# Chapter 4

# Statistics by Simulation (solutions to exercises)

# Contents

# Import Python packages

```
# Import all needed python packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.formula.api as smf
import statsmodels.api as sm
```

# 4.1 Reliability: System lifetime (simulation as a computation tool)

### |||| Exercise 4.1 Reliability: System lifetime (simulation as a computation tool)

In this exercise we will practice to do our own simulation.

You should use Python to solve this exercise and getting help from internet and generative AI is allowed. In an exam situation you should be able to read and understand code that carries out a simulation. Therefore it is also a good idea to study the solution for this exercise (after you have made your own solution) - you do not need to make your own code in the exact same way, but it is good to practice reading and understanding code from solutions (and from the book in general).

A device comprising of three components; "A", "B" and "C", is only functioning as long as all three components are functioning individually. The lifetime (in months) of the individual components are assumed to follow exponential distributions: The average lifetime of components of type "A" is 2 months, the average lifetime of components of type "B" is 3 months and the average lifetime of components of type "C" is 5 months. Hence we can describe the lifetime of each component type with the random variables, $X_A$, $X_B$ and $X_C$ that follow exponential distributions:

$$X_A \sim Exp(\lambda = 1/2)$$
$$X_B \sim Exp(\lambda = 1/3)$$
$$X_C \sim Exp(\lambda = 1/5)$$

Describing the overall lifetime of devices is however not as straight forward. We will therefore simulate the situation.

a) Simulate a large number of devices (at least 1000 – go for 10000 or 100000 if your computer is up for it). For each device you will need to simulate the lifetimes of component A, B and C. Store this information in a DataFrame. In your simulation, how long does the first three devices stay functioning?

b) To check our simulation, compute the average lifetime of each component type. Do these values match your expectations?

We now have a simulated dataset that we can play with. We can compute the overall lifetime for each simulated device, and investigate how this *simulated random variable* behaves:

c) For each device calculate the device lifetime and store this value in a new column in your DataFrame. Make a histogram of the device lifetimes (hence investigating the distribution of this *simulated random variable*).

(hint: consider how the random variable $Y$ = "device lifetime" is a function of the three component lifetimes: is it the sum, the mean, the median, the minimum, the maximum, the range or something else?)

d) Estimate the mean device lifetime from the simulated data.

e) Estimate the standard deviation of system lifetimes from the simulated data.

f) Estimate the probability that the system fails within 1 month.

g) Estimate the median system lifetime from the simulated data.

h) Estimate the 10th percentile of system lifetimes from the simulated data.

i) What seems to be the distribution of system lifetimes? (histogram etc)

## 4.2 Basic bootstrap CI

### |||| Exercise 4.2 Basic bootstrap CI

(Can be handled without using R) The following measurements were given for the cylindrical compressive strength (in MPa) for 11 prestressed concrete beams:

$$38.43, 38.43, 38.39, 38.83, 38.45, 38.35, 38.43, 38.31, 38.32, 38.48, 38.50.$$

1000 bootstrap samples (each sample hence consisting of 11 measurements) were generated from these data, and the 1000 bootstrap means were arranged on order. Refer to the smallest as $\bar{x}^*_{(1)}$, the second smallest as $\bar{x}^*_{(2)}$ and so on, with the largest being $\bar{x}^*_{(1000)}$. Assume that

$$\begin{aligned}
\bar{x}^*_{(25)} &= 38.3818,\\
\bar{x}^*_{(26)} &= 38.3818,\\
\bar{x}^*_{(50)} &= 38.3909,\\
\bar{x}^*_{(51)} &= 38.3918,\\
\bar{x}^*_{(950)} &= 38.5218,\\
\bar{x}^*_{(951)} &= 38.5236,\\
\bar{x}^*_{(975)} &= 38.5382,\\
\bar{x}^*_{(976)} &= 38.5391.
\end{aligned}$$

a) Compute a 95% bootstrap confidence interval for the mean compressive strength.

b) Compute a 90% bootstrap confidence interval for the mean compressive strength.

## 4.3 Various bootstrap CIs

### |||| Exercise 4.3 Various bootstrap CIs

Consider the data from the exercise above. These data are entered into Python as:

```
x = np.array([38.43, 38.43, 38.39, 38.83, 38.45, 38.35,
       38.43, 38.31, 38.32, 38.48, 38.50])
```

Now generate $k = 1000$ bootstrap samples and compute the 1000 means (go higher if your computer is fine with it)

a) What are the 2.5%, and 97.5% quantiles (so what is the 95% confidence interval for $\mu$ without assuming any distribution)?

b) Find the 95% confidence interval for $\mu$ by the parametric bootstrap assuming the normal distribution for the observations. Compare with the classical analytic approach based on the $t$-distribution from Chapter **??**.

c) Find the 95% confidence interval for $\mu$ by the parametric bootstrap assuming the log-normal distribution for the observations. (Help: To use the `stats.lognorm.rvs` function to simulate the log-normal distribution, we face the challenge that we need to specify the mean and standard deviation on the log-scale and not on the raw scale, so compute mean and standard deviation for log-transformed data for this Python-function)

d) Find the 95% confidence interval for the lower quartile $Q_1$ by the parametric bootstrap assuming the normal distribution for the observations.

e) Find the 95% confidence interval for the lower quartile $Q_1$ by the non-parametric bootstrap (so without any distributional assumptions)

## 4.4 Two-sample TV data

### |||| Exercise 4.4 Two-sample TV data

A TV producer had 20 consumers evaluate the quality of two different TV flat screens - 10 consumers for each screen. A scale from 1 (worst) up to 5 (best) were used and the following results were obtained:

| TV screen 1 | TV screen 2 |
|---|---|
| 1 | 3 |
| 2 | 4 |
| 1 | 2 |
| 3 | 4 |
| 2 | 2 |
| 1 | 3 |
| 2 | 2 |
| 3 | 4 |
| 1 | 3 |
| 1 | 2 |

a) Compare the two means without assuming any distribution for the two samples (non-parametric bootstrap confidence interval and relevant hypothesis test interpretation).

b) Compare the two means assuming normal distributions for the two samples - without using simulations (or rather: assuming/hoping that the sample sizes are large enough to make the results approximately valid).

c) Compare the two means assuming normal distributions for the two samples - simulation based (parametric bootstrap confidence interval and relevant hypothesis test interpretation – in spite of the obviously wrong assumption).

## 4.5 Non-linear error propagation

### |||| Exercise 4.5 Non-linear error propagation

Solved question (a)-(c) without using Python, but use Python to simulate in question (d).

The pressure $P$, and the volume $V$ of one mole of an ideal gas are related by the equation $PV = 8.31T$, when $P$ is measured in kilopascals, $T$ is measured in kelvins, and $V$ is measured in liters.

a) Assume that $P$ is measured to be 240.48 kPa and $V$ to be 9.987 L with known measurement errors (given as standard deviations): 0.03 kPa and 0.002 L. Estimate $T$ and find the uncertainty in the estimate.

b) Assume that $P$ is measured to be 240.48kPa and $T$ to be 289.12K with known measurement errors (given as standard deviations): 0.03kPa and 0.02K. Estimate $V$ and find the uncertainty in the estimate.

c) Assume that $V$ is measured to be 9.987 L and $T$ to be 289.12 K with known measurement errors (given as standard deviations): 0.002 L and 0.02 K. Estimate $P$ and find the uncertainty in the estimate.

d) Try to answer one or more of these questions by simulation (assume that the errors are normally distributed).

## 4.6 Simulation of statistical power

### |||| Exercise 4.6 Simulation of statistical power

This exercise should be solved without using Python.

A company wants to measure the effect of a newly developed coating, that may improve shelf-life of their product. Without the coating the product has an average shelf-life of 27 days. They decide that the new coating will be worth the extra expense if the average shelf-life is increased by 1.5 days or more. But they also know that the shelf-life of their products are subject to rather large variability and they expect the standard deviation to be of order $\sim 5$ days.

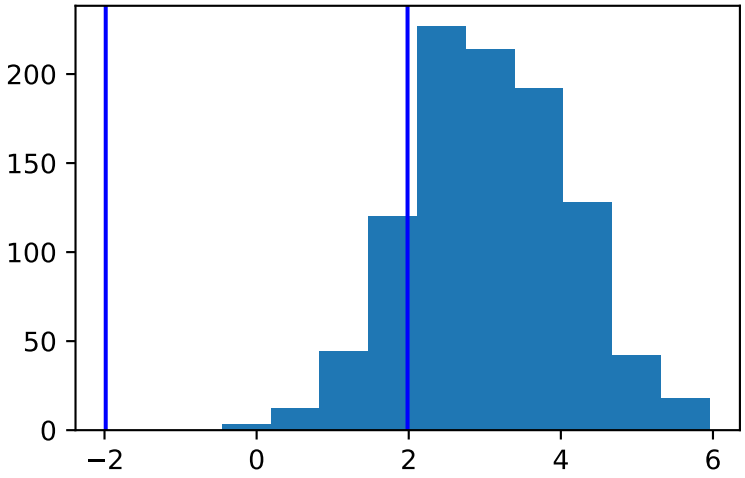
The company plans to make an experiment where they apply the new coating to 100 products and measure the shelf-life of these. But before they initiate the actual experiment they decide to simulate the risk of not detecting any significant improvement, if the true improvement is only 1.5 days (or less). They assume that the measured shelf-life, $X$, follows a normal distribution:

$$X \sim N(\mu = 28.5, \sigma^2 = 5^3)$$

The company carries out the following simulation in Python:

```
np.random.seed(412)

# Number of simulations
k = 1000

# sample size
n = 100

# Simulation
shelf_life_samples = stats.norm.rvs(loc=28.5, scale=5, size=(k,n))

# Simulation statistics:
sim_means = shelf_life_samples.mean(axis=1)
sim_SEM = shelf_life_samples.std(ddof=1, axis=1)/np.sqrt(n)

# Simulation results
sim_tobs = (sim_means - 27)/sim_SEM
tcritical = stats.t.ppf(0.975, df=n-1)
```

```
reject_null_hyp = np.abs(sim_tobs) >= tcritical

plt.hist(sim_tobs)
plt.axvline(-tcritical, linestyle='-', color="blue", ymin=0, ymax=1)
plt.axvline(tcritical, linestyle='-', color="blue", ymin=0, ymax=1)
plt.show()
```




```
n_rejected_null_hyp = np.sum(reject_null_hyp)
print(n_rejected_null_hyp)

858
```

Hint: Notice the variable `reject_null_hyp` is a boolean variable (`True` or `False`). Here we print the first 10 values in the variable `reject_null_hyp` and the first 10 values of the variable `sim_tobs`:

```
print(reject_null_hyp[:10])

[ True  True  True  True  True  True  True False False  True]

print(sim_tobs[:10])

[2.0167269  2.44449889 2.86036866 2.37276581 3.84951488 2.97859893
 2.83082468 1.67866428 1.37293584 3.29996005]
```

a) How many experiments (samples) were simulated?

b) What is "`shelf_life_samples`"? (what are the dimensions of this variable?)

c) What is "`sim_SEM`"?

d) Explain the hypothesis test carried out in the simulation (and the histogram).

e) What is the conclusion from the simulation - regarding the risk of not detecting any significant improvement?

f) Perform the theoretical calculation of needed sample size if the company wants a power of 0.90 (and using $\alpha = 0.05$).

For this calculation you will need one or more of the following values:

```
print(stats.norm.ppf(0.025))

-1.9599639845400545

print(stats.norm.ppf(0.05))

-1.6448536269514729

print(stats.norm.ppf(0.10))

-1.2815515655446004

print(stats.norm.ppf(0.80))
```

```
0.8416212335729143

print(stats.norm.ppf(0.90))

1.2815515655446004

print(stats.norm.ppf(0.95))

1.6448536269514722

print(stats.norm.ppf(0.975))

1.959963984540054
```