Experimental and Statistical Methods in Biological Sciences — Exam 1
Department of Mathematics and Systems Analysis — 11.12.2018
Aalto University — J. Virta

Guidelines: The exam has 4 problems, each worth 6 points. Write complete sentences and motivate your answers properly. Each answer sheet should contain:

- Course name
- LASTNAME and FIRSTNAMES (in block letters)
- Student number
- Study program and year
- Date and signature

Allowed equipment: Writing equipment, an A4-sized note (hand-written, text only on one side, own name in the upper right corner, no need to return)

P1 (Types of studies) Explain the following terms and give a concrete example of each.

a) Observational study **(1.5p)**

b) Controlled experiment **(1.5p)**

c) Simulation study **(1.5p)**

d) Survey **(1.5p)**

P2 (Confidence intervals)

a) Let $x_1, x_2, \ldots, x_n$ be an independent and identically distributed (i.i.d.) sample from a distribution F_x and let θ be an unknown parameter of the distribution F_x. Let $\hat{\theta}$ be an estimate of the parameter θ calculated from the sample $x_1, x_2, \ldots, x_n$. Explain how to construct a 90% bootstrap confidence interval for θ. **(4p)**

b) When and why is it better to use bootstrap confidence intervals rather than exact parametric confidence intervals? **(1p)**

c) When and why is it better to use exact parametric confidence intervals rather than bootstrap confidence intervals? **(1p)**

P3 (Location testing) Researchers have collected pairs of data, (x_i, y_i), $i = 1, \ldots, n$, from a total of $n = 15$ subjects. The variable x_i describes the skill of the ith subject in a particular task before an intervention and the variable y_i the skill of the ith subject after the intervention. The researchers are interested in studying whether the intervention has an effect on the skill and plan to use either paired t-test or paired sign test.

a) State the assumptions and hypotheses (two-sided alternative) of the paired t-test. **(2p)**

b) State the assumptions and hypotheses (two-sided alternative) of the paired sign test. **(2p)**

c) Give at least three plausible reasons why the conclusions of the two tests might differ when applied to the researchers' data. **(2p)**

P4 (Linear regression)

a) Explain what is the *variance inflation factor* (VIF) of an explanatory variable. **(1p)**

b) Explain what is a *residual* in a linear model. **(1p)**

Given 100 observations of three variables, y, x_1, x_2, a linear regression model $y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon$ was fitted. The model summary from R is given in the table below and the plot on the bottom of the page shows the model residuals against the fitted values.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.0226	0.1917	20.98	0.0000
x1	2.2714	0.2804	8.10	0.0000
x2	0.9225	0.1974	4.67	0.0000

(The values in the last column of the table are so small that R rounded them down to zeroes.) Additionally, the variance inflation factors (VIF) of the two explanatory variables were 1.802 and 1.802. The coefficient of determination of the model was $R^2 = 0.7193$.

c) Give an interpretation for the estimated regression coefficient 2.2714 of the variable x_1 in the model summary. **(1p)**

d) Based on the model results (model summary and the residual plot), is the fitted model *good*? Should something be done or are the results satisfactory? Discuss at least three different aspects of the model results. **(3p)**