STATISTICAL METHODS FOR DATA SCIENCE

MASTER BIG DATA UNIPI - 2024

FABRIZIO LILLO, VALENTINA MACCHIATI

Instructions

Create a notebook **Nome_Cognome.ipynb** in which you put your answers and upload this file on Moodle platform. The total number of points for the present test is **40**; however, you will receive the maximum score with **30**.

1. Exercise (4.5 point total. 1.5 each)

1.1. **You are renovating your house and you need a sofa. You go to the PoltroneSofa shop and find a sofa whose declared length is 180 cm. You take a tape and measure the sofa's length 18 times. The average of the measurements you obtain is 182 cm. You need to test if the true height is 180 cm. Which test should you use?**

(1) Kolmogorov-Smirnov test
(2) t-test
(3) F-test
(4) chi-squared test
(5) None of the above

1.2. **You also need two pillows of different size. Each pillow is measured 40 times. Which test should you use to test if the two variances are the same?**

(1) Kolmogorov-Smirnov test
(2) t-test
(3) F-test
(4) chi-squared test
(5) None of the above

1.3. **A group of researchers is investigating the effectiveness of three different teaching methods on improving students' performance. They conducted an experiment with 90 students randomly divided into three groups, with 30 students in each group. After the experiment, they recorded whether each student passed or failed the final exam, resulting in a contingency table. The researchers want to test the null hypothesis that the teaching method and students' exam performance are independent. Which test should they use?**

(1) Kolmogorov-Smirnov test
(2) t-test
(3) F-test
(4) chi-squared test
(5) None of the above

Time available: 75 min
Date: 9th April 2024.

2. Exercise (12 points total)

Given the results of a chemical analysis of wines grown in the same region in Italy (derived from three different cultivars), a researcher wants to analyze the relationship between the 13 constituents found. Dataset: `wine.csv`

(1) Describe the dataset; compute the average, variance, and quartiles of 'total_phenols' and 'color_intensity'. Plot them in a scatterplot. **(1.5 points)**
(2) compute the covariance between all the columns and plot the results in a heatmap **(1.5 points)**
(3) are 'total_phenols' and 'color_intensity'; 'total_phenols' and 'alcohol' correlated according to the Spearman correlation? Is it significant (95%)? Write explicitly the answer. **(2 points)**
(4) Use the linear regression model (WITH intercept) to describe the data above: y='alcohol', x= 'color intensity' then plot (scatterplot). Has the predictor x a significant influence on the response, at the 95% confidence level? Motivate your answer. **(3 points)**
(5) use the linear regression model (WITH intercept) to describe the data above:
1- y='alcohol', x= ['color_intensity','proline'].
2- y='alcohol', x= ['color_intensity','proline','magnesium'] Formally, which predictors have a significant influence on the response, at the 95% confidence level? Motivate your answer. Which model performs better? Discuss in detail the quantities that need to be considered (and those that do not). **(4 points)**

3. Exercise (10 points total)

The dataset `diagnoses.csv` contains features computed from digitized images of fine needle aspirate (FNA) of a breast mass. The features describe various characteristics X of cell nuclei present in the images. Each instance in the dataset represents a sample from a patient, and the task is to predict whether the sample is benign ($y = 0$) or malignant ($y = 1$).

(1) Compute the Pearson correlation coefficient between 'smoothness1' and 'compactness1'; 'smoothness1' and 'texture1'. Then discuss if they are significant at the 95% confidence level. **(2 points)**
(2) Write down the logistic regression model: X=['texture1', 'area1', 'compactness1', 'concave_points1']], y='diagnosis'. Fit WITHOUT the constant. **(3 points)**
(3) Look at the output of the logistic regression model. Formally, which predictors have a significant influence on the response, at the 95% confidence level? Motivate your answer. **(3 points)**
(4) Estimate the probability for $y = 1$ with X_test1=[18,500,0.2,0.05] and X_test2=[10,40,0.14,0.06]. What would be your prediction for y in these cases? **(2 points)**

4. Exercise (7 points total)

Given the dataset `grades.csv`

(1) Test if `write` is normally (i.e. Gaussian) distributed by performing a Kolmogorov-Smirnov test, with a confidence level of 95%. Discuss the results. **(2 points)**
(2) plot a QQ plot; `write` against Gaussian quantiles to check whether the sample is normal (i.e. Gaussian) **(2 points)**
(3) Compute the confidence interval, with a confidence level of 95%, of the mean of the variable `read` and median of the variable `science` with a number of bootstrap samples equal to 500. **(3 points)**

5. Exercise (6.5 points total)

Suppose we observe n realizations $y_1, y_2, ..., y_n$ of n independent random variables, $Y_1, Y_2, ..., Y_n$ all having a normal (i.e. Gaussian) distribution with mean μ_Y and variance σ_Y^2. Consider the following alternative estimator for σ_Y^2 , the variance of the Y_i, where c > 0: $Z = \frac{1}{(n-c)^2} \sum_{i=1}^{n} (Y_i - \mu_Y)^2$

We want to illustrate that this estimator is a biased estimator for σ_Y^2. Fix $c = 4$ and implement the estimator above.

(1) Randomly draw 20 observations from the normal (i.e. Gaussian) distribution with mean 5 and variance 16 and compute an estimate using Z. Repeat this procedure 5000 times. **(2 points)**
(2) Plot a histogram and cumulative distribution of Z. **(2 points)**
(3) A simple computation shows that $E[Z] = \frac{n}{(n-4)^2} \sigma_Y^2$ Is Z biased or unbiased? Is Z asymptotically unbiased? Motivate your answer. **(2.5 points)**