Experimental and Statistical Methods in Biological Sciences
Department of Mathematics and Systems Analysis
Aalto University

Exam 2
18.2.2019
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 (**Sampling**) Explain how the following forms of sampling work. For each, give also a concrete example of its use.

a) Random sampling **(2p)**

b) Systematic sampling **(2p)**

c) Cluster sampling **(2p)**

P2 (**t-tests**) Consider the two-sample t-test and the paired t-test.

a) Give the statistical assumptions of each of the tests. **(2p)**

b) State the null hypothesis and the two-sided alternative hypothesis of each of the tests. **(2p)**

c) Explain, with examples, in which situations one should apply each of the tests. **(2p)**

P3 (**Two-way analysis of variance**) Two-way analysis of variance was performed to assess the effects of *sex* (men, women) and *age* (three age groups: 1, 2, 3) on performance in several tests (A, B, C, D, E). The corresponding line graphs for the sample means are given in the figure on the third page.

a) In which of the five tests there is one and only one main effect? B, C **(2p)**

b) In which of the five tests there is an interaction effect? D, E **(2p)**

c) In which of the five tests there are two main effects and an interaction effect? E **(2p)**

Justify your answer in each case.

P4 (**Logistic regression**)

a) Explain why the linear regression model,

$$E(y_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip},$$

is not suitable when the response variable y_i has a Bernoulli distribution. Explain also how the above model is altered in logistic regression to avoid this issue. **(2p)**

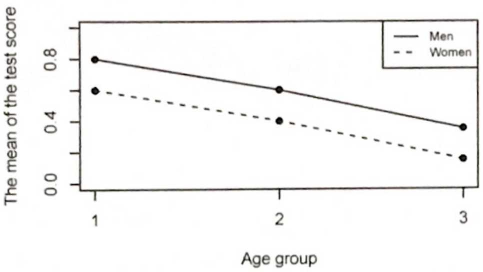
b) Explain what *odds ratio* (OR) measures and give a simple example of its use. **(2p)**

c) A data set contains the variables *Sex* (male, female) and *Survived* (yes = 1, no = 0) measured for 714 passengers onboard Titanic. We are interested in studying the relationship between *Survived* (response) and *Sex* (explanatory variable). Fitting a logistic regression model, $logit(P(Survived_i = 1)) = b_0 + b_1 Sex_i$, to the data gives the following output:

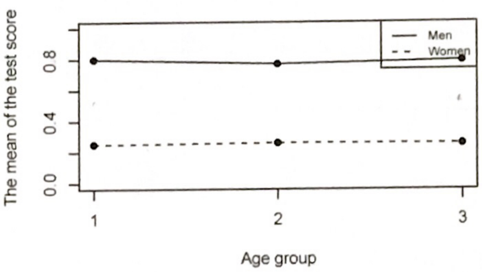
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1243	0.1439	7.81	0.0000
Sexmale	-2.4778	0.1850	-13.39	0.0000

Give an interpretation for the `Sexmale`-coefficient -2.4778 through odds ratios and state the null hypothesis related to the p-value 0.0000 in the lower right corner of the output table. *Hint:* $\exp(2.4778) \approx 11.9$. **(2p)**

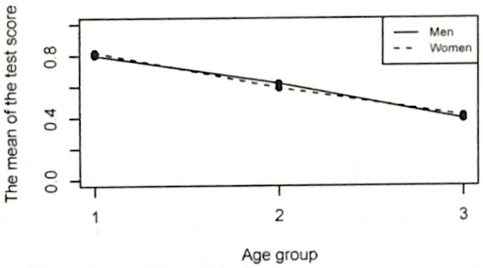
A


The mean of the test score
0.0
0.4
0.8
1
2
3
Age group
Men
Women

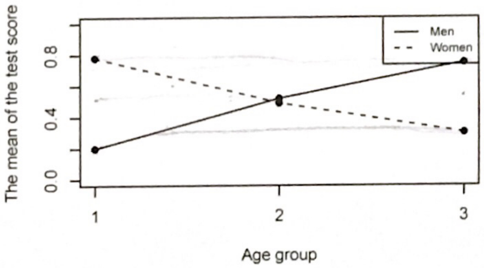
B


The mean of the test score
0.0
0.4
0.8
1
2
3
Age group
Men
Women

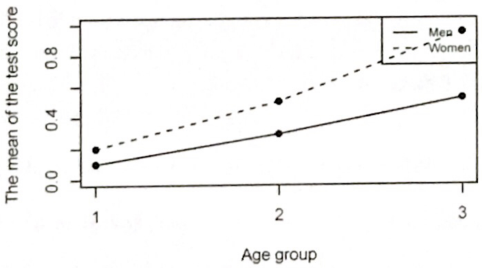
C


The mean of the test score
0.0
0.4
0.8
1
2
3
Age group
Men
Women

D


The mean of the test score
0.0
0.4
0.8
1
2
3
Age group
Men
Women

E


The mean of the test score
0.0
0.4
0.8
1
2
3
Age group
Men
Women