MIOT H5014

Statistical Analysis for Engineers

Worksheet 5 on Hypothesis Testing

Semester 2, 2016/7

Question 1

A lecturer is seeking to prove that attendance is a strong predictor of final marks for a group of students. For 14 students, the attendance over the course of the semester and the final mark were recorded. The correlation between the two variables will be investigated with an NHBST.

1. It is proposed to test the Null Hypothesis that $\rho = 0$ with a transformation of r to a variable that follows the Student t-distribution. Before gathering any data, set up the Null and alternative Hypothesis and other elements of the test and state the assumptions made about the model and data.
2. The data has been collected and the values are given in the table shown. Calculate r and carry out the test.
3. Repeat the two previous steps with the Fisher transformation of ρ.
4. Use the inverse of the Fisher transformation to calculate a confidence interval for ρ.

Data:

Att.:	29	84	91	78	24	42	65
Mark:	47	73	76	33	32	57	64
Att.:	57	38	58	48	69	76	46
Mark:	54	38	65	50	75	75	48

Question 2

In the scenario presented in the previous question, it transpires that only the first four data points are reliable. The correlation between the two variables will be investigated with an NHBST.

1. It is proposed to test the Null Hypothesis by carrying out a permutation test. Write down the permutations for the numbers 1 to 4.
2. With just 4 points, identify a scenario when you can you're your calculations because the Null Hypothesis has already been rejected or will not be rejected.
3. Calculate r for each permutation, noting the original value, and carry out the test by stating how many values are above the correct value.

Question 3

Let X and Y be two random variables with means μ_X, μ_Y and standard deviations σ_X, σ_Y respectively. Define the covariance of X and Y as

$$\text{cov}(X, Y) = E[\ (X - \mu_X)(Y - \mu_Y)\ ] = E[\ XY\ ] - \mu_X\mu_Y.$$

The correlation coefficient is then defined by

$$\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}.$$

It is required to link the variable Y to X by a linear equation $Y = \alpha + \beta X$.

1. Values of the coefficients α and β will be estimated by finding a and b respectively such that the quantity Q, defined by

$$Q = E[(Y - a - bX)^2]$$

is minimised with respect to a and b. Show that this leads to the equations

$$\mu_Y = a + b\mu_X \text{ and } b = \frac{\rho_{XY}\sigma_Y}{\sigma_X}.$$

2. Show further that the value of Q at this minimum point is the value

$$E[(Y - a - bX)^2] = \sigma_Y^2(1 - \rho_{XY}^2).$$

Question 4

It is to be determined whether the fruit yield of tomato plants depends on the amount used of two additives for the soil. The regression model shown here will be applied:

$$Y = \alpha + \beta_1 X_1 \ + \beta_2 X_2,$$

where Y is the weight of fruit yielded, X_i is the amount in appropriate units of additive i, $i = 1, 2$.

1. Write out the elements $X^{\mathrm{T}}X$ and $X^{\mathrm{T}}Y$ in the equation for the pseudo-inverse OLS estimator for $(\alpha, \beta_1, \beta_2)$. These elements should be written in terms of the Y_i and $X_{i,1}$, $X_{i,2}$.
2. Use these results to calculate the OLS estimator b for the data given below.
3. Set up the Null and alternative Hypothesis and other elements of the test to investigate whether the last variable is required to 'explain' Y. Carry out the test with the data given and using your previous work. Your test statistic will be

$$\frac{\|Xb\|^2 - \|Xb_s\|^2}{q} \Big/ \frac{\|e\|^2}{n-p}.$$

4. Repeat this work to investigate whether both variables are required to ‘explain’ Y.

The data is given in the following table:

Y	X_1	X_2
4.75	0.58	6.98
17.42	8.39	0.71
9.29	4.67	2.47
3.29	2.78	1.99
4.89	4.14	7.57
0.33	1.15	4.71
5.99	0.95	0.09
17.7	7.44	0.31
15.87	8.68	0.53
-4.72	0.99	9.22
7.00	4.36	7.42
11.83	6.87	1.87
12.88	8.77	5.55
10.65	3.81	5.83
10.16	7.84	2.91
6.62	6.89	7.49
21.02	8.3	0.91
8.35	8.27	6.96
11.92	5.57	0.19
-1.36	0.64	6.21
19.29	9.57	2.83
3.71	6.05	9.52
20.68	9.42	2.95