Статистика

или

Сказка о волшебной лопате

О статистике вообще

- *There are three kinds of lies: lies, damned lies, and statistics*

Anonymous

Основные понятия

- Переменные - это то, что можно измерять, контролировать или что можно изменять в исследованиях.
- Независимыми переменными называются те, которые варьируются исследователем, тогда как зависимые переменные - те, которые измеряются.
- Выборка – совокупность переменных.

Основные понятия

- Математическое ожидание, среднее арифметическое
- Дисперсия
- Стандартное отклонение
- Стандартная ошибка
- Нормальное распределение
- Корреляция
- Гистограмма

Неосновные понятия

- t-критерий Стьюдента
- Критерий Фишера
- Критерий Пирсона (Хи-квадрат)
- Критерий согласия Колмогорова
- Тест Вальда
- U-критерий Манна — Уитни
- Критерий Уилкоксона
- Критерий Краскела — Уоллиса
- Критерий Кохрена
- Критерий Лиллиефорса

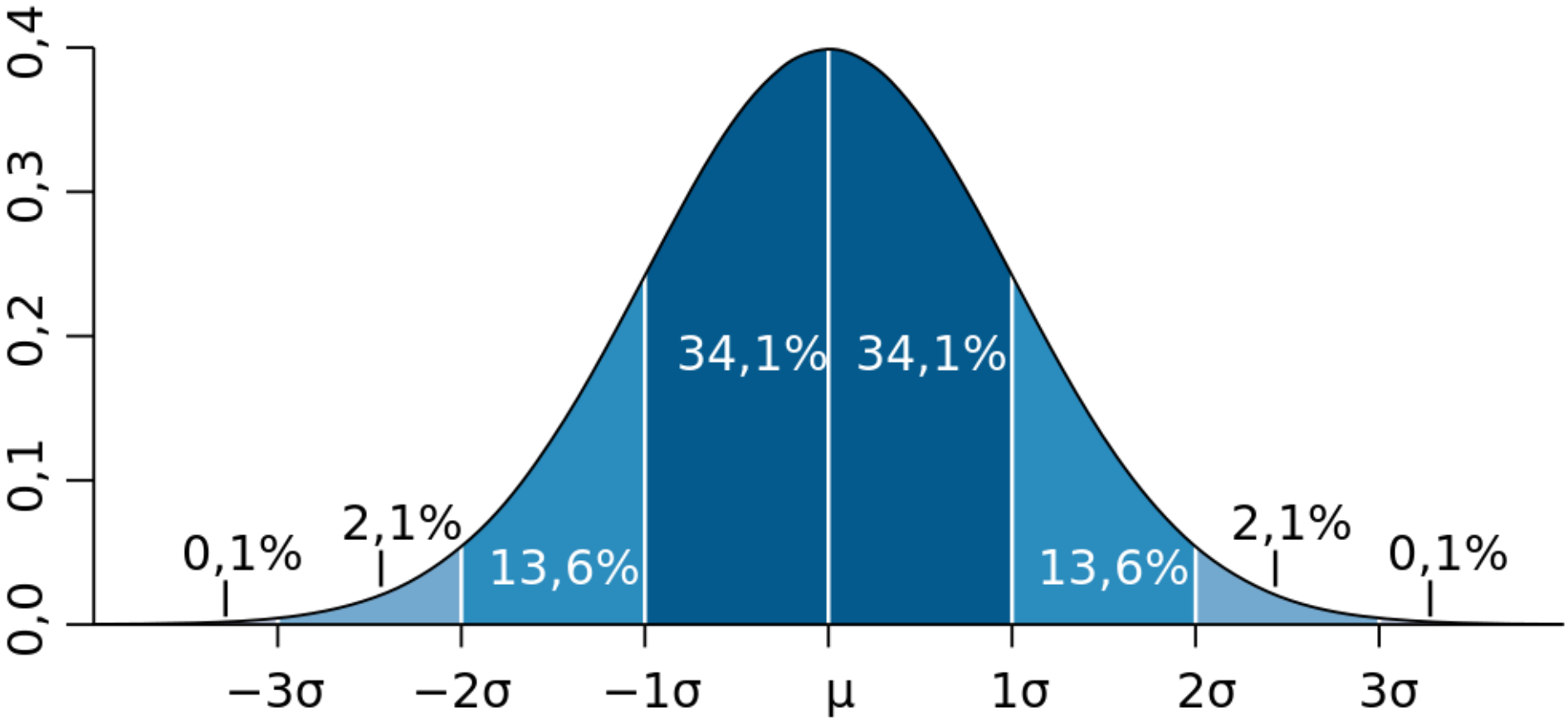
Формулки

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

$$SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}.$$

Нормальное распределение и правило 3 сигм (два в одном слайде)



Критерий Стьюдента

Уи́льям Си́ли Го́ссет
(он же Стьюдент)

Собственно формула

$$t = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{s_1^2}{n1} + \frac{s_2^2}{n2}}}$$

# $p < 0.05$, $p < 0.001$

- Это вероятность (шанс), что мы правы или ошибаемся, делая статистическое предположение

Что почем

- STATISTICA Base for Windows v.10 Russian
  http://statsoft.ru/
  цена за лицензию ~~54 206,46 руб~~. 60 965,50 руб.
- IBM SPSS Statistics Standard
  http://www-01.ibm.com/software/analytics/spss/products/statistics/
  цена за лицензию ~~$5 430~~ $5 590

- GNU SPSS http://www.gnu.org/software/pspp/
- R-project http://www.r-project.org/

STATISTICA - [статистика* - Categ. Scatterplot: средний балл vs. Сумма баллов ЕГЭ]

Home | Edit | View | Insert | Format | Statistics | Data Mining | Graphs | Workbook | Help | Options

Cut | Copy | Delete | Paste | Screen Catcher | Clipboard

Controls | Grid | Status Bar | View

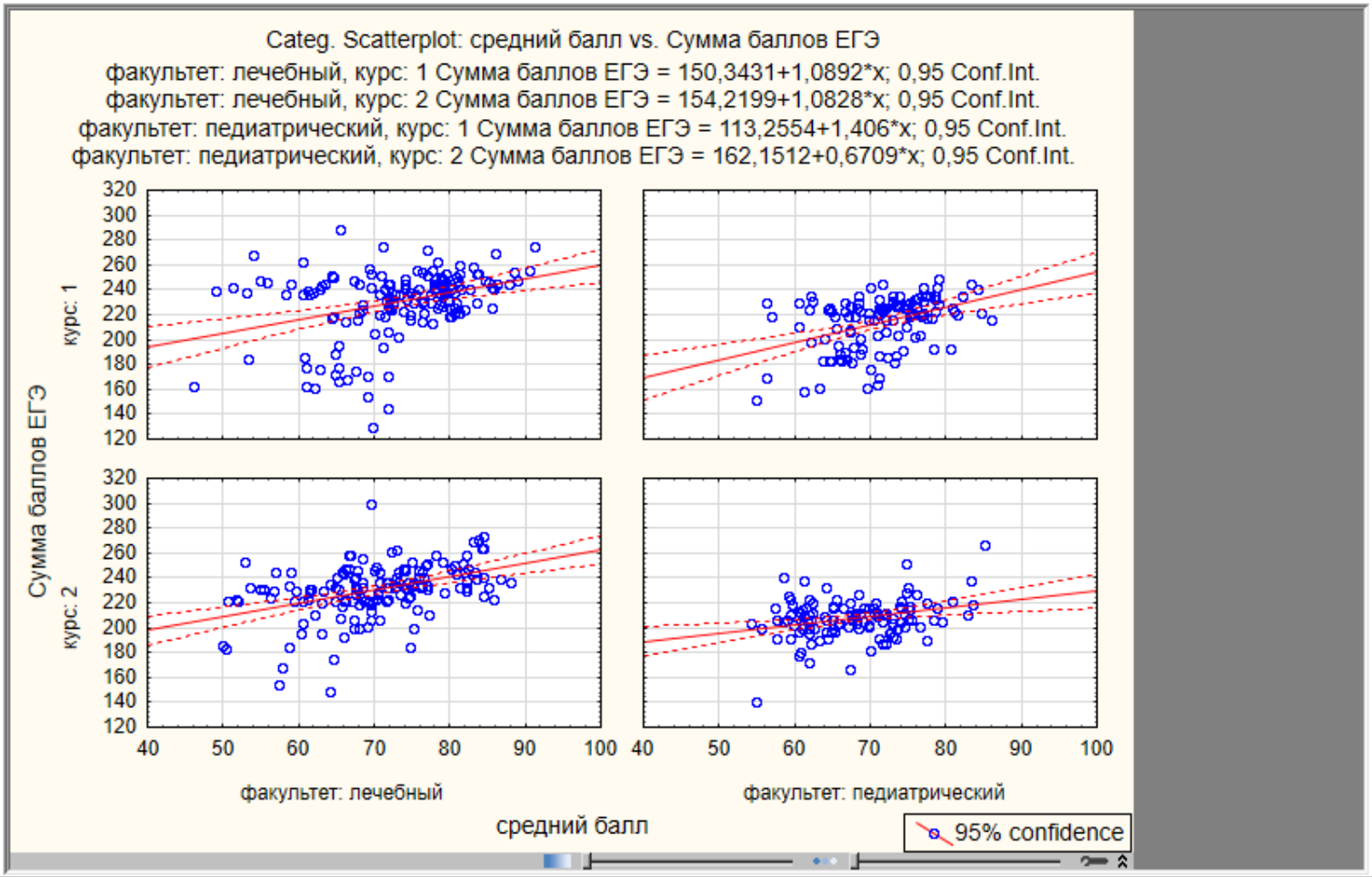
Normal Graph [mo... | Customize Graph

Selection | Graph | Options

MultiDoc Style | Graph Data | Zoom subgraph | Lock Graph | Events | Tools

OLE | Object | Links

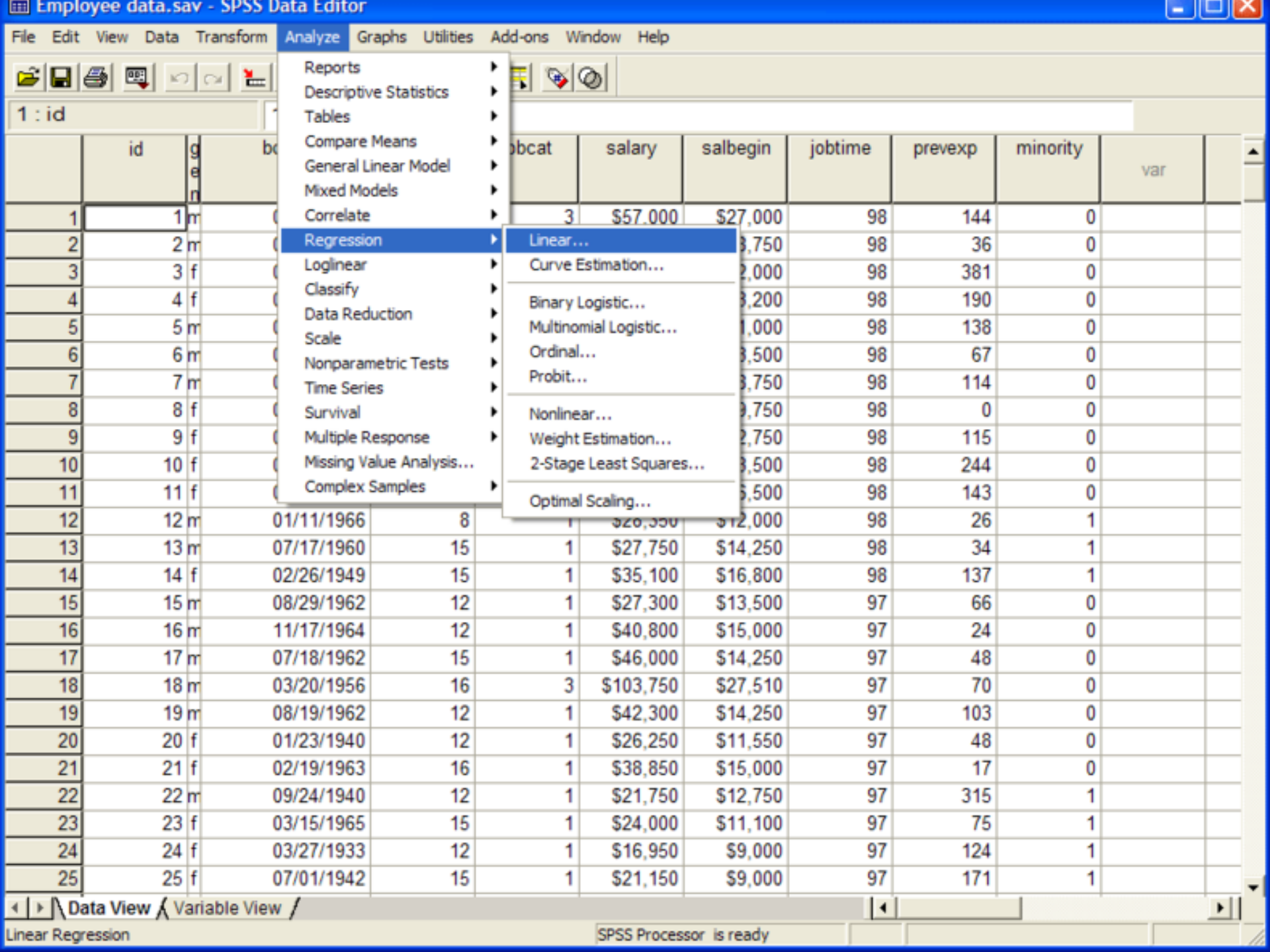
- Histogram: средни
- Histogram: колич
- Histogram: Сумма
- Descriptive Statisti
- Basic Statistics/Tables (Ли
  - Breakdown results dia
    - Breakdown Table
    - Within-Group Cor
    - Within-Group Cor
    - Within-Group Cor
    - Within-Group Cor
    - Categ. Scatterplot:
- Basic Statistics/Tables (Ли
  - Breakdown results dia
    - Within-Group Cor
    - Within-Group Cor
    - Within-Group Cor
    - Within-Group Cor
- 2D Box Plots (Лист1 in ста
  - Box Plot of средний б
  - Box Plot of количеств
  - Box Plot of Сумма ба
- Basic Statistics/Tables (Ли
  - Correlations dialog
    - Correlations (Лист
    - Correlations (Лист
- Basic Statistics/Tables (Ли
  - T-test for independen
    - T-tests; Grouping:

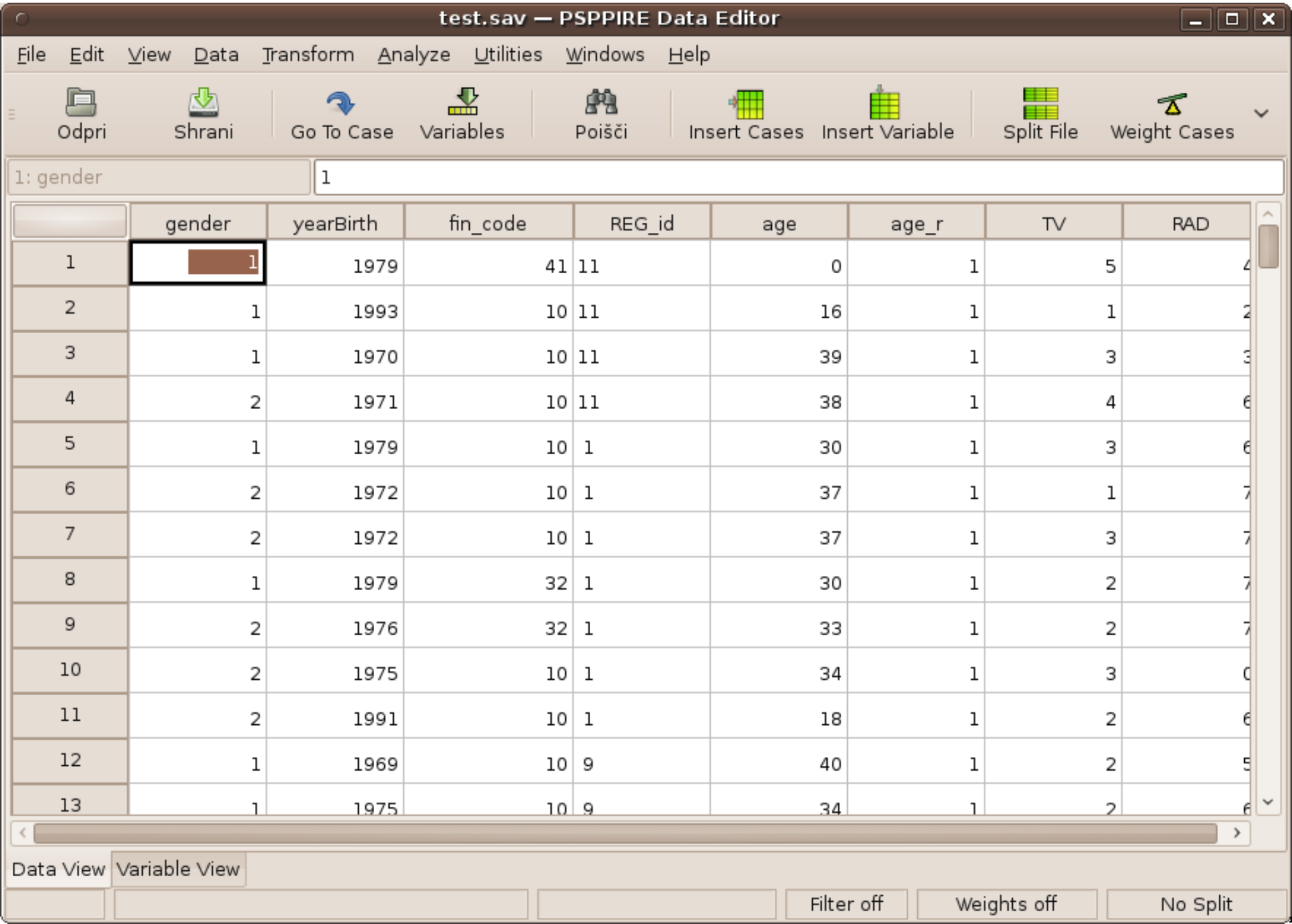


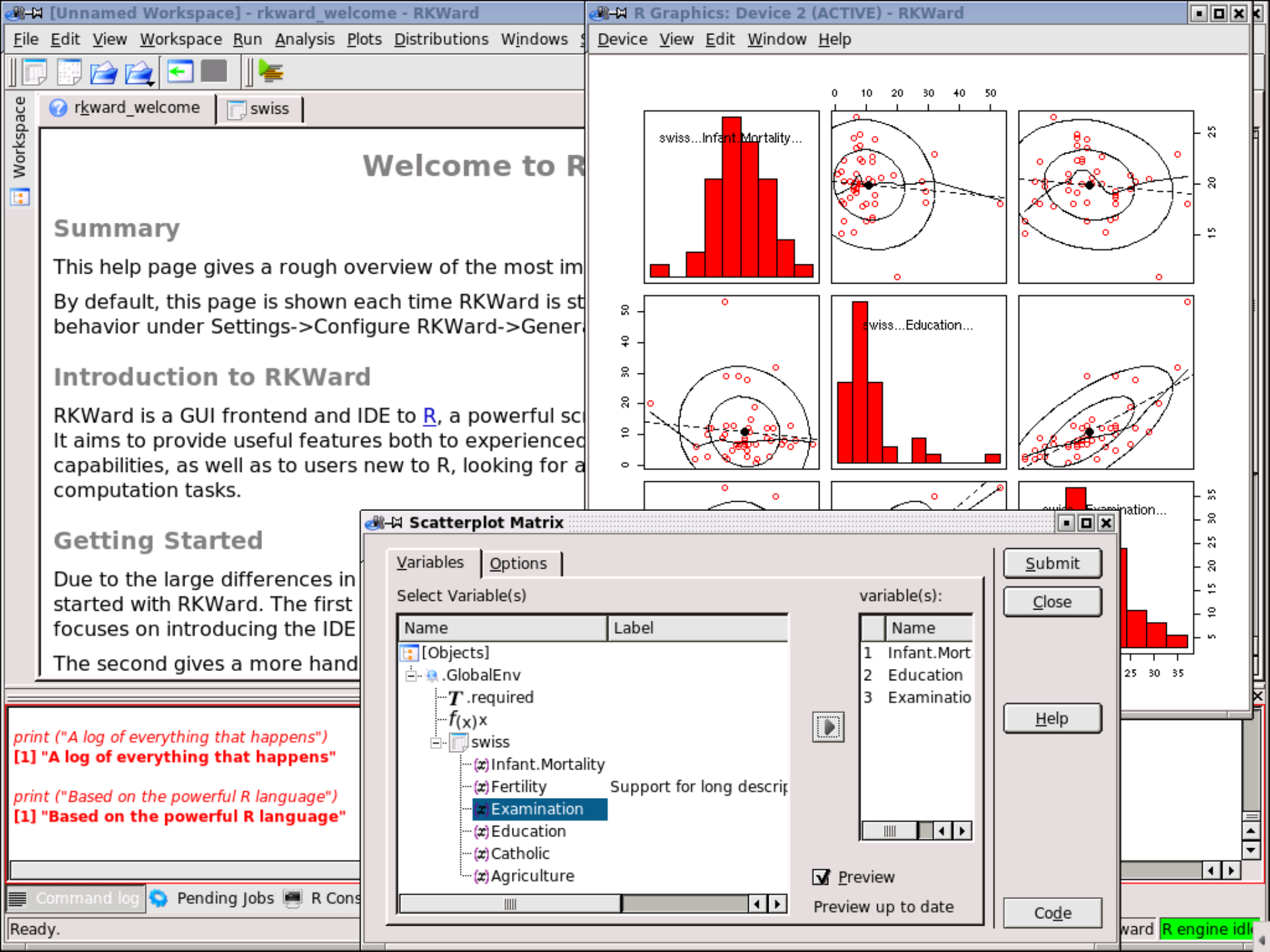
Within-Group Correlations (Лист1 in статистика) | Categ. Scatterplot: средний балл vs. Сумма баллов ЕГЭ

Descriptive Statistics:...

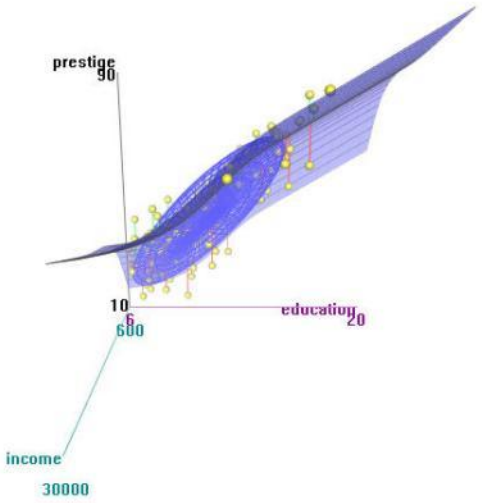
Ready | Лист1 in стат | CAP | NUM | ЗАП


Employee data.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
1 : id
id
gender
bdate
educ
jobcat
salary
salbegin
jobtime
prevexp
minority
var
1 1 m 3 $57,000 $27,000 98 144 0
2 2 m 3,750 98 36 0
3 3 f 2,000 98 381 0
4 4 f 3,200 98 190 0
5 5 m 1,000 98 138 0
6 6 m 3,500 98 67 0
7 7 m 3,750 98 114 0
8 8 f 9,750 98 0 0
9 9 f 2,750 98 115 0
10 10 f 3,500 98 244 0
11 11 f 5,500 98 143 0
12 12 m 01/11/1966 8 1 $26,550 $12,000 98 26 1
13 13 m 07/17/1960 15 1 $27,750 $14,250 98 34 1
14 14 f 02/26/1949 15 1 $35,100 $16,800 98 137 1
15 15 m 08/29/1962 12 1 $27,300 $13,500 97 66 0
16 16 m 11/17/1964 12 1 $40,800 $15,000 97 24 0
17 17 m 07/18/1962 15 1 $46,000 $14,250 97 48 0
18 18 m 03/20/1956 16 3 $103,750 $27,510 97 70 0
19 19 m 08/19/1962 12 1 $42,300 $14,250 97 103 0
20 20 f 01/23/1940 12 1 $26,250 $11,550 97 48 0
21 21 f 02/19/1963 16 1 $38,850 $15,000 97 17 0
22 22 m 09/24/1940 12 1 $21,750 $12,750 97 315 1
23 23 f 03/15/1965 15 1 $24,000 $11,100 97 75 1
24 24 f 03/27/1933 12 1 $16,950 $9,000 97 124 1
25 25 f 07/01/1942 15 1 $21,150 $9,000 97 171 1
Data View
Variable View
Linear Regression
SPSS Processor is ready


test.sav — PSPPIRE Data Editor
File Edit View Data Transform Analyze Utilities Windows Help
Odpri
Shrani
Go To Case
Variables
Poišči
Insert Cases
Insert Variable
Split File
Weight Cases
1: gender
1
gender yearBirth fin_code REG_id age age_r TV RAD
1 1 1979 41 11 0 1 5 4
2 1 1993 10 11 16 1 1 2
3 1 1970 10 11 39 1 3 3
4 2 1971 10 11 38 1 4 6
5 1 1979 10 1 30 1 3 6
6 2 1972 10 1 37 1 1 7
7 2 1972 10 1 37 1 3 7
8 1 1979 32 1 30 1 2 7
9 2 1976 32 1 33 1 2 7
10 2 1975 10 1 34 1 3 0
11 2 1991 10 1 18 1 2 6
12 1 1969 10 9 40 1 2 5
13 1 1975 10 9 34 1 2 6
Data View
Variable View
Filter off
Weights off
No Split


[Unnamed Workspace] - rkward_welcome - RKWard
File Edit View Workspace Run Analysis Plots Distributions Windows
Workspace
rkward_welcome
swiss
Welcome to R
Summary
This help page gives a rough overview of the most im
By default, this page is shown each time RKWard is st
behavior under Settings->Configure RKWard->Gener
Introduction to RKWard
RKWard is a GUI frontend and IDE to R, a powerful sc
It aims to provide useful features both to experienced
capabilities, as well as to users new to R, looking for a
computation tasks.
Getting Started
Due to the large differences in
started with RKWard. The first
focuses on introducing the IDE
The second gives a more hand
print ("A log of everything that happens")
[1] "A log of everything that happens"
print ("Based on the powerful R language")
[1] "Based on the powerful R language"
Command log
Pending Jobs
R Cons
Ready.
R Graphics: Device 2 (ACTIVE) - RKWard
Device View Edit Window Help
swiss...Infant.Mortality...
swiss...Education...
Scatterplot Matrix
Variables
Options
Select Variable(s)
Name
Label
[Objects]
.GlobalEnv
.required
f(x) x
swiss
Infant.Mortality
Fertility
Support for long descrip
Examination
Education
Catholic
Agriculture
variable(s):
Name
1 Infant.Mort
2 Education
3 Examinatio
Preview
Preview up to date
Submit
Close
Help
Code
R engine idl
ward

RGL device 1 [Focus]



R Commander

File Edit Data Statistics Graphs Models Distributions Tools Help

Data set: Prestige | Edit data set | View data set | Model: LinearModel.1

Script Window

```
library(mgcv, pos=4)
scatter3d(Prestige$education, Prestige$prestige, Prestige$income,
  fit="additive", residuals=TRUE, bg="white", axis.scales=TRUE, grid=TRUE,
```

Linear Model

Enter name for model: LinearModel.2

Variables (double-click to formula)

census
education
income
prestige

Model Formula: + * : / %in% - ^ ( )

prestige ~ education + income + type

Subset expression

<all valid cases>

OK | Cancel | Reset | Help

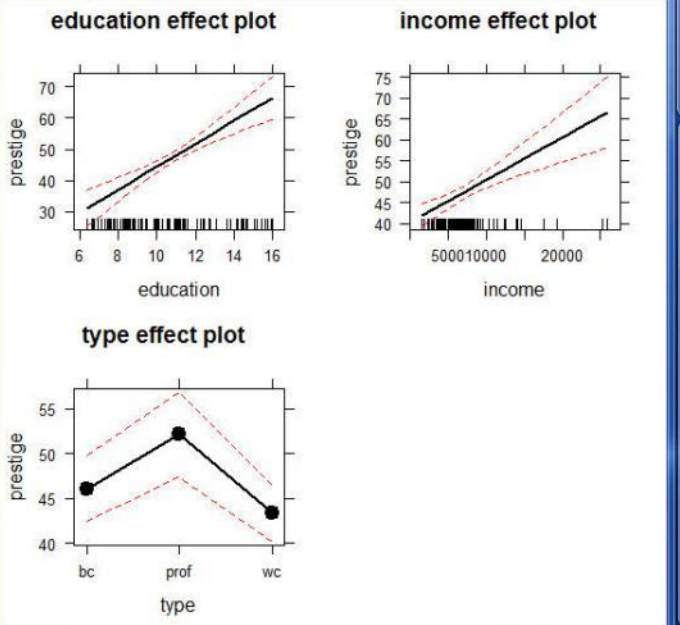
```
> library(effects, pos=4)

> trellis.device(theme="col.whitebg")

> plot(allEffects(LinearModel.1), ask=FALSE)
```

Messages

```
[1] NOTE: R Commander Version 1.9-1: Sat Sep 08 12:11:18 2012
[2] NOTE: The dataset Prestige has 102 rows and 6 columns.
```

R Graphics: Device 2 (ACTIVE)

File History Resize



Prestige

	education	income	women	prestige	census	type
gov.administrators	13.11	12351	11.16	68.8	1113	prof
general.managers	12.26	25879	4.02	69.1	1130	prof
accountants	12.77	9271	15.70	63.4	1171	prof
purchasing.officers	11.42	8865	9.11	56.8	1175	prof
chemists	14.62	8403	11.68	73.5	2111	prof
physicists	15.64	11030	5.13	77.6	2113	prof
biologists	15.09	8258	25.65	72.6	2133	prof
architects	15.44	14163	2.69	78.1	2141	prof
civil.engineers	14.52	11377	1.03	73.1	2143	prof
mining.engineers	14.64	11023	0.94	68.8	2153	prof
surveyors	12.39	5902	1.91	62.0	2161	prof
draughtsmen	12.30	7059	7.83	60.0	2163	prof
computer.programers	13.83	8425	15.33	53.8	2183	prof
economists	14.44	8049	57.31	62.2	2311	prof
psychologists	14.36	7405	48.28	74.9	2315	prof
social.workers	14.21	6336	54.77	55.1	2331	prof
lawyers	15.77	19263	5.13	82.3	2343	prof
librarians	14.15	6112	77.10	58.1	2351	prof
vocational.counsellors	15.22	9593	34.89	58.3	2391	prof
ministers	14.50	4686	4.14	72.8	2511	prof
university.teachers	15.97	12480	19.59	84.6	2711	prof
primary.school.teachers	13.62	5648	83.78	59.6	2731	prof
secondary.school.teachers	15.08	8034	46.80	66.1	2733	prof

Немного экстрима

```
students <- read.csv("students.csv")
students$course <- factor(students$course)
students$course_spec <- paste(students$course, "курс,", students$speciality, sep = " ")
students$course_spec <- factor(students$course_spec)
sink("Описательные статистики.txt")
for(i in 1:length(levels(students$course_spec))){
  tmp <- subset(students, students$course_spec == levels(students$course_spec)[i])
  summary <- lapply(tmp[3:5], summary)
  print(levels(students$course_spec)[i])
  cat('\n')
  print(summary)
}
sink()
for(i in 3:5){
  png(filename = paste("Диаграмма размаха ", gsub('\\.', ' ', names(students)[i]), '.png', sep=''))
  par(las = 2, mar = c(12.1, 5.1, 4.1, 2.1))
  boxplot(students[,i] ~ students$course_spec, ylab = gsub('\\.', ' ', names(students)[i]))
  dev.off()
}
```

Результат

Текст

[1] "1 курс, лечебный"

$mean

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
46.00	67.89	74.09	73.17	79.78	91.14

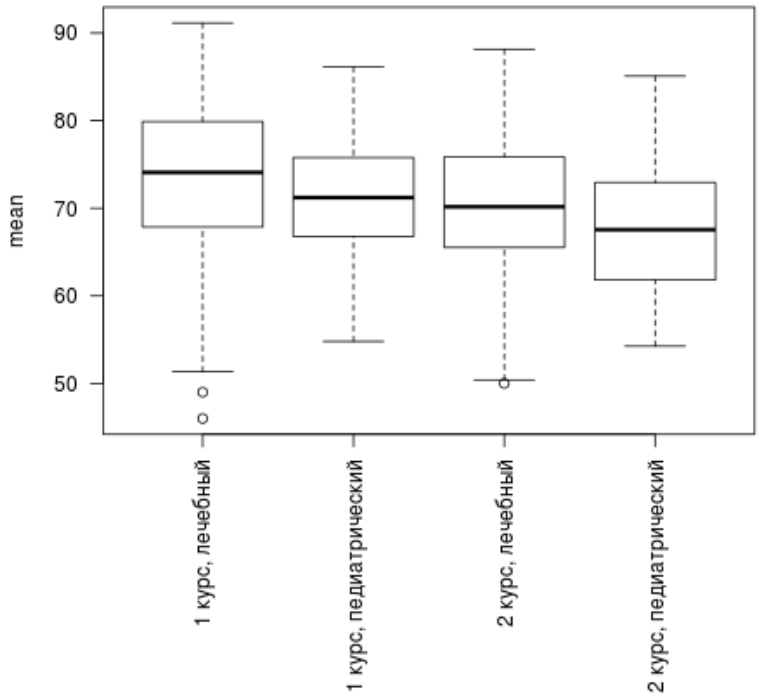
$count

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.00	40.00	43.00	41.08	46.00	56.00

$sum

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
130	222	238	230	246	288	1

Диаграммы



Попробуем на примере

Один бьётся в горячке, другой остывает в морге, а средняя температура по больнице 36,6 °C

Виктор Шендерович

Если мой сосед бьет свою жену ежедневно, а я — никогда, то в свете статистики мы оба бьем свою жену через день.

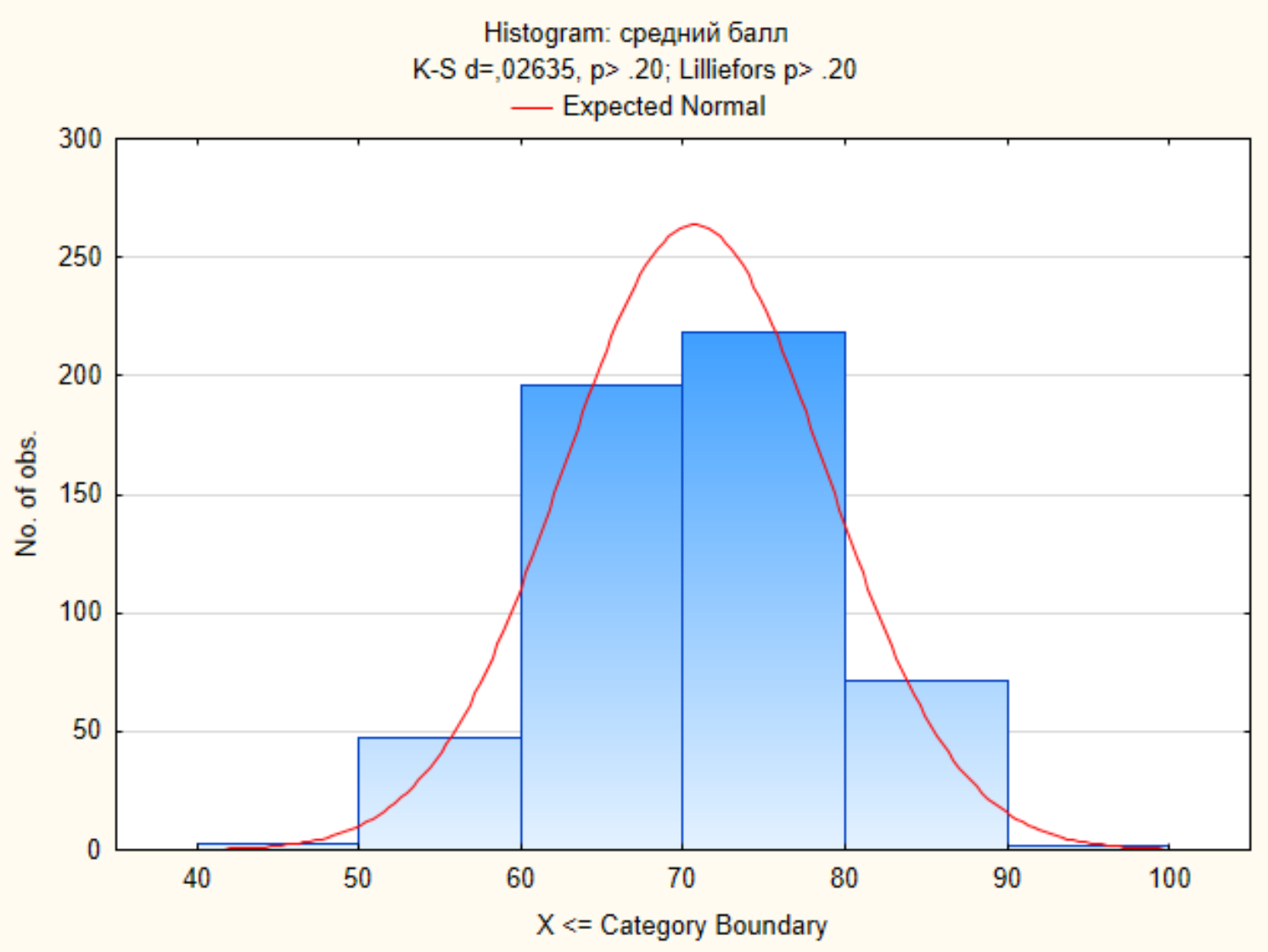
Бернард Шоу

Исходные данные

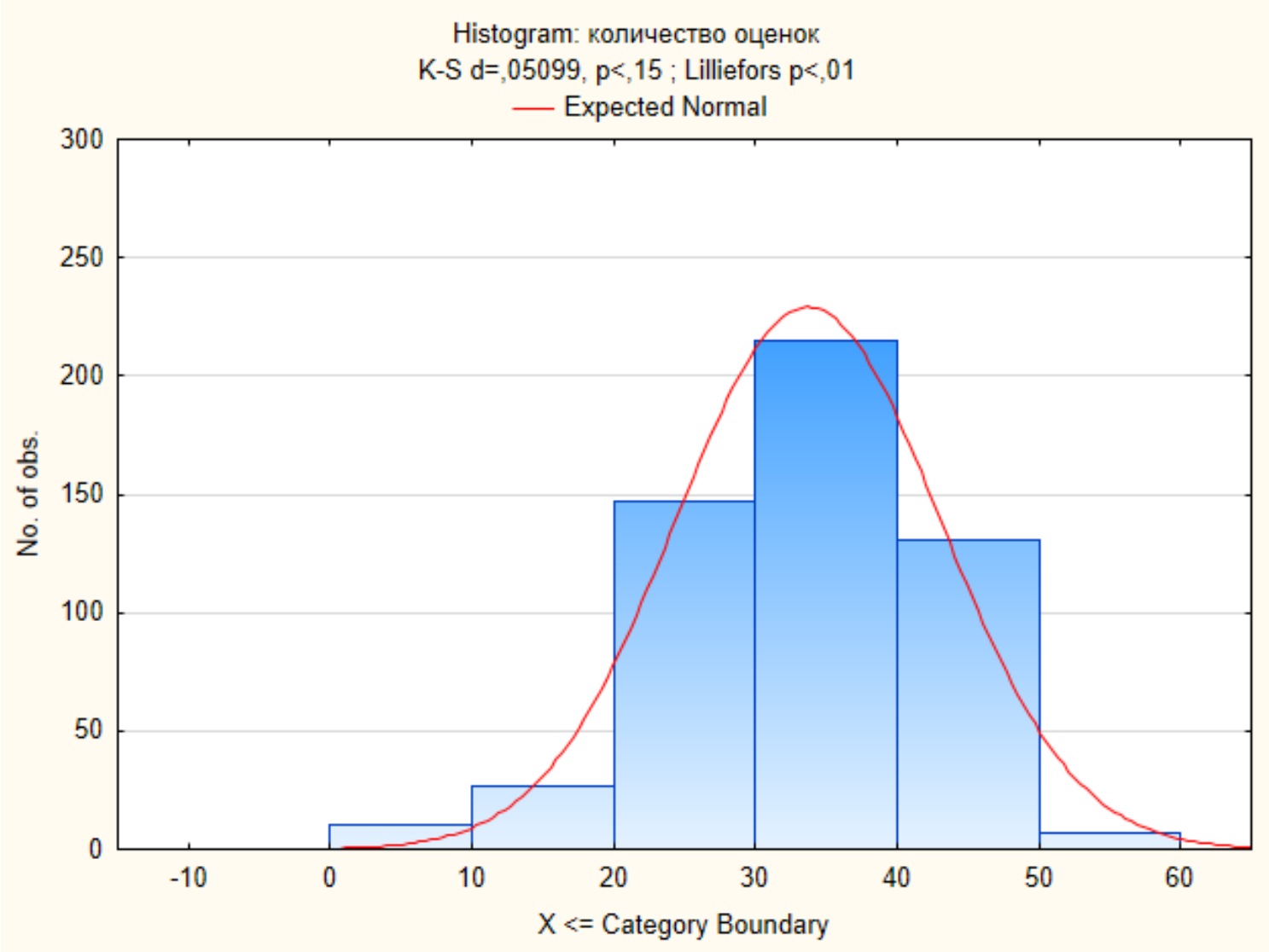
Лист1

	2 номер зачетки	3 факультет	4 курс	6 средний балл	7 количество оценок	8 Сумма баллов ЕГЭ
1	39106	педиатрич	2	57,318	22	216
2	39471	лечебный	1	61,929	14	238
3	38385	лечебный	2	52,769	13	253
4	38865	лечебный	2	61,677	31	231
5	39486	лечебный	1	55,794	34	246
6	39107	педиатрич	2	58,48	25	240
7	39076	педиатрич	2	67,865	37	212
8	38970	лечебный	2	82,147	34	230
9	38850	лечебный	2	62,143	28	210
10	38955	лечебный	2	73,267	30	244
11	38910	лечебный	2	62,964	28	220
12	38340	лечебный	2	57,5	16	155
13	39395	лечебный	1	77,864	44	240
14	39396	лечебный	1	64,429	21	252

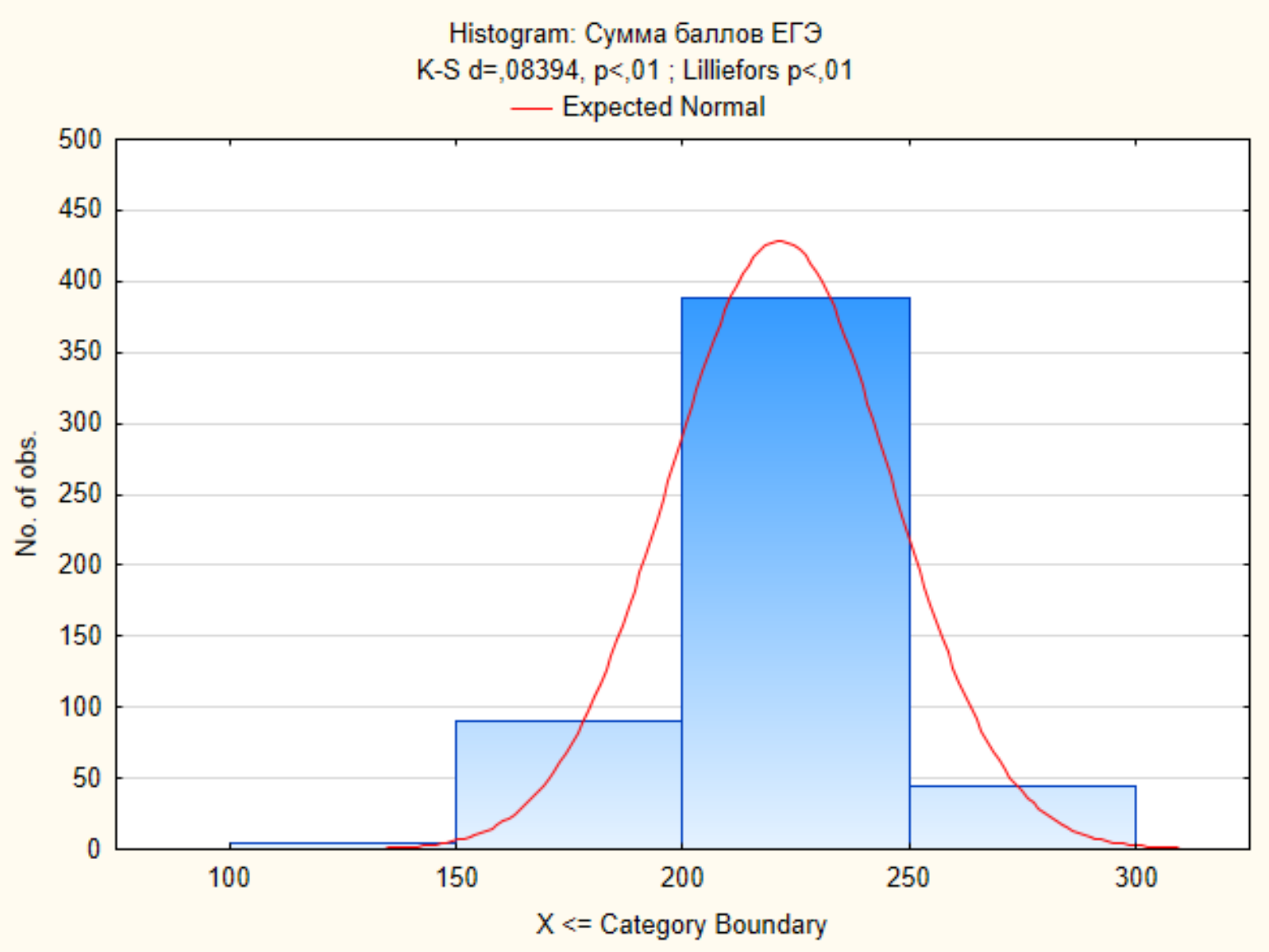
Проверка на нормальность



Проверка на нормальность



Проверка на нормальность



Описательная статистика

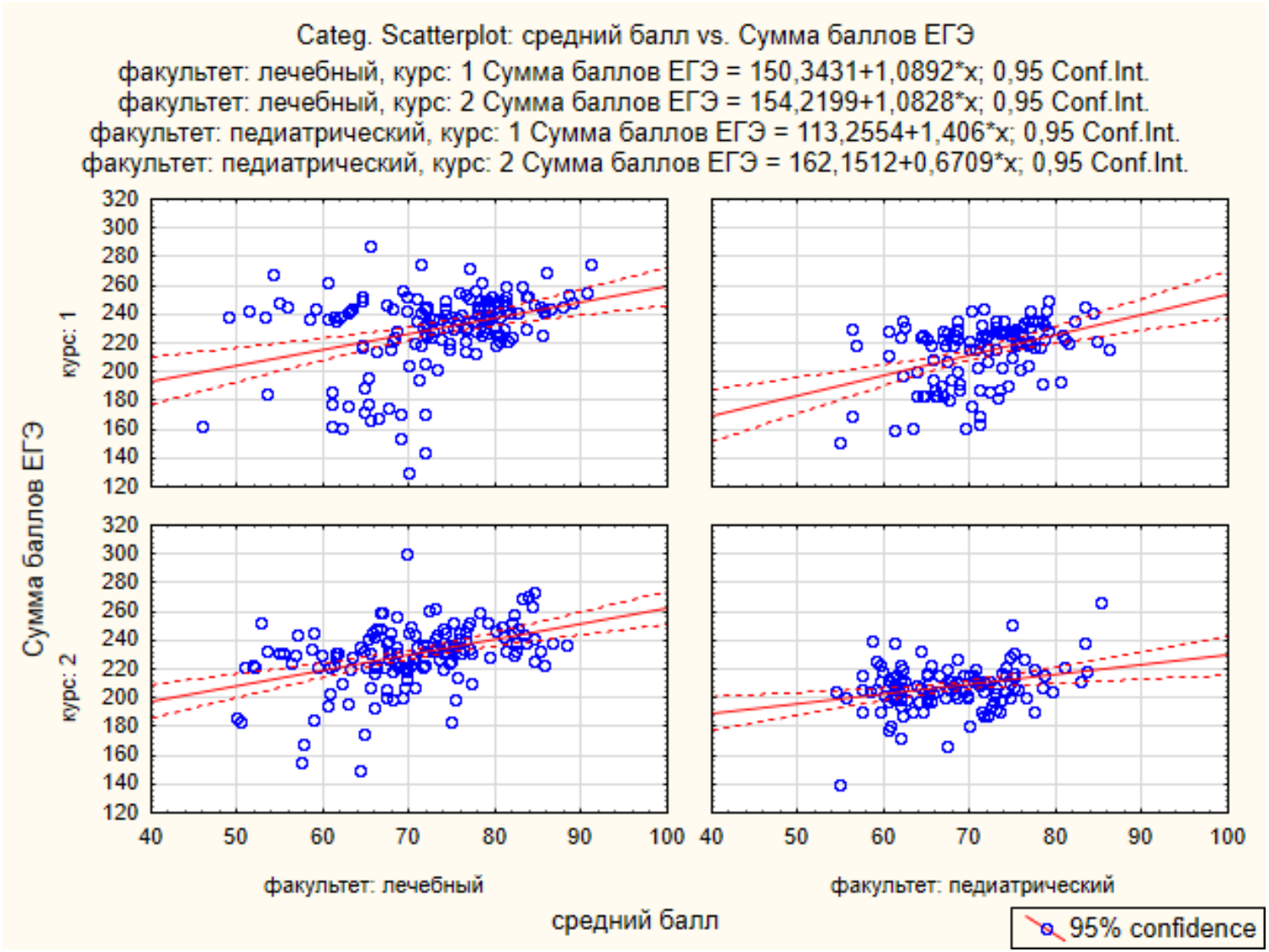
Variable	Descriptive Statistics (Лист1 in статистика)					
	Valid N	Mean	Minimum	Maximum	Std.Dev.	Standard Error
средний балл	537	70,6991	46,0000	91,1450	8,12655	0,350687
количество оценок	537	33,7356	1,0000	56,0000	9,35000	0,403482
Сумма баллов ЕГЭ	528	221,4905	130,0000	300,0000	24,57008	1,069275

Описательная статистика, продолжение

Breakdown Table of Descriptive Statistics (Лист1 in статистика)
Smallest N for any variable: 528

курс	факультет	количество оценок Means	количество оценок Std.Dev.	Сумма баллов ЕГЭ Means	Сумма баллов ЕГЭ Std.Dev.	средний балл Means	средний балл Std.Dev.
1	лечебный	41,08000	9,231926	229,9530	27,74116	73,17233	8,894486
1	педиатрический	34,64463	6,832349	213,5000	21,53858	71,16892	6,390091
2	лечебный	27,96644	7,222727	230,2603	21,83198	70,28835	8,640205
2	педиатрический	30,72650	7,558535	207,5913	16,00652	67,56538	6,912787
All Groups		33,73557	9,349999	221,4905	24,57008	70,69907	8,126554

Корреляция в картинках



Корреляция в цифрах

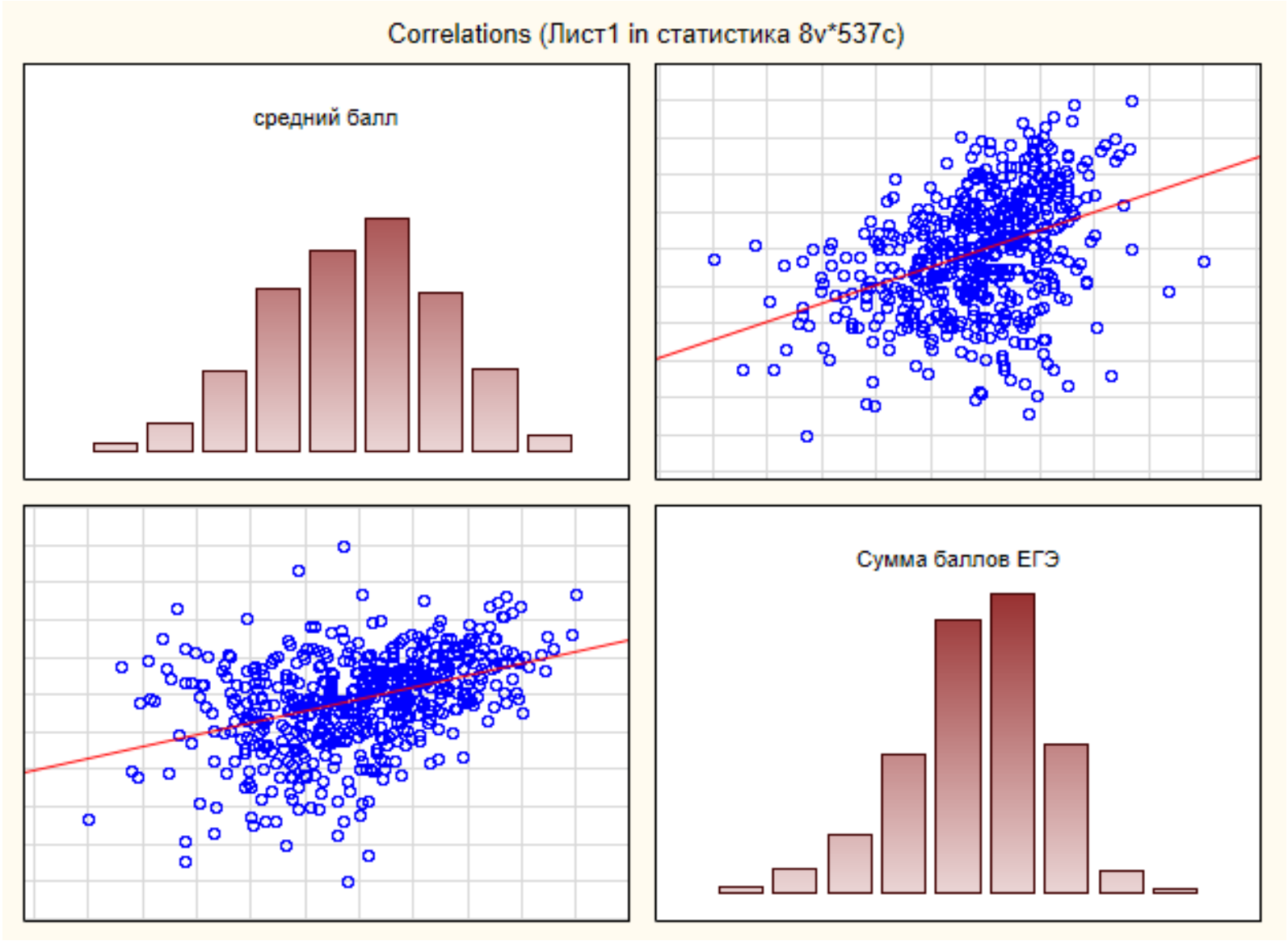
Variables	Within-Group Correlations (Лист1 in статистика) Group: курс:1 факультет:лечебный Marked correlations are significant at p < ,05000 **средний балл**	Сумма баллов ЕГЭ
средний балл	1,000000	0,347987
Сумма баллов ЕГЭ	0,347987	1,000000

Variables	Within-Group Correlations (Лист1 in статистика) Group: курс:1 факультет:педиатрический Marked correlations are significant at p < ,05000 **средний балл**	Сумма баллов ЕГЭ
средний балл	1,000000	0,416506
Сумма баллов ЕГЭ	0,416506	1,000000

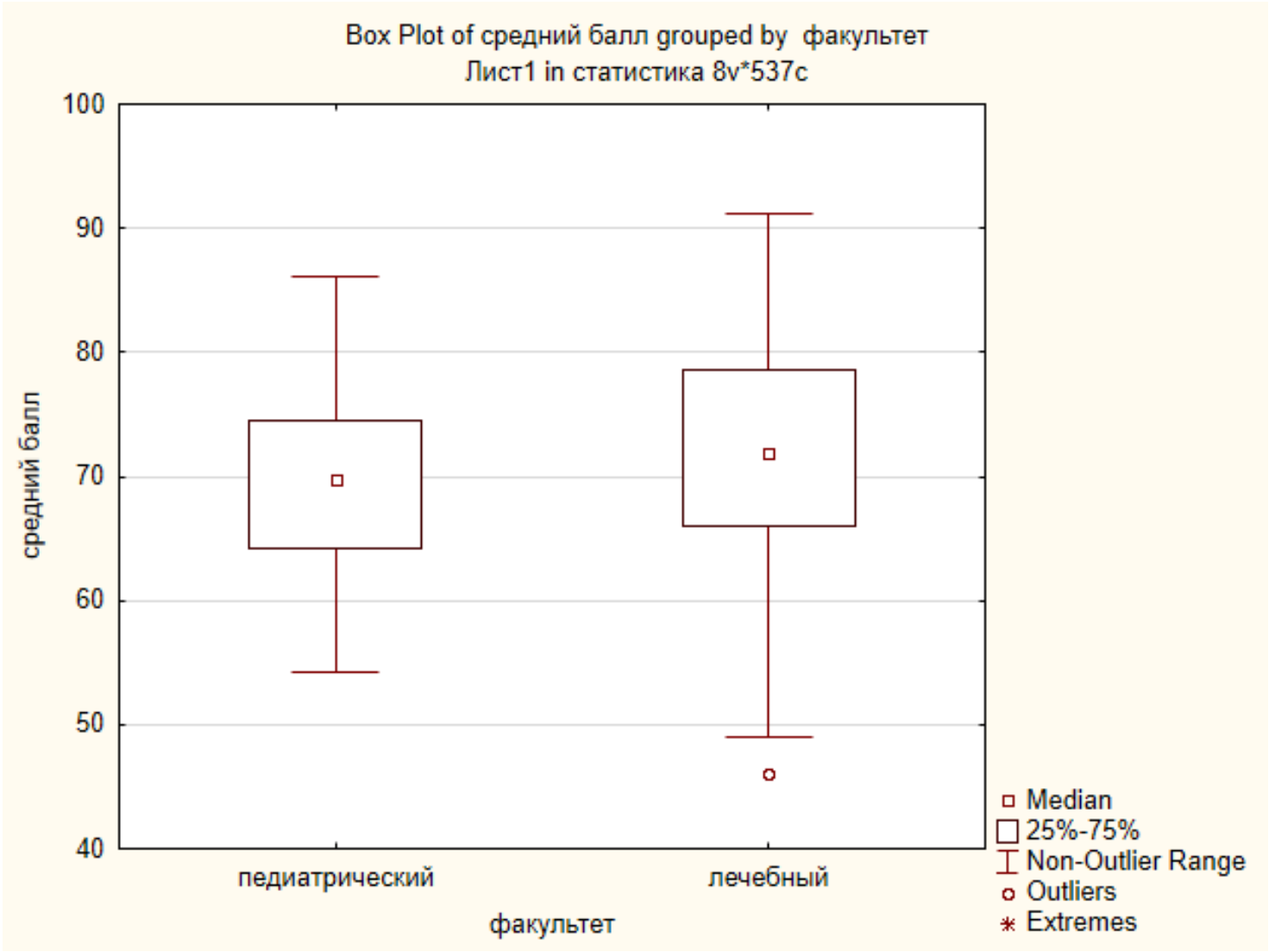
Variables	Within-Group Correlations (Лист1 in статистика) Group: курс:2 факультет:лечебный Marked correlations are significant at p < ,05000 **средний балл**	Сумма баллов ЕГЭ
средний балл	1,000000	0,431266
Сумма баллов ЕГЭ	0,431266	1,000000

Variables	Within-Group Correlations (Лист1 in статистика) Group: курс:2 факультет:педиатрический Marked correlations are significant at p < ,05000 **средний балл**	Сумма баллов ЕГЭ
средний балл	1,000000	0,287644
Сумма баллов ЕГЭ	0,287644	1,000000

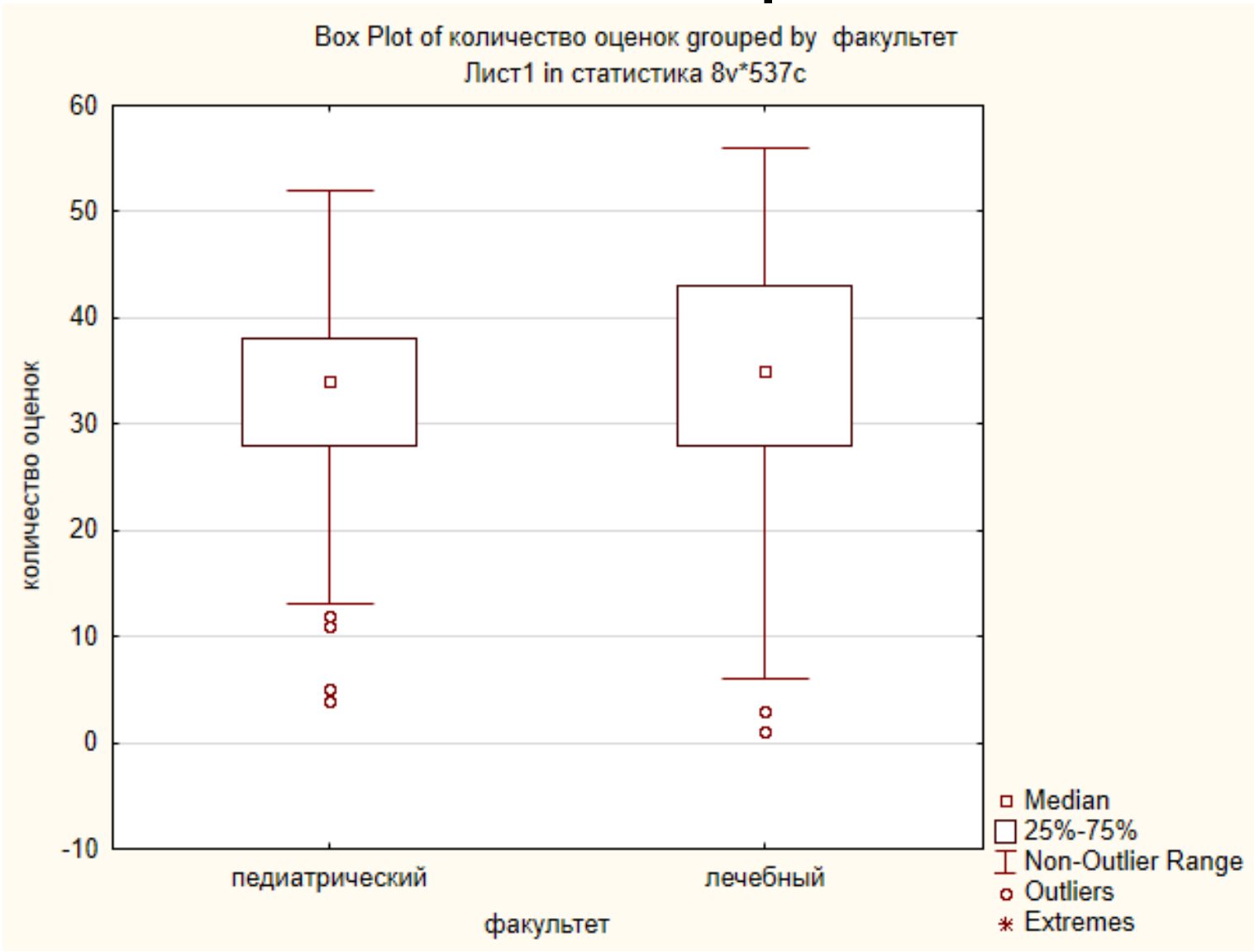
Корреляция всего



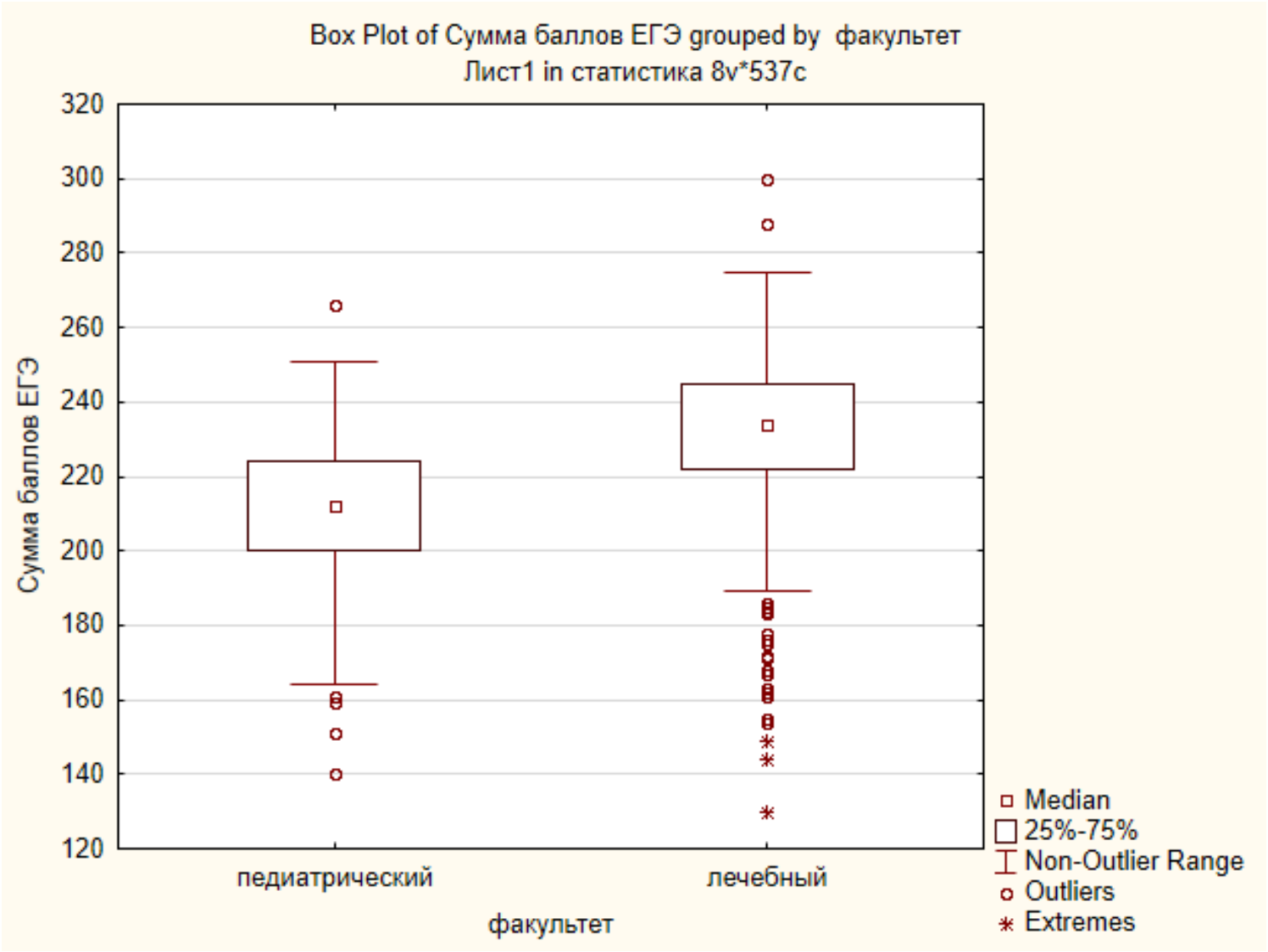
Ищем выбросы



Больше выбросов



Очень много выбросов



Лечебники против педиатров

Variable	T-tests; Grouping: факультет (Лист1 in статистика) Group 1: лечебный Group 2: педиатрический Mean лечебный	Mean педиатрический	t-value	p
средний балл	71,73516	69,39743	3,342783	0,000887

Variable	T-tests; Grouping: факультет (Лист1 in статистика Group 1: лечебный Group 2: педиатрический Mean лечебный	Mean педиатрический	t-value	p
Сумма баллов ЕГЭ	230,1051	210,5837	9,857704	0,000000

Курс на курс

T-tests; Grouping: курс (Лист1 in статистика)
Group 1: 1
Group 2: 2

Variable	Mean 1	Mean 2	t-value	p
средний балл	72,27782	69,09065	4,629789	0,000005

Идеальное исследование глазами медицины доказательств

- Многоцентровое
- Двойное
- Слепое
- Рандомизированное
- Плацебоконтролируемое

МЕТА-АНАЛИЗ

Ошибки и злоупотребления

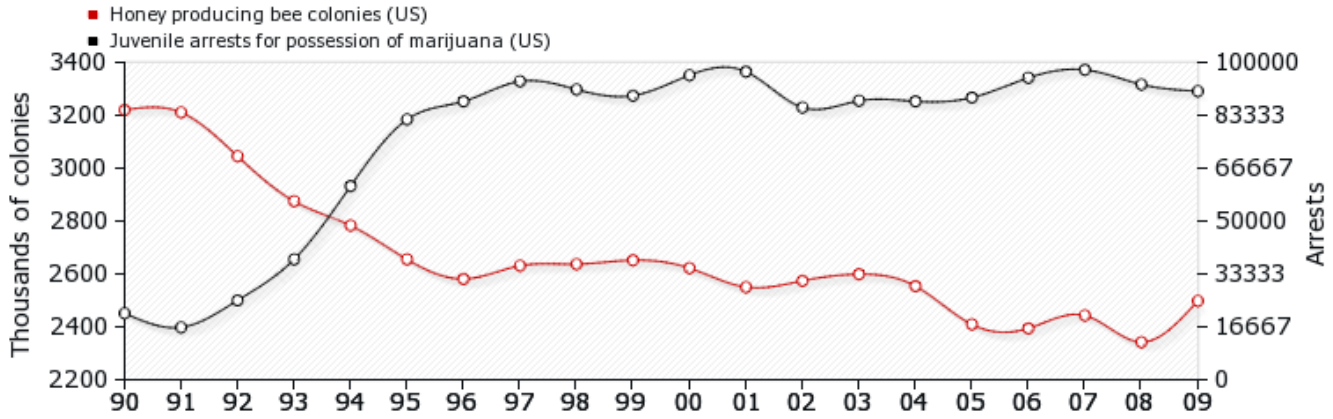
Только та статистика надёжна, которую сфальсифицировали вы сами

Народная мудрость

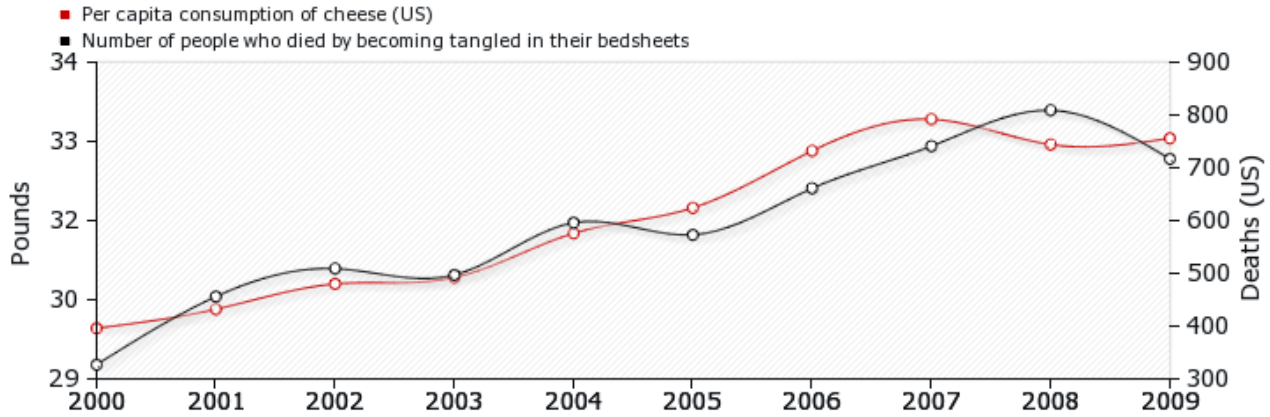
Основные ошибки

- Неправильные критерии включения и исключения
- Недостаточная рандомизация
- Недостаточное количество наблюдений
- Неверный выбор группы методов
- Использование неприменимых методов
- Неправильная интерпретация результатов

Невероятные корреляции

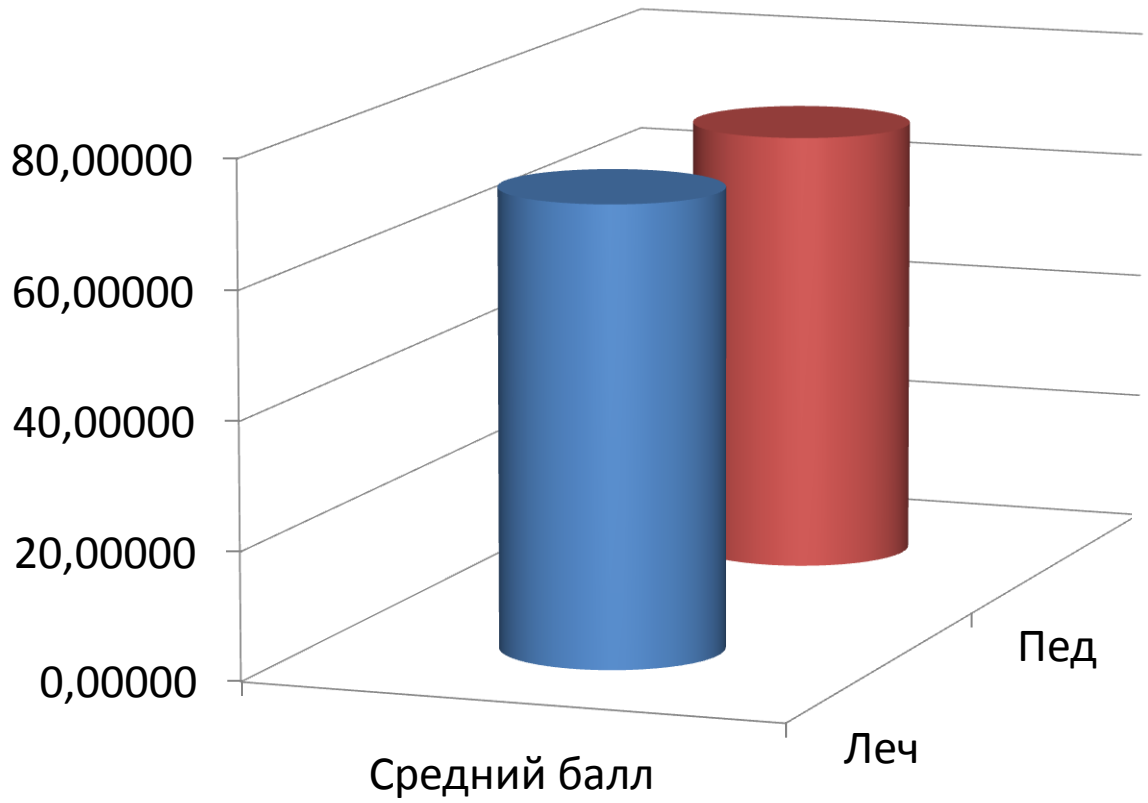


Correlation: -0.933389

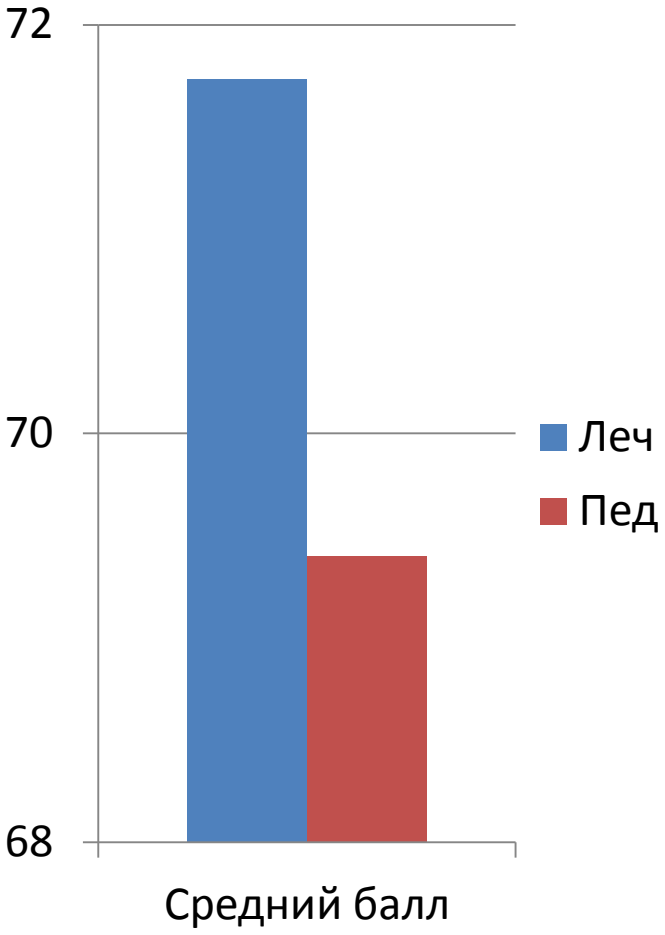


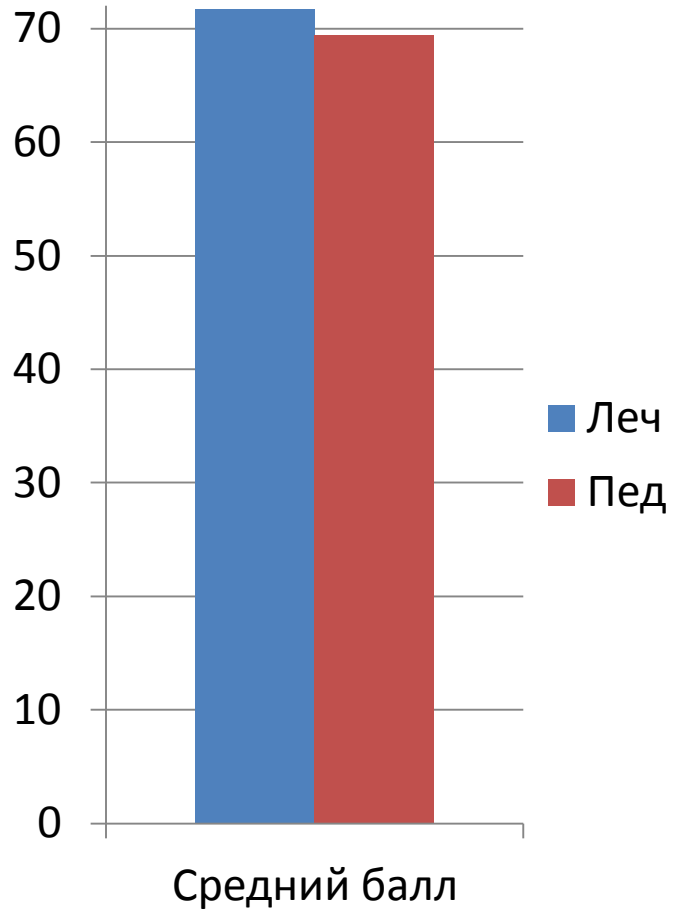
Correlation: 0.947091

Манипуляция визуализацией

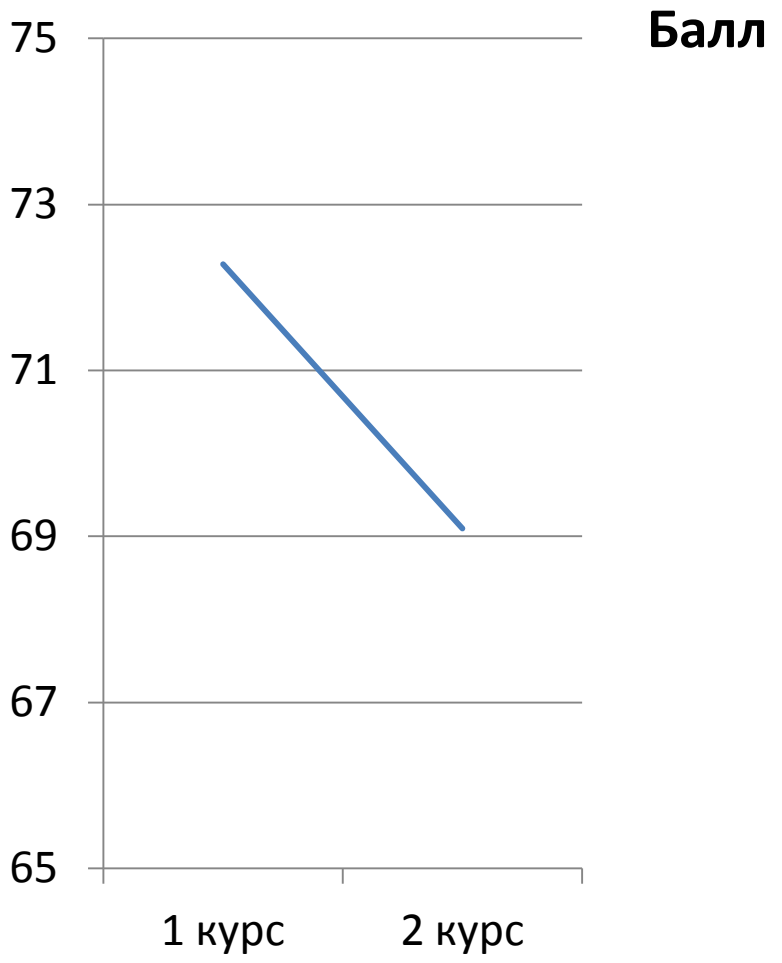


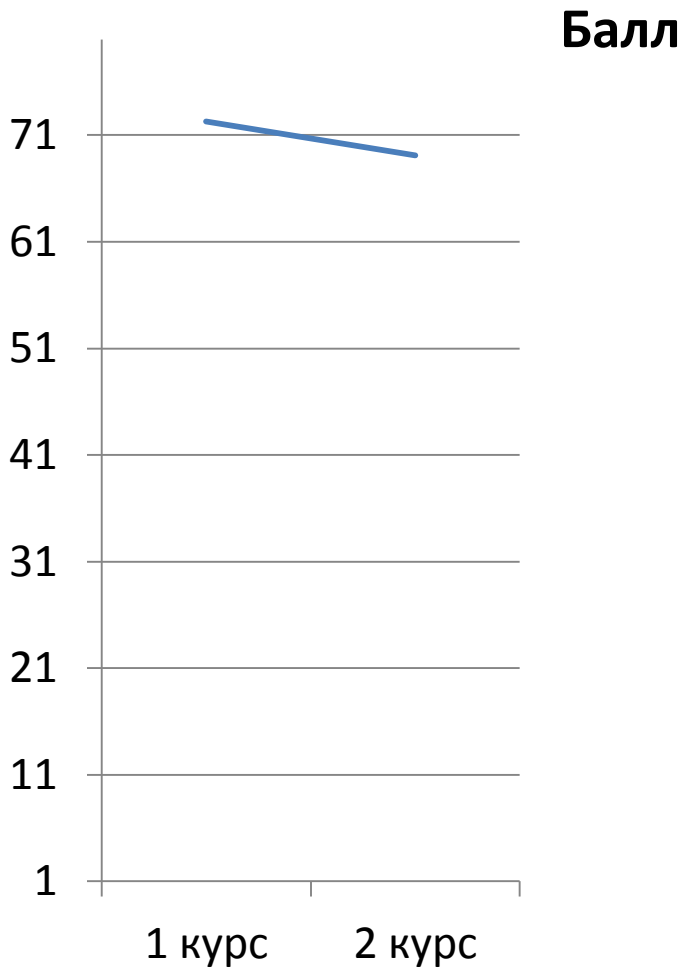
Продолжаем манипулировать



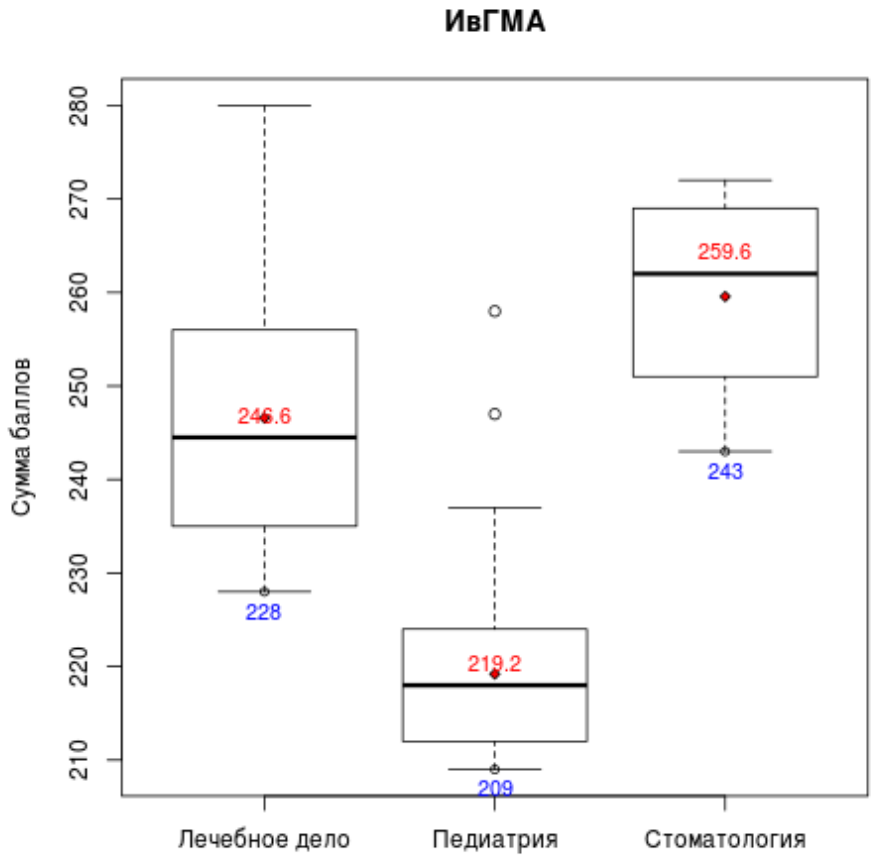


Обманчивые тенденции





Честная графика



Благодарю за внимание

Вопросы?