LLL–Seminari u okviru TEMPUS projekta

Naziv projekta: **511140 – TEMPUS – JPCR "Master programe in Applied Statistics - MAS"**

Broj projekta: **511140**

Nosilac projekta: **Departman za matematiku i informatiku, PMF Novi Sad**

Rukovodilac: **Prof. dr Andreja Tepavčević**

Vreme trajanja: **15.10.2010. – 14.10.2013.**

Finansiranje: **Projekat finansira EU**

STATISTIČKA ANALIZA PODATAKA, IZBOR I OBRADA UZORKA I TUMAČENJE REZULTATA ISTRAŽIVANJA

Četvrta etapa - **STATISTIČKA ANALIZA PODATAKA**

U procesu statističkog istraživanja na nekoj populaciji, polazi se od pojedinačnih vrednosti obeležja, što znači da operišemo velikim brojem podataka. Često se u praksi traži brza informacija, pa je potrebno definisati neku karakteristiku obeležja koja će u većoj ili manjoj meri dobro predstavljati to obeležje. Drugim rečima, treba seriju podataka zameniti malim brojem nekih novih veličina.

Ti brojevi koji na neki način prezentuju posmatrano obeležje nazivaju se *parametri obeležja*. Oni su pokazatelji (*mere*) rasporeda vrednosti posmatranog obeležja na uzorku i populaciji. U tu svrhu, tj. za dobijanje više informacija o statističkim serijama, koriste se:

(1) *srednje vrednosti* (mere centriranosti, mere centralne tendencije),

(2) *mere odstupanja*, i

(3) *mere oblika*.

(1) SREDNJE VREDNOSTI

Srednje vrednosti nose zajedničke karakteristike svih vrednosti obeležja na posmatranom statističkom skupu. Pojam srednje vrednosti posmatranog obeležja može se uvesti po dva osnova, pa razlikujemo:

- *računske srednje vrednosti* (izračunate iz podataka uzorka ili populacije, a to su: *aritmetička sredina, geometrijska sredina, harmonijska sredina, sredina kvadrata*, i dr.) i

- *pozicione srednje vrednosti* (one su određene pozicijom koju zauzimaju u seriji podataka, a to su: *medijana* i *modus* (*mod*)).

ARITMETIČKA SREDINA

Neka su $x_1, x_2, ..., x_n$ vrednosti numeričkog obeležja X uzete u uzorak.
Aritmetička sredina vredosti obeležja X na ovom uzorku je vrednost:

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

U slučaju intervalno prikazanih podataka određuju se sredine intervala (klasa) $x_{s_i} = \frac{a_i + a_{i+1}}{2}$, $(i = 1, ..., k)$, pa je aritmetička sredina $\overline{x}_n^*$ data sa:

$$\overline{x}_n^* = \frac{1}{n}\sum_{i=1}^{k} f_i x_{s_i} \text{, za } \sum_{i=1}^{k} f_i = n.$$

Aritmetička sredina se smatra najvažnijom merom obeležja i ima veliki značaj u ozbiljnim statističkim analizama. Aritmetička sredina je osetljiva na ekstremne vrednosti, ali je zato saglasna sa pojavom koja se ponaša linearno. Zbir odstupanja svake pojedinačne vrednosti obeležja X od aritmetičke sredine jednak je nuli, tj.

$$\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right) = 0.$$

Zbir kvadrata odstupanja svake pojedinačne vrednosti obeležja od aritmetičke sredine je minimalan, tj.

$$\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2 \le \sum_{i=1}^{n}\left(x_i-a\right)^2,$$

gde je a proizvoljan broj.

PRIMER 9. Za uzorak obrađen u Primeru 7 aritmetička sredina je $\overline{x}_n = 7.4425$, dok za te iste podatke grupisane i prikazane u Tabeli 5 je $\overline{x}_n^* = 7.44$. Jasno je da vrednost $\overline{x}_n$ preciznije izračunata, jer se grupisanjem podataka izgubilo na preciznosti, u ovom slučaju zanemarljivo malo.

GEOMETRIJSKA SREDINA

Geometrijska sredina G od brojeva $x_1, x_2, ..., x_n$ data je sa:

$$G = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n}.$$

PRIMER 10. Plata radnika posle prve godine staža bila je 16 000 dinara, posle druge 20 000 dinara i posle treće godine 24 000 dinara. Koliko iznosi prosečno povećanje ove plate.

Rešenje. Ovaj problem se rešava preko geometrijske sredine, pa je

$$G = \sqrt{x_1 \cdot x_2} = \sqrt{\frac{20000}{16000} \cdot \frac{24000}{20000}} = \sqrt{\frac{3}{2}}.$$

Zaista, $16000 \cdot \sqrt{\frac{3}{2}} \cdot \sqrt{\frac{3}{2}} = 24000$, tj. plata se prosečno povećavala $\sqrt{\frac{3}{2}}$ puta, tj. u odnosu na prvu godinu plata se povećala 1.5 puta.

HARMONIJSKA SREDINA

Harmonijska sredina se koristi kod obrnuto proporcionalnih veličina i izračunava se za vrednosti različite od nule.

Harmonijska sredina predstavlja recipročnu vredost aritmetičke sredine recipročnih vrednosti obeležja X, pa za uzorak $(x_1, x_2, ..., x_n)$ harmonijska sredina je:

$$H = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}\right)} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Odnos prethodnih triju sredina je sledeći:

$$H \leq G \leq \overline{x}_n .$$

PRIMER 11. Dva radnika rade na dve iste mašine. Za 1h prvi radnik proizvede 6 proizvoda, a drugi radnik 4 proizvoda. Naći:

a) prosečan broj proizvoda po radniku na toj mašini;

b) prosečno vreme za jedan proizvod na toj mašini.

Rešenje.

a) Prosečan broj proizvoda po radniku na toj mašini je aritmetička sredina $\overline{x}_2 = \frac{x_1+x_2}{2} = \frac{6+4}{2} = 5.$

b) Za jedan proizvod prvom radniku treba 10 minuta, a drugom 15 minuta. Zato prosečno vreme za jedan proizvod na toj mašini je harmonijska sredina te dve vrednosti:

$$H = \frac{2}{\frac{1}{10}+\frac{1}{15}} = \frac{60}{5} = 12, \text{ a ne } \frac{10+15}{2} = 12.5 \text{ minuta.}$$

MEDIJANA

Medijana je *poziciona* srednja vrednost koja je određena pozicijom koju zauzima u rastućem nizu vredosti obeležja X. Kako je aritmetička sredina $\overline{x}_n$ osetljiva na ekstremne vredosti obeležja, uticaj tih ekstremnih vrednosti se isključuje upotrebom medijane ili moda (modusa).
Medijana je ona vrednost obeležja koja varijacioni niz vrednosti deli na dva jednaka dela. Zato definicija pojma medijane zavisi od toga da li je broj n paran ili neparan.

Neka je $x_1, x_2, ..., x_n$ varijacioni niz uzoračkih vrednosti obeležja X.

Medijana se definiše kao vrednost x_{med} data sa:

$$x_{med} = \begin{cases} x_{(n+1)/2}, & \text{ako je } n \text{ neparan broj;} \\ \dfrac{x_{n/2} + x_{n/2+1}}{2}, & \text{ako je } n \text{ paran broj.} \end{cases}$$

U slučaju kada je n – paran broj, može se dogoditi da ta vrednost ne pripada varijacionom nizu $x_1, x_2, ..., x_n$.

Geometrijski gledano, medijana je tačka na x-*osi* koja deli histogram na dva dela jednakih površina.

PRIMER 12. Medijana za uzorak iz Primera 6 određuje se uvidom u Tabelu 3 i kolonu kumulativ ispod iz te tabele.

Tabela 3

potr. (x_i)	br. dom. (f_i)	rel. fr. [%]	kum. ispod	kum. iznad	kum.fr. [%]
7	8	20%	8	40	20%
9	4	10%	12	32	30%
12	10	25%	22	28	55%
15	6	15%	28	18	70%
16	7	17,5%	35	12	87.5%
20	5	12,5%	40	5	100%
$\sum$	40	100%			

Treći član u toj koloni je broj 22, što znači da se u tom kumulativu prvi put sadrži 20, tj. $n/2$ vrednosti obeležja iz rastućeg varijacionog niza. To znači da vrednost obeležja $x_3 = 12\text{kg} = x_{med}$ predstavlja medijalnu vrednost mesečne potrošnje voća u zimskom periodu za jednu porodicu. U ovom primeru, $\overline{x}_{40} = 12.85\text{kg}$ što znači da su u ovom slučaju vrednosti aritmetičke sredine i medijane približne.

PRIMER 13. Medijana za uzorak iz Primera 7 dobija se kao aritmetička sredina 20. i 21. vrednosti obeležja (prosečna ocena), jer je obim uzorka 40. Zato od uzoračkih vrednosti moramo prvo napraviti varijacioni niz, a on je sledeći niz vrednosti:

6.53	6.58	6.60	6.71	6.77	6.80	6.80	6.86	6.87	6.89
6.90	6.94	7.06	7.10	7.11	7.12	7.20	7.31	7.33	7.34
7.34	7.48	7.50	7.53	7.53	7.54	7.57	7.57	7.87	7.90
7.95	7.97	8.00	8.11	8.20	8.48	8.50	8.57	8.60	8.67

$$x_{med} = \frac{x_{20} + x_{21}}{2} = \frac{7.34 + 7.34}{2} = 7.34 \neq 7.44 = \bar{x}_{40}.$$

Kod intervalnih serija distribucije frekvencija, izračuna se kumulativ ispod i odredi se medijalni interval $[a_m, a_{m+1})$, $(m = 1,...,k-1)$. To je prvi u nizu rastućih intervala koji sadrži $n/2$ vrednosti posmatranog obeležja. Ako je f_m frekventnost toga intervala, onda je

$$x^*_{med} = a_m + \frac{n/2 - F_{m-1}}{f_m} \cdot d,$$

gde je $F_{m-1} = \sum_{j=1}^{m-1} f_j$ kumulativna frekvencija predmedijalnog intervala i

d – dužina medijalnog intervala.

Za Primer 7, prikazan Tabelom 5, medijalni interval je $[7.2,7.6)$ i

$$x^*_{med} = 7.2 + \frac{20-16}{12} \cdot 0.4 = 7.33 \neq 7.44 = \bar{x}_{40}.$$

Vrednosti x_{med} i x^*_{med}, su približno jednake i obe se bitno razlikuju od aritmetičke sredine. Da li je ta razlika zaista bitna, može se proveriti metodom testiranja parametarske hipoteze o uzoračkoj srednjoj vrednosti. Kada se neke izračunate vrednosti iz običnog rasporeda uzorka i iz intervalno sređenih podataka tog istog uzorka bitno razlikuju, koriste se takozvane *Šepardove korekcije* za popravku ovih $*$*–vredosti*.

U svakom slučaju, ta razlika ukazuje na činjenicu da u svakom konkretnom uzorku treba dobro razmisliti koju meru centriranosti treba koristiti.
Prava $x = 7.33$ deli histogram na Slici 2 na dva dela jednakih površina.

MODUS (MOD)

Modus ili *mod* je ona *poziciona* vrednost obeležja koja se najčešće pojavljuje u realizovanom uzorku. Takvih vrednosti može biti jedna ili više. Kod neprekidnih numeričkih obeležja merenih intervalnom skalom ili skalom odnosa, nema velike potrebe za ovom vrednošću. U tom slučaju, aritmetička sredina je najbolja mera centriranosti.

PRIMER 14. U Primeru 6, modus je $x_{\text{mod}} = 12\text{kg}$, što znači da je najčešće slučaj da domaćinstva troše 12kg voća mesečno u zimskom periodu. U ovom primeru je $x_{med} = x_{\text{mod}} = 12\text{kg}$.

PRIMER 15. U Primeru 7, postoje četiri modalne vrednosti: $x_{\text{mod}(1)} = 6.80$, $x_{\text{mod}(2)} = 7.34$, $x_{\text{mod}(3)} = 7.53$, $x_{\text{mod}(4)} = 7.57$.

U slučaju intervalno prikazanih podataka, prvo se odredi modalni interval, tj. interval koji ima najveću frekventnost. Modus je vrednost iz toga intervala, npr. $[a_m, a_{m+1})$, koja se izračunava po formuli:

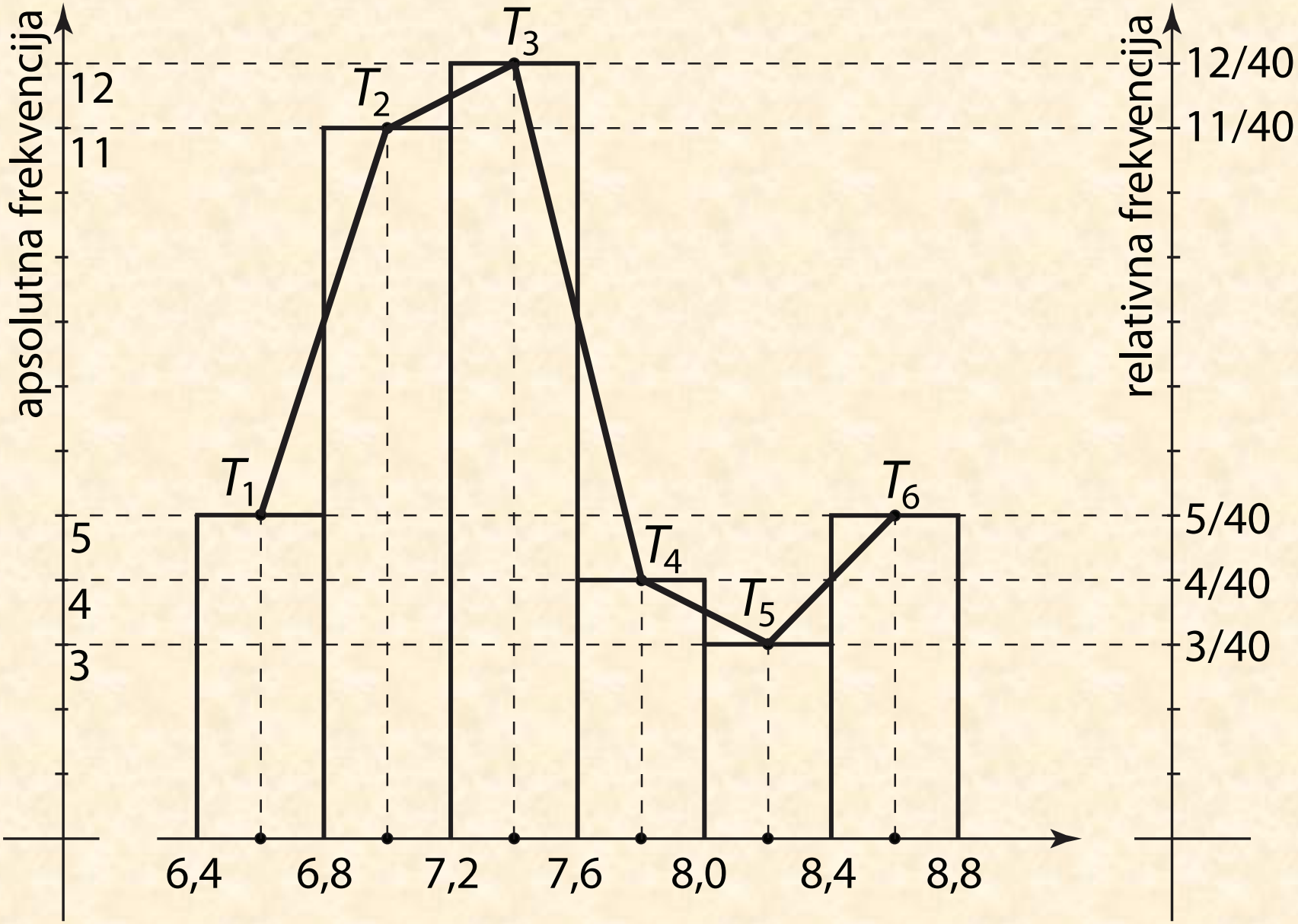
$$x^*_{\text{mod}} = a_m + \frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \cdot d,$$

gde su f_m, f_{m-1} i f_{m+1} frekvencije modalnog, predmodalnog i postmodalnog intervala i d - dužina intervalnih klasa.

PRIMER 16. Za Primer 7, čiji su grupisani podaci prikazani u Tabeli 5, modalni interval je $[7.2, 7.6)$, pa je:

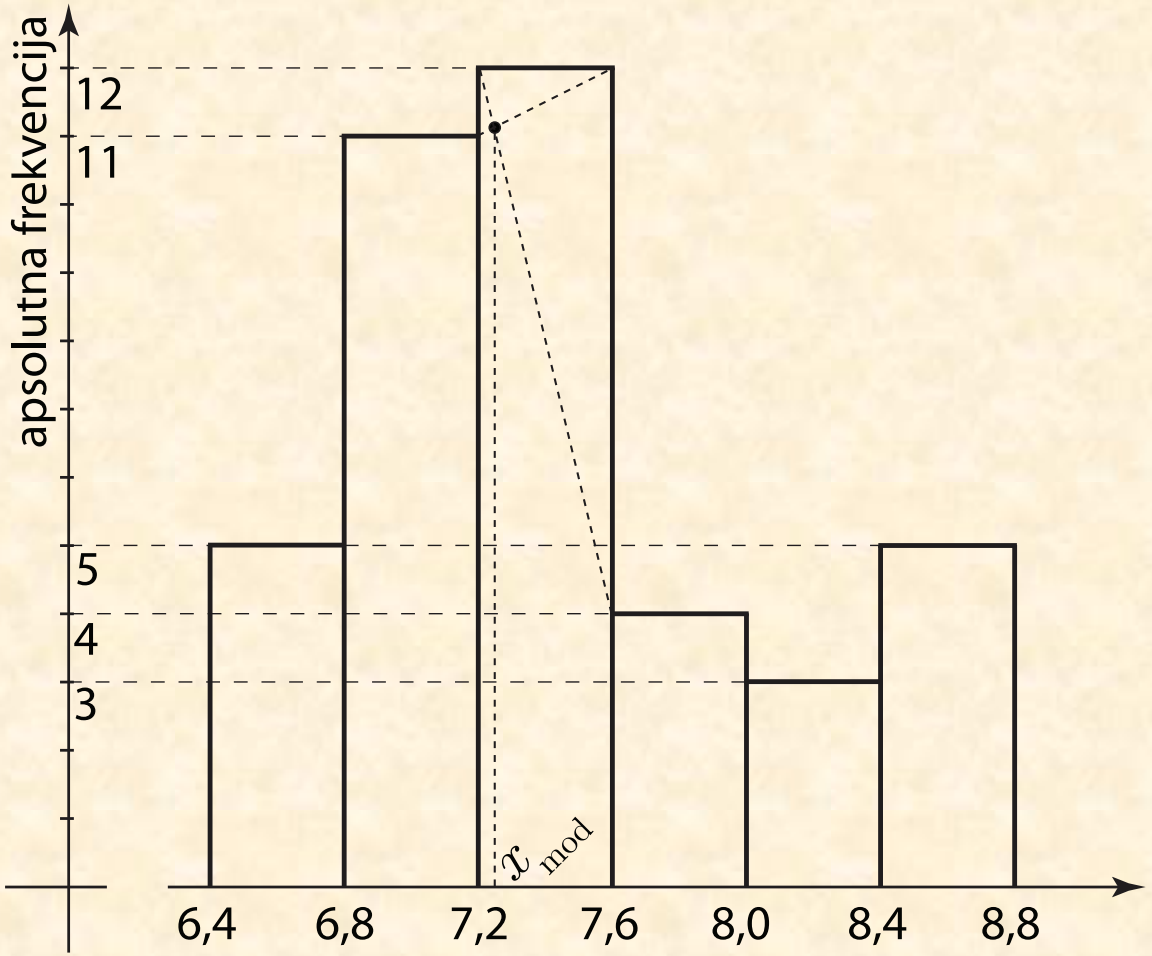
$$x^*_{\text{mod}} = 7.2 + \frac{12 - 11}{(12 - 11) + (12 - 4)} \cdot 0.4 = 7.24 < 7.33 = x_{med}.$$

Geometrijski, modus je vrednost na $x-osi$ za koju poligon frekvencija dostiže maksimum (tačka T_3 na Slici 2).



Slika 2. Histogram i poligon apsolutne i relativne frekvencije prosečne ocene studenata.

Približna vrednost modusa može se odrediti i na histogramu: dijagonalno se spoje krajnja i početna vrednost histograma na modalnom intervalu sa krajnjom tačkom histograma predmodalnog i početnom tačkom histograma postmodalnog intervala. Apscisa tačke preseka ovih duži je modalna tačka (Slika 49).



Slika 49. Modus određen iz histograma frekvencija.

(2) MERE ODSTUPANJA

Mere centriranosti, kao što su aritmetička sredina i druge, ne daju dovoljno informacija o obeležju.

PRIMER 17. Posmatrajmo ocene dva studenta.

Prvi: $7,8,8,9,7,8,9,8,8,9 \Rightarrow \bar{x}_{10} = 8.1, x_{med} = 8$ i $x_{\text{mod}} = 8.$

Drugi: $7,6,8,6,10,8,10,6,10,10 \Rightarrow \bar{y}_{10} = 8.1,\ y_{med} = 8$ i $y_{\text{mod}} = 10.$

Primećujemo da se njihove prosečne ocene i medijane poklapaju, ali kod prvog studenta dužina intervala varijacije $([7,9])$ je 2, dok je kod drugog taj interval $([6,10])$ dužine 4, što ukazuje na veće varijacije u kvalitetu pripreme ispita. Te uočene razlike mogu se oceniti merama rasturanja ili merama odstupanja (varijacije).

RASPON VARIJACIJE

Raspon varijacije je određen dužinom intervala varijacije obeležja i predstavlja razliku između najveće i najmanje vrednosti obeležja:

$$R = x_{\max} - x_{\min.}$$

Ova mera nije uvek dobar pokazatelj rasturanja vrednosti obeležja jer zavisi od ekstremnih vrednosti, koje su često izuzetci i kao takve ih obično isključujemo iz uzorka.

SREDNJE APSOLUTNO ODSTUPANJE

Ova mera odstupanja pokazuje prosečno odstupanje svake vrednosti obeležja od aritmetičke sredine uzorka, tj.

$$sao = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \bar{x}_n\right|.$$

Ako se radi o intervalnoj distribuciji frekvencija, onda je:

$$sao^* = \frac{1}{n}\sum_{i=1}^{k} f_i \left|x_{s_i} - \bar{x}_n\right|, \text{ gde je } x_{s_i} \text{ sredina klse.}$$

U Primeru 17, $sao_{\text{prvi}} = 0.54$ i $sao_{\text{drugi}} = 1.52$ što ukazuje na značajno veće rasturanje oko srednje vrednosti uzorka.

SREDNJE KVADRATNO ODSTUPANJE

Rad sa apsolutnim vrednostima valjalo bi, kad god je to moguće, izbeći. To se jednostavno postiže kvadriranjem prethodno posmatranih razlika $\left|x_i - \bar{x}_n\right|$.

Na taj način došlo se do jedne nove mere odstupanja, koja je istovremeno najvažnija i najviše se koristi.

Uzoračka disperzija (ili *disperzija uzorka* ili *srednje kvadratno odstupanje*) definiše se sa:

$$\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}_n^2.$$

U praksi i u računarskim paketima dosta se koristi tzv. *korigovana* ili *popravljena vrednost uzoračke disperzije*:

$$\bar{s}_n'^2 = \frac{n}{n-1}\bar{s}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2.$$

Kada je obim uzorka n veliki, svejedno je koju od ove dve mere ćemo koristiti.

Kako bi se ova mera odstupanja izražavala u istim jedinicama kao i vrednosti obeležja, posmatra se kvadratni koren iz uzoračke disperzije. Ta vrednost $\bar{s}_n = \sqrt{\bar{s}_n^2}$ se naziva *standardna devijacija*.

Navedene formule važe i za intervalno prikazane podatke, stim što se za popravku vrednosti $\bar{s}_n^{2*}$ koristi *Šepardova korekcija za uzoračku disperziju* (objavljena prvi put u jednom časopisu iz 1898.) koja je data formulom:

$$\bar{s}_n^2 = \bar{s}_n^{2*} - \frac{d^2}{12},$$ pri čemu su svi intervali iste dužine d.

Uopšte, prosek odstupanja vrednosti obeležja od aritmetičke sredine uzet na nekom stepenu k, naziva se *centralni momenat reda k*:

$$c_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^k, \text{ gde je } c_2 = \bar{s}_n^2.$$

Za njihovo izračunavanje koriste se *obični momenti* k-tog reda:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k, \text{ gde je } m_1 = \bar{x}_n.$$

Te veze za neke centralne momente su sledeće:

$$c_1 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^1 = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}n\bar{x}_n = 0$$

$$c_2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x}_n + \bar{x}_n^2\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\frac{\bar{x}_n}{n}\sum_{i=1}^{n}x_i + \frac{1}{n}\sum_{i=1}^{n}\bar{x}_n^2$$

$$= m_2 - 2m_1m_1 + \frac{1}{n}\cdot nm_1^2 = m_2 - m_1^2,$$

$$c_3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^3 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^3 - 3x_i^2\bar{x}_n + 3x_i\bar{x}_n^2 - \bar{x}_n^3\right) = m_3 - 3m_2m_1 + 2m_1^2,$$

$$c_4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^4 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^4 - 4x_i^3\bar{x}_n + 6x_i^2\bar{x}_n^2 - 4x_i\bar{x}_n^3 + \bar{x}_n^4\right)$$

$$= m_4 - 4m_3m_1 + 6m_2^2m_1^2 - 4m_1^4 + m_1^4 = m_4 - 4m_3m_1 + 6m_2^2m_1^2 - 3m_1^4,$$

$$c_5 = m_5 - 5m_4m_1 + 4m_1^5.$$

KOEFICIJENAT VARIJACIJE

Koeficijenat varijacije se koristi za upoređivanje *varijabilnosti* dvaju statističkih skupova i računa se na sledeći način:

$$c_V = \frac{\bar{s}_n}{\bar{x}_n}.$$

Ponekad se izražava u % i ima smisla samo ako je $\bar{x}_n \neq 0$.

PRIMER18. Za uzorak iz Primera 17, pregled nekih mera dat je u Tabeli 7.

Tabela 7.

ocene stud. (x_i)	6	7	8	9	10	$\bar{x}_{10}$	x_{med}	x_{mod}	$\bar{s}_{10}$	*sao*	$c_V 100\%$
prvi stud.	0	2	5	3	0	8.1	8	8	0.7	0.54	9%
drugi stud.	3	1	2	0	4	8.1	8	10	1.7	1.52	21%

Sve ove izračunate mere pokazuju da je uspeh prvog studenta bolji.

(3) MERE OBLIKA

Sve do sada navedene mere odstupanja **mogu se primeniti** samo na obeležja za koja je **moguće izračunati aritmetičku sredinu**. To su obeležja čija je merna skala intervalna, skala odnosa i ponekad ordinalna skala. U slučaju nominalne skale, aritmetička sredina ili medijana se ne može izračunati, ali i za ovu vrstu obeležja treba definisati neku meru odstupanja. U tu svrhu se koriste pokazatelji različitosti i time se bavi *klaster analiza*.

Podaci o vrednostima numeričkog obeležja na jedinicama statističkog skupa obično nisu pravilno i simetrično raspoređeni oko svojih srednjih vrednosti. Mere odstupanja ukazuju na veličinu odstupanja od srednje vrednosti, ali ne i na smer odstupanja. Zato se koriste *mere asimetrije* i *spljoštenosti*.

Neka je $(x_1, x_2, ..., x_n)$ realizovani prost slučajan uzorak. Značajnu informaciju o uzorku daju nam uzorački momenti:

- običan uzorački momenat *k*-tog reda: $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$, i

- centralni uzorački momenat *k*-tog reda: $c_k = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}_n\right)^k$.

Uzoračka srednja vrednost $\bar{x}_n = m_1$, a uzoračka disperzija $s_n^{-2} = c_2$.
Dodatnu, vrlo važnu informaciju o uzorku dobijamo izračunavajući neke odnose između običnih i centralnih momenata odgovarajućeg reda.

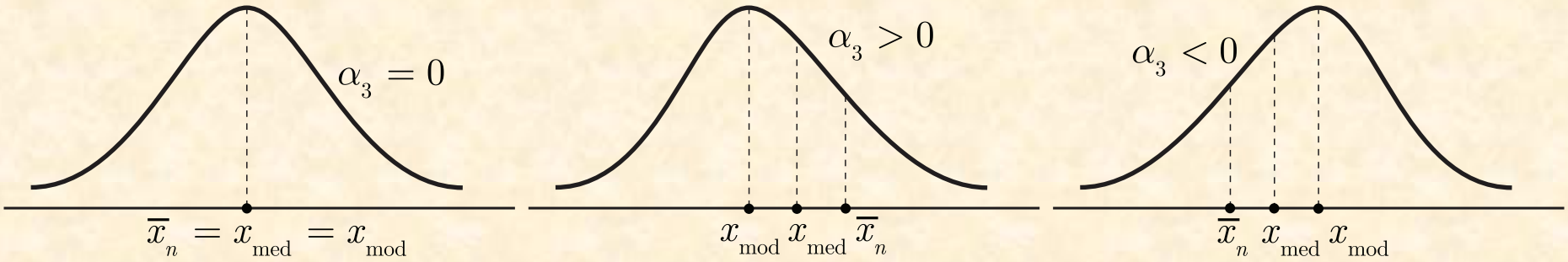
Uzoračkи koeficijenat asimetrije je: $\alpha_3 = \frac{c_3}{s_n^{-3}} = \frac{c_3}{\left(\sqrt{c_2}\right)^3}$

Uzoračkи koeficijenat spljoštenosti (ekscesa) je: $\alpha_4 = \frac{c_4}{s_n^{-4}} - 3.$

Ako je $\alpha_3 = 0$, onda je $\bar{x}_n = x_{med} = x_{\text{mod}}$ i raspodela je simetrična u odnosu na aritmetičku sredinu uzorka.

Ako je $\alpha_3 > 0$, onda je $\bar{x}_n > x_{med} > x_{\text{mod}}$ i imamo asimetriju udesno.

Ako je $\alpha_3 < 0$, onda je $\bar{x}_n < x_{med} < x_{\text{mod}}$ i imamo asimetriju u levo (Slika 50).

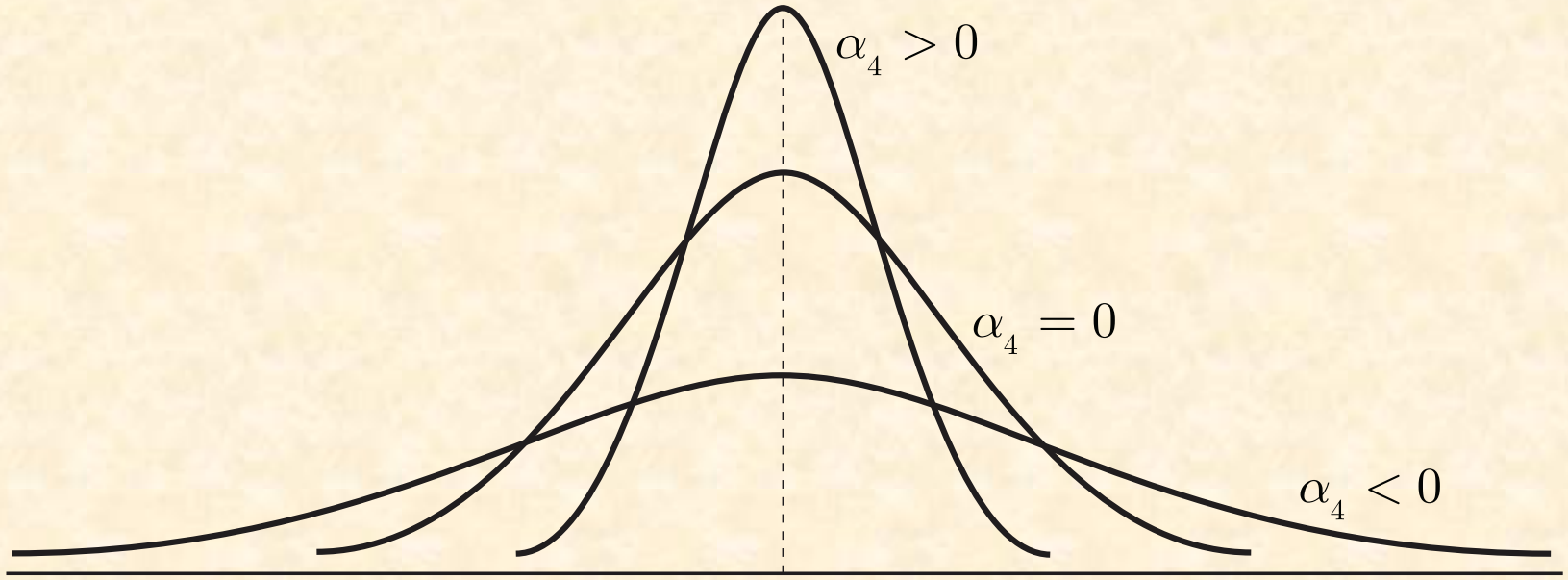


Slika 50. Mera asimetrije

Što je α_3 po apsolutnoj vrednosti veće, asimetrija je veća. Smatra se da je asimetrija umerena kada je $|\alpha_3| < 0.5$ (za $|\alpha_3| < 0.1$ nema asimetrije; $0.1 \leq |\alpha_3| < 0.25$ asimetrija je mala i $0.25 \leq |\alpha_3| < 0.5$ asimetrija je srednja) . Ukoliko je $|\alpha_3| \geq 0.5$ asimetrija je jaka.

Spljoštenost je proporcionalna parnim momentima. Za normalnu raspodelu obeležja, koeficijenat spljoštenosti $\alpha_4 = 0$. Ako je $\alpha_4 > 0$, onda je veća koncentracija vrednosti obeležja oko aritmetičke sredine i kriva raspodele je izduženog oblika (spljoštenost krive je manja nego kod funkcije gustine normalne raspodele).

Ako je $\alpha_4 < 0$, onda je spljoštenost krive raspodele veća od spljoštenosti normalne krive, tj. kriva je niža od normalne krive (Slika 51).



Slika 51. Mera spljoštenosti

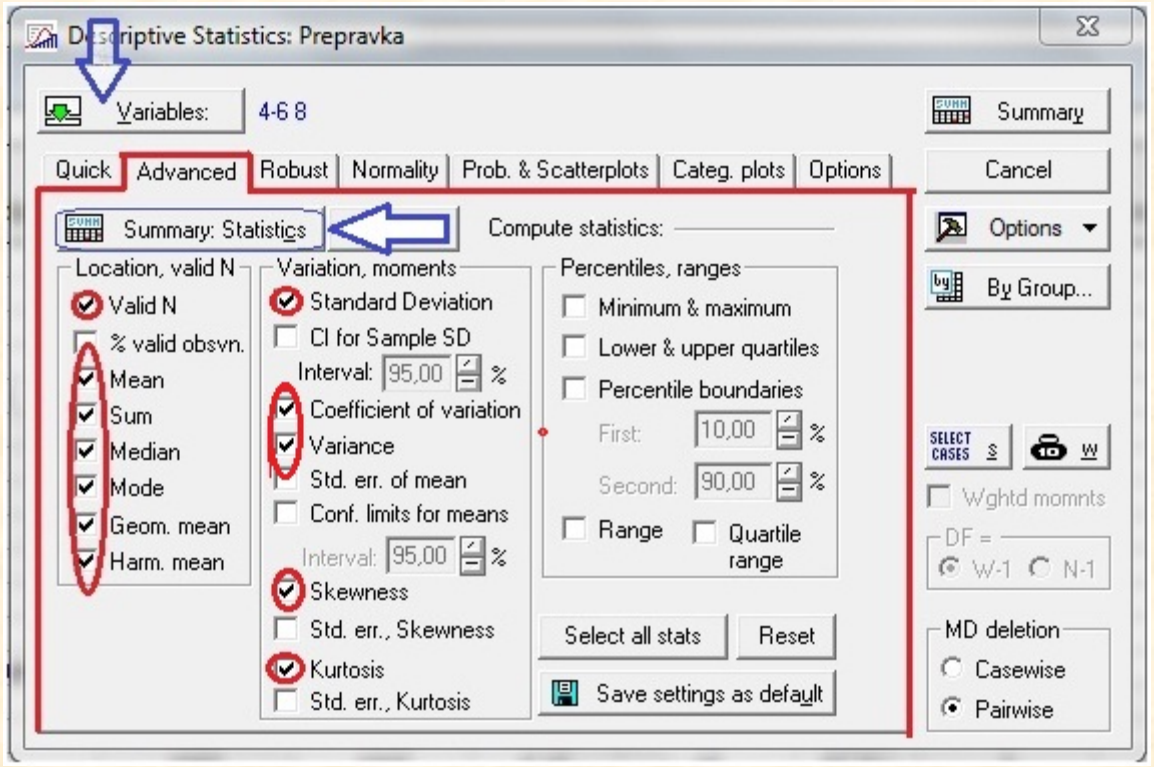
PRIMER 19. Za uzorke iz Primera 17, važi:

Za prvog studenta : $\alpha_3 = \frac{c_3}{\frac{-3}{s_{10}}} = \frac{-0.048}{0.7^3} = -0.14$ i $\alpha_4 = \frac{c_4}{\frac{-4}{s_n}} - 3 = -0.96.$

Slično izračunavamo i za drugog studenta: $\alpha_3 = -0.03$ i $\alpha_4 = -1.35.$

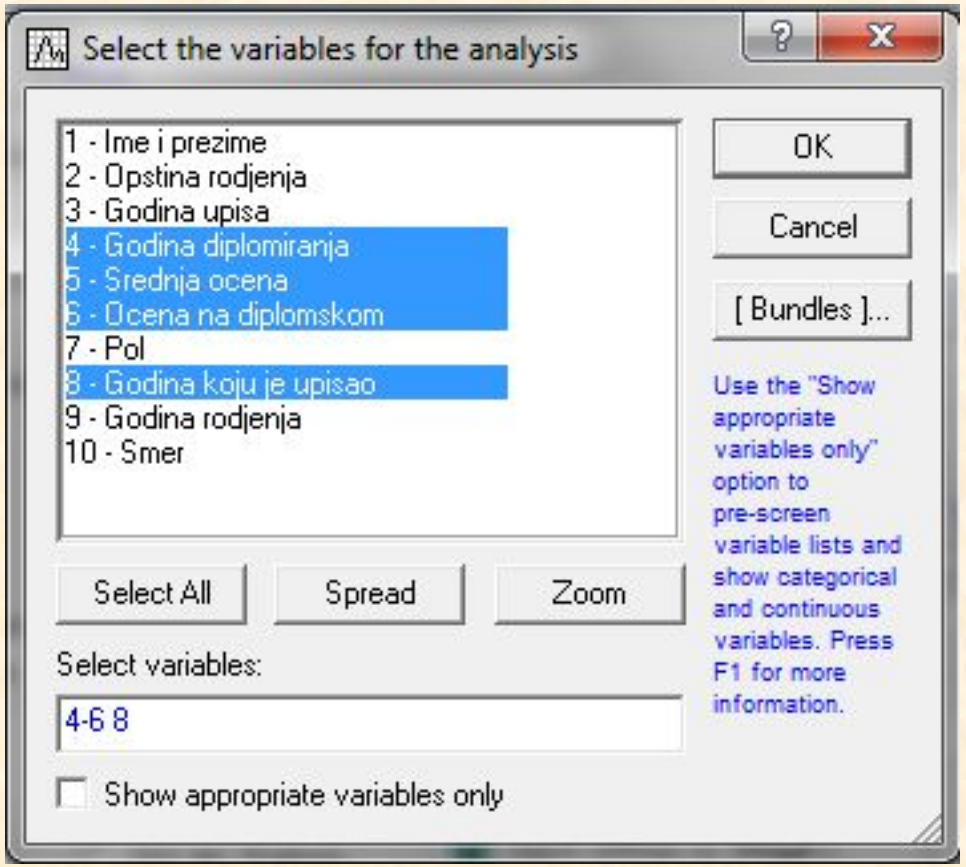
U oba slučaja raspodela frekvencija je simetrična u odnosu na srednju ocenu studenta i u oba slučaja spljoštenost krive je veća od spljoštenosti normalne krive, ali je kod drugog studenta to više izraženo.

Sada ćemo u paketu ***Statistica 8*** pokazati postupak dobijanja srednjih vrednosti. Iz menija *Statistics* biramo *Basic Statistics/Tables*, pa onda odaberemo *Descriptive statistics*. Na dugmetu *Variables* odaberemo obeležja za koja računamo srednje vrednosti. U našem slučaju to su obeležja: „Srednja ocena“, „Ocena na diplomskom“ i druga. Onda prelazimo u karticu *Advanced* i biramo sve srednje vrednosti koje nam trebaju (Slika 52).



Slika 52. Izgled prozora za izbor vrednosti koje računamo.

Nakon štikliranja određenih opcija prelazimo na odabir varijabli *(eng. Variables)* i dobijamo sledeći prozor (Slika 53) i potvrdom na *Ok*, vraćamo se u prethodni prozor gde nam preostaje samo da izaberemo opciju "*Summary: Statistics*".



Slika 53. Prozor za izbor varijable

Prethodno opisanim koracima dobijamo rezultate prikazane u sledećim tabelama.

Variable	Descriptive Statistics (Prepravka)												
	Broj studenta	Srednja vrednost	Geometrijska sredina	Harmonijska sredina	Medijana	Modus	Frekvencija modusa	Zbir	Varijansa	Srednje kvadratno odstupanje	Koeficijenat varijacije	Koeficijenat asimetrije	Koeficijenat spljostenosti
Godina upisa	56	1986,339	1986,334	1986,329	1988,000	1990,000	10	111235,0	20,62825	4,541833	0,22865	-1,73237	2,912691
Srednja ocena	56	7,196	7,176	7,157	7,085	Multiple	3	403,0	0,30734	0,554383	7,70416	1,25154	1,956527
Ocena na diplomskom	56	9,768	9,752	9,734	10,000	10,00000	46	547,0	0,29058	0,539059	5,51870	-2,30740	4,482751
Godina koju je upisao	56	1,357	1,223	1,143	1,000	1,000000	45	76,0	0,56104	0,749025	55,19134	1,72240	1,121945

Slika 54. Prikaz tabele sa računskim i pozicionim vrednostima za studente koji su diplomirali 1995. god.

Variable	Descriptive Statistics (Prepravka)												
	Broj studenta	Srednja vrednost	Geometrijska sredina	Harmonijska sredina	Medijana	Modus	Frekvencija modusa	Zbir	Varijansa	Srednje kvadratno odstupanje	Koeficijenat varijacije	Koeficijenat asimetrije	Koeficijenat spljostenosti
Godina diplomiranja	504	1997,331	1997,331	1997,330	1998,000	1998,000	141	1006655	1,7369	1,3179	0,0660	-0,3085	-1,0891
Srednja ocena	504	7,177	7,159	7,141	7,070	7,000000	15	3617	0,2850	0,5339	7,4379	1,2365	2,7013
Ocena na diplomskom	504	9,710	9,689	9,665	10,000	10,00000	398	4894	0,3771	0,6141	6,3245	-2,1103	3,6749
Godina koju je upisao	504	1,782	1,5404	1,358	1,000	1,000000	302	898	0,9384	0,9687	54,3673	0,4618	-1,7486

Slika 55. Prikaz tabele sa računskim i pozicionim vrednostima za studente koji su diplomirali u periodu 1995-1999. god.

Variable	Descriptive Statistics												
	Broj studenta	Srednja vrednost	Geometrijska sredina	Harmonijska sredina	Medijana	Modus	Frekvencija modusa	Zbir	Varijansa	Srednje kvadratno odstupanje	Koeficijenat varijacije	Koeficijenat asimetrije	Koeficijenat spljostenosti
Godina diplomiranja	22	1997,136	1997,136	1997,136	1997,000	1997,000	6	43937,00	1,7424	1,3179	0,0660	-0,3085	-1,0891
Srednja ocena	22	7,277	7,252	7,228	7,070	8,050000	2	160,10	0,3925	0,5339	7,4379	1,2365	2,7013
Ocena na diplomskom	22	9,727	9,711	9,694	10,000	10,00000	17	214,00	0,3030	0,6141	6,3245	-2,1103	3,6749
Godina koju je upisao	22	1,727	1,491	1,320	1,000	1,000000	14	38,00	0,9697	0,9687	54,3673	0,4618	-1,7486

Slika 56. Prikaz tabele sa računskim i pozicionim vrednostima za studente iz slučajnog uzorka cele populacije 1995-1999. god.

Nakon klika na *Summary: Statistics* iz populacije studenata diplomiranih od 2000-2004. godine dobijamo sledeće tabele sa srednjim vrednostima.

Variable	Descriptive Statistics (Diplomirali 2001)											
	Br. stud.	Srednja vrednost	Geom. sredina	**Harmon. sredina**	Medijana	Modus	Frekvenc. modusa	Varijansa	Srednje kv. odst.	Koef. varijacije	Koef. asimetrije	Koef. spljostenosti
Prosek	62	7.076290	7.064106	7.052363	7.010000	7,380000	4	0.182352	0.427026	6.034609	1.203681	2.913796

Slika 57. Tabela sa srednjim vrednostima za uzorak diplomiranih 2001. god.

Variable	Descriptive Statistics (Spreadsheet)											
	Br. stud.	Srednja vrednost	Geom. sredina	Harmon. sredina	Medijana	Modus	Frekvenc. modusa	Varijansa	Srednje kv. odst.	Koef. varijacije	Koef. asimetrije	**Koef. spljostenosti**
Prosek	561	7.285401	7.262148	7.239950	7.170000	7,000000	17	0.355986	0.596646	8.189609	1.043066	1.039616

Slika 58. Tabela sa srednjim vrednostima za celu populaciju (diplomirali 2000-2004).

Variable	Descriptive Statistics (Slucajan uzorak studenata)											
	Br. stud.	Srednja vrednost	Geom. sredina	Harmon. sredina	Medijana	Modus	Frekvenc. modusa	Varijansa	Srednje kv. odst.	Koef. varijacije	Koef. asimetrije	**Koef. spljostenosti**
Prosek	43	7.429767	7.404880	7.381498	7.270000	Multiple	2	0.403226	0.635001	8.546713	1.334866	1.530622

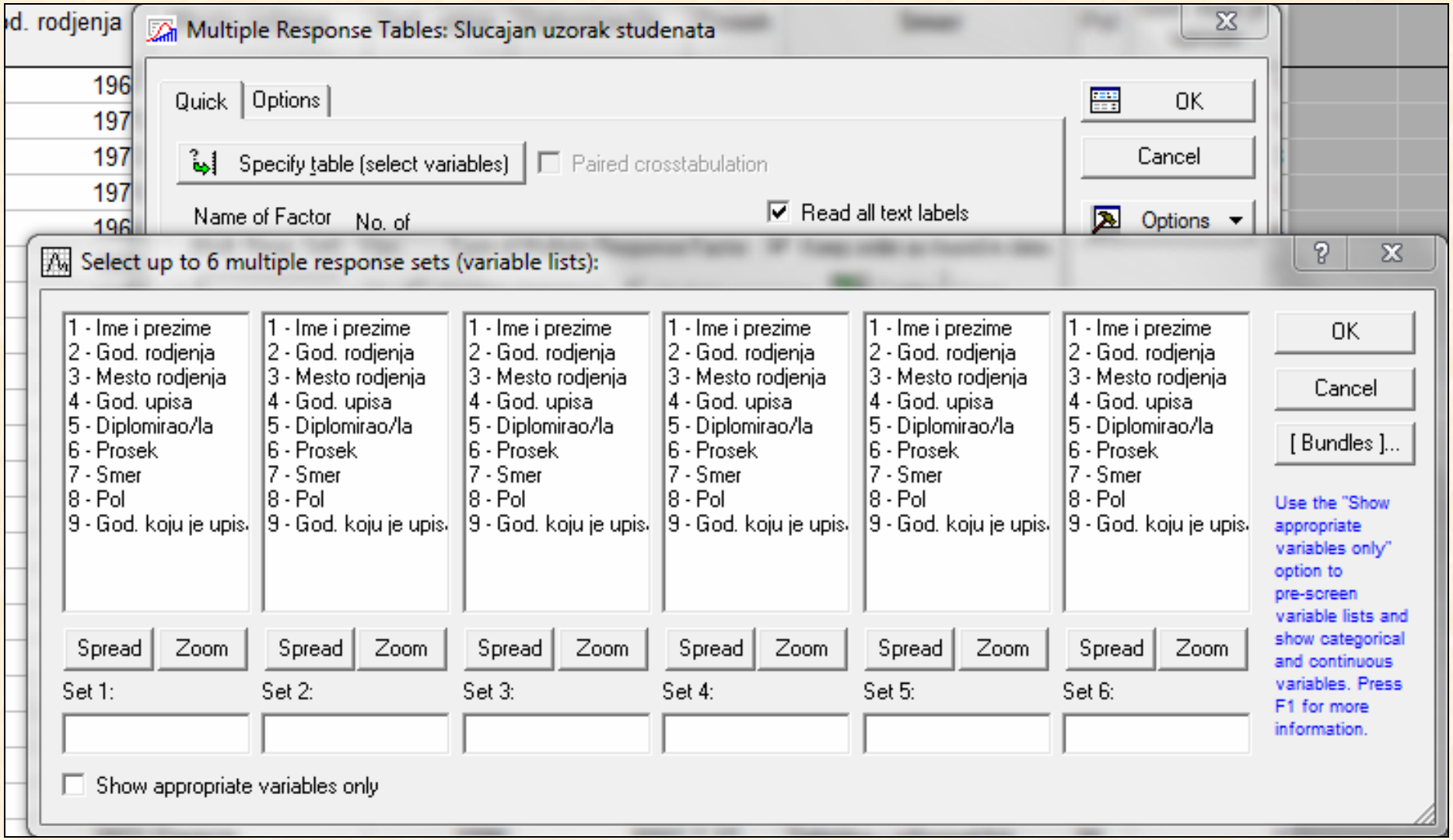
Slika 59. Tabela sa srednjim vrednostima za slučajan uzorak studenata iz populacije 2000-2004.

TABLICE KONTIGENCIJE

Programski paket *Statistica 8* ima mogućnost kreiranja tzv. *tablica kontigencije* koje su veoma pogodne za praćenje nekih statističkih podataka. Na primer, možemo pratiti brojno stanje studenata u odnosu na dva obeležja – godine upisa i godine završetka studija.

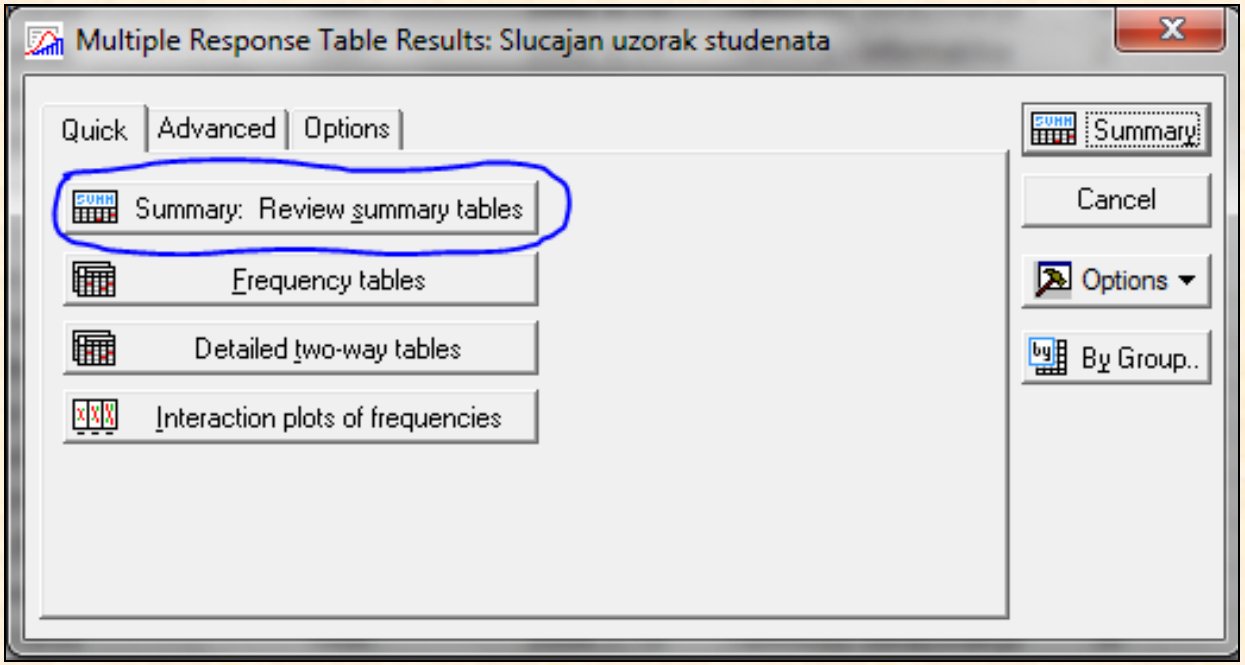
Postupak je sledeći:

Najpre iz menija *Statistics* biramo *Basic Statistics/Tables*, i nakon toga biramo *Multiple response tables*. Dobićemo novi prozor i u njemu biramo *Specify Table* kao na Slici 60.



Slika 60. Kreiranje tablice kontigencije za slučajni uzorak iz 2000-2004. god.

Sledeće što treba da uradimo je da odaberemo obeležja koja želimo da pratimo. To su u našem slučaju „God. upisa“ i „Diplomirao/la, što selektujemo i potvrdimo sa *OK*. Tada nam se otvara novi prozor u kome treba da kliknemo na *Summary: Review summary tables* (Slika 61).



Slika 61. Poslednji korak pri kreiranju tabele kontigencije.

Na taj način dobijamo željene rezultate na Slici 62.

N=43 God. upisa	Diplomirao/la 2000,	Diplomirao/la 2002,	Diplomirao/la 2004,	Diplomirao/la 2003,	Diplomirao/la 2001,	Row Totals
1996,	3	1	1	1	1	7
1997,	1	2	0	2	0	5
2003,	0	0	3	0	0	3
1999,	0	0	0	2	0	2
2002,	0	0	4	3	0	7
2001,	0	1	2	4	0	7
1994,	0	1	1	0	0	2
1975,	0	0	0	0	1	1
1992,	0	0	1	0	0	1
1998,	1	0	1	2	0	4
1995,	2	1	0	0	0	3
1990,	0	1	0	0	0	1
All Grps	7	7	13	14	2	43

Slika 62. Tabela kontigencije za slučajni uzorak iz populacije 2000-2004.

Primenom navedenog postupka obrađujemo studente koji su diplomirali 1991, 1992, 1993 i 1994 (Slika 63).

N=214 Godina upisa	**Godina završetka 1991**	Godina završetka 1992	Godina završetka 1993	Godina završetka 1994	Row Totals
1976	5	2	0	0	7
1983	8	3	1	2	14
1977	3	1	4	2	10
1979	4	1	0	0	5
1978	5	1	2	0	8
1980	4	5	0	0	9
1975	1	1	0	0	2
1986	8	10	9	6	33
1981	3	2	1	1	7
1987	3	2	8	10	23
1984	11	6	3	7	27
1982	2	2	1	1	6
1985	16	7	0	6	29
1988	0	1	4	13	18
1989	0	0	2	10	12
1990	0	0	0	3	3
1992	0	0	0	1	1
All Grps	73	44	35	62	214

Slika 63. Tabela kontigencije između godine upisa i godine diplomiranja za studenate koji su diplomirali od 1991-1994. god.

Sa slike jasno možemo videti da je najviše studenata diplomiralo 1991. godine i to njih 73.

Obrada još jedne populacije (2008–2011.) prikazana je na Slici 64.

N=1000 Godina upisa	Godina završetka studija 2008,	Godina završetka studija 2009,	Godina završetka studija 2010,	Godina završetka studija 2011,	**Row Totals**
2007,	17	55	71	37	180
1999,	1	8	6	2	17
2004,	5	23	33	14	75
2002,	9	19	26	12	66
2003,	4	20	37	16	77
2000,	3	11	4	2	20
2001,	7	22	16	10	55
2005,	4	38	58	17	117
1996,	1	3	0	0	4
1998,	1	4	4	2	11
2006,	3	73	60	39	175
1994,	0	1	0	0	1
1995,	0	1	1	1	3
2008,	0	33	84	44	161
1990,	0	0	1	0	1
2009,	0	0	24	9	33
1989,	0	0	1	0	1
1997,	0	0	0	1	1
1993,	0	0	0	1	1
2010,	0	0	0	1	1
All Grps	55	311	426	208	1000

Slika 64. Prikaz broja studenata u odnosu na dva obeležja – godina upisa i godina diplomiranja.

STATISTIČKO ZAKLJUČIVANJE METODOM UZORAKA

UVOD

Proučavanje osnovnog statističkog skupa retko se vrši na celom tom skupu (jer je neekonomično, neracionalno, nemoguće i slično). Uzorak je deo te populacije koji treba da obezbedi što tačnije informacije o osnovnom skupu. Taj zahtev se realizuje kroz *reprezentativnost uzorka*. Sam izbor elemenata iz populacije u reprezentativan uzorak može se izvršiti na više načina (tablica slučajnih brojeva, uz pomoć računara i drugo). U statističkoj teoriji u svakoj konkretnoj situaciji, zavisno od željene tačnosti i stanja u osnovnom statističkom skupu, može se odrediti i obim uzorka za zadate ocene parametara.

Na taj način, uzorak je umanjena slika osnovnog skupa i njegova aritmetička sredina, uzoračka disperzija, mera asimetrije ili mera ekscesa su procene, tj. ocene ovih parametara osnovnog skupa. Pored toga, raspodela frekvencija posmatranog statističkog obeležja na uzorku predstavlja aproksimativnu raspodelu slučajne promenljive u osnovnom skupu.

Neka je $(X_1,...,X_n)$ prost slučajan uzorak od n slučajnih promenljivih $X_1,...,X_n$. Za svaki takav uzorak možemo formirati neku novu slučajnu promenljivu $Y=f(X_1,...,X_n)$ koju nazivamo *statistika*.

Formule po kojima se računaju realizovane vrednosti navedenih numeričkih karakteristika obeležja (aritmetička sredina, uzoračka disperzija, mod, medijana, koeficijenat varijacije, mera asimetrije, mera ekscesa, uzorački koeficijenat korelacije i drugo) mogu poslužiti da se za svaku od navedenih karakteristika napiše odgovarajuća statistika, po analogiji sa tim formulama, i nazivi tih statistika poklapaju se sa nazivima tih numeričkih karakteristika.

Pri proučavanju obeležja na osnovu uzorka srećemo se sa dva osnovna zadatka.

Prvi zadatak ima za cilj da na osnovu uzorka ustanovi kakva je raspodela obeležja u populaciji, kao i da oceni parametre te raspodele izračunavajući realizovane vrednosti pogodno odabranih statistika. To su *problemi ocene parametara raspodele*.

Drugi zadatak sastoji se u tome da na osnovu procenjenih karakteristika osnovnog skupa iz uzorka, formulišemo i proverimo neku pretpostavku (hipotezu) u vezi sa nekom karakteristikom osnovnog skupa. To su *problemi testiranja (verifikacije) hipoteza*. Te pretpostavke mogu se odnositi na parametre raspodele (*parametarske hipoteze*) ili na same raspodele obeležja (*neparametarske hipoteze*).

Inače, u statističkoj teoriji se smatra da su uzorci čiji obim prelazi 30 jedinica osnovnog statističkog skupa veliki uzorci i na njih se primenjuje teorija zasnovana na normalnoj raspodeli. Za uzorak do 30 jedinica osnovnog statističkog skupa kažemo da su mali uzorci i njih razmatramo u okviru teorije studentove t – raspodele.

OCENE PARAMETARA RASPODELE

Koriste se dve vrste ocena.

Tačkasta ocena je ocena nekog parametra realnim brojem koji se izračunava na osnovu uzorka i predstavlja realizovanu vrednost odabrane statistike. Taj broj je tačka na realnoj osi (otuda i naziv) i on služi kao aproksimacija nepoznate vrednosti parametra raspodele.

Intervalne ocene se daju preko intervala, koji se nazivaju *intervalima poverenja*, jer sa unapred zadatom pouzdanošću prekrivaju nepoznati parametar. Ta pouzdanost se zove *nivo poverenja* i obeležava sa β.

TAČKASTE OCENE PARAMETARA RASPODELE

Neka je $(X_1,...,X_n)$ prost slučajan uzorak i θ nepoznati parametar obeležja X u populaciji. Za ocenu ovog parametra biramo statistiku:

$$\widehat{\theta} = f(X_1,...,X_n).$$

Za realizovani uzorak $(x_1,...,x_n)$ izračunamo broj

$$\widehat{v} = f(x_1,...,x_n)$$

koji predstavlja jednu ocenu parametra θ. Obično se zahteva da ta ocena bude *nepristrasna*, tj.

$$E\left(\widehat{\theta}\right) = E\left(f(X_1,...,X_n)\right) = \theta.$$

U slučaju pristrasnosti, mera pristrasnosti je razlika $\left|E\left(\widehat{\theta}\right) - \theta\right|$.

TAČKASTA OCENA SREDNJE VREDNOSTI I DISPERZIJE IZ UZORKA

Neka obeležje X u osnovnom skupu ima srednju vrednost m i disperziju σ^2. Elementi uzorka $(X_1,...,X_n)$ u slučaju izbora elemenata sa vraćanjem iz konačnog osnovnog skupa (ili bez vraćanja iz beskonačnog skupa) imaju ista očekivanja $E(X_i)=m$ i disperzije $D(X_i)=\sigma^2$, $(i=1,...,n)$.

Statistika

$$\overline{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i$$

je nepristrasna ocena za m, bez obzira na raspodelu obeležja X.

Standardna greška pri oceni sredine populacije je

$$\sigma_{\overline{x}}=\frac{\sigma}{\sqrt{n}}.$$

U slučaju da uzorak formiramo iz konačne populacije ali bez vraćanja, standardna greška je

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

gde je N broj elemenata osnovnog statističkog skupa.
Zaključujemo, da se srednja vrednost m populacije aproksimira aritmetičkom sredinom uzorka $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, tj.

$$\widehat{m} = \bar{x}_n,$$

a to je običan momenat prvog reda.

Slično, disperzija σ^2 osnovnog skupa aproksimira se centralnim uzoračkim momentom drugog reda $\bar{s}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_n\right)^2$, tj.

$$\widehat{\sigma}^2 = \bar{s}_n^2.$$

Ova metoda za određivanje tačkastih ocena parametara statističkog skupa, naziva se *metoda momenata*.

Pored ove, za određivanje tačkastih ocena, često se koristi *metod najmanjih kvadrata* i *metod maksimalne verodostojnosti*.

INTERVALNE OCENE PARAMETARA RASPODELE

Pored tačkastih, u praksi možda i češće, koriste se tzv. *intervalne ocene parametra* θ. Suština ovih ocena svodi se na određivanje intervala $[\hat{\theta}_1, \hat{\theta}_2]$ koji sadrži nepoznati parametar θ sa verovatnoćom $100 \cdot \beta\%$.

Dakle, problem se svodi na to da se odrede dve statistike $\hat{\theta}_1 = f_1(X_1, \ldots, X_n)$ i $\hat{\theta}_2 = f_2(X_1, \ldots, X_n)$ takve da je

$$P\{\hat{\theta}_1 \leq \hat{\theta}_2\} = 1 \text{ i } P\{\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2\} = \beta,$$

gde je β zadata verovatnoća koja se zove *nivo poverenja* za θ.

Interval $[\hat{\theta}_1, \hat{\theta}_2]$ je slučajan interval – *interval poverenja* za θ.

PRIMER 20. Odrediti tačkastu ocenu prosečne ocene m prvog i drugog studenta, kao i intervalnu ocenu za parametar m kod oba studenta sa nivoom poverenja $\beta = 0.90$ i $\beta = 0.95$ (Primer 17).

Rešenje. Za tačkastu ocenu prosečne ocene prvog i drugog studenta koristimo aritmetičku sredinu uzorka i ona je u oba slučaja ista:

$$\widehat{m}_{10} = \overline{x}_{10} = \overline{y}_{10} = 8.1$$

Prvi student:

– 90% interval poverenja za prosečnu ocenu je: $[7.67, 8.53]$;

– 95% interval poverenja za prosečnu ocenu je: $[7.57, 8.63]$.

Drugi student:

– 90% interval poverenja za prosečnu ocenu je: $[7.06, 9.14]$;

– 95% interval poverenja za prosečnu ocenu je: $[6.82, 9.38]$.

Očigledno je da povećanje nivoa poverenja β povećava širinu intervala I_m, tj. povećanje garancija za ocenjeni parametar smanjuje preciznost ocene. Dalje, dužina intervala poverenja za prvog studenta je manja za bilo koji nivoa poverenja β.

PRIMER 21. Naći intervalnu ocenu standardne devijacije prosečne ocene studenata iz Primera 17, za $\beta = 0.90$ i $\beta = 0.950$, kao i tačkastu ocenu toga parametra.

Rešenje. Tačkasta ocena,90% i 95% intervalna ocena, redom, standardne devijacije prosečne ocene je :

– za prvog studenta: $\widehat{\sigma} = \bar{s}_n = 0.7$, $I_\sigma = [0, 0.57]$ i $I_\sigma = [0, 0.54]$;

– za drugog studenta: $\widehat{\sigma} = \bar{s}_n = 1.7$, $I_\sigma = [0, 1.40]$ i $I_\sigma = [0, 1.37]$.

Interval poverenja za očekivanu vrednost primenom paketa *Statistica*

Postupak dobijanja intervala poverenja za neku očekivanu vrednost je isti kao i prilikom dobijanja računskih i pozicionih vrednosti u paketu Statistica, samo što umesto tada selektovanih vrednosti u ovom slučaju aktiviramo *"Confirm limits for means"*. Idemo na padajući meni *"Statistics"* izabiramo *"Basic Statistics/Tabeles"*.

Zatim odabiramo *"Descriptive Statistics"* i potvrdimo na *"Ok"*. Na dobijenom prozoru odaberemo karticu *"Advaced"* i aktivnost *"Confirm limits for means"*. Na kraju se odabere koeficijent pouzdanosti (najčešće 95%) i dobija se sledeća tabela potvrdom na "*Summary: Descriptive statistics*".

Variable	Descriptive Statistics (Interval poverenja "Confidence")					
	Broj studenata	Aritmetička sredina	Confidence -95,000%	Confidence 95,000	Minimum	Maximum
Godina diplomiranja	22	1997,136	1996,551	1997,722	1995,000	1999,000
Srednja ocena	22	7,277	7,000	7,555	6,260	8,800
Ocena na diplomskom	22	9,727	9,483	9,971	8,000	10,000
Godina koju je upisao	22	1,727	1,291	2,164	1,000	3,000

Slika 65. Intervali poverenja nekih obeležja za slučajni uzorak iz populacije 1995-1999 pri koeficijentu pouzdanosti od 95% (1995-1999).

Variable	Descriptive Statistics (Interval poverenja cele populacije)					
	Broj studenata	Aritmetička sredina	Confidence -95,000%	Confidence 95,000	Minimum	Maximum
Godina diplomiranja	504	1997,331	1997,216	1997,447	1995,000	1999,000
Srednja ocena	504	7,177	7,131	7,224	6,260	9,830
Ocena na diplomskom	504	9,710	9,657	9,764	7,000	10,000
Godina koju je upisao	504	1,782	1,697	1,867	1,000	4,000

Slika 66. Intervali poverenja posmatranih obeležja za celu populaciju 1995-1999 pri koeficijentu pouzdanosti od 95% (1995–1999).

Variable	Descriptive Statistics (Interval poverenja diplomiranih 1995)					
	Broj studenata	Aritmeticka sredina	Confidence -95,000%	Confidence 95,000	Minimum	Maximum
Godina diplomiranja	56	1995,000			1995,000	1995,000
Srednja ocena	56	7,196	7,047428	7,344358	6,320	9,200
Ocena na diplomskom	56	9,768	9,623496	9,912218	8,000	10,000
Godina koju je upisao	56	1,357	1,156553	1,557733	1,000	3,000

Slika 67. Intervali poverenja nekih obeležja za diplomirane studente 1995god. pri koeficijentu pouzdanosti od 95%.

Iz baze podataka u paketu *Statistica* za populaciju studenata diplomiranih u periodu 2000-2004 dobijeni su rezultati prikazani na slikama 68, 69 i 70.

Variable	Descriptive Statistics (Diplomirali 2001)					
	Br. stud.	Srednja vrednost	Confidence -95,000%	Confidence 95,000	Minimum	**Maximum**
Prosek	62	7.076290	6.967846	7.184735	6.290000	8.720000

Slika 68. Interval poverenja za prosečnu ocenu studenata diplomiranih 2001. pri koef. pouzdanosti 95%.

Variable	Descriptive Statistics (Slucajan uzorak studenata)					
	Br. stud.	Srednja vrednost	Confidence -95,000%	Confidence 95,000	Minimum	**Maximum**
Prosek	43	7.429767	7.234343	7.625192	6.500000	9.370000

Slika 69. Interval poverenja za prosečnu ocenu studenata na osnovu slučajnog uzoraka iz populacije 2000-2004 pri koef. pouzdanosti 95%.

Variable	Descriptive Statistics (Spreadsheet)					
	Br. stud.	Srednja vrednost	Confidence -95,000%	Confidence 95,000	Minimum	**Maximum**
Prosek	561	7.285401	7.235922	7.334880	6.200000	9.600000

Slika 70. Interval poverenja dobijen iz cele populacije pri koef. pouzdanosti 95%.

Slede rezultati vezani za populaciju 2008-2011.

Variable	Descriptive Statistics (Godisnjaci sredjeni)					
	Valid N	Mean	Confidence -95,000%	Confidence 95,000	Minimum	Maximum
Godina upisa	1000	2005,101	2004,927	2005,275	1989,000	2010,000
Godina završetka studija	1000	2009,787	2009,735	2009,839	2008,000	2011,000
Prosečna ocena	1000	7,483	7,444	7,521	6,170	9,880

Slika 71. Intervali poverenja za obeležja populacije2008.-2011. sa nivoom poverenja 95%.

Variable	Descriptive Statistics (Spreadsheet2)					
	Valid N	Mean	Confidence -95,000%	Confidence 95,000	Minimum	Maximum
Godina upisa	57	2005,614	2004,932	2006,296	1998,000	2009,000
Godina završetka studija	57	2009,947	2009,721	2010,174	2008,000	2011,000
Prosečna ocena	57	7,596	7,424	7,769	6,600	9,590

Slika 72. Intervali poverenja za obeležja slučajno odabranog uzorka iz populacije2008–2011. sa nivoom poverenja 95%.

TESTIRANJE STATISTIČKIH HIPOTEZA

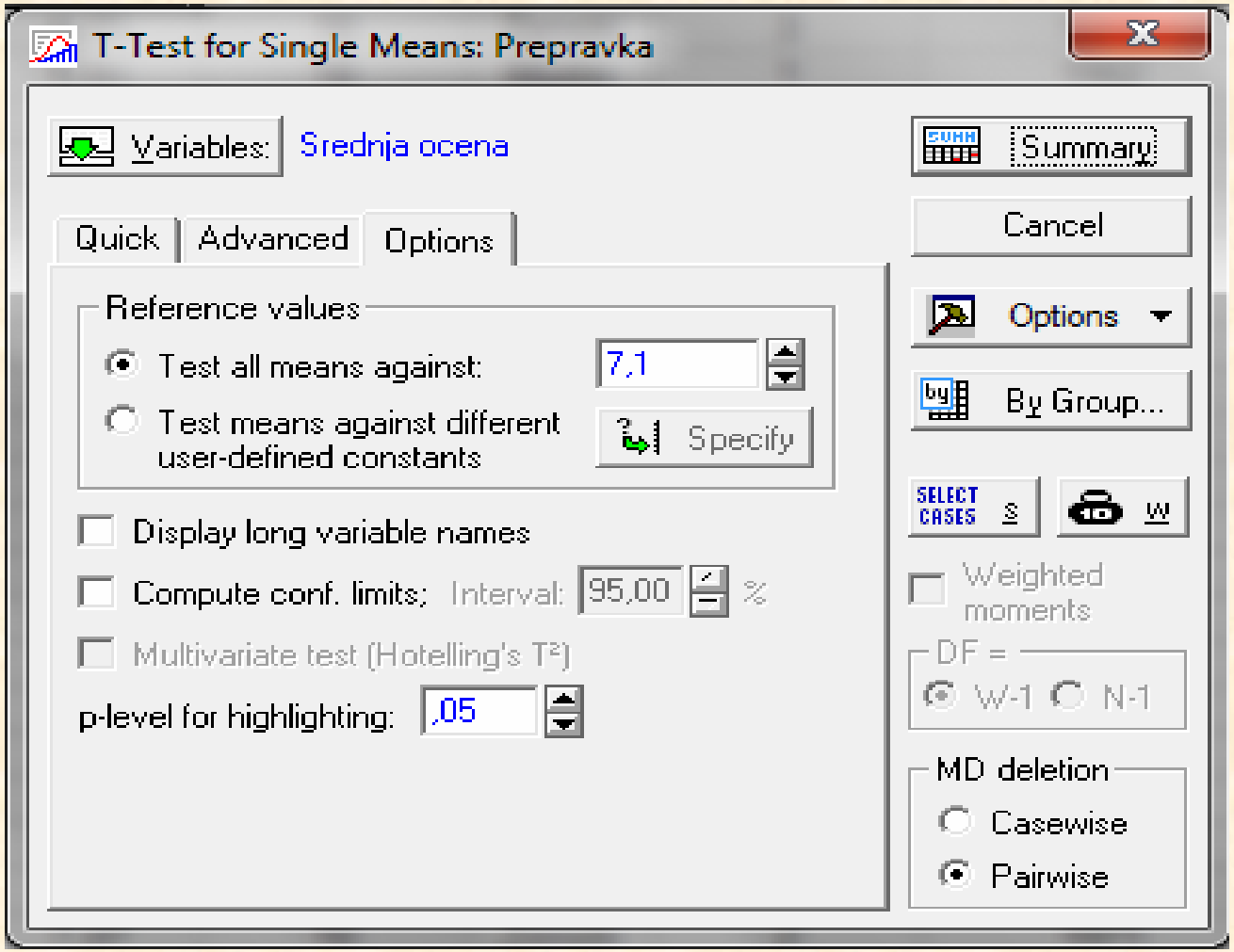
Postupak testiranja hipoteze se izvodi se kroz nekoliko koraka:

1) Definišu se nulta i alternativna hipoteza;
2) Izbor modela teorijske raspodele;
3) Određuje se nivo značajnosti testa α, odnosno verovatnoća $(1-\alpha)$;
4) Definisanje uzorka;
5) Izračunavanje statistike testa na osnovu uzorka;
6) Iz tablice teorijske raspodele očitava se tablična vrednost (kriterijum);
7) Upoređivanje statistike testa sa tabličnom vrednošću;
8) Odluka o prihvatanju ili odbacivanju formulisane hipoteze.

Definisanje hipoteza

Metode ili testovi provere hipoteze omogućavaju da se donese sud o tačnosti hipoteze sa verovatnoćom β dovoljno bliskoj jedinici. Provera statističkih hipoteza naziva se verifikacija statističkih hipoteza. Osnovni zadatak u teoriji provere statističkih hipoteza je određivanje pravila po kome se na osnovu uzoraka može rešiti pitanje da li se postavljena hipoteza prihvata ili odbacuje.

Postupak testiranja hipoteze u paketu Statistica je sledeći: Na meniju *Statistics* izaberemo *Basic Statistics/Tables*, zatim izaberemo *t-test*, *single sample*. Potvrdom na *OK* otvara nam se prozor kao na Slici 73.



Slika 73. Prozor za unos vrednosti m_0.

Klikom na polje *Varibles* odabira se obeležje, a zatim na jezičku *Options* u polje pored teksta *Test all means against* unosimo vrednost za m_0. Vrednost m_o je vrednost koja je približna aritmetičkoj sredini obeležja.

Pri korišćenju statističkog paketa kao rezultat dobija se izračunata *t*-vrednost i odgovarajuća „*p*“ vrednost za izračunato „*t*”. Na osnovu dobijene “*p*” vrednosti zaključci se donose na sledeći nalčin.

– Ako je $p > 0.05$ tada sa pouzdanošću 95% prihvatamo hipotezu H_0.

– Ako je $p \leq 0.05$ tada sa pouzdanošću od 95% odbacujemo hipotezu H_0.

Nivo poverenja se može menjati u zavisnosti od potrebe.

NEJČEŠĆE TESTIRANE HIPOTEZE

Testiranje hipoteze $H_0(m = m_0)$ **protiv** $H_1(m \neq m_0)$, **kada je** σ^2 **poznato**

Možemo vršiti sledeća testiranja:

(1) Uzeti da je $m_0 = \bar{x}_n$, gde je $\bar{x}_n$ prosečna ocena diplomiranih studenata iz uzorka u koji su uzeti studenti upisani na prvu godinu studija pri upisu Fakulteta, a $\sigma^2 = \bar{s}_N^2$ iz populacije iste kategorije studenata. Za prag značajnosti uzeti $\alpha = 0.01$, $\alpha = 0.05$ ili $\alpha = 0.10$.

(2) Uzeti da je $m_0 = \bar{x}_n$, gde je $\bar{x}_n$ prosečna ocena diplomiranih studenata iz uzorka u koji su uzeti studenti upisani na 3. godinu pri upisu na Fakultet, a $\sigma^2 = \bar{s}_N^2$ iz populacije toga uzorka. Za prag značajnosti uzeti $\alpha = 0.01$, $\alpha = 0.05$ ili $\alpha = 0.10$.

Testiranje hipoteze $H_0(m = m_0)$ protiv $H_1(m \neq m_0)$,

kada σ^2 nije poznato

Ovde je $m_0 = \bar{x}_n$, gde je $\bar{x}_n$ prosečna ocena svih diplomiranih studenata iz uzorka, a $\sigma^2 = \bar{s}_n^2$ iz uzorka.

Testiranje hipoteze o jednakosti dve aritmetičke sredine,

tj. $H_0(m_M = m_Z)$ protiv $H_1(m_M \neq m_Z)$

Uzećemo da je $m_M = \bar{x}_n$, gde je $\bar{x}_n$ prosečna ocena studenata iz slučajnog uzorka, a $m_Z = \bar{y}_n$, gde je $\bar{y}_n$ prosečna ocena studentkinja iz tog istog uzorka, za neki prag značajnosti uzeti α.

Testiranje hipoteze o razlici dve populacije,

tj. $H_0(p_1 = p_2)$ protiv $H_1(p_1 \neq p_2)$

Ovde se posmatraju dve populacije studenata: prvu čine studenti upisani na prvu godinu Fakulteta, a drugu populaciju čine studenti upisani na neku stariju godinu-prelaznici sa visokih škola ili nekih drugih fakulteta. Parametar p_1 odnosi se na procenat studenata koji studiraju 5 i više godina, a upisani su u 1. godinu Fakulteta, a parametar p_2 odnosi se na procenat studenata koji studiraju 3 i više godina, a upisani su u 3. godinu Fakulteta. Za prag značajnosti uzeti $\alpha = 0.01$, $\alpha = 0.05$ ili $\alpha = 0.10$.

Na sledećim slikama videćemo rezultate testiranja hipoteze u vezi sa obeležjem Srednja ocena za različite uzorke. Nulta hipoteza tvrdi da je vrednost srednje ocene cele populacije studenata diplomiranih od 1995. do 1999. god., $m_0 = 7.17746$ sa pragom značajnosti $\alpha = 0.05$. Alternativna hipoteza je $H_1(m \neq m_0)$.

Pri korišćenju statističkog paketa kao rezultat dobija se izračunata *t*-vrednost i odgovarajuća „*p*“ vrednost za izračunato „*t*”. Na osnovu dobijene “*p*” vrednosti zaključci se donose na sledeći način:

- Ako je $p > 0.05$ tada sa pouzdanošću 95% prihvatamo hipotezu H_0.
- Ako je $p \leq 0.05$ tada sa pouzdanošću od 95% odbacujemo hipotezu H_0.

Variable	Test of means against reference constant (value) (Prepravka)							
	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
Srednja ocena	7,177460	0,533852	504	0,023780	7,100000	3,257420	503	0,001200

Slika 74. Rezultati testiranja hipoteze o srednjoj oceni cele populacije

Na osnovu dobijene vrednosti *p* koja je manja od 0,05 zaključujemo da se nultna hipoteza odbacuje za prag značajnosti $\alpha = 0{,}05$.

Variable	Test of means against reference constant (value) (tabela diplomiranih 199							
	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
Srednja ocena	7,195893	0,554383	56	0,074083	7,100000	1,294405	55	0,200934

Slika 75. Rezultati testiranja hipoteze
o srednjoj oceni studenata diplomiranih 1995. god.

Vrednost $p = 0{,}200934 > 0{,}05$ pa nemamo razloga da odbacimo nultu hipotezu, odnosno prihvatamo nultu hipotezu da je prosečna ocena diplomiranih studenata 1995. godine upravo 7,195893 sa pragom značajnosti $\alpha = 0{,}05$.

Variable	Test of means against reference constant (value) (Slucajni uzorak 22 stu							
	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
Srednja ocena	7,277273	0,626473	22	0,133564	7,100000	1,327244	21	0,198678

Slika 76. Dobijeni rezultati prilikom testiranja hipoteze o srednjoj oceni studenata na osnovu slučajnog uzorka.

Kao što vidimo iz tabele dobijena vrednost $p = 0{,}198678 > 0{,}05$ i prihvatamo nultu hipotezu, što znači da je prosečna ocena studenata diplomiranih u periodu od 1995–1999., približno 7.277273 sa nivoom poverenja 95%.

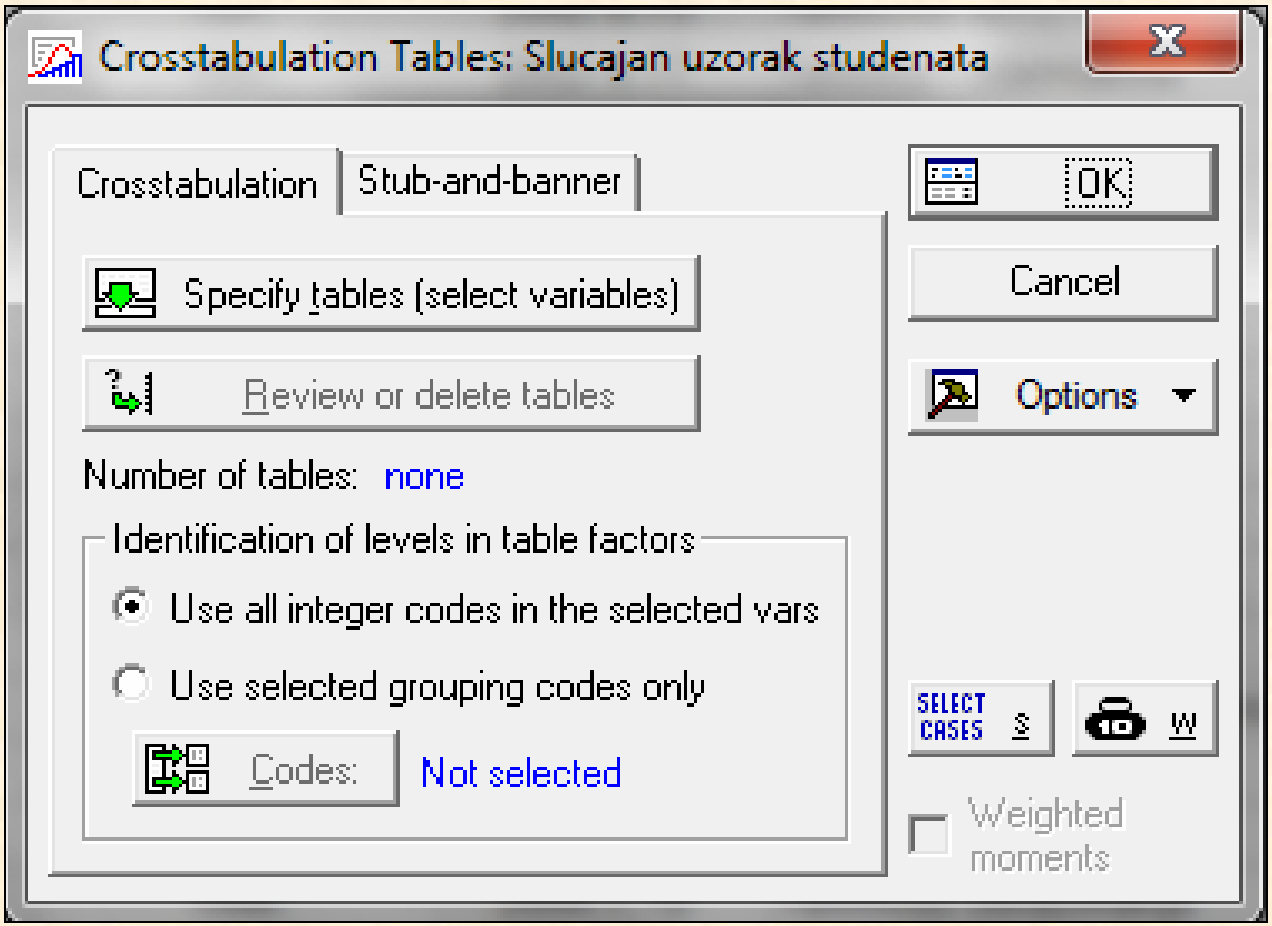
Primetimo da program *Statistica* u slučaju odbacivanja nulte hipoteze u prozoru sa rezultatima koristi crvenu boju za izračunate vrednosti, što je slučaj i na prethodnim slikama. Na taj način smo i vizuelno opomenuti da nultu hipotezu treba odbaciti.

χ^2–TEST NEZAVISNOSTI

Koristeći *χ^2-test* mogu se odrediti verovatnoće povezanosti između dva obeležja ali ne i jačina te povezanosti. Jačina povezanosti može se odrediti primenom koeficijenata kontigencije. χ^2 se primenjuje kada je potrebno utvrditi da li se neke realizovane frekvencije razlikuju od frekvencije koje bismo očekivali pod određenom hipotezom.

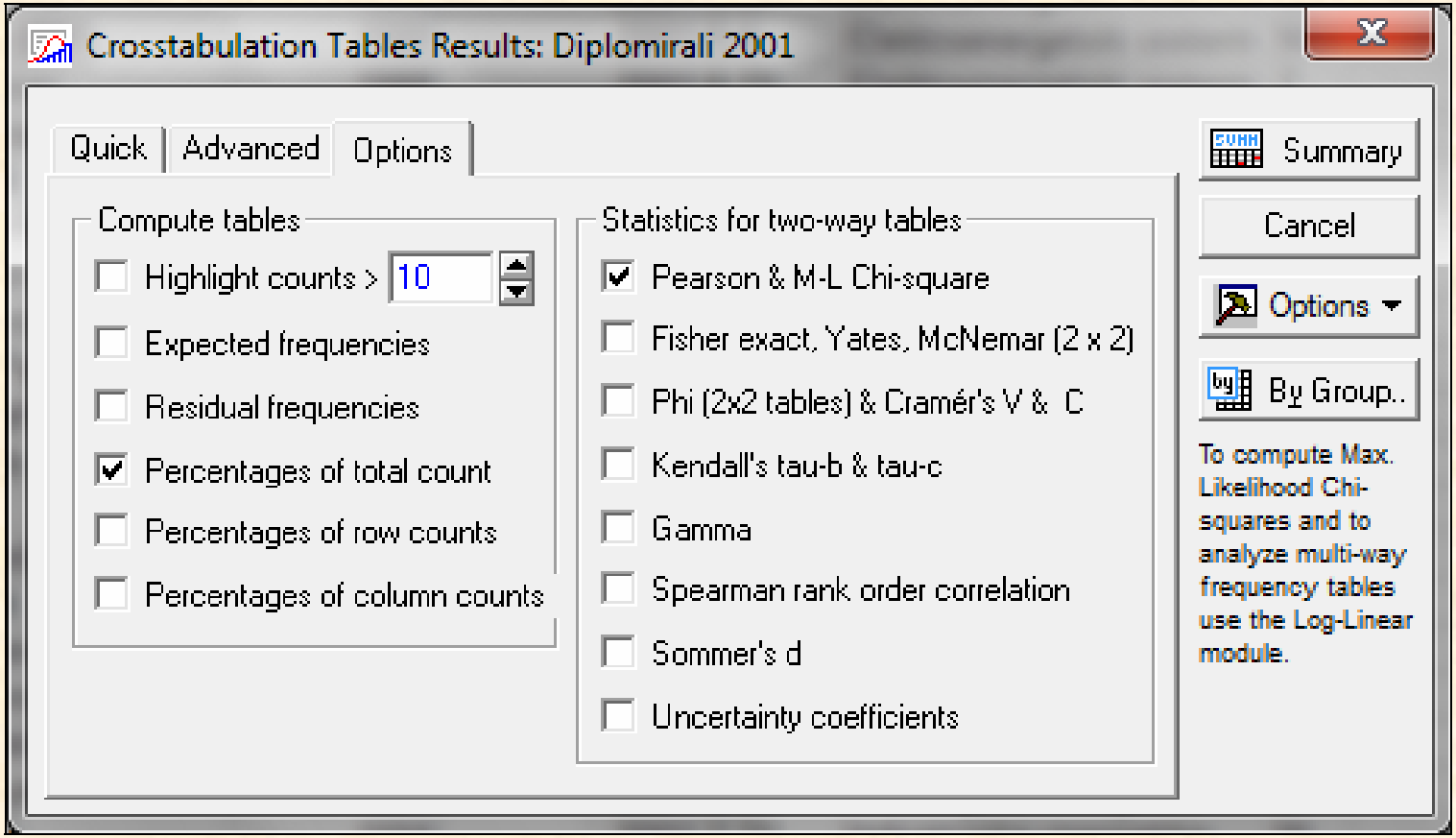
U paketu *Statistica*, χ^2-test se može realizovati tako što iz menija *Statistics*, izaberemo *Basic Statistics/Tables*, a zatim *Tables and Banners*.

Dobijamo prozor kao na Slici 77.



Slika 77. Izgled prozora *Crosstabulation Tables*.

Izbor obeležja vršimo preko dugmeta *Specify tables* i nakon toga potvrdimo dvaput *OK*. Sada dobijamo novi prozor prikazan na Slici 78.



Slika 78. Prozor za izbor vrste analize.

Po izboru analize koju želimo da vršimo, vraćamo se na karticu *Advanced* i kliknemo na opciju *Detailed two-way tables*.

Neke moguća testiranja:

– Ispitati preko uzorka da li su POL i DUŽINA STUDIRANJA zavisna obeležja.

– Ispitati preko uzorka da li su PROSEČNA OCENA i DUŽINA STUDIRANJA zavisna obeležja.

U sledećem primeru ispitivali smo zavisnost između dva obeležja iz populacije studenata diplomiranih u periodu 1995-1999. godine na Tehničkom fakultetu u Čačku.

Ispituje se da li postoji zavisnost između pola studenata i ocene dobijene na diplomskom radu (Tabele 79 i 80).

Pol	2-Way Summary Table: Observed Frequencies (Prepravka)				
	Ocena na diplomskom 7	Ocena na diplomskom 8	Ocena na diplomskom 9	Ocena na diplomskom 10	**Row Totals**
Muški	1	17	38	284	340
Total %	0,20%	3,37%	7,54%	56,35%	67,46%
Ženski	2	17	31	114	164
Total %	0,40%	3,37%	6,15%	22,62%	32,54%
Totals	**3**	**34**	**69**	**398**	**504**
Total %	0,60%	6,75%	13,69%	78,97%	100,00%

Slika 79. Tabela kontigencije cele populacije 1995–1999. god.

Statistic	Statistics: Pol(2) x Ocena na diplomskom(4) (		
	Chi-square	df	**p**
Pearson Chi-square	13,89005	df=3	p=,00306
M-L Chi-square	13,27645	df=3	p=,00408

Slika 80. Rezultat χ^2-testa cele populacije

Na osnovu dobijenih rezultata u programu *Statistica* prikazanih na Slikama 79 i 80 možemo zaključiti sledeće:

H_0 – hipoteza koja tvrdi da su pol i ocena studenata na diplomskom ispitu nezavisna obeležja na celoj populaciji diplomjranih studenata u periodu 1995–1999.

Kako je $p = 0{,}00306 < 0{,}05$ to znači da sa pouzadnošću od 95% odbacujemo nultu hipotezu i zaključujemo da postoji zavisnost između pola studenata i ocena na diplomskom za posmatranu populaciju.

Međutim prema odrađenom slučajnom uzorku te populacije dobijamo sledeće rezultate (Slika 81).

Statistic	Statistics: Pol(2) x Ocena na diplomskom		
	Chi-square	df	p
Pearson Chi-square	5,540441	df=2	p=,06265
M-L Chi-square	5,792301	df=2	p=,05524

Slika 81. Tabela χ^2-testa za slučajni uzorak.

Kako je $p = 0{,}06265 > 0{,}05$ to znači da na osnovu uzorka, sa pouzdanošću od 95%, prihvatamo nultu hipotezu i zaključujemo da su pol i ocena na diplomskom ispitu nezavisna obeležja.

Za subjektivno odabrani uzorak iz populacije 1995–1999., tj. za studente diplomirane 1995. dobijamo sledeće rezultate (Slika 82).

Statistic	Statistics: Pol(2) x Ocena na		
	Chi-square	df	p
Pearson Chi-square	6,609574	df=2	p=,03671
M-L Chi-square	5,578685	df=2	p=,06147

Slika 82. Tabela χ^2-testa za studente koji su diplomirali 1995. god.

Sada je $p = 0{,}03671 < 0{,}05$ što znači da sa pouzadnošću od 95% odbacujemo nultu hipotezu i zakljucujemo da postoji zavisnost između pola studenata i ocene na diplomskom ispitu.

Za generaciju studenata diplomiranih 2001. godine ispitaćemo zavisnost između pola studenta i smera koji je student upisao, za $\alpha = 0{,}05$.

Pol	2-Way Summary Table: Observed Frequencies (Diplomirali 2001)							
	Smer Elektroenergetski sistemi	Smer Industrijska energetika	Smer Mehatronika	Smer Tehnika i informatika	Smer Tehnicko obrazovanje	Smer Inženjer mašinstva	Smer Profesor mašinstva	Row Totals
Z	2	2	0	12	3	0	1	20
Total %	3.23%	3.23%	0.00%	19.35%	4.84%	0.00%	1.61%	32.26%
M	7	17	3	5	7	2	1	42
Total %	11.29%	27.42%	4.84%	8.06%	11.29%	3.23%	1.61%	67.74%
Totals	9	19	3	17	10	2	2	62
Total %	14.52%	30.65%	4.84%	27.42%	16.13%	3.23%	3.23%	100.00%

Slika 83. Tabela kontigencije za uzorak studenata diplomiralih 2001.

Statistic	Statistics: Pol(2) x Smer(7) ([		
	Chi-square	df	p
Pearson Chi-square	18,64315	df=6	p=,00481
M-L Chi-square	20,06258	df=6	p=,00270

Slika 84. Tabela χ^2-testa za uzorak studenata diplomiralih 2001.

Kako je $p = 0{,}00481 < 0{,}05$ to znači da odbacujemo nultu hipotezu,
pa sa pouzdanošću od 95% možemo reći da smer na koji se student upisao zavisi od pola za generaciju studenata Tehničkog fakulteta u Čačku koji su diplomirali 2001. godine.