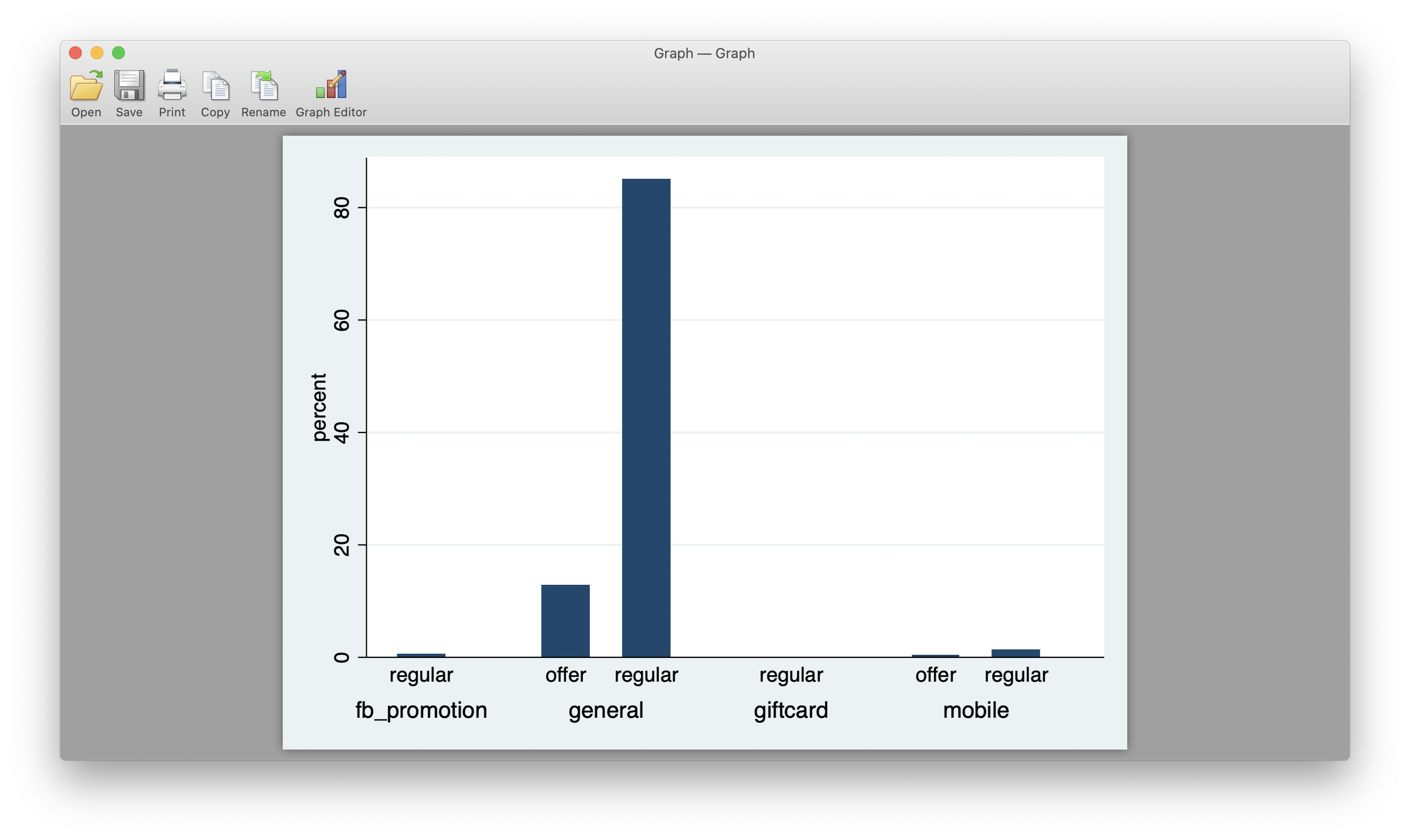
Course: Data Analysis (task № 3)

Student’s Name and Surname: Pavel Drankov

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: Yes, H0: Does payment\_type and payment\_method are independent? , Pearson chi2(3) = 118.8332,  
  Conclusion: Yes, they are independent

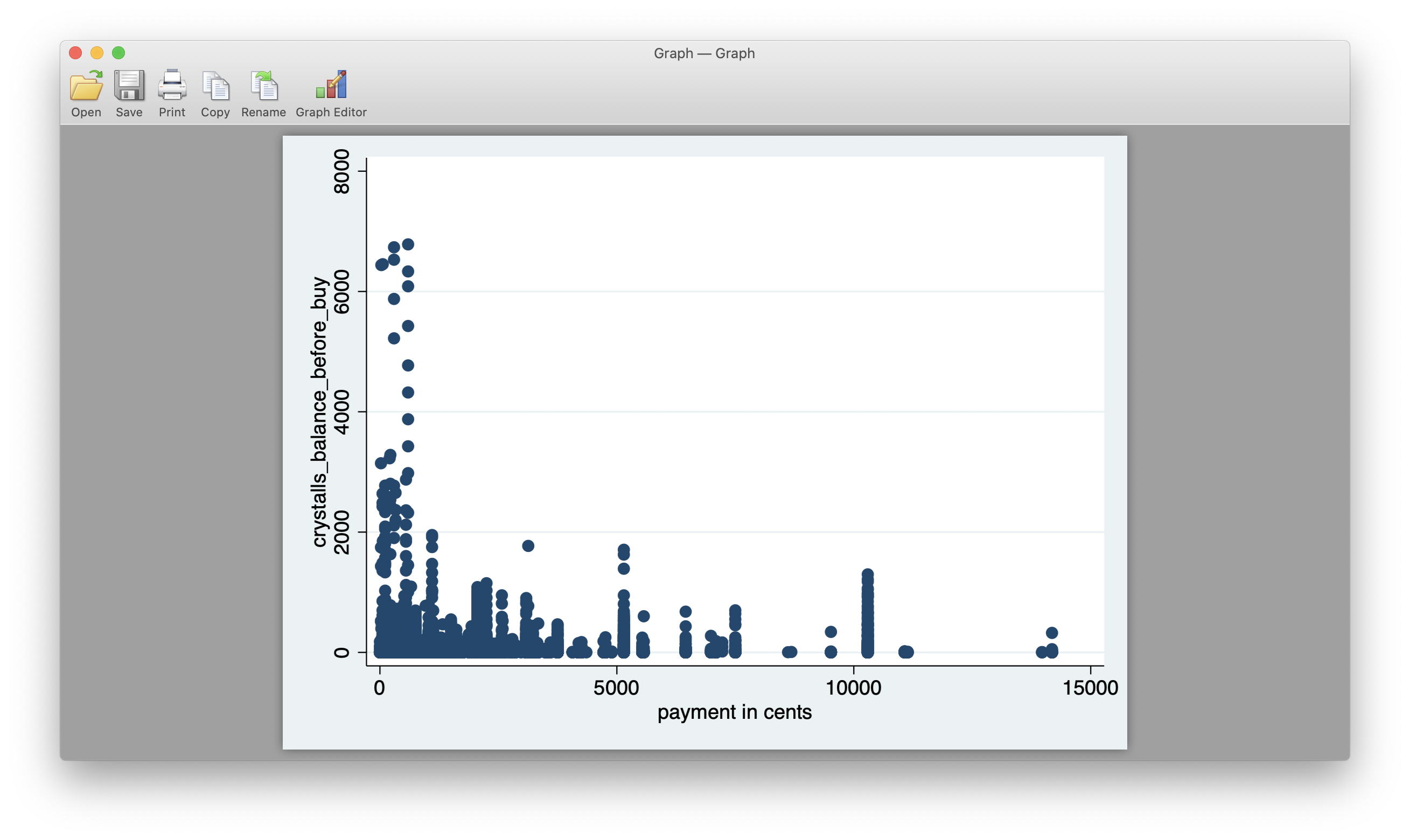


1.2. Analyze the relationship between **payment\_type** and **crystalls\_balance\_before\_buy** using Chi-square statistical test. Is the Chi-square test applicable for this pair of variables? If yes, formulate hypotheses, interpret the results of analysis and make conclusions. Create a suitable graph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

Answer: No, not applicable, **crystalls\_balance\_before\_buy** has too may values.

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.

One-sample Kolmogorov-Smirnov test against theoretical distribution

normprob((payment-payment\_mu)/payment\_s)

Smaller group D P-value

-----------------------------------

payment: 0.2527 0.000

Cumulative: -0.2033 0.000

Combined K-S: 0.2527 0.000

H0: is payment variable distribution similar to the normal?  
Conclusion: No, not even close to be similar.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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 | pwcorr | Variable are dependent | Correlation coef = 0.0565  Almost no relationship | + | 0.05 |
| crystalls\_balance\_before\_buy and crystalls\_bought | pwcorr | Variable are dependent | Correlation coef = 0.1140  Very weak correlation | + | 0.05 |
| crystalls\_bought and payment | pwcorr | Variable are dependent | Correlation coef = 0.8511  Strong correlation | + | 0.05 |

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

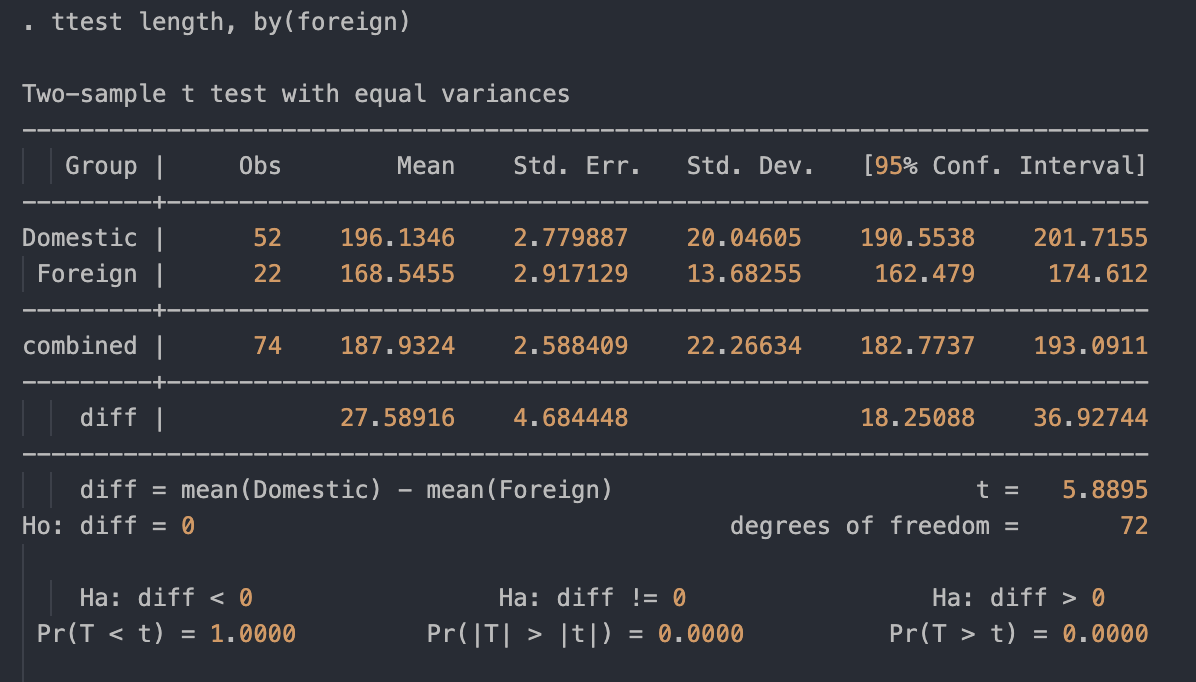
Calculate paired correlation coefficient between **funding** and **disease** variables.   
  
Correlation coef = 0.73, p = 0.000

Now calculate the correlation coefficient between the same pair of variables controlling for the number of visits (**visits** variable).

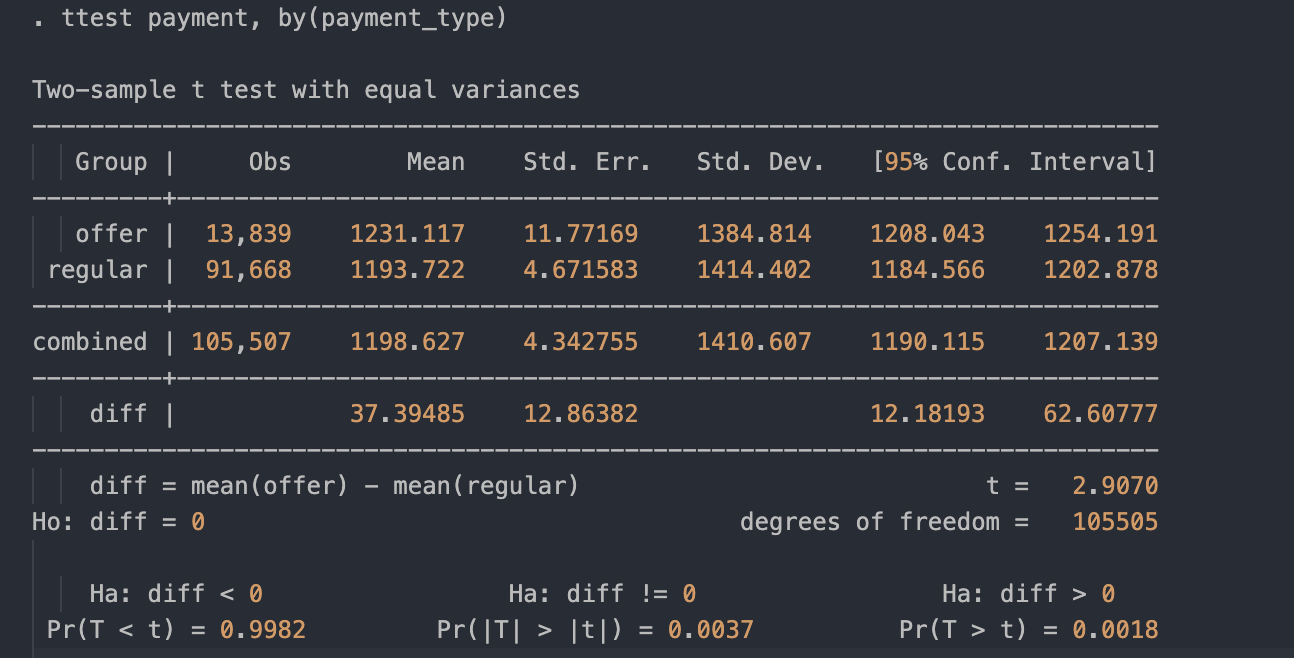
Partial correlation is 0.0133 on significance level = 0.9275, which is a non significant level.

Interpret the results: funding and disease variables strongly correlated.

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.

T-test is exactly what we need. H0: is mean length of car is dependenct of foreign status. Answer: Yes, foreign cars are shorter, and the difference between means is 27 inches.

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.   
H0: there is a difference in payments between different payment types. On the level of significance 0.05> 0.0037 we can accept the hypotheses.

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.



H0: there is a difference in payments between different payment methods. On the level of significance 0.05> 0.0001 we can accept the hypotheses.

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

- this MS Word file with answers,

- the do-file with the corresponding commands.