Chapter 4

Statistics by Simulation (solutions to exercises)

Contents

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 R-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 <- cbind(xA,xB,xC)
```

And just as an example, remember from the examples in the chapter that the way to easily compute e.g. the mean of the three values for each of all the k rows of this matrix is:

```
simmeans <- apply(x, 1, mean)
```

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?).

|||| Solution

Note that the lifetime can be seen as the minimal value of the three random component lifetimes:

$$\text{"Lifetime"} = \min(X_A, X_B, X_C).$$

First, note that the generated solution below has been generated with this seed in order to get the same result each time. Note, that when a simulation analysis is carried out, this number should only be set once and set randomly (potentially it is possible to find a seed (see Remark 2.12) that gives a rare simulation result and thus showing a "wrong" result, however if k is high enough this is very hard). The solution below has been generated with the following seed

```
## You might want to set the seed to achieve a particular result
set.seed(82719)
```

The following R-code generates 10.000 simulated system lifetimes:

```
## Number of simulations
k <- 10000
## Generating k component A lifetimes
xA <- rexp(k,1/2)
## Checking the mean of these
mean(xA)

[1] 2.018


## Generating k component B lifetimes
xB <- rexp(k,1/3)
## Checking the mean of these
mean(xB)

[1] 2.998


## generating k component C lifetimes
xC <- rexp(k,1/5)
## Checking the mean of these
mean(xC)

[1] 5.046


# Putting these three sets of k lifetimes together into a
# single k-by-3 matrix:
x <- cbind(xA,xB,xC)

# Finding the minimum value of the three components
# in each of the k situations:
lifetimes <- apply(x,1,min)
```

|||| Solution

Let us have a look at these simulated lifetimes:

```
## Histogram of the simulated lifetimes
hist(lifetimes, col = "blue", nclass = 30)
```

b) Estimate the mean system lifetime.

|||| Solution

```
## The estimated mean lifetime
mean(lifetimes)

[1] 0.974
```

c) Estimate the standard deviation of system lifetimes.

|||| Solution

```
## The estimated std. dev. of the lifetime
sd(lifetimes)

[1] 0.9842
```

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

|||| Solution

We need to count how often the lifetimes are smaller than or equal to 1 month – this can in R be achieved by use of a logical operator:

```
## The fraction of times the simulated lifetime was below or equal 1
mean(lifetimes <= 1)

[1] 0.6437
```

In R `FALSE` is a 0 and a `TRUE` is a 1 - this is why we can simply apply the `mean` function directly on the vector of TRUES and FALSES like this.

e) Estimate the median system lifetime

|||| Solution

```
## The estimated median lifetime
median(lifetimes)

[1] 0.6731
```

f) Estimate the 10th percentile of system lifetimes

|||| Solution

```
## The estimated 10% quantile
quantile(lifetimes, 0.10)

   10%
0.1007
```

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

|||| Solution

We already made the histogram above. It appears that the minimum of the three exponential variables also has a distribution that looks like an exponential. In fact, there is a theoretical result (beoynd the syllabus of this course) that states that the distribution of the minimum of these three exponential distributions is again an exponential distribution but now with

$$\lambda_{min} = \lambda_A + \lambda_B + \lambda_C = 1/2 + 1/3 + 1/5 = 31/30.$$

Note how this matches nicely with the found mean above!

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.

|||| Solution

Looking at Method box 4.10, we see that we need to find the 2.5%, and 97.5% quantiles of the 1000 bootstrap samples. According to the rule for finding the 2.5% quantile this should be the average of the 25th andn the 26th observation:

$$q_{0.025} = \frac{\bar{x}^*_{(25)} + \bar{x}^*_{(26)}}{2} = 38.3818,$$

and similarly

$$q_{0.975} = \frac{\bar{x}^*_{(975)} + \bar{x}^*_{(976)}}{2} = \frac{38.5382 + 38.5391}{2} = 38.5387,$$

and hence the 95% bootstrap confidence band is:

$$[38.3818; 38.5387].$$

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

|||| Solution

As above we get:

$$q_{0.05} = \frac{\bar{x}^*_{(50)} + \bar{x}^*_{(51)}}{2} = \frac{38.3909 + 38.3919}{2} = 38.3914,$$

and similarly:

$$q_{0.95} = \frac{\bar{x}^*_{(950)} + \bar{x}^*_{(951)}}{2} = \frac{38.5218 + 38.5236}{2} = 38.5227,$$

and hence the 90% bootstrap confidence band is:

$$[38.3914; 38.5227].$$

4.3 Various bootstrap CIs

|||| Exercise 4.3 Various bootstrap CIs

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

```
x <- c(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)?

|||| Solution

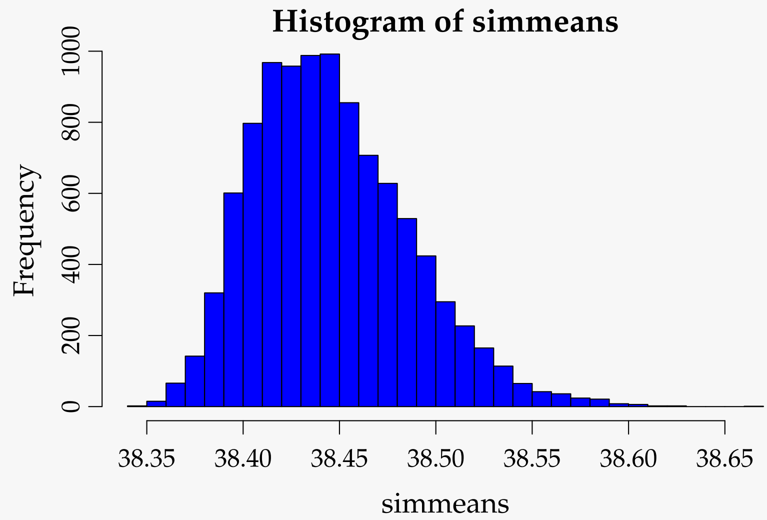
The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(6287)
```

```
x <- c(38.43, 38.43, 38.39, 38.83, 38.45, 38.35,
       38.43, 38.31, 38.32, 38.48, 38.50)
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))

 2.5% 97.5%
38.38 38.54
```

```
hist(simmeans, col="blue", nclass=30)
```



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.

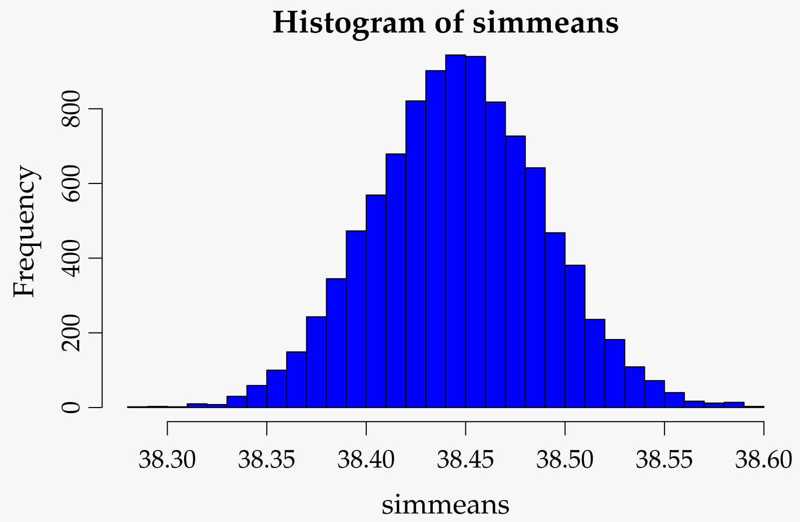
Solution

First we do the parametric bootstrap:

```
k <- 10000
n <- length(x)
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))
```

```
 2.5% 97.5%
38.36 38.53
```

```
hist(simmeans, col="blue", nclass=30)
```



And the classic t-based approach (without simulation):

```
t.test(x)


One Sample t-test

data:  x
t = 900, df = 10, p-value <2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 38.35 38.54
sample estimates:
mean of x
    38.45
```

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

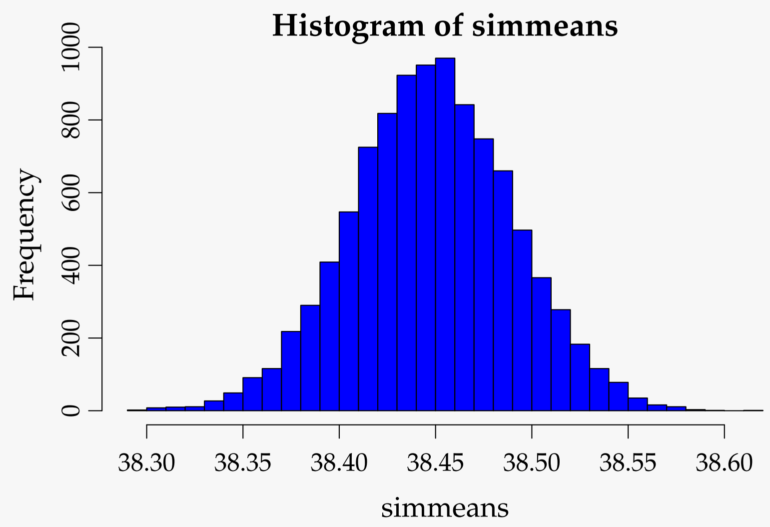
|||| Solution

We do the parametric bootstrap using the log-normal distribution.

```
k <- 10000
simsamples <- replicate(k, rlnorm(n, mean(log(x)), sd(log(x))))
simmeans <- apply(simsamples, 2, mean)
quantile(simmeans, c(0.025, 0.975))

 2.5% 97.5%
38.37 38.53
```

```
hist(simmeans, col="blue", nclass=30)
```



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

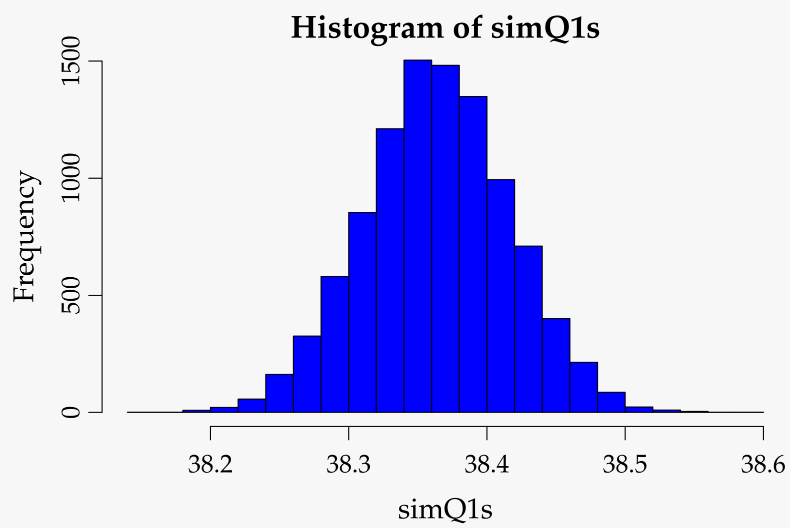
|||| Solution

We do the parametric bootstrap of lower quartile Q_1 using the normal distribution by first making a $Q1$-function in R, and then the usual stuff:

```
Q1 <- function(x){ quantile(x, 0.25) }
k <- 10000
simsamples <- replicate(k, rnorm(n, mean(x), sd(x)))
simQ1s <- apply(simsamples, 2, Q1)
quantile(simQ1s, c(0.025, 0.975))

 2.5% 97.5%
38.26 38.47

hist(simQ1s, col="blue", nclass=30)
```



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

|||| Solution

We simply substitute the sampling line with the non-parametric version:

```
k <- 10000
simsamples <- replicate(k, sample(x, replace = TRUE))
simQ1s <- apply(simsamples, 2, Q1)
quantile(simQ1s, c(0.025, 0.975))

 2.5% 97.5%
38.31 38.43
```

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).

|||| Solution

The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(98273)
```

```
x1 <- c(1, 2, 1, 3, 2, 1, 2, 3, 1, 1)
x2 <- c(3, 4, 2, 4, 2, 3, 2, 4, 3, 2)
## Number of simulated (bootstrapped) samples
k = 10000
## Simulated samples of TV1 group
simx1samples = replicate(k, sample(x1, replace = TRUE))
## Simulate samples of TV2 group
simx2samples = replicate(k, sample(x2, replace = TRUE))
simmeandifs = apply(simx1samples, 2, mean) - apply(simx2samples, 2, mean)
## The quantiles giving the 95% CI
quantile(simmeandifs, c(0.025,0.975))

 2.5% 97.5%
 -1.9  -0.5
```

We reject the null hypothesis of $\mu_1 = \mu_2$, since zero is not included in the CI of the differences.

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).

Solution

```
t.test(x1, x2)


Welch Two Sample t-test

data:  x1 and x2
t = -3.2, df = 18, p-value = 0.005
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.9987 -0.4013
sample estimates:
mean of x mean of y
      1.7       2.9
```

We reject the null hypothesis of $\mu_1 = \mu_2$.

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).

Solution

```
simx1samples <- replicate(k, rnorm(n, mean(x1), sd(x1)))
simx2samples <- replicate(k, rnorm(n, mean(x2), sd(x2)))
simmeandifs = apply(simx1samples, 2, mean) - apply(simx2samples, 2, mean)
quantile(simmeandifs, c(0.025,0.975)) # percentiles

   2.5%   97.5%
-1.9066 -0.5006
```

We reject the null hypothesis of $\mu_1 = \mu_2$.

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.

|||| Solution

This is a almost direct copy of the rectangle example ($A = XY$) (Example 4.5), since $T = PV/8.31$, so since: To use the approximate error propagation rule, we must differentiate the function $f(x,y) = xy/8.31$ with respect to both x and y:

$$\frac{\partial f}{\partial x} = y/8.31 \quad \frac{\partial f}{\partial y} = x/8.31.$$

We get the result: $\hat{T} = 240.48 \cdot 9.987/8.31 = 289.0101$, and the uncertainty is:

$$\sigma_{\hat{T}} = \sqrt{9.987^2 \times 0.03^2 + 240.48^2 \times 0.002^2}/8.31 = 0.0682.$$

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.

|||| Solution

$$V = f(P,T) = 8.31T/P.$$

So:

$$\frac{\partial f}{\partial T} = 8.31/P \quad \frac{\partial f}{\partial P} = -8.31\frac{T}{P^2},$$

and hence:

$$\hat{V} = 8.31 \cdot 289.12/240.48 = 9.9908.$$

and

$$\sigma_{\hat{V}} = 8.31\sqrt{1/240.48^2 \times 0.02^2 + 289.12^2/240.48^4 \times 0.03^2} = 0.00143.$$

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.

|||| Solution

Since

$$P = f(V,T) = 8.31T/V,$$

we can simply change the roles of P and V in the above and find similarly

$$\frac{\partial f}{\partial T} = 8.31/V \quad \frac{\partial f}{\partial V} = -8.31\frac{T}{V^2},$$

and hence

$$\hat{P} = 8.31 \cdot 289.12/9.987 = 240.5715,$$

and

$$\sigma_{\hat{P}} = 8.31\sqrt{1/9.987^2 \times 0.02^2 + 289.12^2/9.987^4 \times 0.002^2} = 0.0510.$$

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

|||| Solution

Let's look at 3. The following R-code will do the job:

The solution below has been generated with the following seed (see Remark 2.12)

```
## You might want to set the seed to achieve a particular result
set.seed(28973)
```

```
k <- 10000
Vs <- rnorm(k, 9.987, sd = 0.002)
Ts <- rnorm(k, 289.12, sd = 0.02)
Ps <- 8.31*Ts/Vs
sd(Ps)

[1] 0.05124
```

Rerunning this a few times will show that 0.051 is the proper result. This additional re-running gives a feeling of the error in the simulation - rather small here. Alternatively increase k.

Similarly 2. can be handled as:

```
k <- 10000
Ps <- rnorm(k, 240.28, sd = 0.03)
Ts <- rnorm(k, 289.12, sd = 0.02)
Vs <- 8.31*Ts/Ps
sd(Vs)

[1] 0.001432
```

Providing again basically the same answer as above: 0.0014.