Statistical inference
MS-C1620 Exam
Department of Mathematics and Systems Analysis 14.4.2021
Aalto University J Kohonen

Instructions: Answer in English. Write clearly and give reasons for your answers. A number only as an answer does not yield points. The exam has 4 problems, each worth 6 points.

Write your solutions by hand, clearly, on paper (or a tablet computer), and send your solutions in PDF form to the return box on the course page. Make sure that every page shows: course code, last name, first name, student number and date.

P1 In each question, answer TRUE or FALSE. In this problem you do not need to write reasons. 1 point per item for correct answer, maximum amount of points obtainable is 6.

(a) If a distribution is symmetric around its mean, it has a positive skewness coefficient.

(b) The two-sample proportion test can be used even if the two samples have different sizes.

(c) In linear regression, correlation coefficient is the slope of the regression line.

(d) Confidence level indicates how much confidence you have in the model assumptions.

(e) A normal distribution is symmetric around its mean, but there are also other distributions that are symmetric.

(f) Bonferroni correction is a method for using linear regression with nonlinear data.

(g) The chi-squared (χ^2) distribution is skewed to the right.

(h) In hypothesis testing, the null hypothesis is rejected when getting a p-value smaller than the significance level.

P2

(a) You are testing the following hypotheses with some method of normality testing.

H_0 : The sample $x_1, \ldots, x_n$ comes from a normal distribution.
H_1 : The sample $x_1, \ldots, x_n$ does not come from a normal distribution.

Describe what it means to conduct Type I and Type II errors in this context (do not give the general definitions of Type I and II errors but instead state what they mean for this specific pair of hypotheses). **(2p)**

(b) Draw two examples of quantile-quantile (Q-Q) plots: (i) one where the sample clearly comes from a normal distribution, and (ii) one where it clearly does not. **(2p)**

(c) Name two ways besides Q-Q plot for checking/testing the normality of a sample. **(1p)**

(d) A researcher wants to model her data with Model X that makes a normality assumption. For this, she tests her data for normality and gets a p-value of 0.055 (for the hypotheses given in part a). Based on the p-value, she decides to use Model X. Can the researcher fully trust the results of the model? Explain why or why not. **(1p)**

P3 In each of the following scenarios you have an i.i.d. (independent and identically distributed) sample $x_1, \ldots, x_n$ from some distribution F. Describe (in 3–5 sentences, and including also the assumptions that your chosen methods make) how you would investigate the following research questions.

(a) Does the sample come from a distribution with median equal 0? **(2p)**

(b) Does the sample come from the standard normal distribution? **(2p)**

(c) Does the sample come from a distribution whose standard deviation is 5? **(2p)**

P4

(a) How does backward selection conduct variable selection? List also two of its drawbacks. **(3p)**

(b) Why is simply picking the model which gives the largest value of R^2 not a good idea with respect to variable selection? **(1p)**

(c) The following plot shows the LASSO coefficient profiles in a regression problem with a response variable y_i and the explanatory variables $x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}$. (If the colors are not well visible, note that the order of the profiles from top to bottom in the left end of the plot is `x5`, `x4`, `x2`, `x3`, `x1`). The optimal value of $\log(\lambda)$ given by cross-validation is shown as a vertical line.
(i) Which variable does LASSO hold as the most important one? Which is the second most important? **(1p)**
(ii) Write down (approximately) the estimated coefficients of the five predictors in the model corresponding to the optimal value of $\log(\lambda)$, and explain which variables (if any) have been left out of the model by this stage. **(1p)**