MS-C1620 Statistical Inference – Exam

Allowed equipment: Writing equipment and one-sided A4 of hand-written notes.

18.04.2019

---

1. Answer either TRUE or FALSE (**1p** per item for correct answer, maximum amount of points obtainable is 6).
   a. In two-sample proportion test the sample sizes of the two groups need not be the same.
   b. In the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the error term is usually assumed to have expected value one, $\mathrm{E}(\varepsilon_i) = 1$.
   c. The aim of descriptive statistics is to draw conclusions about a population based on a sample.
   d. The two-sample rank test (Wilcoxon rank-sum test) makes the assumption that the medians of the distributions of the two samples are the same.
   e. Bartlett's test is a normality test (that is, used to test whether a sample comes from a normal distribution).
   f. Median is a measure of scatter.
   g. LASSO can be used for variable selection.
   h. In bootstrap, the number of observations in each of the bootstrap samples is the same as the number of observations in the original sample.

---

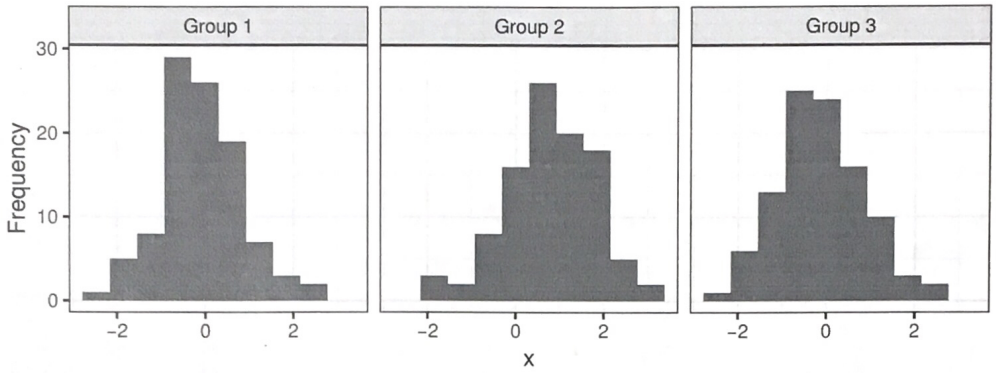
2. a. Assume you are testing the following pair of hypotheses with some method of normality testing,

   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 an example of a quantile-quantile (Q-Q) plot where:
      i. the sample clearly comes from a normal distribution, **(1p)**
      ii. the sample clearly does not come from a normal distribution. **(1p)**
   c. Name two different 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)**

---

3. Consider analysis of variance (ANOVA) on a sample of three groups with 50 observations in each of them (assume that the groups are independent and that the observations are i.i.d. within each group). On the next page are shown the histograms of the groups, ANOVA summary and the results of Bartlett's test.
   a. State the null hypothesis and the alternative hypothesis of ANOVA for this three-group case. **(1p)**
   b. What would you conclude based on the ANOVA results? **(1p)**
   c. Describe how one can check whether the assumptions of ANOVA are satisfied. Are they satisfied in the current example? **(2p)**
   d. The next step in the analysis would be to conduct pair-wise testing between the groups. Bonferroni correction is often used in this context. Why is this? Describe also how the Bonferroni correction is applied. **(2p)**




```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## group         2  57.83  28.913   31.61 3.57e-13 ***
## Residuals   297 271.63   0.915
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## ##############################################################
##
##  Bartlett test of homogeneity of variances
##
## data:  x by group
## Bartlett's K-squared = 0.85135, df = 2, p-value = 0.6533
```

---

4. a. Explain the difference between errors and residuals in a linear regression model. **(1p)**
   b. Give two uses for the residuals of a linear regression model. **(2p)**
   c. What does multicollinearity mean? **(1p)**
   d. Consider a drug experiment where
   - the continuous response y_i is the change in the amount of a specific antigen in the ith patient's blood one day after receiving the drug (higher is better),
   - the binary predictor x_{i1} describes which drug the ith patient received ($x_{i1} = 0$ for placebo, $x_{i1} = 1$ for the new experimental substance),
   - the continuous predictor x_{i2} describes the amount of the drug (placebo or the new experimental substance) the ith patient received.

   To study whether the new experimental substance is more efficient than placebo in increasing the amount of the antigen in blood, we fit the linear regression model

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}.$$

   Note: "placebo" refers to a sugar pill or something similar which should have no effect on the patient.

   i. Which null hypothesis (concerning the model coefficients $\beta_0, \beta_1, \beta_2, \beta_{12}$) should we test to determine whether the new experimental substance and the placebo are equally effective in increasing the amount of the antigen in blood? **(1p)**

   ii. It seems reasonable to assume that for those patients who received placebo, the amount of placebo received has no effect on the outcome. State this observation in terms of the model coefficients $\beta_0, \beta_1, \beta_2, \beta_{12}$. **(1p)**