Statistics with R – Exercise 1

Note: Carefully read the *Guidelines* document on the exercises first. Solve the tasks below using R and upload your solutions until Friday, the 6th of November (group registration until 31st of October). We will shortly discuss the solutions in the lecture on the 13th of November.

Task 1[1] **– Sequences** ($1 + 1 + 2 = 4$ points)

1. Create the vector `x`:

```
x

 [1]  0  2  4  6  8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40
```

2. Create a second vector `y`, which contains exactly the same elements with the same multiplicities as vector `x`, but in a *random* order. One possible solution could be

```
y

 [1] 26 38 36 28  8  4 22 14 18  0 24  6 40 10  2 34 30 20 16 32 12
```

3. On how many and which position(s) do the values of `x` and `y` agree? In the above example the correct answer would be 2 and in the positions 5 and 8 both vectors have the same values.

Task 2 – Sequences ($2 + 2 + 2 = 6$ points)
Some famous mathematicians found nice-looking approximation formulas for π. The goal of this exercise is to verify these formulas.

1. John Wallis (1616-1703):

$$\prod_{i=1}^{\infty} \left(\frac{2i}{2i-1} \cdot \frac{2i}{2i+1} \right) = \frac{2}{1} \cdot \frac{2}{3} \cdot \frac{4}{3} \cdot \frac{4}{5} \cdot \frac{6}{5} \cdot \frac{6}{7} \cdot \frac{8}{7} \cdot \frac{8}{9} \cdot \ldots = \frac{\pi}{2}$$

2. Gottfried Leibnitz (1646-1716):

$$\sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{2i-1} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \ldots = \frac{\pi}{4}$$

[1] Simply typing in the numbers on the keyboard is not a valid solution! See also the *guidelines* document.

3. Leonhard Euler (1707-1783):

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \ldots = \frac{\pi^2}{6}$$

As you can see, the formulas above involve infinite sums or products (which is of course impossible to do – even with R ☺). Instead, we choose a large $n = 10000$ and calculate the finite sum $\sum_{i=1}^{n}$ and product $\prod_{i=1}^{n}$. Which of the above formulas shows the smallest relative deviation from π for this value of n?

Task 3 – Vectors ($1 + 1 + 1 + 1 + 1 = 5$ points)

1. Create a data vector `x` containing the natural numbers $1, 2, \ldots, 100$ and a second data vector `y` containing a sample of size $n = 70$ from the set of natural numbers $1, 2, \ldots, 150$ with replacement.
2. Determine those elements of `x`, which are not contained in `y`. How many elements are these?
3. Are there duplicate entries in your vector `y`? If yes, create a new vector `z`, with these duplicates. If no, just take `y` as vector `z`.
4. How many elements of `z` are multiples of 3?
5. Revert the vector `y` without (!) using the function `rev()`, i.e. if `y` were the vector $(5, 20, 81)$, the result should be the vector $(81, 20, 5)$.

Task 4 – Point Estimation ($1 + 2 + 1 + 1 = 5$ points)
Assume you have a normally distributed population $\mathcal{N}(\mu, \sigma^2)$ (expectation μ and variance σ^2 are unknown). We would like to estimate these unknown parameters and therefore take a sample $x_1, \ldots, x_n$ of size n from this population. As you have learned in your elementary statistics lecture, you can estimate the population mean μ with the sample mean (*Stichprobenmittel*) $\overline{x}$ and the population variance σ^2 with the empirical variance s^2

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

In this exercise we simulate this situation in R.

1. Draw a reproducible sample of size $n = 30$ from a normal distribution with $\mu = 5$ and $\sigma^2 = 4$.
2. Estimate μ and σ^2 on the basis of your sample using the above formulas, i.e. without (!) using the functions `mean()`, `var()` and `sd()`.

3. Compare your results with the output of the functions `mean()` and `var()`.

4. Are your estimates close to the population values? Repeat the steps 1 and 3 from above with a sample of size $n = 3000$. What do we learn?

Task 5 – Interval Estimation $(1 + 3 + 1 = 5$ points)
Continuing from the last exercise, we also learned in the elementary statistics lecture that – instead of the above so-called *point estimates* (*Punktschätzer*) $\overline{x}$ and s^2 for μ and σ^2 – we can calculate confidence intervals for these population parameters. The formulas are

$$\text{Confidence interval for } \mu: \quad \left[\overline{x} - t_{1-\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}};\ \overline{x} + t_{1-\alpha/2;n-1} \cdot \frac{s}{\sqrt{n}}\right]$$

and

$$\text{Confidence interval for } \sigma^2: \quad \left[\frac{(n-1)\cdot s^2}{\chi^2_{1-\alpha/2;n-1}};\ \frac{(n-1)\cdot s^2}{\chi^2_{\alpha/2;n-1}}\right],$$

where $\overline{x}$, s^2 and n have the same meaning as in the last exercise, $1 - \alpha$ is the level of confidence (*Konfidenzniveau*) and $t_{\alpha;n}$ and $\chi^2_{\alpha;n}$ are the α-quantiles of the t and χ^2-distributions with n degrees of freedom. Also this situation is simulated in R:

1. Draw a sample of size $n = 30$ from a normal distribution with $\mu = 5$ and $\sigma^2 = 4$.

2. Calculate a confidence interval for μ and σ^2 for $\alpha = 0.05$ (hence the confidence level is $1 - \alpha = 0.95$). R-functions such as `mean()`, `sd()` and `var()` are allowed.

3. Do the true parameters lie in your confidence intervals?[2] If yes, is this always the case? If no, why not?

[2] Also this simple question shall be answered using R-functions, NOT by just looking at the results.