about the homogeneity of the three variances.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Hint 3.10: Look up **?leveneTest** *in R's help function (it is now also present because you have loaded the* **car** *package).*

R code 3.5d:

Answer 3.5d: