

## Вариант 1

**Problem 1.** 44 out of  $n_1 = 150$  customers said "Yes" to a new brand in neighborhood 1. 70 out of  $n_2 = 150$  customers said "Yes" to a new brand in neighborhood 2. Work out the 90% confidence interval for the difference between population proportions of the customers who liked the new product in neighborhoods 1 and 2.

*Вопрос 1.* Find the required confidence interval: 1) (0.0526908, 0.293976) 2) (0.0626688, 0.283998) 3) (0.0826247, 0.264042) 4) (0.103488, 0.243179) 5) Верного ответа нет, указываю свой...

*Вопрос 2.* What conclusion should be drawn?: 1) Nothing can be concluded based on the data provided 2)  $\pi_1 > \pi_2$  3) sample proportion 1 is less than sample proportion 2 4)  $\pi_2 > \pi_1$  5)  $\pi_2 + \pi_1 < 0$  6) Верного ответа нет, указываю свой...

*Вопрос 3.* What is  $\pi_1$ ,  $\pi_2$ ? 1) Population means 2) Sample proportions 3) Sample means 4) Population proportions 5) Верного ответа нет, указываю свой...

**Problem 2.** Use command `read.csv("chi2test1.csv", header=F, sep=",")` to load the data. Create a two-way table (command `table` followed by `as.matrix`) and perform the  $\chi^2$ -test. Answer the following questions:

*Вопрос 1.* State the null ( $H_0$ ) hypothesis: 1) The means of  $V_1$  and  $V_2$  are equal 2) Variables  $V_1$  and  $V_2$  are not independent 3) Variables  $V_1$  and  $V_2$  are independent 4) Variables  $V_1$  and  $V_2$  are independent 5) The means of  $V_1$  and  $V_2$  are not equal 6) Верного ответа нет, указываю свой...

*Вопрос 2.* Based on the p-value and using a 5% significance level, decide whether or not to reject  $H_0$ : 1) Depends on the sample size 2) Reject 3) Impossible to decide based on the available information 4) Not reject 5) Верного ответа нет, указываю свой...

**Problem 3.** Logistic regression. Use command `read.csv("Pred1.csv", header=F, sep=",")` to load the table. The first column stands for the binary response variable, two subsequent columns contain the explanatory variables. Divide the dataset into the training subset (first 1498 points) and the test subset (last 200 points). Fit the logistic regression model to the data. For fitting, use the training subset only. Make predictions on the test subset only. Answer the following questions:

*Вопрос 1.* Find the logistic regression model for the data: 1)  $\frac{e^{0.0707674+0.401542L_1}}{1+e^{0.0707674+0.401542L_1}}$  2)  $e^{0.0707674+0.401542L_1+0.632297L_2}$  3)  $\frac{e^{0.401542+0.0707674L_1+0.632297L_2}}{1+e^{0.401542+0.0707674L_1+0.632297L_2}}$  4)  $\frac{e^{0.401542+0.632297L_1+0.0707674L_2}}{1+e^{0.401542+0.632297L_1+0.0707674L_2}}$  5)  $\frac{e^{0.401542+0.632297L_1+0.0707674L_2}}{1+e^{0.401542+0.632297L_1+0.0707674L_2}}$  6)  $\frac{e^{0.0707674+0.401542L_1+0.632297L_2}}{1+e^{0.0707674+0.401542L_1+0.632297L_2}}$  7) Верного ответа нет, указываю свой...

*Вопрос 2.* Provide a 95% confidence interval for the coefficient  $\beta_1$ . Discuss whether  $\beta_1$  significantly differs from zero: 1) CI=(-0.167674, 0.167674), stat. signif. 2) CI=(-0.167674, 0.167674), stat. insignif. 3) CI=(-0.085548, 0.085548), stat. signif. 4) CI=(-0.261779, 0.261779), stat. signif. 5) CI=(-0.167674, 0.167674), stat. insignif. 6) Верного ответа нет, указываю свой...

*Вопрос 3.* Obtain the predicted probability for the last point in the test set: 1) 0.232 2) 0.607 3) 0.133 4) 0.494 5) Верного ответа нет, указываю свой...

*Вопрос 4.* Compute the ratio of correct predictions over the test set: 1) 0.444 2) 0.650 3) 0.472 4) 0.556 5) 0.372 6) 0.611 7) 0.261 8) Верного ответа нет, указываю свой...

*Вопрос 5.* Switch to linear discriminant analysis and compute the ratio of correct predictions: 1) 0.355 2) 0.610 3) 0.398 4) 0.249 5) 0.530 6) 0.424 7) 0.636 8) Верного ответа нет, указываю свой...

*Вопрос 6.* Classify the last point of the test set (as 0 or 1) using the linear discriminant analysis 1) 1.000 2) 2.000 3) 1.500 4) 0.000 5) Верного ответа нет, указываю свой...

*Вопрос 7.* Switch to the quadratic discriminant analysis and compute the ratio of correct predictions: 1) 0.256 2) 0.545 3) 0.600 4) 0.409 5) 0.436 6) 0.655 7) 0.365 8) Верного ответа нет, указываю свой...

*Вопрос 8.* Classify the last-but-one point of the test set (as 0 or 1) using the quadratic discriminant analysis 1) 1.500 2) 2.000 3) 1.000 4) 0.000 5) Верного ответа нет, указываю свой...

**Problem 4.** Use command `read.csv("PCA1.csv", header=F, sep=",")` to load the data. The matrix contains a number of observations (rows) of multiple variables (columns). Perform the PCA analysis and answer the following questions:

*Вопрос 1.* Using the 2D representation, find a pair of most dissimilar individuals: 1) 18 and 17 2) 17 and 10 3) 2 and 7 4) 26 and 2 5) 32 and 39 6) 37 and 42 7) Верного ответа нет, указываю свой...

*Вопрос 2.* The dissimilarity between the individuals observed in the 2D plane implies...: 1) Similarity in the scores of all of the original variables 2) Difference in the scores of all of the original variables 3) Different positions in

the ordered list of individuals 4) Difference in the scores of some specific original variable 5) Верного ответа нет, указываю свой...

*Вопрос 3.* Obtain the absolute value of the coordinate of the most distant individual (with respect to the origin) along the first PC direction (choose the closest answer): 1) 1.322 2) 3.271 3) 4.154 4) 0.277 5) 1.905 6) Верного ответа нет, указываю свой...

*Вопрос 4.* Compute the quality of representation of the 22-th(nd) individual along the 2-st(nd) PC direction. Draw a conclusion: 1) 0.326478, poor 2) 0.223233, poor 3) 0.131149, acceptable 4) 0.279041, poor 5) 0.306945, poor 6) 0.186958, poor 7) 0.326478, good 8) Верного ответа нет, указываю свой...

*Вопрос 5.* Compute the percentage of explained variance along the 2-st(nd) PC direction: 1) 0.235 2) 0.110 3) 0.258 4) 0.157 5) 0.200 6) 0.275 7) 0.188 8) Верного ответа нет, указываю свой...

*Вопрос 6.* The explained variance from Question 5 is ... 1) The variance computed over the projections of data points onto the respective PC direction 2) The variance computed over the projections of data points onto the axis corresponding to the first original variable 3) The variance computed over the projections of data points onto the plane spanned on the axes corresponding to the first two original variables 4) The variance computed over the projections of data points onto the plane spanned on the first two PC directions 5) Верного ответа нет, указываю свой...

*Вопрос 7.* Based on the 2D representation, find the most correlated variables: 1)  $V_2$  and  $V_3$  2)  $V_1$  and  $V_4$  3)  $V_5$  and  $V_3$  4)  $V_1$  and  $V_2$  5)  $V_2$  and  $V_4$  6)  $V_3$  and  $V_4$  7) Верного ответа нет, указываю свой...

*Вопрос 8.* Consider the 2D plane describing individuals. What does the origin of the coordinate frame correspond to? 1) The mean individual (over all individuals) with respect to scores of all variables 2) The mean variable with respect to all variables 3) The individual with the least variance 4) The most important individual (over all individuals) 5) Верного ответа нет, указываю свой...