Chapter 4

Statistics by Simulation (solutions to exercises)

Contents

Import Python packages

```python
# 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)

A system consists of three components A, B and C serially connected, such that A is positioned before B, which is again positioned before C. The system will be functioning only so long as A, B and C are all functioning. The lifetime in months of the three components are assumed to follow exponential distributions with means: 2 months, 3 months and 5 months, respectively (hence there are three random variables, X_A, X_B and X_C with exponential distributions with $\lambda_A = 1/2$, $\lambda_B = 1/3$ and $\lambda_C = 1/5$ resp.). A little Python-help: You will probably need (or at least it would help) to put three variables together to make e.g. a $k \times 3$-matrix – this can be done by the `cbind` function:

```
x = np.column_stack((xA,xB,xC))
```

And just as an example, in Python we can easily compute e.g. the mean of the three values for each of all the k rows of this matrix by using the "axis" argument in the `np.mean` function. This argument specifies the axis along which the means are computed, so for axis=1, we take the mean for the three columns in the x. Many functions in Python have this argument, so it is a good idea to get familiar with it. Example for mean:

```
simmeans = np.mean(x, axis=1)
```

a) Generate, by simulation, a large number (at least 1000 – go for 10000 or 100000 if your computer is up for it) of system lifetimes (hint: consider how the random variable Y = System lifetime is a function of the three X-variables: is it the sum, the mean, the median, the minimum, the maximum, the range or something even different?).

b) Estimate the mean system lifetime.

c) Estimate the standard deviation of system lifetimes.

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

e) Estimate the median system lifetime

f) Estimate the 10th percentile of system lifetimes

g) 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 μ without assuming any distribution)?

b) Find the 95% confidence interval for μ by the parametric bootstrap assuming the normal distribution for the observations. Compare with the classical analytic approach based on the t-distribution from Chapter 2.

c) Find the 95% confidence interval for μ by the parametric bootstrap assuming the log-normal distribution for the observations. (Help: To use the `np.random.lognormal` 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

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).