MV013 Statistics for computer science

May 24, 2022

Instructions:

- Write down your solution on separate sheets of paper.
- All your results need to be explained. The result with no explanation will not be accepted. Google or Wikipedia solutions will not be accepted, as well.
- You may use `R` for computations, but all the relevant results state in your solution. **Do not** write your R-code.
- Time to complete: `100 minutes`.
- **Unreadable solutions will be ignored.**

1. **($3+3+3+4+4+4+4+5=30$ points)**

   (a) Let
   $$15,\ 18,\ 15,\ 19,\ 16,\ 15,\ 13,\ 14,\ 15,\ 18$$
   be a realisation of a random sample. Write down ranks for each observation. In case of ties, use midranks.

   (b) Give an example on two three-dimensional vectors (realisations) for which their sample correlation is 0.

   (c) Can we assume that the following data come from a normal distribution. Give **two** arguments for your opinion. Data:
   $$28,\ 17,\ 8,\ 10,\ 25,\ 11.$$

   (d) For the following data representing people's age indicate and **name** problems and suggest how to clean it. Data:
   $$37,\ 50,\ -2,\ 44,\ 157,\ 22,\ N/A,\ 39,\ 65,\ 0.$$

   (e) Let $u_1 = 0.146$ and $u_2 = 0.588$ be two independent realisations of a random variable with uniform $\mathrm{U}(0,1)$ distribution. Based on u_1 and u_2, generate a realisation of a random variable with normal $\mathcal{N}(-1,4)$ distribution. Explain your approach in detail.

   (f) Let
   $$1,\ 18,\ 9,\ 5,\ 36,\ 17,\ 11,\ 40,\ 20$$
   be a realisation of a random sample. Write down one **parametric** bootstrap sample based on the above realisation (when assuming that the data come from **exponential** distribution). Explain your approach.

(g) Draw two regression diagnostic plots for detection of heteroscedasticity in linear regression model: one indicating homoscedasticity, while the second one indicating heteroscedasticity. Do not forget to describe both axes and explain what you see in the plots.

(h) Let $X_1, \ldots, X_n$ be a random sample from normal distribution $\mathcal{N}(\mu, \sigma^2)$. Derive in detail a formula for **99% left-sided** confidence interval for parameter σ^2.

2. **(10 points)** Let

$$25,\ 1,\ 6,\ 5,\ 93$$

be a random sample from log-normal distribution with the unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. It has the following pdf:

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(\log x - \mu)^2}{2\sigma^2}}, \text{ for } x > 0.$$

(a) Write down log-likelihood function for your data and compute its value for $\mu = 3$ and $\sigma^2 = 2$.

(b) For the given $\sigma^2 = 2$, find the maximum likelihood estimate of μ (analytically or numerically).

3. **(10 points)** We asked 137 people (71 men and 66 women) about their favourite ice cream flavour. 37 women like chocolate, 17 strawberry and 12 vanilla. Among men, 21 like chocolate, 18 strawberry and 32 vanilla. The goal is to determine if there is a difference between preferences for men and women.

(a) Define the mathematical model – random variables and their distributions, null and alternative hypotheses, model assumptions. Explain all the symbols you use in detail.

(b) Is there any significant dependency between gender and ice cream flavor preference? Compute the value of the test statistic and the corresponding critical region, or p-value. You do not need to check the assumptions.

(c) Formulate the conclusion with your own words.

4. **(10 points)** For our problem, we decided to use ANOVA model. First, we checked all the assumptions and performed the analysis. Corresponding results are summarized in the following **incomplete** table:

Source of variation	Sum of squares	Df	MSE	F	p-value
Groups	25.56	2	12.78	2.45	.
Residual	244.87	47	5.21		
Total	270.43				

(a) How many groups (populations) did we compare?

(b) What is the overall sample size n?

(c) Write down **two** estimates of the common variance σ^2. Explain.

(d) Make up your own example for which such an ANOVA model might be used. Do not forget to indicate what do you test.

(e) Compute the corresponding p-value (it was intentionally left blank in the ANOVA table).

(f) What is the conclusion based on this ANOVA model? What does it mean for your example?

Warnings:

- Any suspicion of exam cheating will be reported to Educational Law, which handles such cases on behalf of the deans.
- Test questions are subject to copyright. Any unauthorized reprint or use of these without express written permission from the author represent its violation and can result in legal charges.
- In case of suspicious solutions and results, you might be subject to an additional oral exam.