Статистички софтвер 4

Четврти час

Марија Радичевић

Математички факултет, Београд

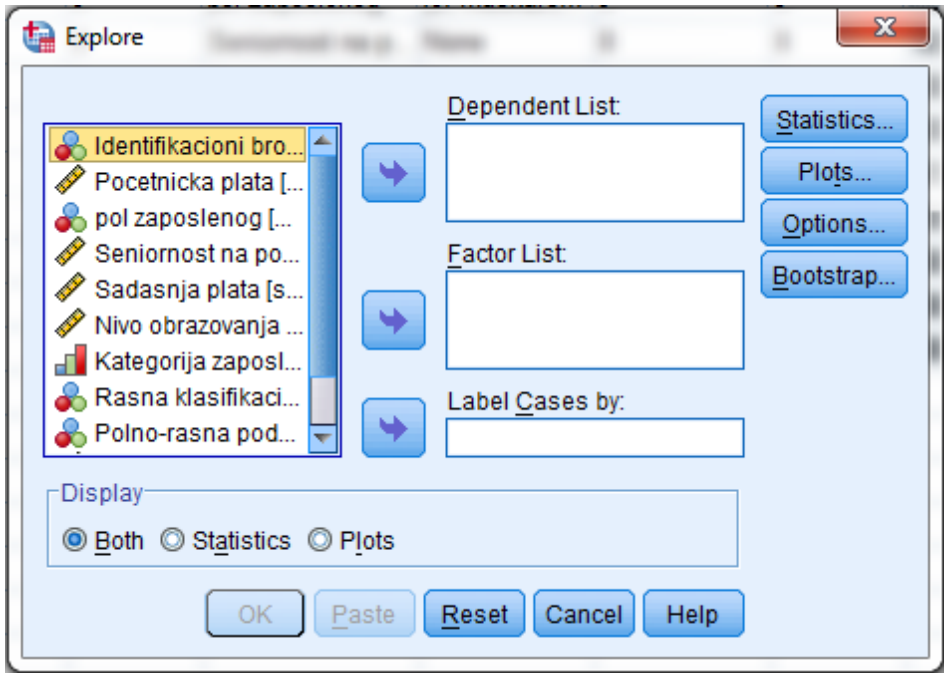
2015.

Садржај

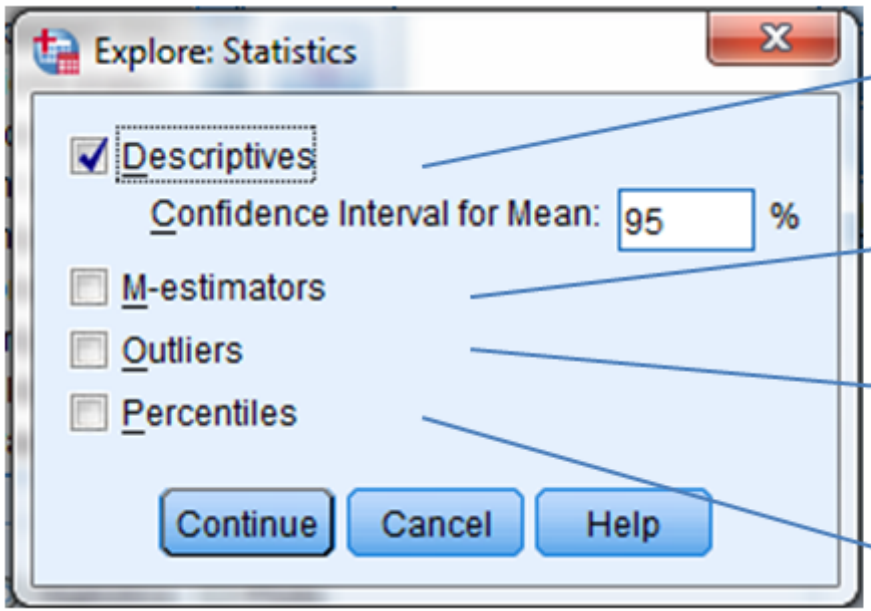
Explore

- информације о расподели података
- робусне оцене централне тенденције појава
- тестове о испуњености претпоставки основних статистичких техника анализе података

Analyze⇒Descriptive Statistics⇒Explore



Статистичке мере



аритметичка средина, интервали поверења за аритметичку средину, медијана, варијанса,...

робусне статистике максималне веродостојности

5 највећих и 5 најмањих вредности промењљиве

5%, 10%, 25%, 50%, 75%, 90%, 95%

Дескриптивне статистике

	Kategorija zaposlenih			Statistic	Std. Error
Pocetnicka plata	sluzbenik	Mean		5733,95	84,423
		95% Confidence Interval for Mean	Lower Bound	5567,59	
			Upper Bound	5900,30	
		5% Trimmed Mean		5661,71	
		Median		5700,00	
		Variance		1617875,502	
		Std. Deviation		1271,957	
		Minimum		3600	
		Maximum		12792	
		Range		9192	
		Interquartile Range		1500	
		Skewness		1,251	,162
		Kurtosis		4,470	,322
	obuceni sluzbenik	Mean		5478,97	80,322
		95% Confidence Interval for Mean	Lower Bound	5320,12	
			Upper Bound	5637,82	
		5% Trimmed Mean		5440,49	
		Median		5400,00	

Робустне оцене централне тенденције

- зависе од једноставних претпоставки основне расподеле података и нису јако осетљиве на њихово нарушавање
- мере на које екстремне опсервације (аутлајери) имају мали ефекат
- робустна статистика је отпорна на грешке резултата, које настају одступањем од претпостављене расподеле
- на приме, медијана је робусна оцена, док средња вредност није

⋆ робустне статистике $\rightsquigarrow$ опис искривљених статистика (са екстремним опсервацијама)

⋆ неробустне статистике $\rightsquigarrow$ опис симетричних расподела

Поткресана средина

- искључује екстремне (нетипичне) вредности
- пондерисана аритметичка средина: 0−нетипичне вредности, 1−опсервације ближе централном делу расподеле

M-процењивачи: заснивају се на стандардизованом одступању d_i вредности појединачних опсервација од оцењене локације

Хуберова оцена до критичне тачке 1.339 пондер је 1, а затим вредност пондера опада

Туркиева оцена вредност пондера се смањује како одступање расте до тачке 4.685, а затим је једнак 0

Хампелова оцена има три критчне тачке: 1.7, 3.4 и 8.5; пондер су: 1, $1.7/d_i$, $\frac{1.7}{d_i}\frac{8.5-d_i}{5.1}$, 0, респективно

Андруова оцена слично Туркијевој, са критичном тачком 1.339π

Робусне оцене централне тенденције

M-Estimators[e]

	Kategorija zaposlenih	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
Pocetnicka plata	sluzbenik	5660,84	5603,36	5630,91	5602,79
	obuceni sluzbenik	5433,99	5439,22	5455,36	5439,36
	sluzbenik obezbedjenja	.	.	.	.
	sluzbenik sa fakult. diplomom	9867,12	9887,87	9927,55	9888,07
	posebna grupa sluzbenika	12850,37	12674,23	12784,55	12674,69
	MBA diploma	13319,54	13557,53	13650,01	13549,47
	tehnicka sluza	18493,91	17657,92	17918,58	17657,65

a. The weighting constant is 1,339.

b. The weighting constant is 4,685.

c. The weighting constants are 1,700, 3,400, and 8,500

d. The weighting constant is 1,340*pi.

e. Some M-Estimators cannot be computed because of the highly centralized distribution around the median.

Перцентили и екстремне вредности

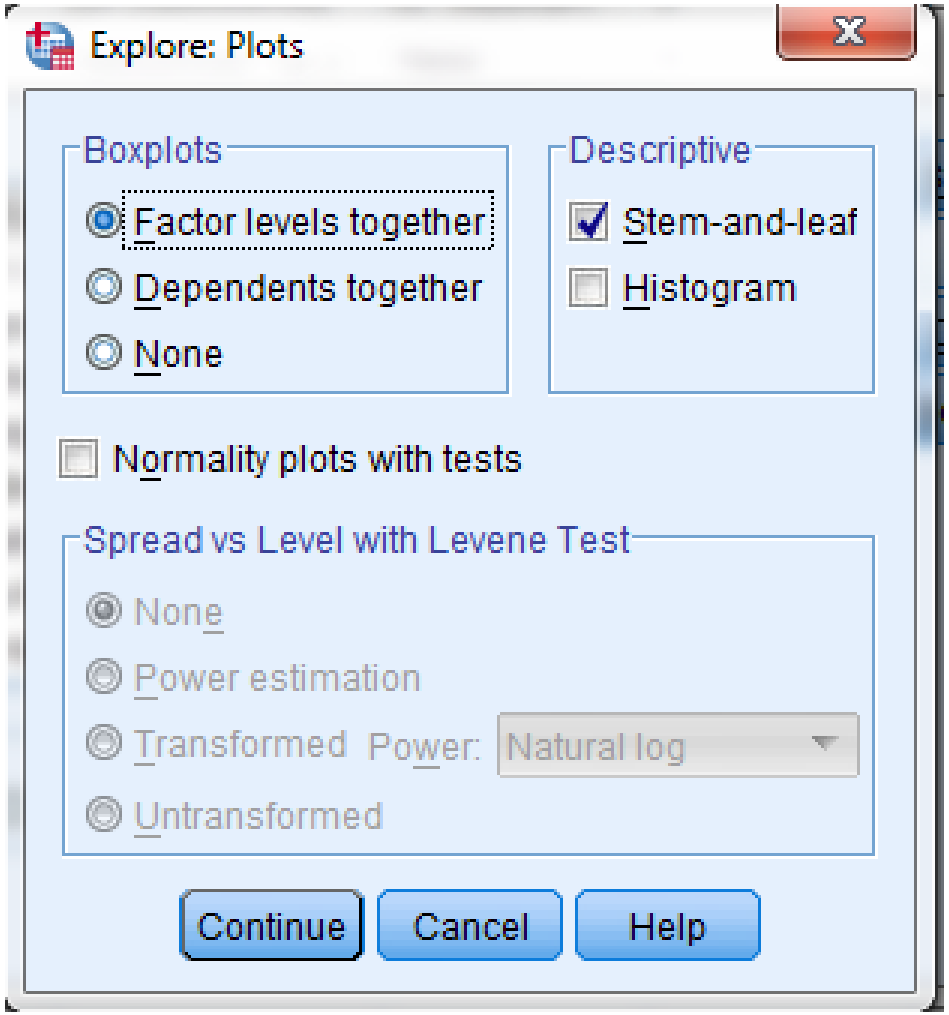
Percentiles

			Percentiles						
		Kategorija zaposlenih	5	10	25	50	75	90	95
Weighted Average (Definition 1)	Pocetnicka plata	sluzbenik	4080,00	4080,00	4800,00	5700,00	6300,00	7200,00	7800,00
		obuceni sluzbenik	4380,00	4422,00	4500,00	5400,00	6300,00	6600,00	7200,00
		sluzbenik obezbedjenja	4320,00	5592,00	6000,00	6300,00	6300,00	6300,00	6300,00
		sluzbenik sa fakult. diplomom	6909,60	7560,00	8400,00	9492,00	11646,00	12801,60	13179,60
		posebna grupa sluzbenika	9143,40	9997,20	10992,00	13098,00	14376,00	17568,00	20747,40
		MBA diploma	7200,00	7200,00	10098,00	12996,00	15498,00	.	.
		tehnicka sluzba	13992,00	13992,00	16242,00	18000,00	23748,00	.	.
Tukey's Hinges	Pocetnicka plata	sluzbenik			4800,00	5700,00	6300,00		
		obuceni sluzbenik			4500,00	5400,00	6300,00		
		sluzbenik obezbedjenja			6000,00	6300,00	6300,00		
		sluzbenik sa fakult. diplomom			8400,00	9492,00	11496,00		
		posebna grupa sluzbenika			10992,00	13098,00	14256,00		
		MBA diploma			12996,00	12996,00	13992,00		
		tehnicka sluzba			16992,00	18000,00	21000,00		

Extreme Values[h]

	Kategorija zaposlenih			Case Number	Identifikacioni broj	Value
Pocetnicka plata	sluzbenik	Highest	1	227	932	12792
			2	186	783	11100
			3	217	925	9300
			4	218	1041	8496
			5	219	997	8400
		Lowest	1	38	1010	3600
			2	22	1096	3600
			3	11	921	3600
			4	50	741	3900
			5	48	945	3900[a]
	obuceni sluzbenik	Highest	1	202	796	7800

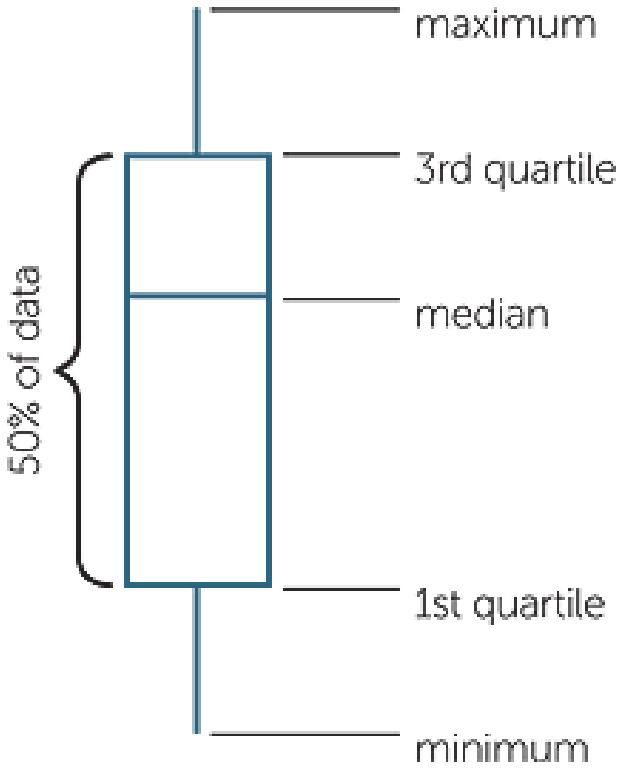
Графички приказ



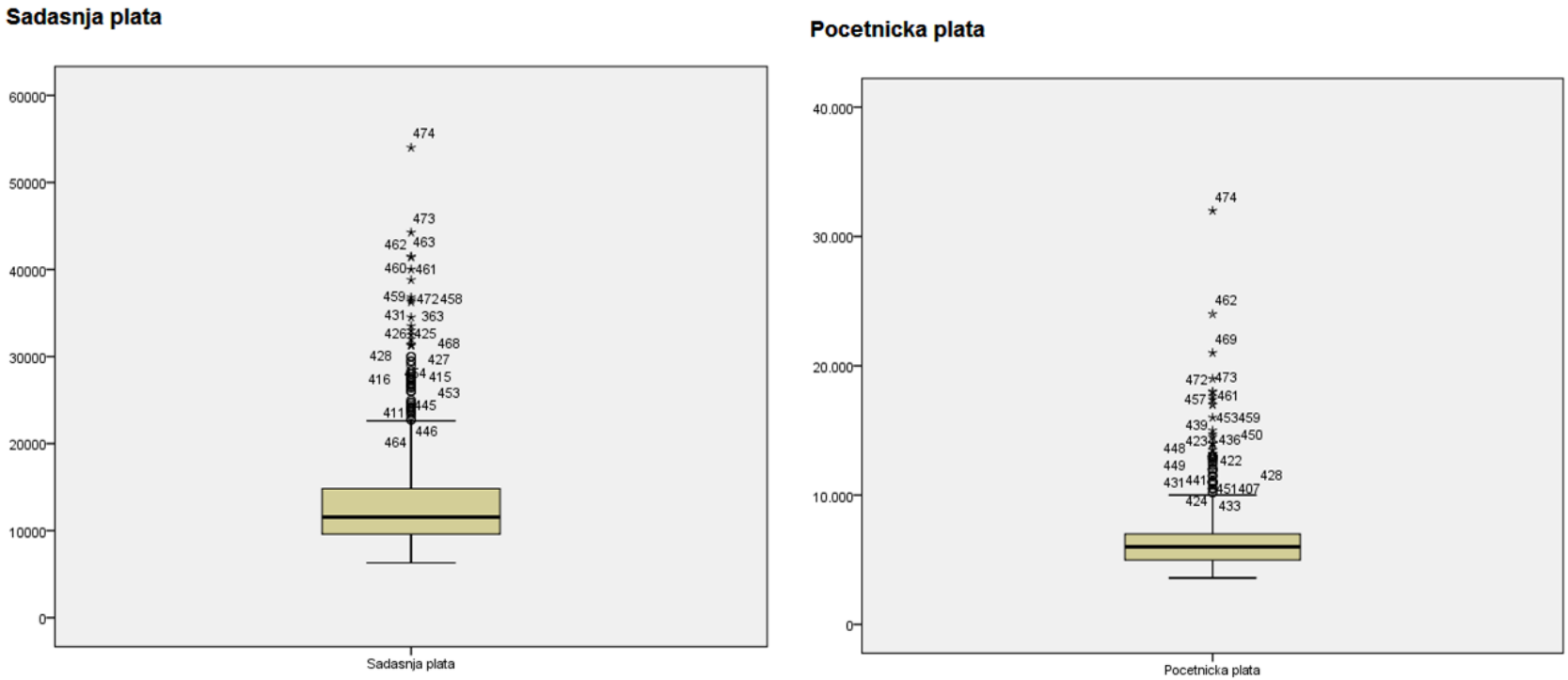
Box plot

Графички приказ распореда сумарних статистичких мера.

- Интерквартилно растојање: $I_{qr} = q_3 - q_1$
- Унутрашње границе узорка: $f_1 = q_1 - 1.5I_{qr}, f_3 = q_3 + 1.5I_{qr}$
- Спољашње границе узорка: $F_1 = q_1 - 3I_{qr}, F_3 = q_3 + 3I_{qr}$

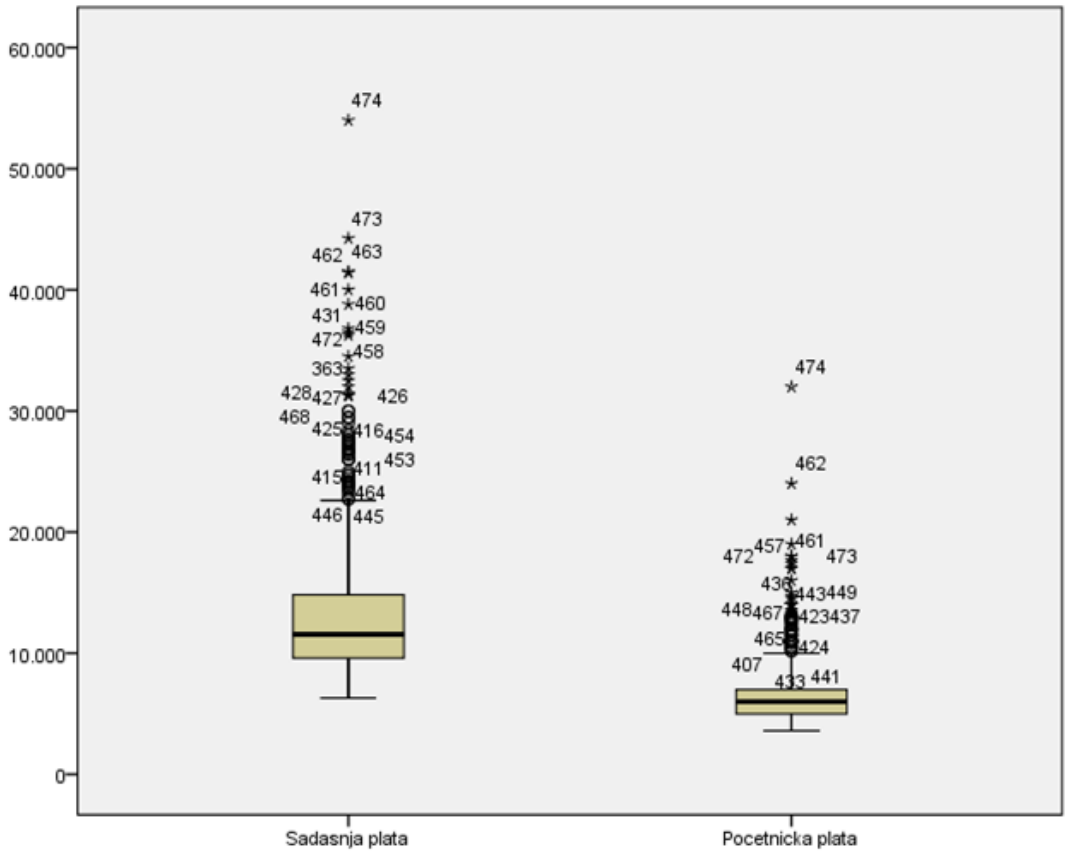


Box plot



Factor Levels together - за сваку групу једне промењиве на једном дијаграму, а ако има више зависних промењивих приказује се посебно дијаграм за сваку зависну промењиву

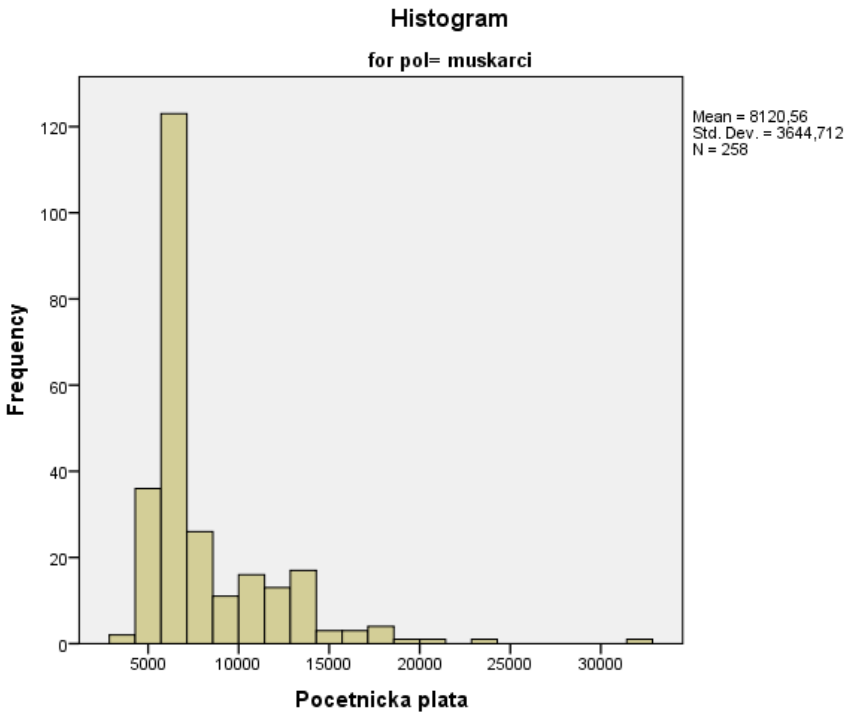
Box plot

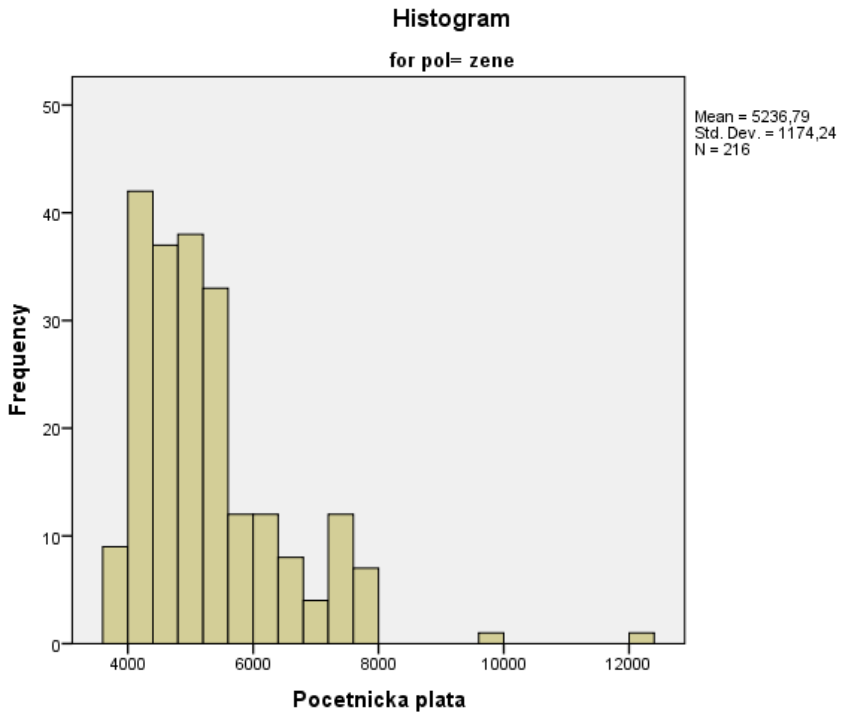


Dependents together

- за све зависне промењљиве и за сваку групу приказује се један дијаграм

Histogram


Histogram
for pol= muskarci
Mean = 8120,56
Std. Dev. = 3644,712
N = 258
Frequency
Pocetnicka plata
0
20
40
60
80
100
120
5000
10000
15000
20000
25000
30000


Histogram
for pol= zene
Mean = 5236,79
Std. Dev. = 1174,24
N = 216
Frequency
Pocetnicka plata
0
10
20
30
40
50
4000
6000
8000
10000
12000

Дијаграм стабљике и листова

```
Pocetnicka plata Stem-and-Leaf Plot

 Frequency     Stem &  Leaf

    11,00        3 .  66999
    46,00        4 .  00000000000022333333344
    62,00        4 .  55555555555556666888888888999
    56,00        5 .  11111111222444444444444445&
    37,00        5 .  55566777777777778
   106,00        6 .  000000000000000000000000013333333333333333333333334&
    38,00        6 .  666666666666699999&
    14,00        7 .  2222222
    20,00        7 .  5558888889
    11,00        8 .  1444&
     6,00        8 .  77&
     5,00        9 .  34&
     1,00        9 .  &
     1,00       10 .  &
    60,00 Extremes    (>=10200)

Stem width:    1000
Each leaf:        2 case(s)
```

фреквенције опсервација

водећа цифра стабла

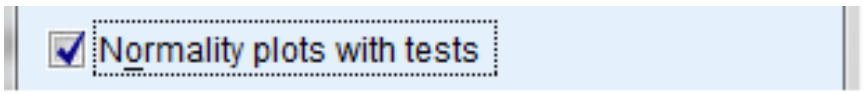
детаљна дистрибуција појава

ширина стабла

број опсервација представљених једним листом

Тестирање нормалности

- једна од најчешћих претпоставки примене статистичких техника је нормалност зависне променљиве
- тестови који се најчешће користе су Колмогоров-Смирнов и Шапиро-Вилк
- за визуелну оцену нормалности се користе хистограм, $Q-Q$ график и $P-P$ график



Колмогоров-Смирнов

- велик обим узорка
- H_0: подаци имају нормалну расподелу
- тест статистика: $D_n = \sup |F_n(x) - G(x)|$

Шапиро-Вилк

- за узорке обима до 2000
- H_0: подаци имају нормалну расподелу
- тест статистика: $W = \frac{(\sum a_i y_{(i)})^2}{\sum (y_i - \bar{y})^2}$

$p-$вредност$< \alpha \Rightarrow$ одбацујемо нулту хипотезу

$p-$вредност$> \alpha \Rightarrow$ прихватамо нулту хипотезу

Tests of Normality

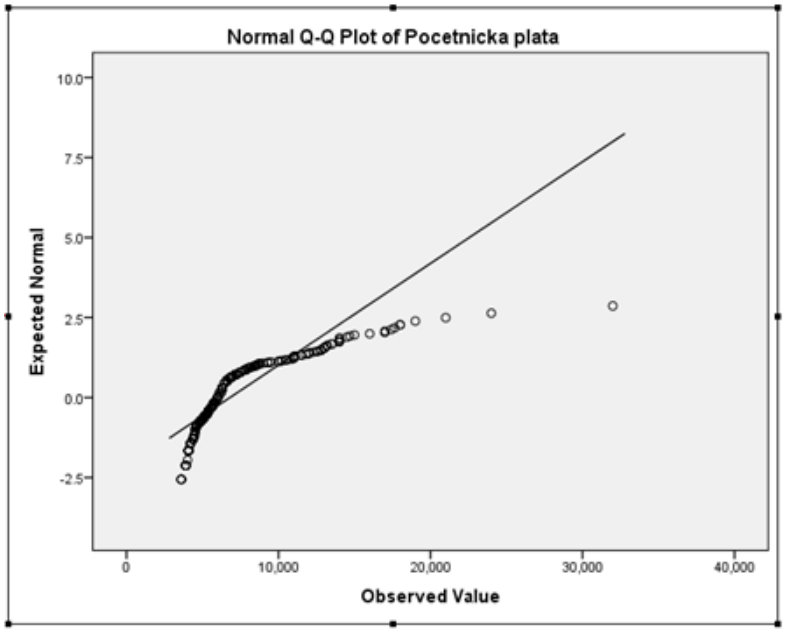
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Pocetnicka plata	.252	474	.000	.715	474	.000

a. Lilliefors Significance Correction

$\Rightarrow$одбацује се H_0

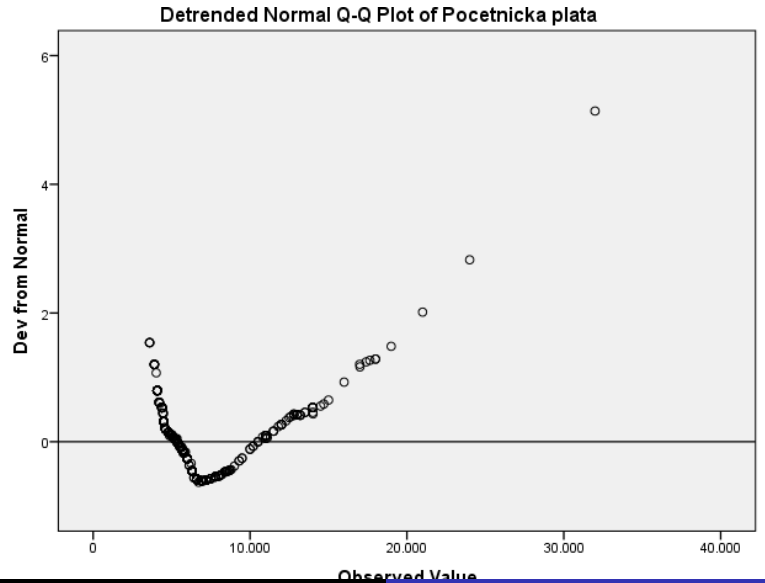
Q-Q plot

- $x-$оса: емпиријске вредности квантила
- $y-$оса: очекиване вредности квантила из нормалне расподеле
- када подаци заиста потичу из нормалне расподеле подаци леже близу праве линије

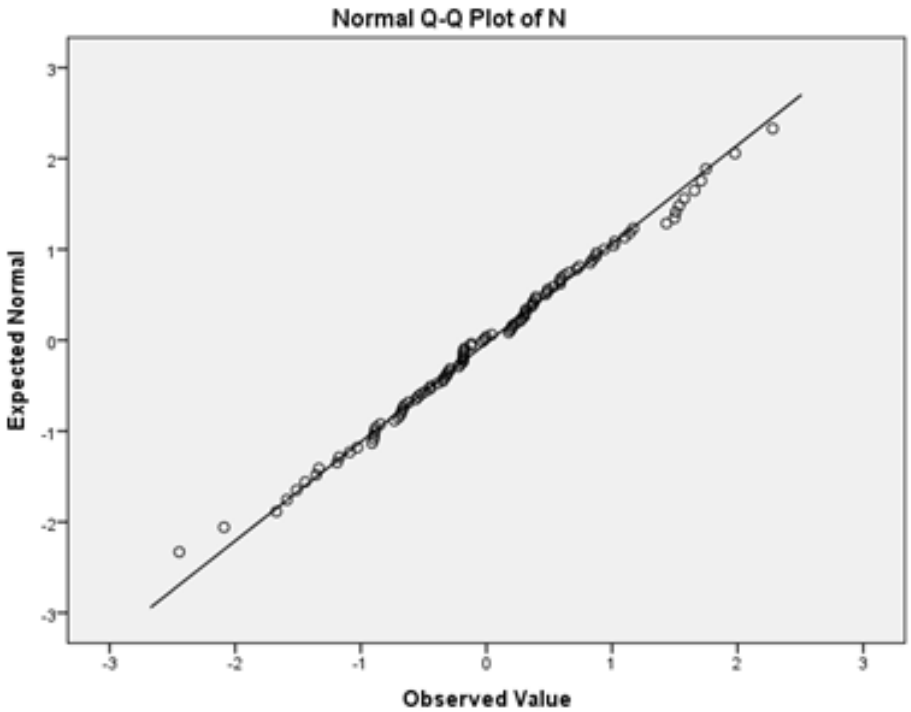


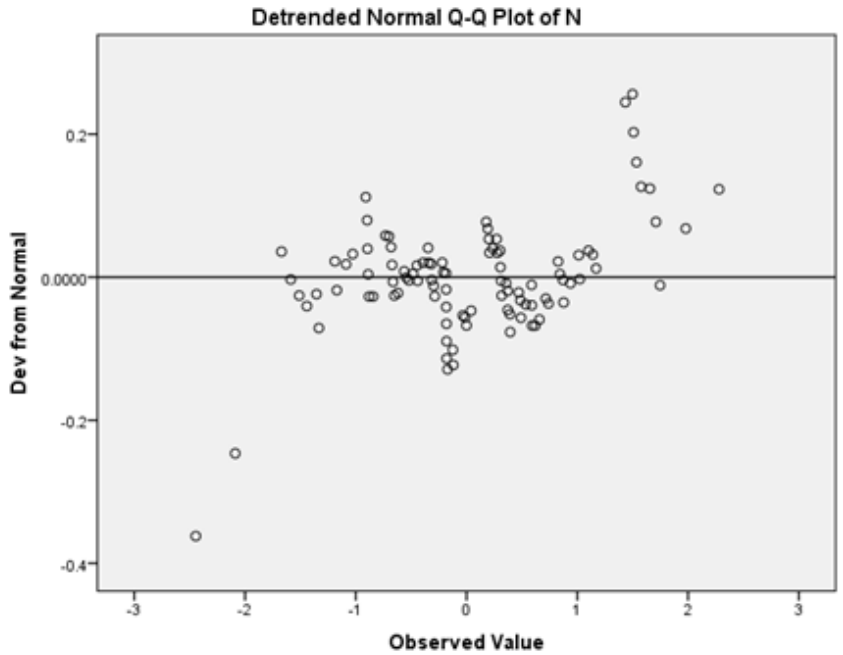
Detrended Q-Q plot

- после искључења нормалног тренда
- нормална расподела је представљена правом хоризонталном линијом из тачке 0
- тачке се добијају одузимањем очекиваног од посматраног квантила
- ако је емпиријска расподела нормална онда су тачке случајно распоређене око праве линије

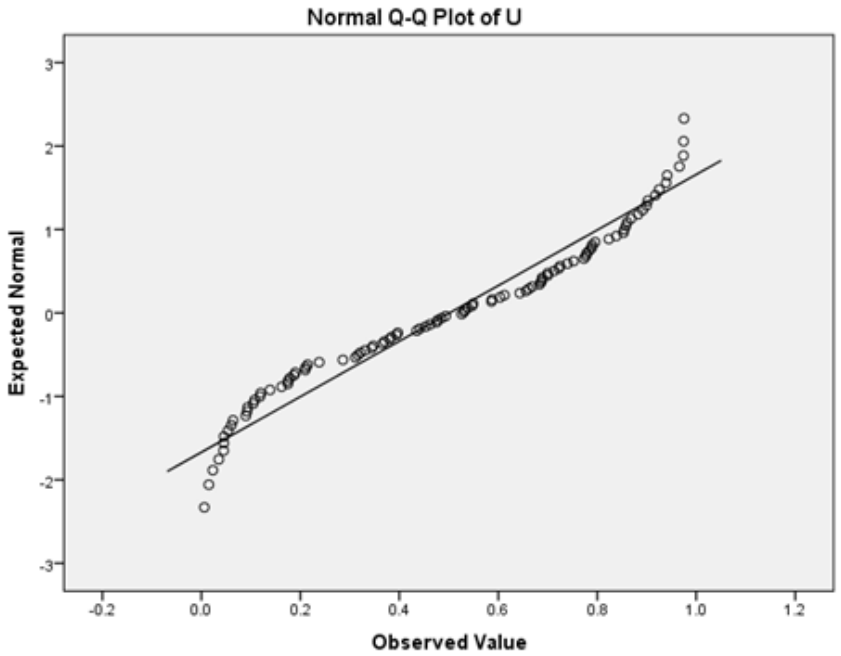


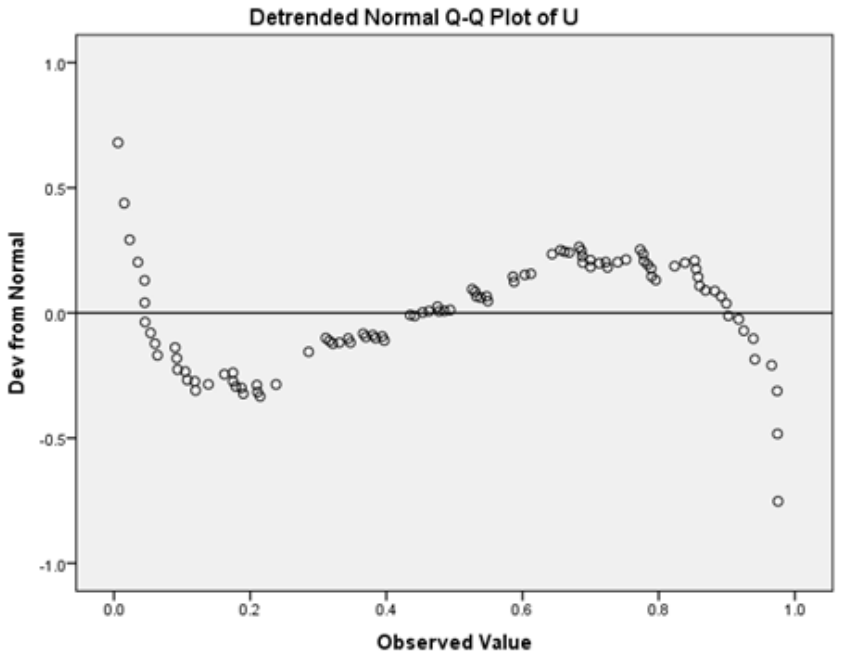
Q-Q plot нормалне расподеле



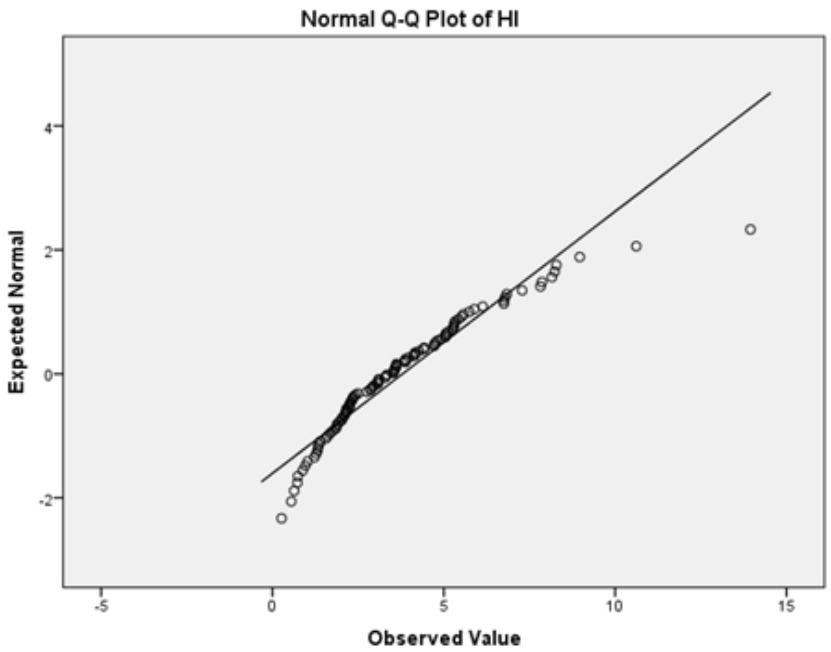


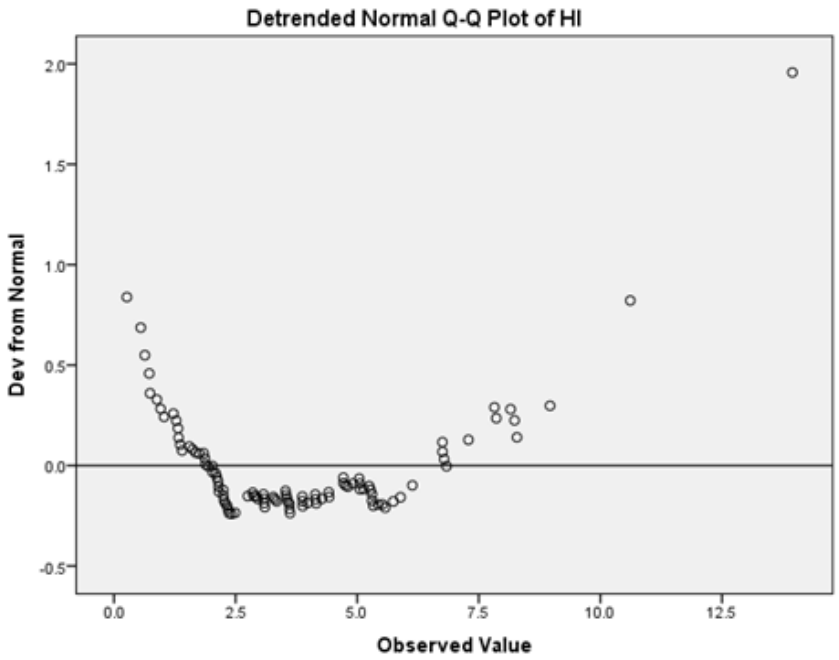
Q-Q plot униформне расподеле



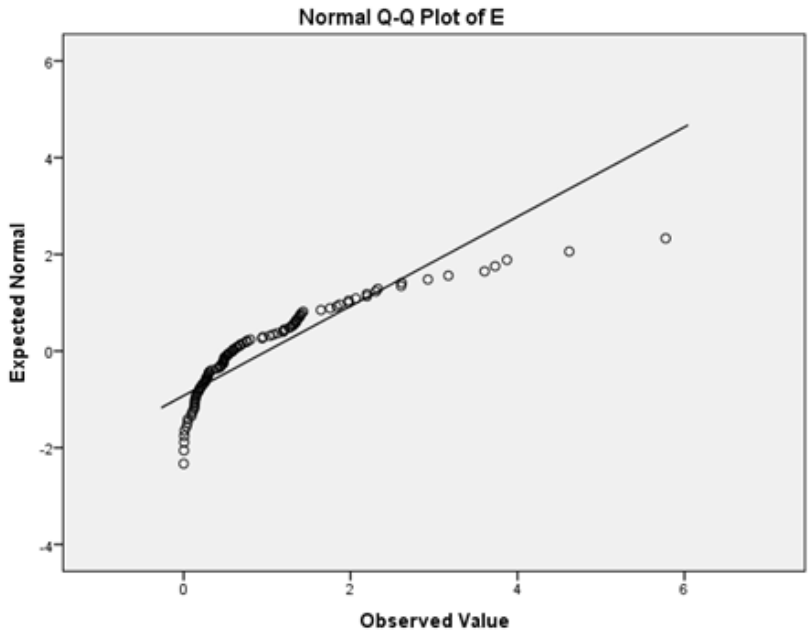


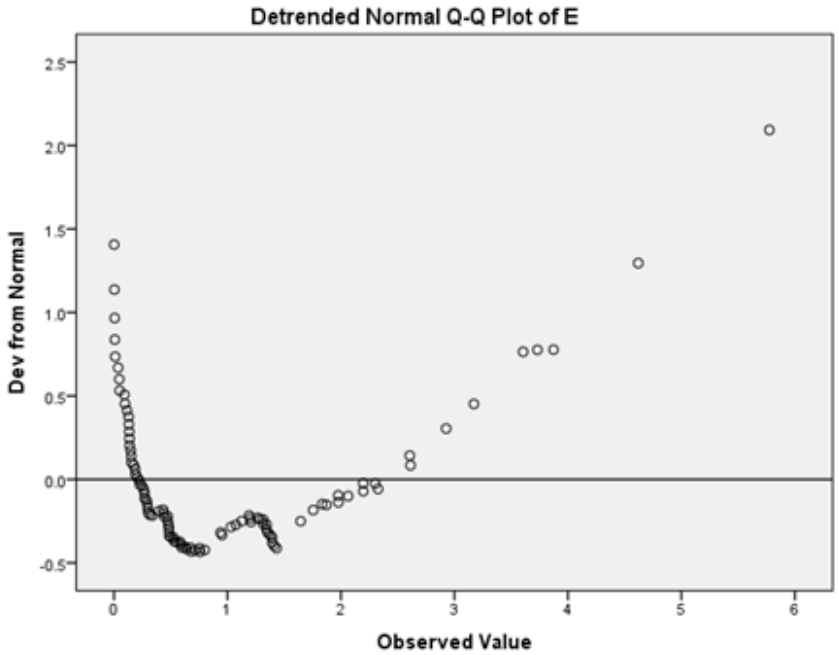
Q-Q plot хи-квадрат расподеле



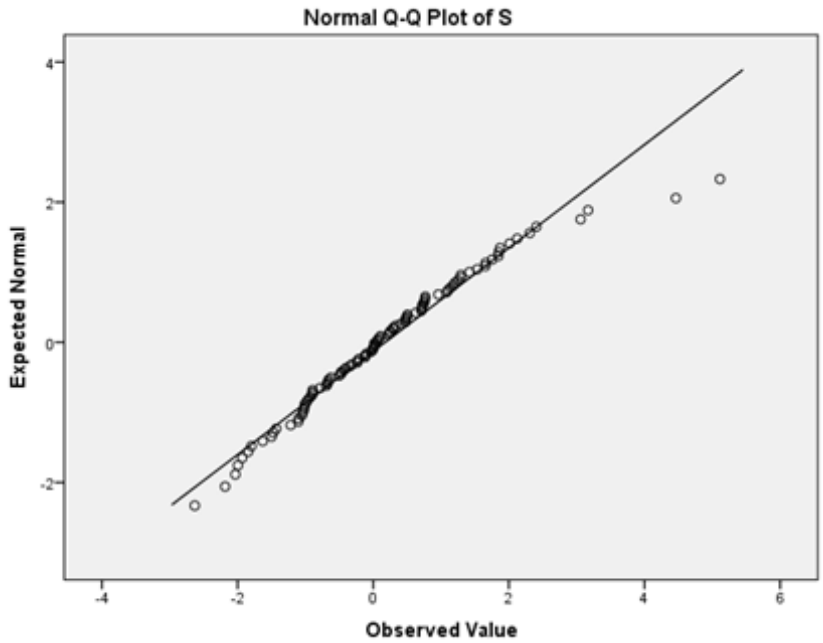


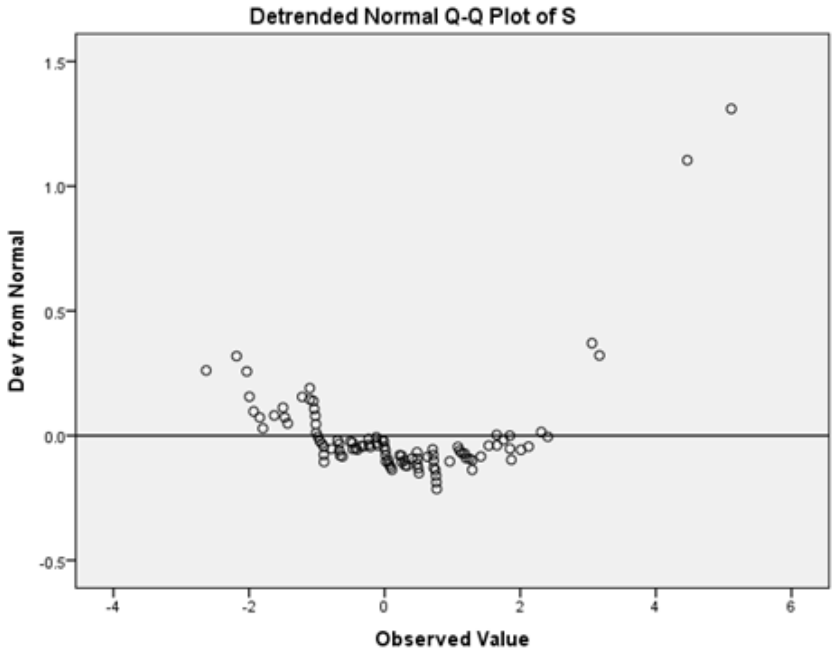
Q-Q plot експоненцијалне расподеле





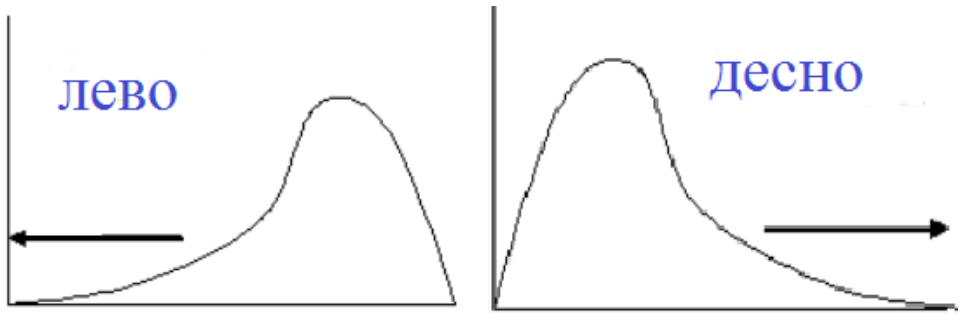
Q-Q plot Студентове расподеле





Трансформације промењљиве ради постизања нормалне расподеле

- приближавање расподеле података, која није нормална, нормалној расподели
- постоје две форме трансформације у зависности да ли је расподела лево или десно искошена
- врсте трансформација промењљиве X:
  1. логаритамска трансформација: $LOG10(X)$
  2. корена трансформација: $SQRT(X)$
  3. инверзна трансформација: $1/X$

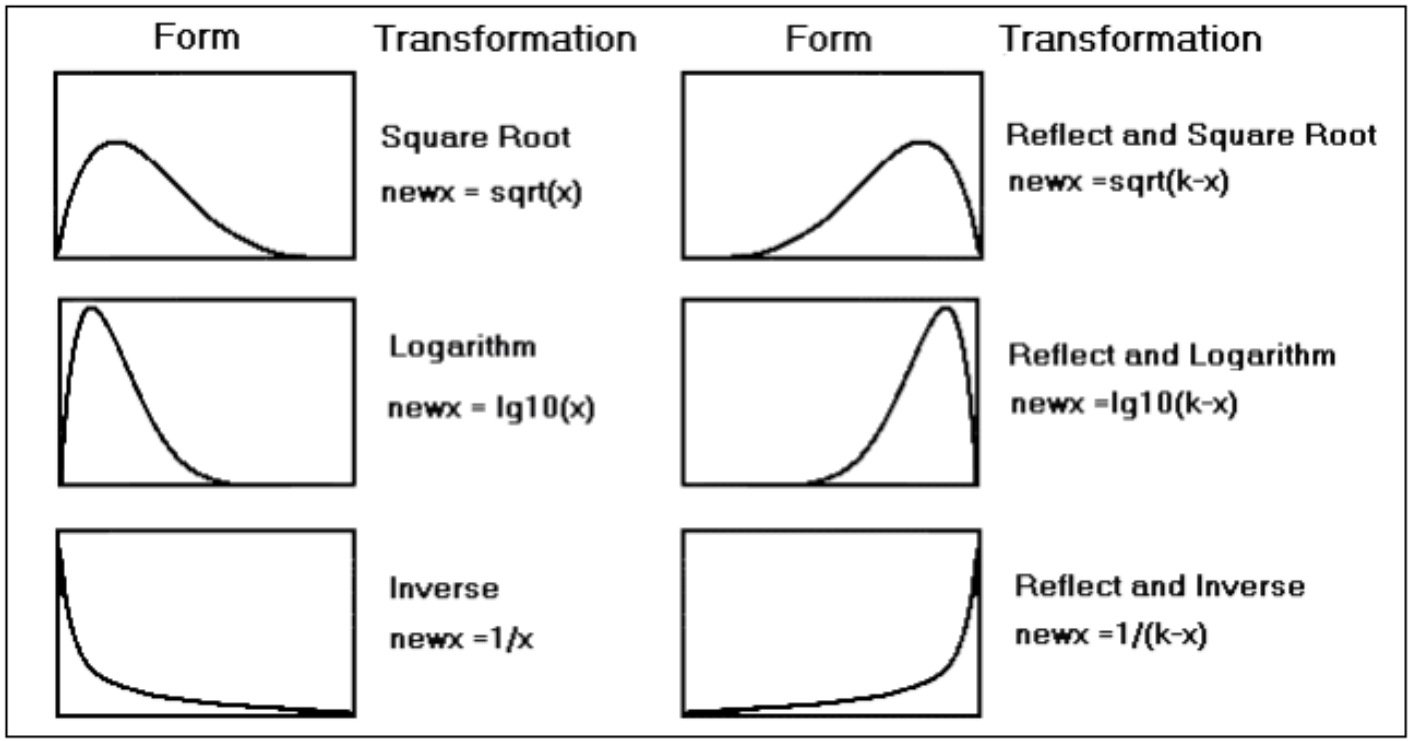


- за све трансформације аргумент мора бити већи од нуле
- ако има негативних вредности онда се врши додатна трансформација

  десно искошене трансформација се врши на основу минимума тако што се свакој вредности дода апсолутна вредност минимума увећана за 1

  лево искошене врши се рефлексија на следећи начин: $1 - vrednost + |\max\_vrednost|$, чиме је добијена десно искошена расподела

Код негативне асиметрије (лева искошеност) прво се изврши рефлексија у позитивну асимертију $(\max_i x_i + 1 - x_i)$, а затим се примени одговарајућа трансформација.



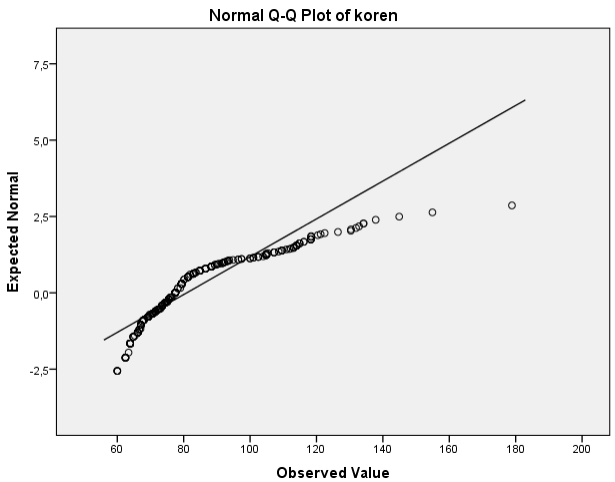
У пракси је препорука да се покуша више трансформација а на основу особина добијених података да се одреди која је најбоља.

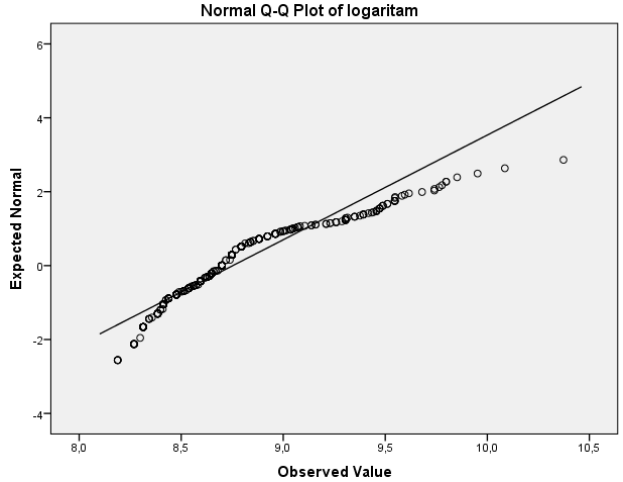
Пример

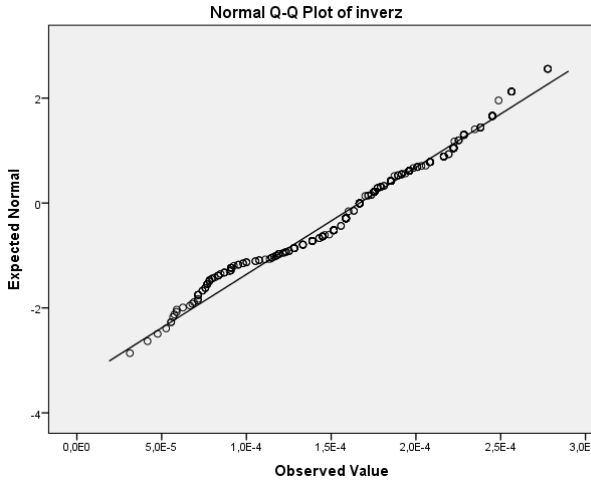
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
koren	,218	474	,000	,818	474	,000
logaritam	,179	474	,000	,897	474	,000
inverz	,104	474	,000	,976	474	,000

a. Lilliefors Significance Correction







Једнакост варијансе

- често је потребно тестирати једнакост (хомогеност) варијансе неке појаве за различите групе (предуслов, на пример, за $t-$тест и *ANOVA*-у)
- најчешће се користи Левене тест, јер се може користити за различите расподеле
  1. рачунање апсолутних разлика вредности промењљиве за сваку опсервацију од групне средине
  2. примена анализе варијансе на тако добијене разлике (ако су резидуали приближно једнаки онда је просечна вредност једнака у свим групама)

- график *Spread vs Level* показује однос између варијабилитета и нивоа појаве за различите групе
- ако су тачке приблизно око хоризонталне линије, не постоји зависност
- ако варијансе нису једнаке, могу се трансформисати подаци у зависности од нагиба (*slope*) праве линије, тј. вредности 1−нагиб

1-нагиб	тип трансформације
3	кубна трансформација
2	квадратна трансформација
1	нема трансформације
$\frac{1}{2}$	квадратни корен
0	логаритамска трансформација
$-\frac{1}{2}$	реципрочна вредност квадратног корена
-1	реципрочна вредност

Опције у Spread vs Level

Power estimation даје график природног логаритма за интерквартиле рангова насупрот природном логаритму медијана свих ћелија

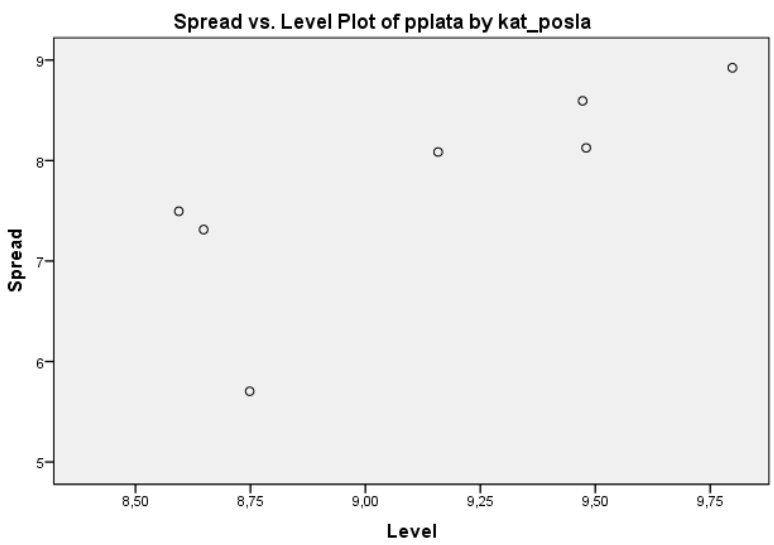
Transformed допушта да се изабере једна од *power alternatives* и даје график трансформисаних података као однос интерквартила и медијане трансформисаних података

- *Natural log* природни логаритам; користи се када је проблем у линеарности
- *1/square root* реципрочна вредност квадратног корена; за веће вредности и смањење позитивног нагиба
- *Reciprocal* реципрочна вредност; јако велике вредности после трансформације постају мале
- *Square root* квадратни корен
- *Square* свака вредност се квадрира
- *Cube* трећи степен сваке вредности

Untransformed црта график оригиналних података (еквивалентно трансформацији са *power* $= 1$)

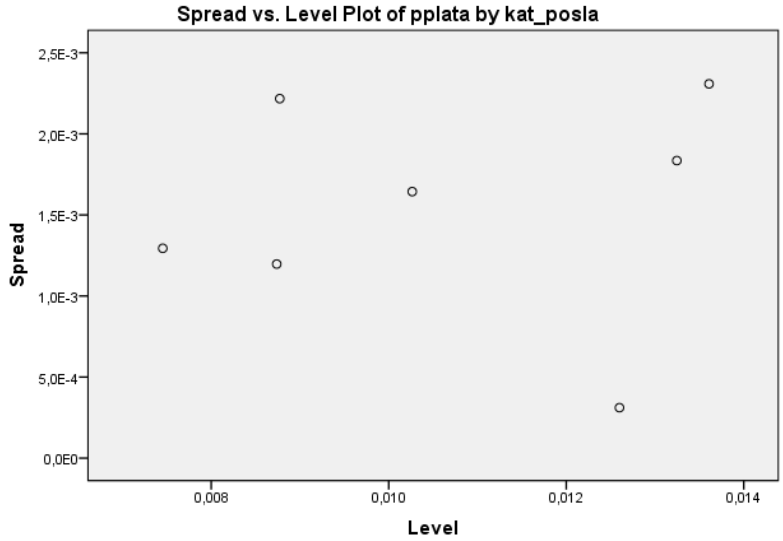
Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Pocetnicka plata	Based on Mean	26,709	6	467	,000
	Based on Median	19,610	6	467	,000
	Based on Median and with adjusted df	19,610	6	61,428	,000
	Based on trimmed mean	25,203	6	467	,000



* Plot of LN of Spread vs LN of Level

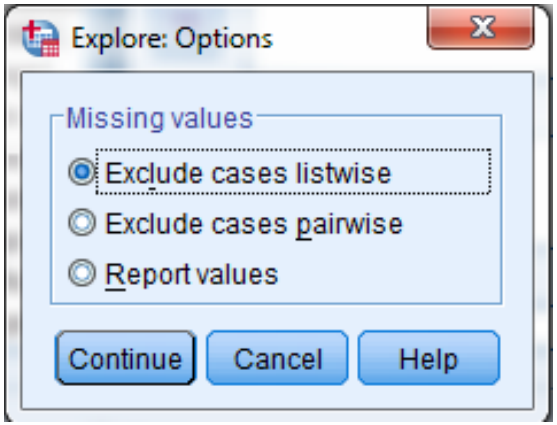
Slope = 1,732 Power for transformation = -,732



* Data transformed using P = -,500

Slope = ,021

Options



Омогућава третман недостајућих вредности.

Excluded Cases Listwise искључује опсервације које имају недостајуће вредности (било у зависној променљивој или у фактор променљивој)

Excluded Cases Pairwise искључује само оне опсервације које имају недостајуће вредности за променљиву која тренутно учествује у анализи

Report Values за фактор променљиву недостајућа вредност се третира као посебна категорија и даје приказ анализе посебно за ту категорију, док се уопсервације са недостајућим вредностима зависне променљиве искључују из анализе

Садржај

Пример 1

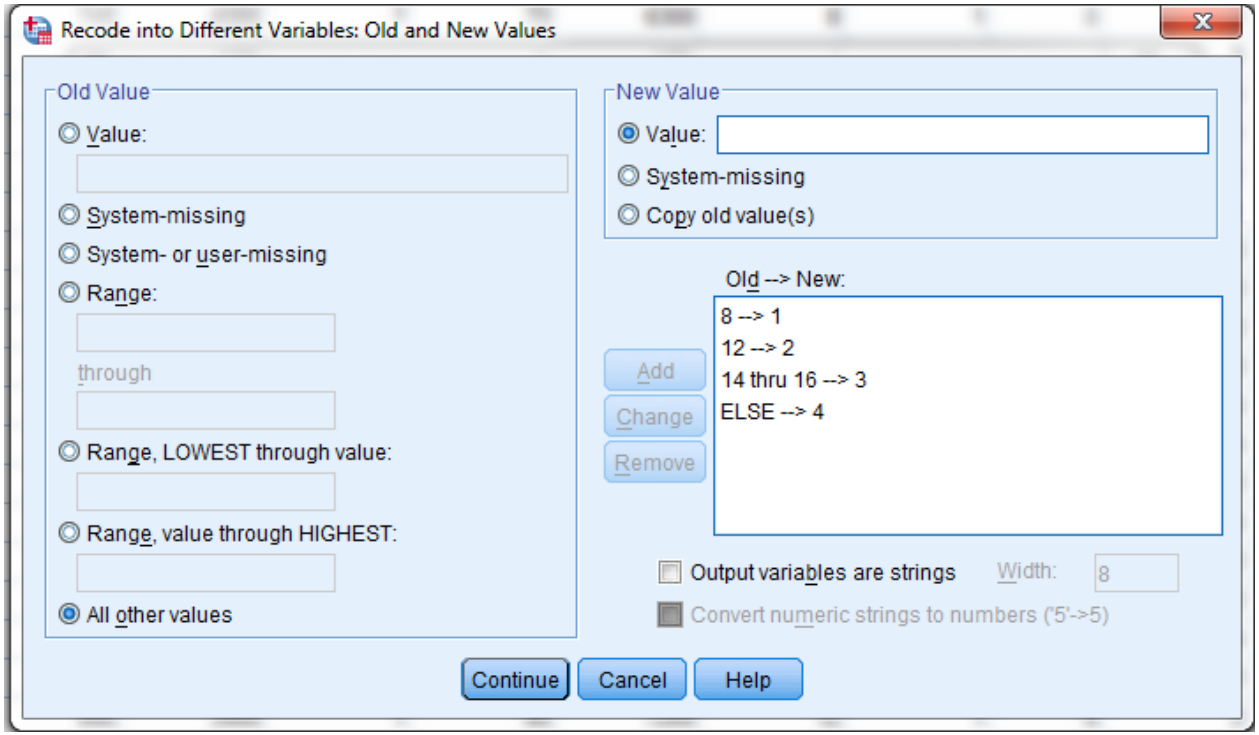
Приказати *Box plot* за променљиву *splata* по категоријама запослених: 1 - завршена основна школа; 2 - завршена средња школа; 3 - завршена виша, висока школа или факултет; 4 - мастер, специјалистичке и докторске студије.

Из табеле фреквенција за променљиву образовање виде се вредности које се појављују у оквиру ове променљиве(*Analyze* ⇒ *Descriptive Statistics* ⇒ *Frequencies*).

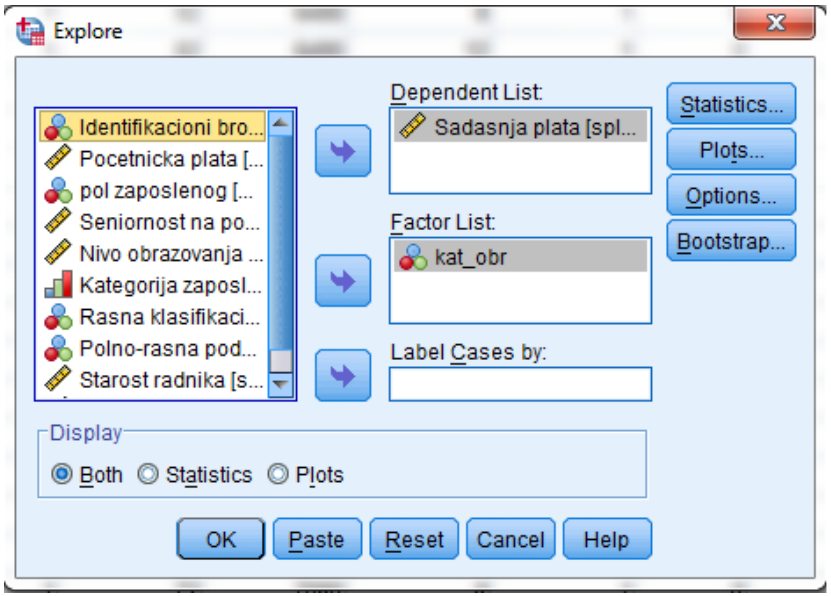
Nivo obrazovanja

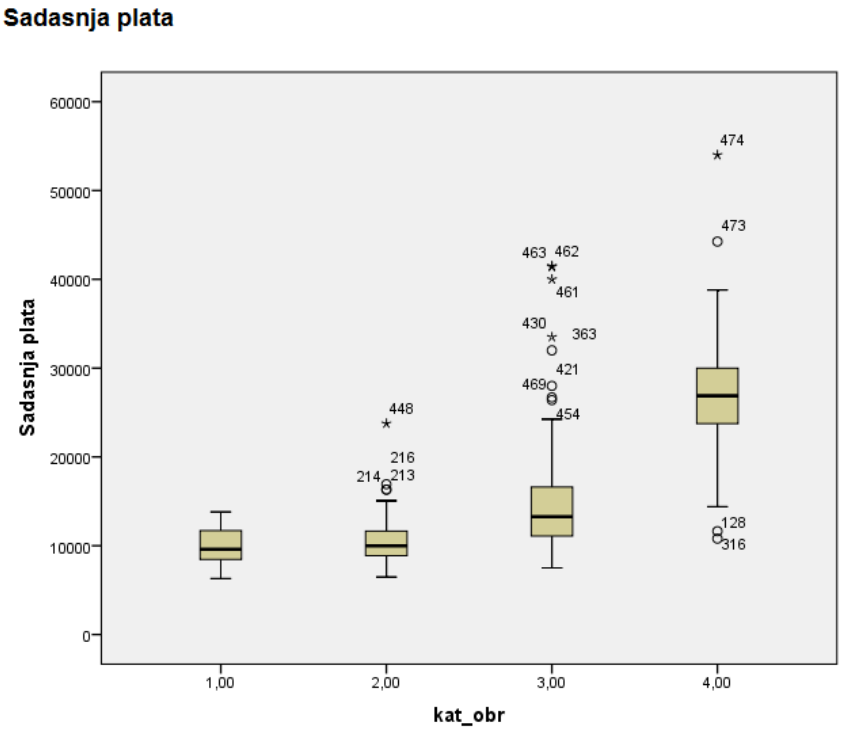
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	8	53	11,2	11,2	11,2
	12	190	40,1	40,1	51,3
	14	6	1,3	1,3	52,5
	15	116	24,5	24,5	77,0
	16	59	12,4	12,4	89,5
	17	11	2,3	2,3	91,8
	18	9	1,9	1,9	93,7
	19	27	5,7	5,7	99,4
	20	2	,4	,4	99,8
	21	1	,2	,2	100,0
	Total	474	100,0	100,0	

Transform $\Rightarrow$ *Recode into Different Variables* - формирање нове променљиве која указује на 4 категорије образовања.



Analyze ⇒ *Descriptive Statistics* ⇒ *Explore*





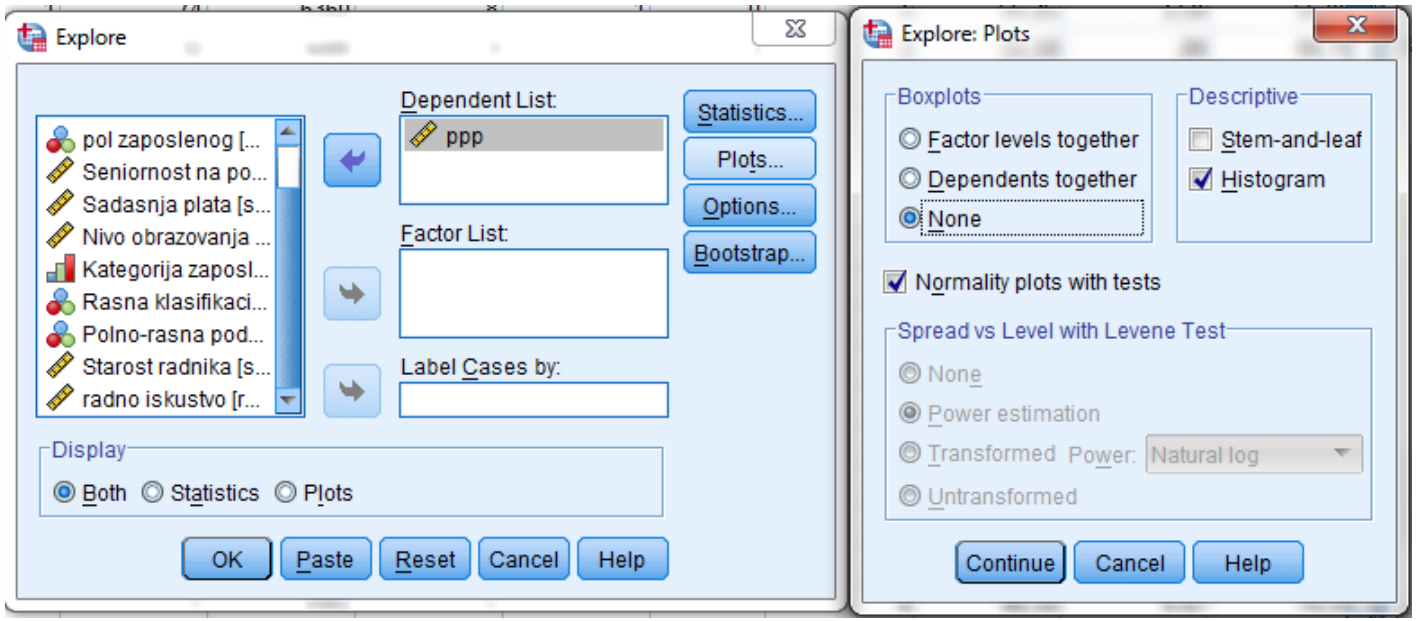
Пример 2

Да ли променљиве *ppp* (проценат пораста плате) има нормалну расподелу? Да ли је закључак исти за сваку групу полно-расне структуре?

Пример 2

Да ли променљиве *ppp* (проценат пораста плате) има нормалну расподелу? Да ли је закључак исти за сваку групу полно-расне структуре?

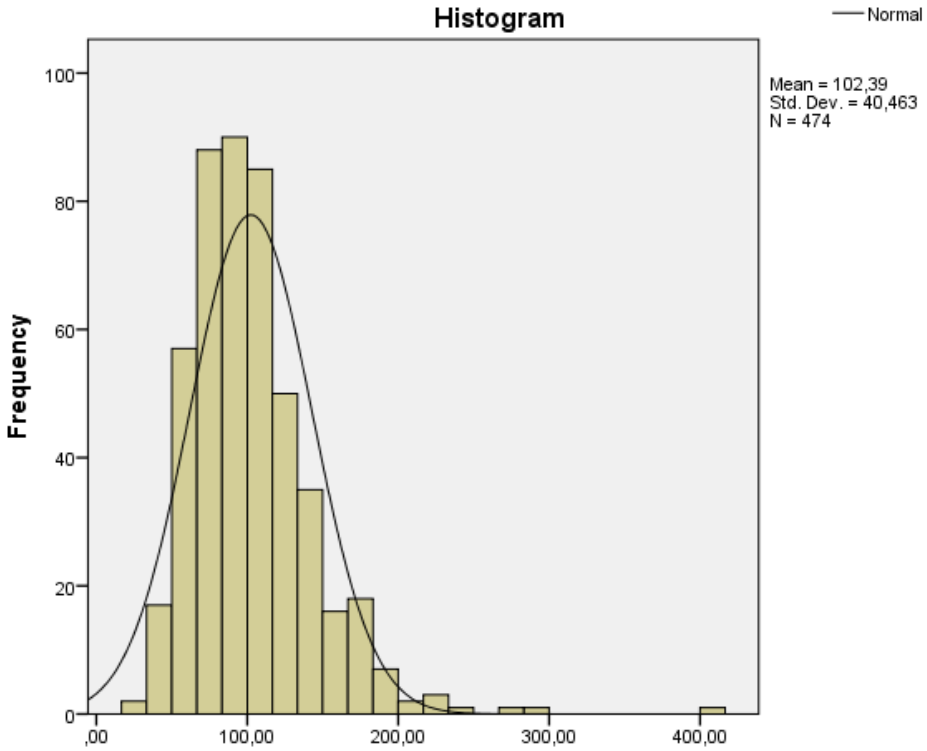
- Transform⇒Compute Variable
- $ppp = (splata - pplata)/pplata * 100$
- Analize⇒Descriptive Statistics⇒Explore

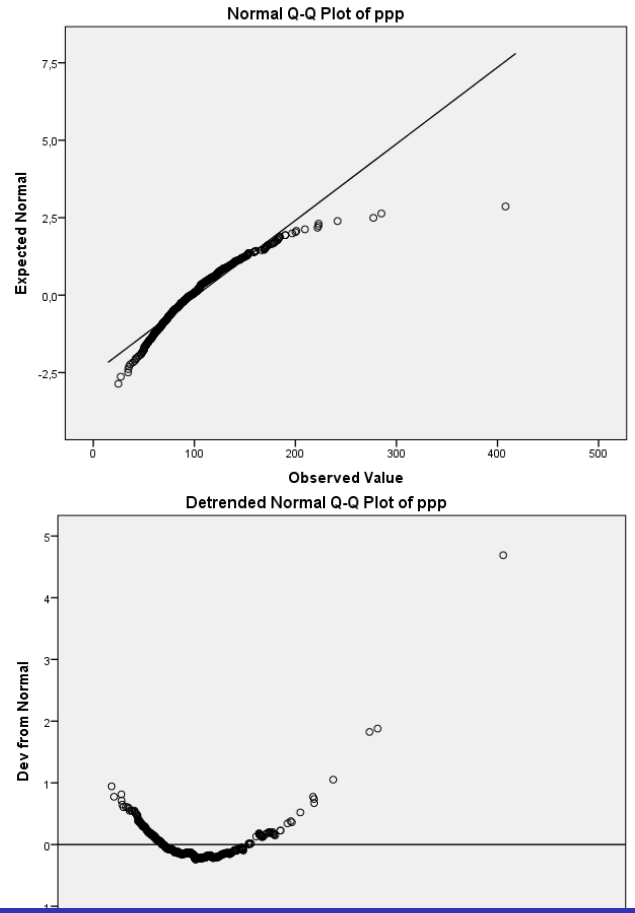


Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
ppp	,095	474	,000	,899	474	,000

a. Lilliefors Significance Correction





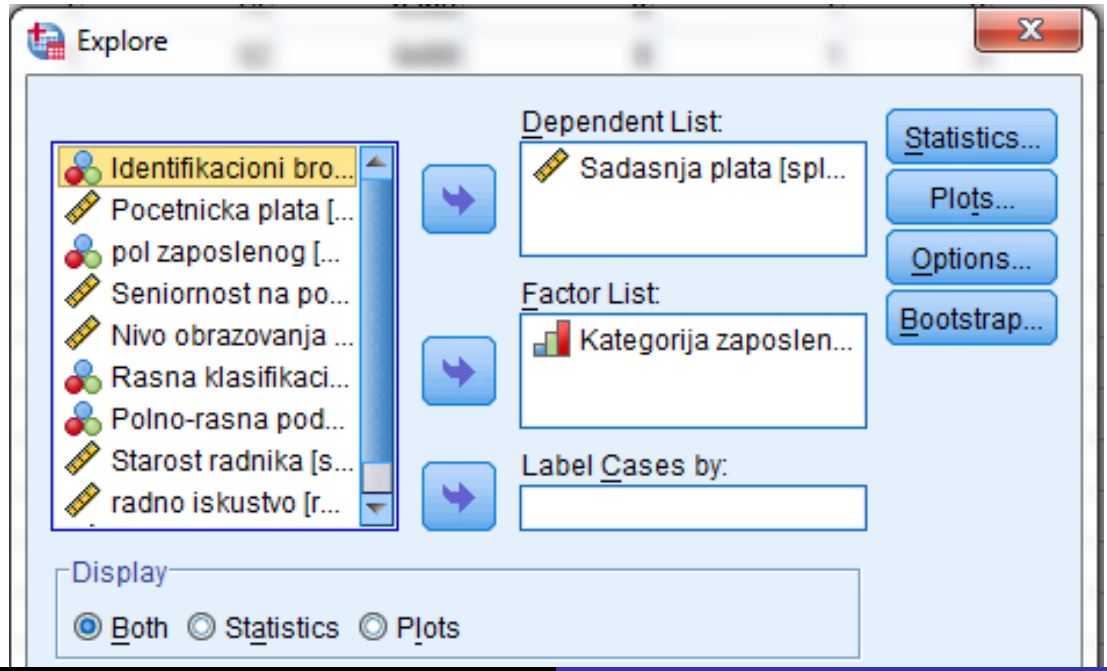
Пример 3

Проверити хомогеност варијабилитета променљиве *splata* за раднике из различитих категорија посла? Ако није хомоген, предложити одговарајућу трансформацију за добијање хомогеног варијабилитета.

Пример 3

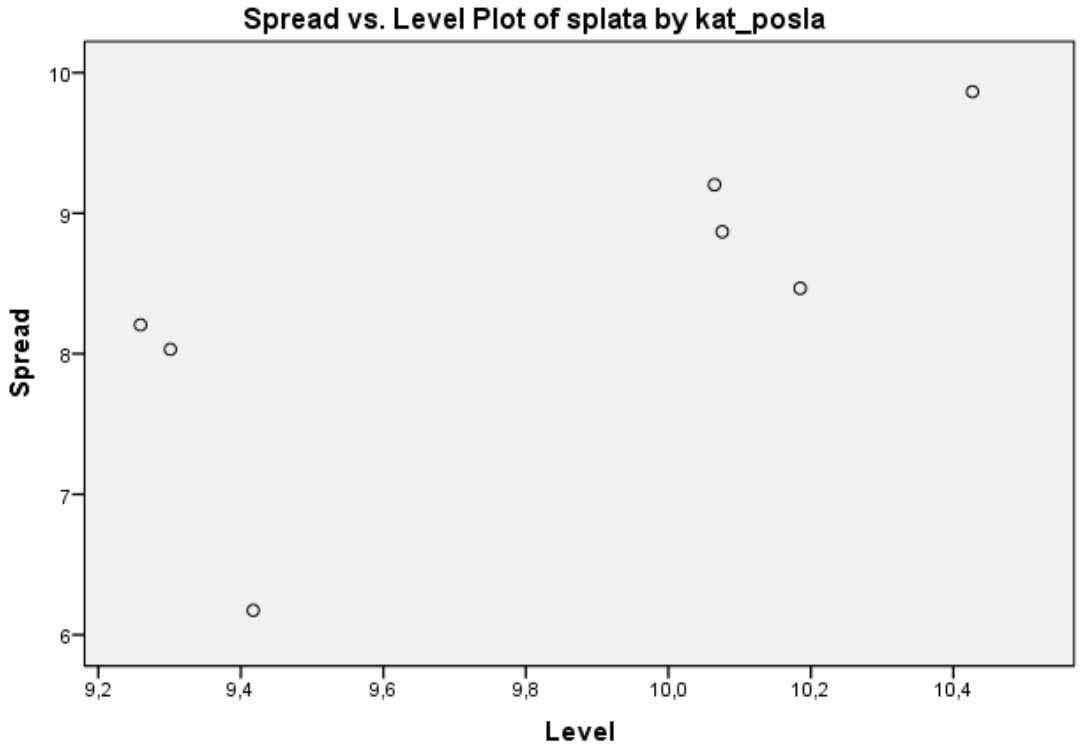
Проверити хомогеност варијабилитета променљиве *splata* за раднике из различитих категорија посла? Ако није хомоген, предложити одговарајућу трансформацију за добијање хомогеног варијабилитета.

- Analize⇒Descriptive Statistics⇒Explore



Test of Homogeneity of Variance

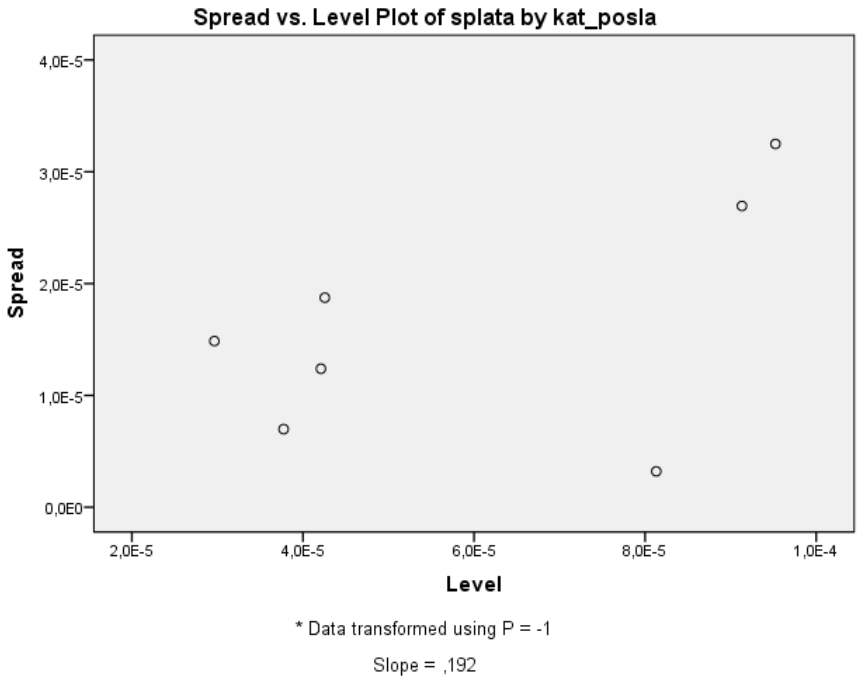
		Levene Statistic	df1	df2	Sig.
Sadasnja plata	Based on Mean	24,840	6	467	,000
	Based on Median	18,604	6	467	,000
	Based on Median and with adjusted df	18,604	6	208,350	,000
	Based on trimmed mean	23,127	6	467	,000



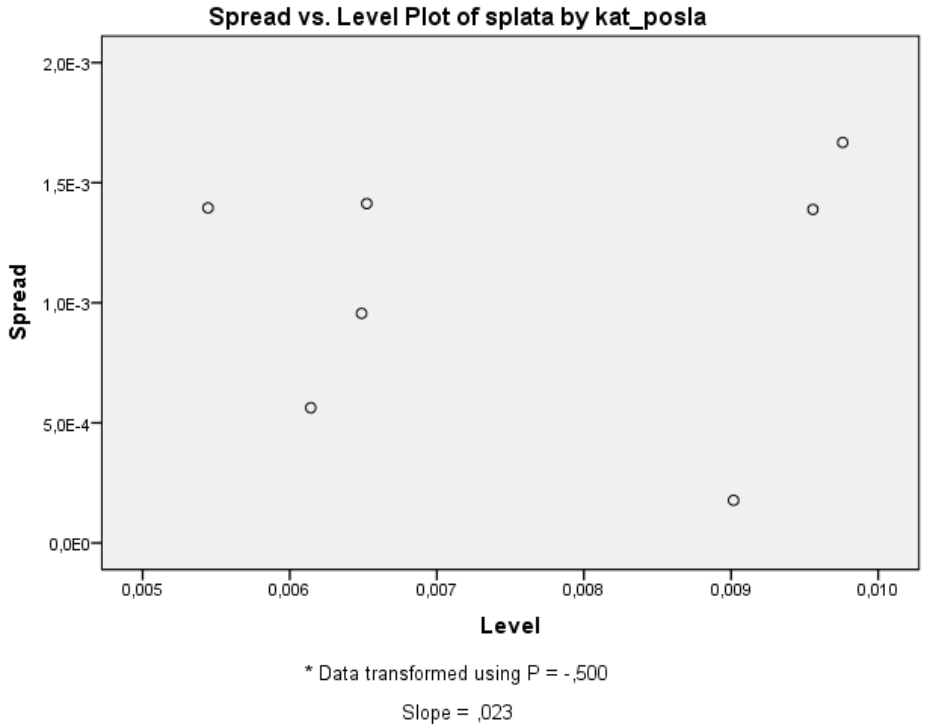
* Plot of LN of Spread vs LN of Level

Slope = 1,768 Power for transformation = -,768

Трансформација: реципрочна вредност



Трансформација: реципрочна вредност квадратног корена



Пример 4

Приказати робусне оцене централне тенденције за променљиву *splata*, посебно за раднике са искуством мањим или једнаким 1 годину, а посебно за остале раднике.

Пример 4

Приказати робусне оцене централне тенденције за промењљиву *splata*, посебно за раднике са искуством мањим или једнаким 1 годину, а посебно за остале раднике.

- Data⇒Select Cases⇒If($rstaz \leq 1$) (када се формира промењљива изгасити опцију)
- Data⇒Split File⇒Compare Groups by filter
- Analize⇒Descriptive Statistics⇒Explore

Case Processing Summary

r_staz <= 1 (FILTER)		Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
Not Selected	Sadasnja plata	380	100,0%	0	0,0%	380	100,0%
Selected	Sadasnja plata	94	100,0%	0	0,0%	94	100,0%

M-Estimators

r_staz <= 1 (FILTER)		Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
Not Selected	Sadasnja plata	12094,20	11297,18	11779,73	11289,84
Selected	Sadasnja plata	10403,22	9882,96	10130,41	9876,03

a. The weighting constant is 1,339.

Хвала на пажњи!