Series 7

1. In this exercise, we consider the Wilcoxon signed rank test. Let $X_1, \ldots, X_m \sim F_X$ and $Y_1, \ldots, Y_m \sim F_Y$ be independent, where (X_i, Y_i) is measured on the same subject i. We let

$$\begin{aligned} H_0: &\quad F_X = F_Y, \text{ i.e. } F_X(x) = F_Y(x-a) \quad \forall x \text{ and } a = 0, \\ H_A: &\quad F_Y \text{ is shifted version of } F_X, \text{ i.e. } F_X(x) = F_Y(x-a) \quad \forall x \text{ and } a > 0. \end{aligned}$$

a) Under H_A, is F_Y shifted to the left or to the right compared to F_X?

b) Let $D_i = X_i - Y_i$, $i = 1, \ldots, m$. Show that H_0 and H_A can be written as

$$\begin{aligned} H_0: &\quad \text{the distribution of } D \text{ is symmetric around } a = 0, \\ H_A: &\quad \text{the distribution of } D \text{ is symmetric around } a > 0. \end{aligned}$$

You can assume that F_X and F_Y are cumulative distribution function of continuous random variables and that the densities f_X and f_Y exist. Indicate in which step you use which assumption.
Hint: Show that $D - a$ is symmetric around 0.

c) The Wilcoxon signed rank test can be viewed as a permutation test with test statistic

$$V = \sum_i \text{rank}(|D_i|) \cdot \mathbb{1}_{D_i > 0}.$$

We will conduct this test fully by hand (i.e. without computer) for the following data

	X	Y
$i = 1$	0.5	0
$i = 2$	3	2

where the test statistic can be computed as follows

	X	Y	D	$\text{rank}(\lvert D \rvert)$	$\mathbb{1}_{D_i > 0}$
$i = 1$	0.5	0	0.5	1	1
$i = 2$	3	2	1	2	1

which results in $V = 1 \cdot 1 + 2 \cdot 1 = 3$.
Under H_0, what can we permute? How many permutations are possible? Write them out and compute the test statistic in each case to obtain the permutation distribution. Verify that new group assignments of subject i correspond to a sign flip of D_i. Compute the permutation p-value for the test.

d) Compare your results to the Wilcoxon signed rank test in R.

2. Consider the two variables / columns `Y1` and `Y2` of the data set `immer`. They measure the yield in the year 1931 and 1932, respectively. We omit the information that each field / observation was assigned to one of the six different locations and that one of the five different varieties of barley was grown. The farmer suspects that the yield was significantly less in the second year.

```
> require("MASS")
> ?immer
```

a) Plot the data to get a feeling for the data.

b) Conduct an approximate Wilcoxon signed rank test (use 100'000 random permutations) to test this hypothesis, programming yourself (i.e., not using the Wilcoxon test in R). What do you conclude?

c) Compare your results to the Wilcoxon signed rank test in R.

3. In this exercise, we consider polynomial regression:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i$$

where $\epsilon_i \sim 15 \cdot \big(\text{Gamma}(2,1) - 2\big)$ with the parameters shape and rate of the Gamma distribution. We want to test the global null hypothesis at level $\alpha = 0.05$:

$$\begin{aligned} H_0: &\quad \beta_1 = \beta_2 = \beta_3 = 0 \\ H_A: &\quad \text{at least one } \beta_1, \beta_2, \beta_3 \neq 0. \end{aligned}$$

a) Use the data set given on the homepage and fit a regression model using a polynomial up to degree three. Plot the data including the fitted values.

b) Compute the p-value of the global F-test. What do you conclude?
Hint: You can calculate the p-value based on the values of `summary(...)$fstatistic`.

c) Conduct a simulation to assess the type I error rate of the global F-test.
Hint: Simulate 250 new data sets under H_0 and use the following x-values:

```
> n <- 20
> x <- seq(from = -25, to = 30, length.out = n)
```

d) Conduct a simulation to assess the power of the test under the alternative $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = -0.003$, and $\beta_3 = 0.0001$.

e) We consider now the permutation test (with 1000 permutations) for the global null hypothesis. Under H_0, what permutations can be performed? Conduct the permutation test on the data set given on the homepage using the F statistic as your test statistic.

f) Conduct a simulation to assess the type I error rate and the power based on the permutation test for the global null hypothesis. Simulate 250 new data sets each for the calculation of the power and type I error rate. (This is computationally more demanding, i.e., it can take up to half an hour.)

Preliminary discussion: Friday, April 12.

Deadline: Friday, May 03.