Task 1

Description to the paper "On the accuracy of statistical procedures in Microsoft Excel 97" (McCullough & Wilson, 1998)

Aim of paper: Option to use Excel 97 with double precision for statistical queries. Are the results satisfying?

Setting: Excel 97 on a Pentium computer equipped with Windows 95

Areas of testing Excel 97:

- Estimation
- statistical distributions
- Random number generator

Often used measure to approximate the number of accurate digits of an estimation:
"log relative error" (LRE):

$$\lambda = \frac{\log_{10}\left(\left|quantity_{estimated} - value_{certified}\right|\right)}{\left|value_{certified}\right|}$$

For evaluation "Statistical Reference Datasets (StRD)" with a look at:

- **Univariate summary statistics:**

  Calculation of the estimated variance of a underlying distribution...

  ...shall be: $\sigma^2_{estimated} = \frac{\sum (x_i - x_{mean})^2}{n-1}$

  ...Excel does: $\sigma^2_{estimated} = \frac{\sum x_i^2 - n x_{mean}^2}{n-1}$, which only fits to samples with few observations and small magnitudes in calculations.

  This error was also found in earlier version Excel 4.0 → No bug fixing!

- **analysis of variance:**

  ANOVA problem. Look at the LRE of the belonging F-statistic.
  Excel: good performance in easy problems but LRE=0 for an average difficult problem
  It follows: Excel uses an unstable algorithm. It should make use of symbolic methods.

- **linear regression:**

  Evaluation method: Look at the LRE for means and standard errors when using different coefficients. Only consider the both smallest/worst LRE mean/standard error values.

  Excel has a problem with the data set "Filip" which contains almost singular data matrices. Excel doesn't recognize it.
  It follows: Excel should be capable of recognizing it and should refuse to calculate.

This error was also found in earlier version Excel 4.0 → No bug fixing!

- **nonlinear regression/data sets:**

  Excel settings for finding minima:
  → Default (forward derivatives; convergence tolerance = E-3): 21 …

  for forward/central numerical derivatives:
  → convergence tolerance = E-7: 17 ...
  → automatic scaling (re-centering variables); convergence tolerance = E-3: 20 …
  → automatic scaling; convergence tolerance = E-7: 14 ...

  … out of 27 data sets got LRE=0 where Excel wrongly thought to be in a local minimum.
  It follows: Excel should recognize when not ending up in a minimum.

Random number generator (RNG):

Important because random numbers are often needed in statistical processes.
Excel is evaluated by:
- ✔ tests by Knuth (1981)
- ✗ test battery “DIEHARD” (Marsaglia, 1993)

It follows: Excel should use a better RNG.

(Task 2 next page)

Task 2
My experience, needs and interests related to statistics

Experience:

In my old studies psychology, I attended two statistics courses. They both had the aim to mediate us the used statistical methods in the field of science. Therefore, for instance we spoke about t-tests, ANOVA and the working with significance levels. The courses took into account the work in the scientific field of psychology in which you collect data yourself and the number of samples and features is seldom very large. However, also general concepts were discussed.
In my current studies Artificial Intelligence, I visited one statistics course. Here we again talked about topics which were also contained in the psychology courses, for example shortly about hypothesis testing. In some of the homeworks I got also the chance to use the programming language “R”. However, I couldn’t consolidate my knowledge about R.

Needs & interests:

- Solidifying knowledge about R without too much rush (easier programming tasks for beginners)
- Talking about statistical concepts which are useful in the working environment. Give examples for the possible usage of concepts in concrete working environments.

(Task 3 next page)

Task 3
Computer arithmetic example

a) Code in R

used code:
```
i <- 1; ende <- 2; final_sum <- 0;
while(i < ende){ende=(ende+1); current_value <- (1/2**(2*i));
if(final_sum==(current_value+final_sum)){print("Final i is: ");print(i); break};final_sum =
(current_value + final_sum); if(is.infinite(final_sum)){print("Final i is: ");print(i); break}; i=(i+1);
print(final_sum)}
```

beautified code (\n instead of “;”):
```
i <- 1
ende <- 2
final_sum <- 0

while(i < ende)
        {
        ende=(ende+1)
        current_value <- (1/2**(2*i));
        if(final_sum==(current_value+final_sum))
                {
                print("Final i is: ")
                print(i)
                break
                }
        final_sum = (current_value + final_sum)
        if(is.infinite(final_sum))
                {
                print("Final i is: ")
                print(i)
                break
                }
        i=(i+1)
        print(final_sum)
        }
```

Outcomes:
final_sum = 0.3333333
i = 28

b) Floating point representation

$$mantissa = \sum_{i=1}^{m} \frac{u_i}{b^i}$$

We have: $\sum_{i=1}^{\infty} \frac{1}{2^{2i}} = \sum_{i=1}^{\infty} \left(\frac{1}{2^2}\right)^i = \sum_{i=1}^{\infty} \left(\frac{1}{4}\right)^i$ geometic series, therefore:

$$\frac{\frac{1}{4}}{1-\frac{1}{4}} = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3} = finalSum\,2$$

Considering: $finalSum\,2 = (-1)^{u_0} b^e \sum_{i=1}^{m} \frac{u_i}{b^i}$ with

$u_0 = 0, b = 4, e = 0, m = 28\,(when\ stopping\ at\ i = 28)\ else\ \infty, u_i = 1, e_{min} < 0, e_{max} > 0$