

ROBUSNE OCENE CENTRALNE TENDENCIJE

Jovana Lisinac, Marija Baltić, Selma
Halimović, Milena Minić

Matematički fakultet,
Beograd, 2015

UVOD

- ◉ Robusne statistike su statistike koje poseduju dobre osobine za podatke koji su iz različitih raspodela verovatnoća, posebno za one koje nisu normalno raspodeljene. Definišemo ih kao mere na koje ekstremne opservacije (autlajeri) imaju mali efekat. Jedan od motiva robusnih statistika je da se napravi statistička metoda koja ne sadrži uvek autlajere. Drugi motiv je da se obezbedi metoda koja poseduje dobre osobine kada postoje mala odstupanja od datih parametara raspodela.

- Robusne statistike nastoje da obezbede metode koje oponašaju standardne statističke metode, ali na koje ne utiču puno autlajeri ili neka druga mala odstupanja od pretpostavljenog modela. U statistici klasične metode ocenjivanja se oslanjaju na pretpostavke koje se ne sreću često u praksi. Na primer, često se pretpostavlja da su greške podataka normalno raspodeljene, makar asimptotski, ili da se CGT oslanja na proizvodnju normalno raspodeljenih ocena parametara. Nažalost, kada postoje autlajeri među podacima, standardni procenitelji često imaju loše osobine.

DEFINICIJA ROBUSNOSTI

- Robusna statistika je otporna na greške rezultata, koje nastaju odstupanjem od pretpostavljene raspodele (npr. od normalnosti). To znači da ako su date pretpostavke samo asimptotski dotaknute, robusni procenitelj će i dalje imati razumnu efikasnost i jako malu pristrasnost, takođe, biće i asimptotski nepristrasan u smislu da pristrasnost teži ka nuli, kada veličina uzorka teži ka beskonačnosti.

NAPOMENA:

- ◉ Medijana je od robusnih mera, a srednja vrednost nije.
- ◉ Medijana apsolutne devijacije i interkvartalnog ranga su robusne mere statističke disperzije, dok standardna devijacija i opseg nisu.
- ◉ Robusne statistike su pogodne za opis iskrivljenih statistikama, tj. onih sa ekstremnim opservacijama, dok su nerobusne statistike pogodne za opis simetričnih raspodela podataka.

- M-procenjivači su generalizacija procenjivača maksimalne verodostojnosti (MLE). Ono što pokušavamo da uradimo sa MLE procenjivačima je da maksimiziramo

$$\prod_{i=1}^n f(x_i),$$

ili ekvivalentno,

$$\sum_{i=1}^n -\log f(x_i).$$

- ◉ 1964.godine Huber je predložio da se ove generalizuje tako da se minimizira

$$\sum_{i=1}^n \rho(x_i),$$

gde je ρ neka funkcija. MLE procenjivača su specijalan slučaj od M-procenjivača, otuda naziv *maximum likelihood type estimators*. Prethodno minimiziranje može biti urađeno diferenciranjem funkcije ρ i rešavanjem jednačine:

$$\sum_{i=1}^n \psi(x_i) = 0,$$

- Predloženo je nekoliko izbora za ρ i ψ funkcije:

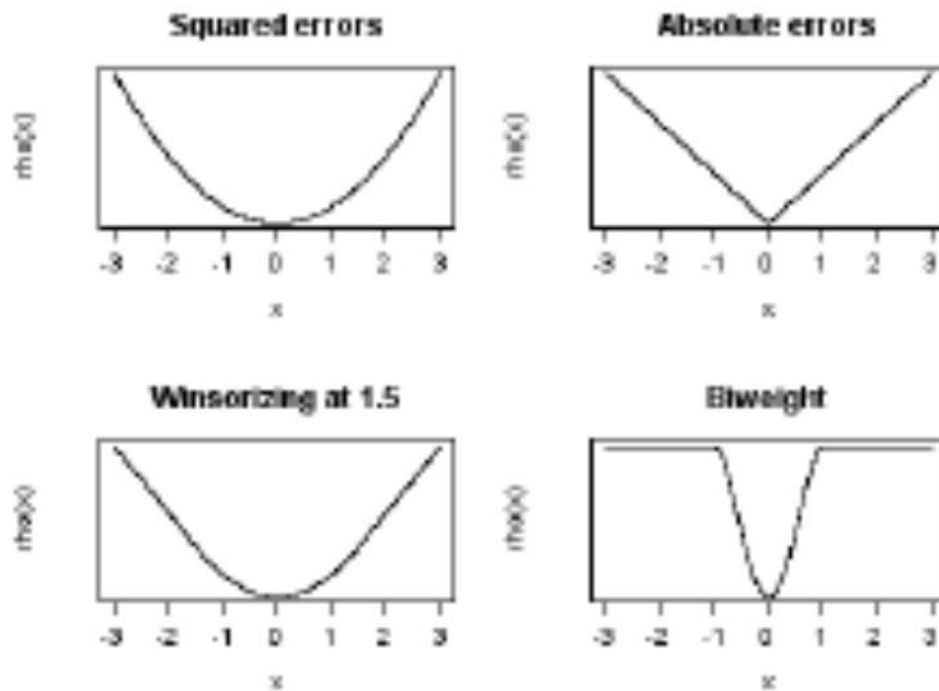


Figure 5: ρ funkcije

- Za kvadratne greške, $\rho(x)$ se povećava po ubrzanoj stopi, dok se za apsolutne greške povećava po konstantnoj stopi. Kada se koristi Vinsorizing, kombinacija ova dva efekta je uvedena na sledeći način: za male vrednosti x , ρ se povećava po kvadratnoj stopi, ali kada se dostigne izabrani prag(kao na slici 5), stopa rasta postaje konstantna. Ova procena je poznata kao Huberova funkcija gubitka.
- Tukey's biweight (poznat i kao bisquare) funkcija ponaša se na sličan način kao funkcija kvadratnih grešaka na početku, ali posle velikih grešaka, funkcija se isključuje.

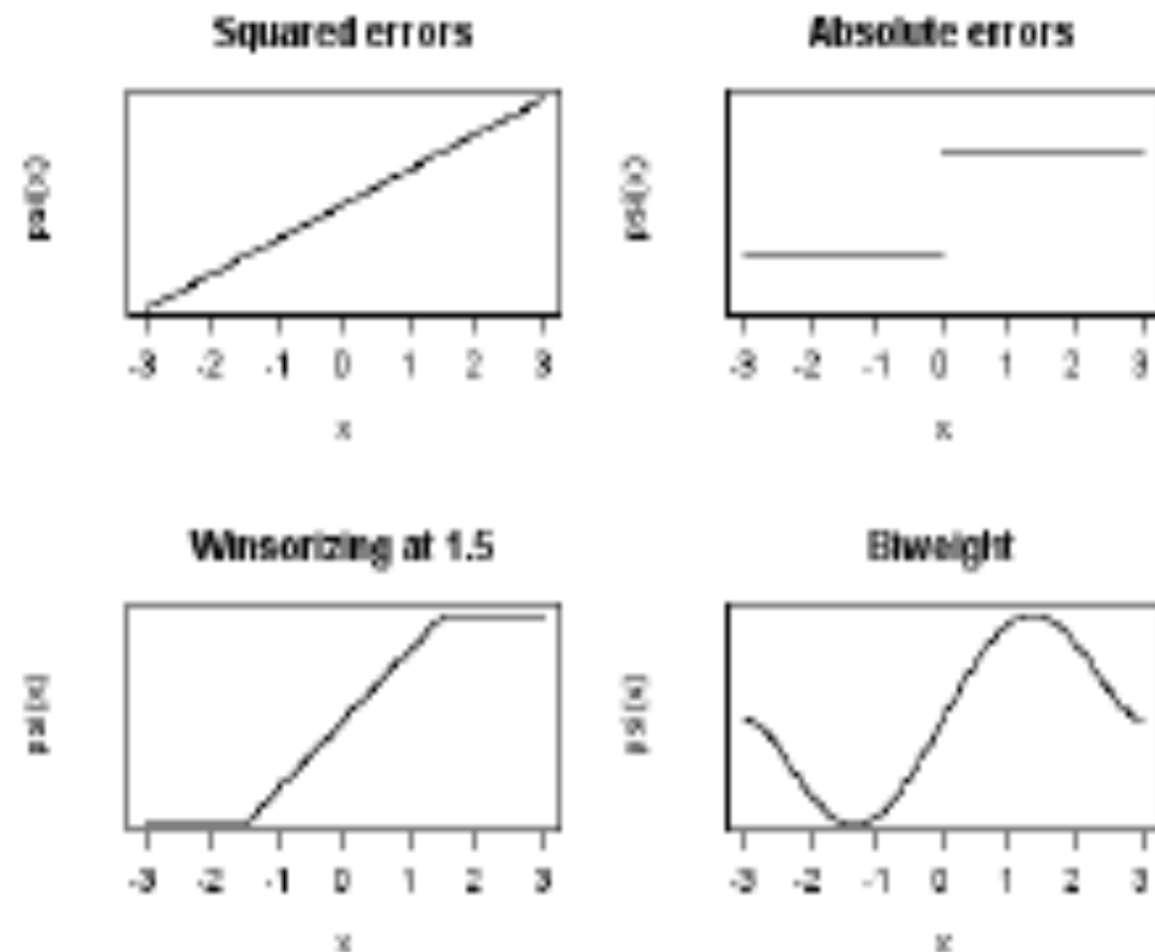


Figure 6: ψ funkcije

IZBOR Ψ I P FUNKCIJE

- ◉ U mnogim praktičnim slučajevima, izbor ψ funkcije nije neophodan za sticanje dobre robusne procene i mnogi izbori će dati slične rezultate koja nude velika poboljšanja, u smislu efikasnosti u pristrasnosti, preko klasičnih procena u prisustvu autlajera.
- ◉ Teorijski, ψ funkcije bi trebalo da budu poželjene, kao i Tukey's biweight funkcija koja je popularan izbor. Maronna et al.(2006) je predložio biweight funkciju sa efikasnošću u normalnom skupu do 85%.

PRIMERI M-PROCENITELJA

- Srednja vrednost odgovara funkciji

$\rho(x) = x^2$, a medijana $\rho(x) = |x|$. Funkcija

$$\psi(x) = \begin{cases} x, & \text{za } |x| < c; \\ 0, & \text{u ostalim slučajevima.} \end{cases}$$

odgovara metričkom skraćivanju i veliki autlajeri nemaju nikakav uticaj na nju.

⊙ Funkcija

$$\psi(x) = \begin{cases} -c, & x < -c; \\ x, & |x| < c; \\ c, & x > c. \end{cases}$$

je poznata kao metrika Vinsorizonovog i donosi ekstremne opservacije za vrednost $\mu \pm c$.

Odgovarajući $-\log f$ je

$$\rho(x) = \begin{cases} x^2, & \text{ako je } |x| < c; \\ c(2|x| - c), & \text{u ostalim slučajevima.} \end{cases}$$

i odgovara gustini sa Gausovom standardnom rapsodelom i dvostruko-eksponencijalnim repovima. Ovo je osmislio Huber.

- Vrednost $c=1.345$ daje 95%-tualnu efikasnost u normalnoj raspodeli.
- Tukey's biweight funkcija je

$$\psi(t) = t \left[1 - \left(\frac{t}{R} \right)^2 \right]_+$$

gde $[]_+$ označava pozitivan deo. Ovo sprovodi „meka“ skraćivanja. Vrednost $R=4,685$ daje 95%-tualnu efikasnost normalne raspodele.

- ◉ Hampleova ψ funkcija ima nekoliko linearnih delov,

$$\psi(x) = \text{sqn}(x) \begin{cases} |x|, & 0 < |x| < a; \\ a, & a < |x| < b; \\ a(c - |x|)/(c - b), & b < |x| < c; \\ 0, & c < |x|. \end{cases}$$

na primer, za vrednostima $a=2.2s$, $b=3.7s$, $c=5.9s$.

Možemo da skaliramo problem za poslednja četiri izbora, jer oni zavise od faktora skale(c , R ili s). Možemo primeniti procenitelje da reskaliramo rezultate, tj.

$$\min_{\mu} \sum_i \rho \left(\frac{y_i - \mu}{s} \right)$$

za faktor razmere s .

- ◉ Alternativno, možemo proceniti s na jednostavan način. MLE procenitelji za datu gustinu $s^{-1} f((x-\mu)/s)$ dovode do sledeće jednačine:

$$\sum_i \psi\left(\frac{y_i - \mu}{s}\right) \left(\frac{y_i - \mu}{s}\right) = n$$

koja nije otporna. Ovo možemo da modifikujemo:

$$\sum_i \chi\left(\frac{y_i - \mu}{s}\right) = (n-1)\gamma$$

za celo x , gde je γ izabrano za doslednost normalne raspodele, pa je $\gamma = E_x(N)$.

- Glavni problem je kada je u pitanju „Hubers
prospal 2“ funkcija sa

$$\chi(x) = \psi(x)^2 = \min(|x|, c)^2$$

- U vrlo malim uzorcima moramo uzeti u obzir
varijabilnost od $\hat{\mu}$ u izračunavanju
Vinsorzonove funkcije.
- Ako je položaj $\hat{\mu}$ poznat, možemo primeniti
ove procenitelje, gde $n-1$ zamenjujemo sa n
da bismo posebno ocenili razmeru s .

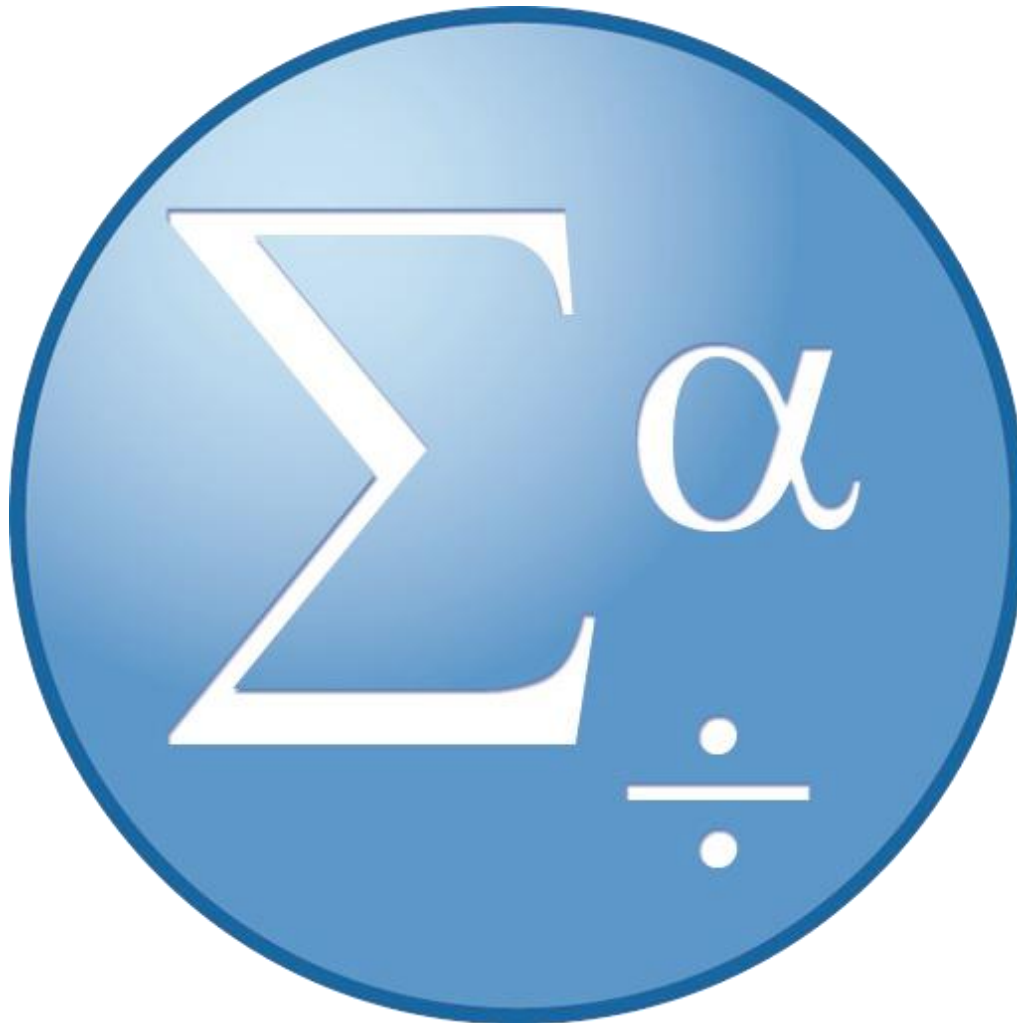
OSOBI NE M-PROCENJIVAČA

- ◉ Primetimo da se M-procenjivači ne odnose uvek na funkciju gustine verovatnoća.
- ◉ Može se pokazati da su M-procenjivači asimptotski normalno raspodeljeni, tako da dok se sve njihove standardne greške mogu izračunati, približan pristup zaključaka je dostupan.

- Pošto su M-procenjivači normalni samo asimptotski, za male veličine uzorka, bilo bi pogodno da se koristi alterativni pristup za zaključavanje, kao što je bootstrap metod. Međutim, M-procene nisu nužno jednostavne(tj. može da postoji više od jednog rešenja koji zadovoljava jednačine). Takođe, moguće je da neki specijalan bootstrap uzorak može da sadrži više autlajera nego procenjivači tačaka preloma. Stoga, nekad treba obratiti posebnu pažnju prilikom dizajniranja bootstrap šema.

- Srednja vrednost je jedino asimptotski normalno raspodeljena i kada su prisutni autlajeri aproksimacija može biti vrlo siromašna, čak i za prilično velike uzorke. Međutim, klasični statistički testovi, uključujući i one bazirane na srednjoj vrednosti, su obično gornje ograničeni kao za normalne veličine uzoraka. Ovo ne važi za M-procenjivači i tipovi prve greške mogu biti znatno iznad nominalnog nivoaa.

PRIMER U SPSS-U



KARTICA EXPLORE

◉ Do ove kartice dolazimo:

Analyze>Descriptive Statistics > Explore

63-baza.SAV [DataSet3] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : pplata 4080

	id	pplata	pol	vreme	splata	obrazova...	kat_posla	rasa	pol_rasa	starost	r_staz	var
1	748	4080									22,92	
2	832	4080									3,58	
3	754	3900									,00	
4	869	4080									,00	
5	969	4080									15,00	
6	825	4080									13,58	
7	1083	4080									24,00	
8	1107	3900									34,33	
9	886	4080									6,33	
10	1127	4080									26,58	
11	921	3600									10,33	
12	935	4080									22,58	
13	831	4080									6,00	
14	940	4080									,00	
15	1128	4080									29,83	
16	749	4080									32,50	
17	647	4080									9,67	
18	826	4080	1	72	7080	8	1	0	3	61,50	15,33	
19	885	3900	1	98	7260	12	1	0	3	62,00	6,00	

Explore

Identifikacioni bro...
Pocetnicka plata [...]
pol zaposlenog [p...
Seniornost na po...
Sadasnja plata [s...
Nivo obrazovanja ...
Kategorija zaposl...
Rasna klasifikacij...
Polno-rasna pod...

Dependent List:

Factor List:

Label Cases by:

Display
☒ Both ☐ Statistics ☐ Plots

OK Paste Reset Cancel Help

KARTICA EXPLORE

63-baza.SAV [DataSet3] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

1 : pplata 4080

	id	pplata	pol	vreme	splata	obrazova...	kat_posla	rasa	pol_rasa	starost	r_staz	var
1	748	4080									22,92	
2	832	4080									3,58	
3	754	3900									,00	
4	869	4080									,00	
5	969	4080									15,00	
6	825	4080									13,58	
7	1083	4080									24,00	
8	1107	3900									34,33	
9	886	4080									6,33	
10	1127	4080									26,58	
11	921	3600									10,33	
12	935	4080									22,58	
13	831	4080									6,00	
14	940	4080									,00	
15	1128	4080									29,83	
16	749	4080									32,50	
17	647	4080									9,67	
18	826	4080	1	72	7080	8	1	0	3	61,50	15,33	
19	995	3900	1	86	7260	12	1	0	3	62,00	6,00	

Explore

Dependent List: Pocetnicka plata [pp...]

Factor List:

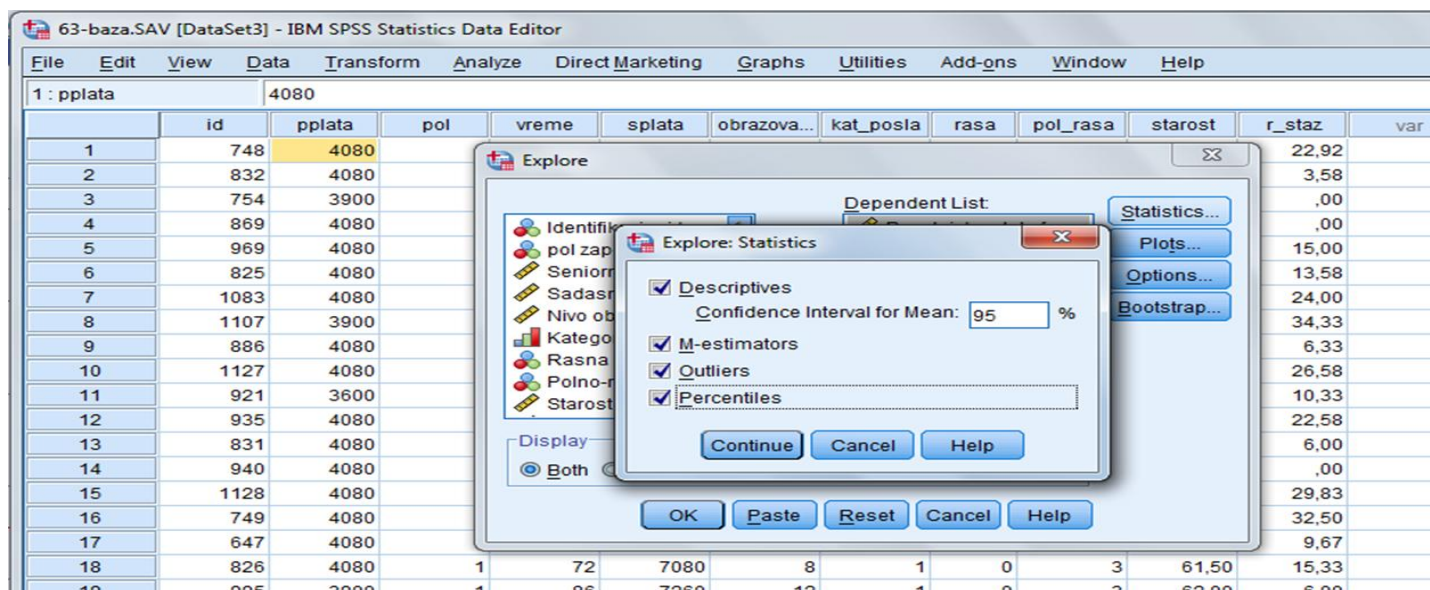
Label Cases by:

Display: ☒ Both ☐ Statistics ☐ Plots

OK Paste Reset Cancel Help

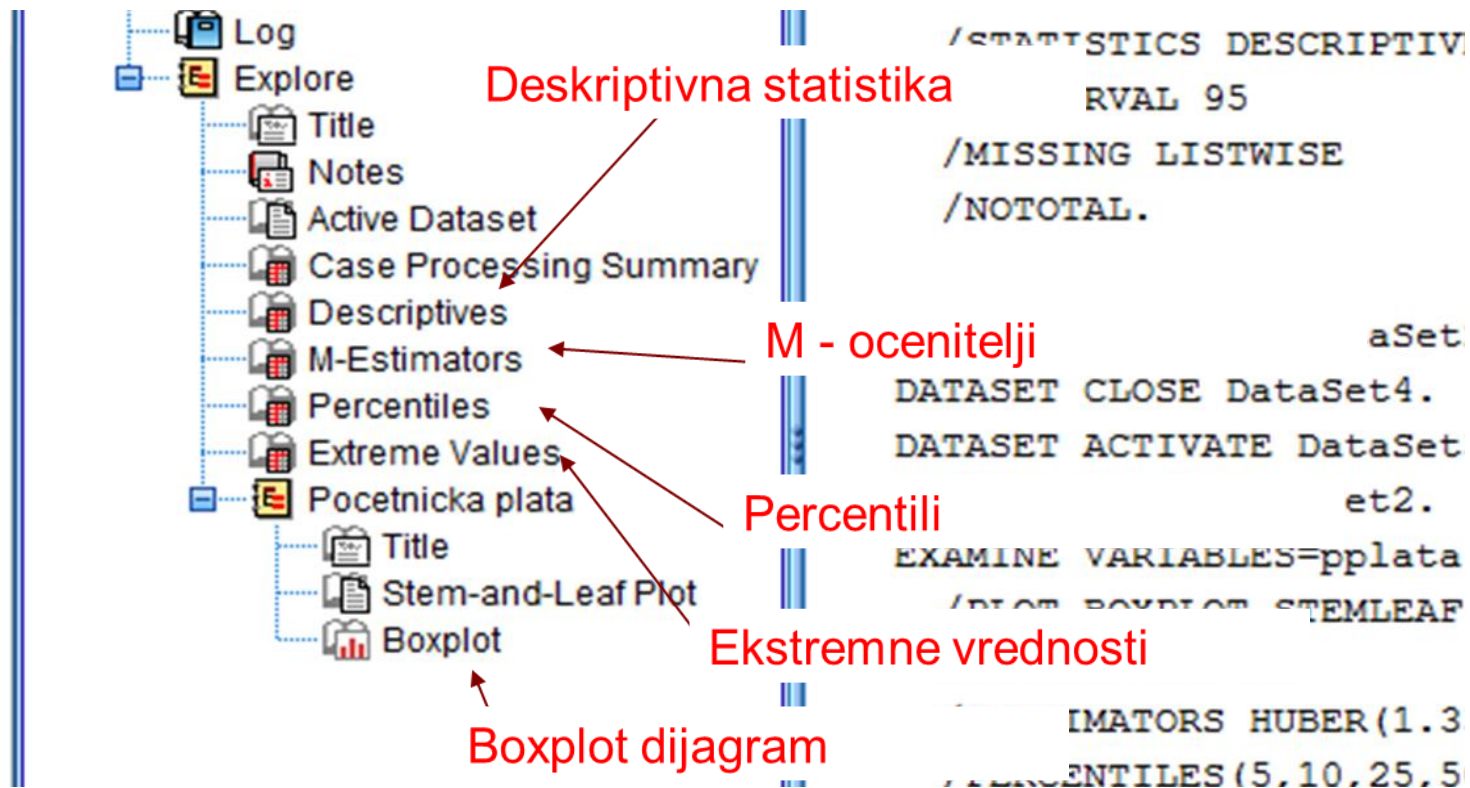
KARTICA EXPLORE

- ◉ Analyze>Descriptive Statistics
> Explore>Statistics>M-estimators



- ◉ Dodatno možemo da izaberemo i opcije Outliers i Percentiles.

ISPIS



- Sa leve strane vidimo strukturu ispisa.

ISPIS: CASE PROCESSING

SUMMARY

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Pocetnicka plata	474	100,0%	0	0,0%	474	100,0%



Broj jedinica u analizi



Broj nedostajućih vrednosti

ISPIS: DESCRIPTIVES

Descriptives

		Statistic	Std. Error
Pocetnicka plata	Mean	6806,43	144,604
	95% Confidence Interval for Mean	Lower Bound	6522,29
		Upper Bound	7090,58
	5% Trimmed Mean	6416,69	
	Median	6000,00	
	Variance	9911511,193	
	Std. Deviation	3148,255	
	Minimum	3600	
	Maximum	31992	
	Range	28392	
	Interquartile Range	2067	
	Skewness	2,853	,112
	Kurtosis	12,390	,224

Početnička plata: deskriptivna statistika

ISPIS: M - OCENITELJI

M-Estimators

	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Pocetnicka plata	5980,36	5695,36	5813,25	5694,43

- 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 $1,340 \cdot \pi$.

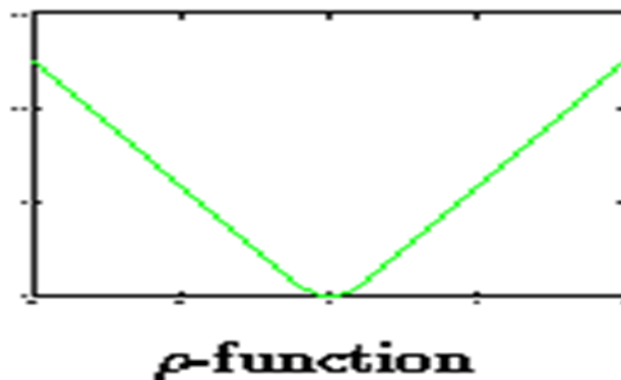
Prikaz 4 različita M- ocenitelja

HUBEROV M-OCENITELJ

M-Estimators

	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Pocetnicka plata	5980,36	5695,36	5813,25	5694,43

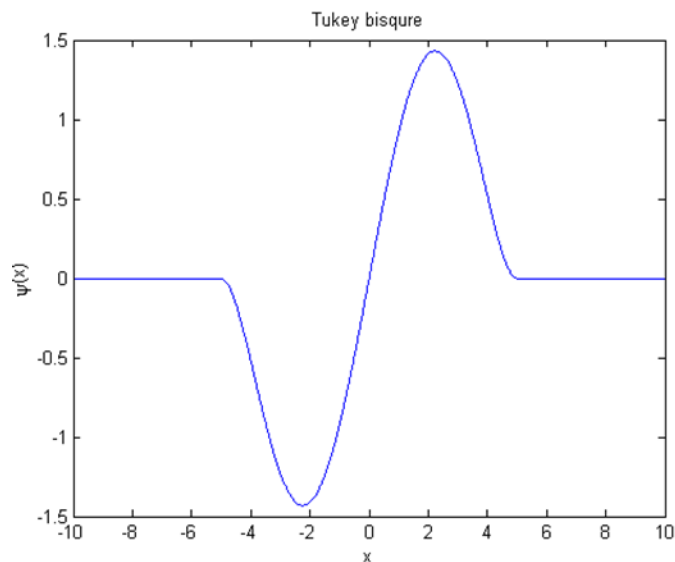
Huber



HUBEROV M-OCENITELJ

M-Estimators

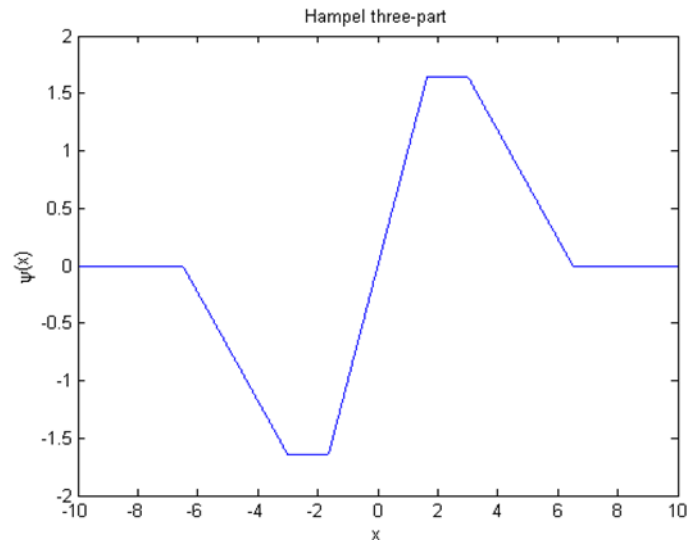
	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Pocetnicka plata	5980,36	5695,36	5813,25	5694,43



HAMPEL'S M-OCENITELJ

M-Estimators

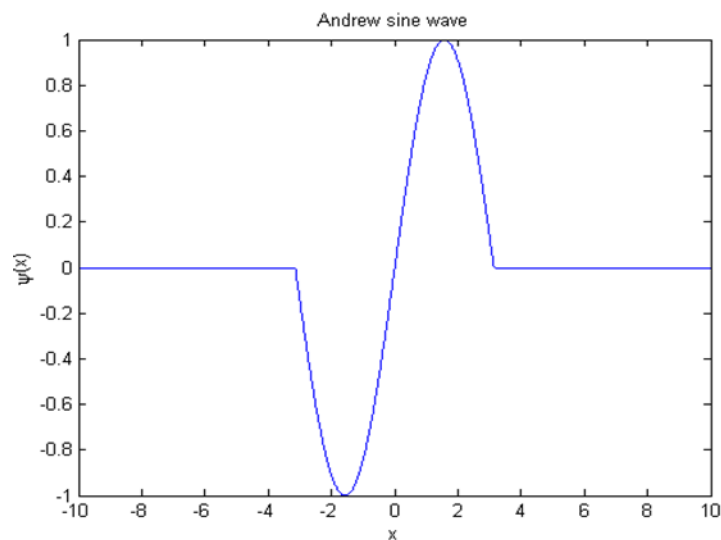
	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Pocetnicka plata	5980,36	5695,36	5813,25	5694,43



ANDREWS' M-OCENITELJ

M-Estimators

	Huber's M-Estimator ^a	Tukey's Biweight ^b	Hampel's M-Estimator ^c	Andrews' Wave ^d
Pocetnicka plata	5980,36	5695,36	5813,25	5694,43



EKSTREMNE VREDNOSTI

Extreme Values

			Case Number	Value
Pocetnicka plata	Highest	1	474	31992
		2	462	24000
		3	469	21000
		4	454	18996
		5	472	18000 ^a
	Lowest	1	380	3600
		2	38	3600
		3	22	3600
		4	11	3600
		5	252	3900 ^b

Broj jedinice

5 najviših vrednosti

5 najnižih vrednosti

a. Only a partial list of cases with the value 18000 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 3900 are shown in the table of lower extremes.

PERCENTILI

Ponderisane
vrednosti

Medijana

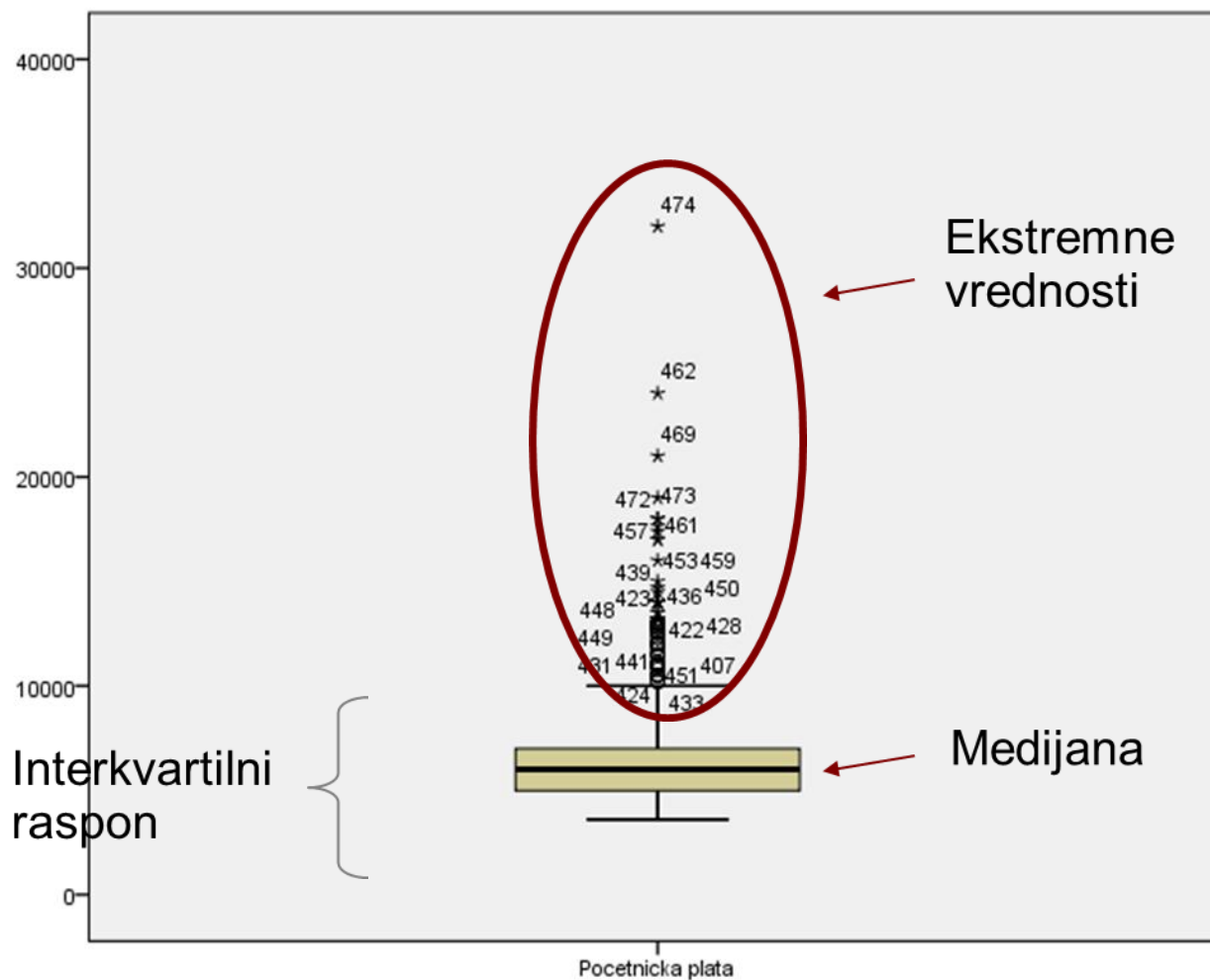
Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Pocetnicka plata	4080,00	4380,00	4980,00	6000,00	7047,00	11004,00	13275,00
Tukey's Hinges	Pocetnicka plata			4980,00	6000,00	6996,00		

Tuckey-jeve
aproksimacije
za I, II i III
kvartil

Interkvartilni
raspon

GRAFIČKI PRIKAZ: BOXPLOT



HVALA NA PAŽNJI!