Course: Data Analysis (task № 3)

Student’s Name and Surname Iurii Aksenov

1. Chi-square test. Use **data\_games.dta** file.

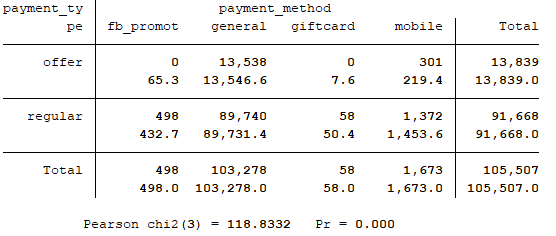
1.1. Analyze the relationship between **payment\_type** and **payment\_method** using Chi-square statistical test. Is the Chi-square test applicable for this pair of variables? If yes, formulate hypotheses, interpret the results of analysis and make conclusions. Create a suitable graph to demonstrate the relationship between these two variables.

**Answer:**

The Chi-square test is applicable for this pair of variables because they are categorical variables.

Ho: There is no relation between the **payment\_type** and **payment\_method**

H1: There is a relation between the **payment\_type** and **payment\_method**.



We conclude that there is a relationship between the variables because the probability (Pr = 0.000) is less than 0.05. So, H0 is rejected, H1 is accepted.



So, we can see that in ‘regular’ payment\_type the ‘general’ payment\_type is more than in ‘offer’ patment\_type.

1.2. Analyze the relationship between **payment\_type** and **crystalls\_balance\_before\_buy** using Chi-square statistical test. Is the Chi-square test applicable for this pair of variables? If yes, formulate hypotheses, interpret the results of analysis and make conclusions. Create a suitable graph to demonstrate the relationship between these two variables.

The Chi-square test is NOT applicable for this pair of variables because one of the variables (**crystalls\_balance\_before\_buy**) is numeric, other (**payment\_type**) is categorical. The Chi-square is used with categorical variables. they are categorical variables.

2. Scatterplot. Normality test. Correlation. Use **data\_games.dta** file.

2.1. Create a scatterplot between **crystalls\_balance\_before\_buy** and **payment**. Copy the scatterplot into this file.



2.2. Run the suitable normality test to conclude whether the distribution of **payment** variable is significantly different from the normal. Formulate hypothesis. Make conclusions.

**Answer:**

H0: the distribution of payment is normal

H1: the distribution of **payment** is different from the normal

Variable | Obs W V z Prob>z

-------------+------------------------------------------------------

payment | 105,507 0.68626 9944.920 25.814 0.00000

Probability is less than 0.05, so the distribution is significantly different from normal => accept H1 hypothesis (the distribution of **payment** is different from the normal).

Calculate an appropriate correlation coefficient between three pairs of variables. Fill in the table below. Interpret the results.

**Answer:**

We use Spearman correlation coefficient because payment is not normal distributed, and Spearman is use for non-normally distributed data. Other variables are non-normally distributed too, so Spearman’s method is used for all pair of variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Type of the appropriate correlation coefficient | Hypotheses | Strength of the relationship | Direction of the relationship | Significance of the relationship |
| crystalls\_balance\_before\_buy and payment | Spearman | H1 dependent variables (p=0.000) | 0.1458 | Positive | Very weak |
| crystalls\_balance\_before\_buy and crystalls\_bought | Spearman | H1 dependent variables (p=0.000) | 0.1451 | Positive | Very weak |
| crystalls\_bought and payment | Spearman | H1 dependent variables (p=0.000) | 0.9065 | Positive | Very strong |

As for result’s interpretation, the represented pairs of variables are dependent, mor over in positive way. However, the variables **crystalls\_bought** and **payment** are very strong correlate instead of other pairs **crystalls\_balance\_before\_buy** and **crystalls\_bought**, **crystalls\_balance\_before\_buy** and **payment**. We can say, that if the people spend more money for payment, they buy more crystals.

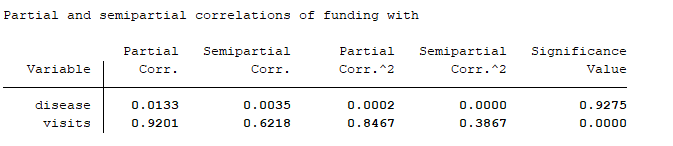
3. Partial correlation. Use **health\_funding.dta** file.

Calculate paired correlation coefficient between **funding** and **disease** variables. Now calculate the correlation coefficient between the same pair of variables controlling for the number of visits (**visits** variable). Interpret the results of analysis.

**Answer:**

Funding and disease have normal distributed data, so we can use Pearson’s correlation method.

Paired correlation coefficient of funding and disease: 0.7371, p-value: 0.000



The partial correlation coefficient (0.0133) between **funding** and **disease** controlling the effect of **visits**. The significance value (0.9275) is more than 0.05 => **funding** and **disease** is independent.

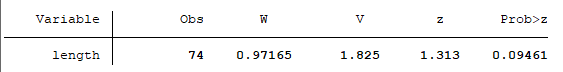
Moreover, the relationship between **funding** and **disease** is actually contained already in the relationship between **funding** and **visits**. So, **disease** does not add a lot of explanatory power above and beyond **visits**. And the reason of the strong relationship between **funding** and **disease** is mostly due to the fact that both of these variables are related to **visits**.

4. T-tests and Nonparametric tests.

4.1. Use **auto.dta** file (example datasets). Select an appropriate test to check if there is a difference in the mean length of foreign and non-foreign cars. Explain you selection. Formulate the hypotheses. Interpret the results of analysis.

**Answer:**

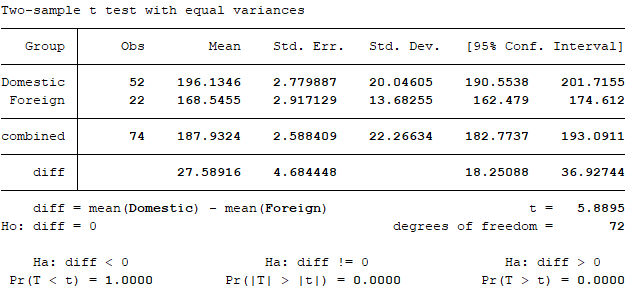
Firstly, understand whether the variable length normal or not.



Probability is more than 0.05, so the distribution is normal and only two categories (foreign and non-foreign) => we can use **t-test (Student’ T-test)**

H0: Mean(d) = M(f) (mean length of domestic cars equals mean length of foreign cars)

Ha: Mean(d) ≠ M(f)



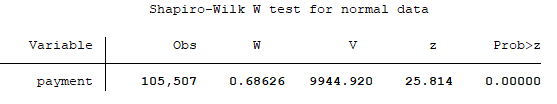
t = 5.8895, p-value = 0.0000

p-value is less than 0.05 then we reject H0 and accept Ha, so the mean length of domestic cars is not equal to the mean length of foreign cars.

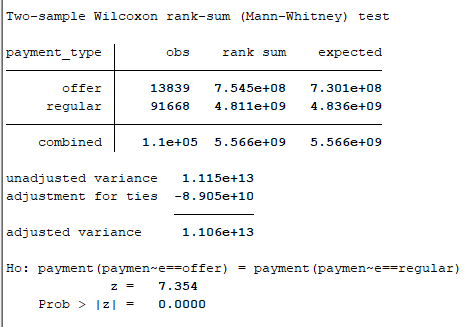
4.2. Use **data\_games.dta** file. Select an appropriate test to understand whether there is a difference in payments between the people who have used different payment types. Explain you selection. Formulate the hypotheses. Interpret the results of analysis.

**Answer:**

Firstly, understand whether the variable length normal or not.



The probability is less than 0.05, so the **payments** is not normal distributed => we can use **Mann-Whitney U Test** to compare the payments of ‘offer’ and ‘regular’ payment\_types. Also, the payment\_methods has only two categories, so, we definitely are able to use **Mann-Whitney U Test**



H0: the payment of offer type is equal payment of regular type

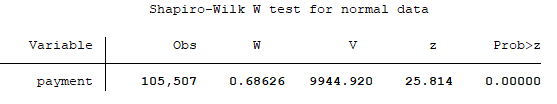
Ha: the payment of offer is not equal payment of regular type

P-value = 0.000 is less than 0.5, so we reject H0 and accept H1. So, there difference between the payments of offer and payments of regular type.

4.3. Use **data\_games.dta** file. Select an appropriate test to understand whether there is a difference in payments between the people who have used different payment methods. Explain you selection. Formulate the hypotheses. Interpret the results of analysis.

**Answer:**

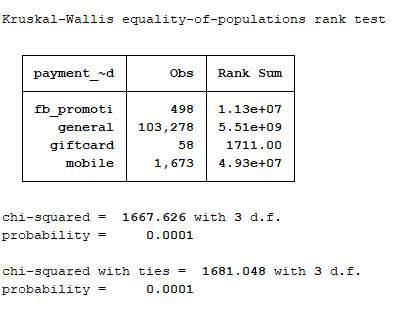
Firstly, understand whether the variable length normal or not.



The probability is less than 0.05, so the **payments** is not normal distributed => we should use non-parametric test to compare the payments of ‘offer’ and ‘regular’ payment\_types.

Secondly, payment\_methods has more than 2 categories (fb\_promotion, giftcard, general, mobile).

We should use non-parametric test with mor than 2 groups => we use **Kruskal-Wallis** test



H0: there is no difference between groups

Ha: there difference between the groups, so means of payments are not equal between different payment methods. Is payment method influent on amount of payments

P-value = 0.0001 is less than 0.05, so we reject H0 and accept Ha. So, the one payment\_method is significantly different from another payment methods. So, the payments in groups are different.

Please send to amelikyan@hse.ru from your personal e-mail:

- this MS Word file with answers,

- the do-file with the corresponding commands.