

Osnovne statističke tehnike obrade i analize podataka

- Statistički skup (populacija) je elementarna statistička kategorija: skup svih slučajeva, elemenata, jedinica koje imaju određenu osobinu; obeležje koje podleže statističkom merenju, obradi analizi; statistički skup može biti definisan u bilo kojoj oblasti života; s obzirom na broj elemenata može biti konačan i beskonačan
- Da bi jedan statistički skup mogao biti realno definisan elementi moraju biti istovrsni, istorodni i istovremeno kvantitativno različiti po istraživanim osobinama (pojam student definiše jasno koja grupacija stanovništva u određenom prostoru i vremenu pripada ovom skupu; studenti se međusobno razlikuju po mnogim osobinama (starosti, dužini studiranja, materijalnom stanju) koje podležu statističkom merenju – time je definisano jedno polje statističkog istraživanja

Obeležja statističkog skupa

- Opisna (kvalitativna) - pol, zanimanje ...
- Numerička (kvantitativna) – starost, lični dohoci...
 - Prekidna – ne može uzeti bilo koju decimalnu vrednost (broj studenata)
 - Neprekidna - može uzeti bilo koju vrednost u datom intervalu što znači da **su moguće bilo koje decimalne vrednosti (starost)**

Statistička serija

- Niz statistički sređenih podataka po obeležju, po vremenskoj ili geografskoj podeli zovemo statističkim serijama
 - Razlikujemo:
 - **Serije strukture**
 - Vremenske serije
 - Geografske serije

Serije strukture

- Prikazuju raspored populacije po modalitetima (kategorijama) opisnog sadržaja tj. prema vrednostima numeričkog obeležja
- Frekvencija kategorija informiše o broju pojavljivanja istog modaliteta tj. vrednosti obeležja. Ako se frekvencija kategorije stavi u odnos prema ukupnom broju elemenata populacije dobija se relativna frekvencija
- **Statističke serije struktura zovu se distribucije ili rasporedi frekvencija**

Tipovi statističkih podataka

- Nominalni
- Ordinalni
- Intervalni
- Podaci odnosa

Nominalni (kategorijalni podaci)

- Nemerljivi, ali prebrojivi podaci; označavaju prisustvo ili odsusutvo nekog svojstva, ali ne i meru intenziteta njegovog ispoljavanja.
- Nominalni su zato što se različitim kategorijama podataka dodeljuju imena; npr. vrste zanimanja poljoprivrednici, nekvalifikovani i kvalifikovani radnici, medicinsko osoblje, administrativni radnici, prosvetni radnici.
- Možemo ih prebrojati i utvrditi tačan broj ljudi koji pripadaju svakoj kategoriji, ali time nismo precizno izmerili neku razliku između njih: ne možemo reći da neko od njih u većoj ili manjoj meri poseduje to svojstvo (npr. pripadnost profesiji) kao što možemo da tvrdimo da je neko stariji (svojstvo broja godina) ili obrazovan (svojstvo obrazovanja), sa višom platom (svojstvo visina plate). Zato za ovakve podatke kažemo da su prebrojivi, ali nemerljivi.

Ordinalni podaci

- grubo merljiva svojstva; možemo samo da tvrdimo da je nešto veće, jednako ili manje, ali ne i da precizno ustanovimo koliko je veće ili manje; rastojanje između dva susedna stepena nije tačno određeno. Stepen obrazovanja društvene klase (viša, srednja gornja, srednja donja, niža; niko ne može da kaže da je rastojanje između niže i srednje donje klase jednako rastojanju između srednje gornje i više.
- Skala omogućava prenosivost nekog uočenog odnosa među podacima; ako je srednje obrazovanje više od osnovnog, a visoko više od srednjeg, onda se može zaključiti i da je visoko obrazovanje više od osnovnog

Intervalni podaci

- podaci merljivi na intervalnim skalama – izražavaju svojstva čija izraženost se može precizno izmeriti;
- na intervalnim skalama razmak između dva podeoka je jednak (Celzijusva skala:-početak skale je tačka mrženja vode – označava nulu, a najveći podeljak je tačka ključanja vode i označena je sa 100.
- razmak između ove dve tačke podeljen je na sto jednakih delova; razlika u temperaturi između 12 i 14 stepeni Celzijusovih jednaka je razlici između 43 i 45 stepeni

Podaci odnosa – racio podaci

- Podaci merljivi na racio-skalama – podaci koji imaju apsolutnu tj. prirodnu nulu – visina prihoda ljudi u dinarima, broj prodatih primeraka neke robe, broj posetilaca sajta itd.

Sređivanje statističkih podataka

primer podaci o ocenama iz srpskog u jednom odeljenju

- Originalni podaci

- 1, 2, 4, 3, 5, 5, 2, 3, 2, 4, 3, 1, 1, 2, 3, 3, 3, 2, 2, 1, 5, 3, 1, 2, 3, 5, 5,

- Uređena statistička serija

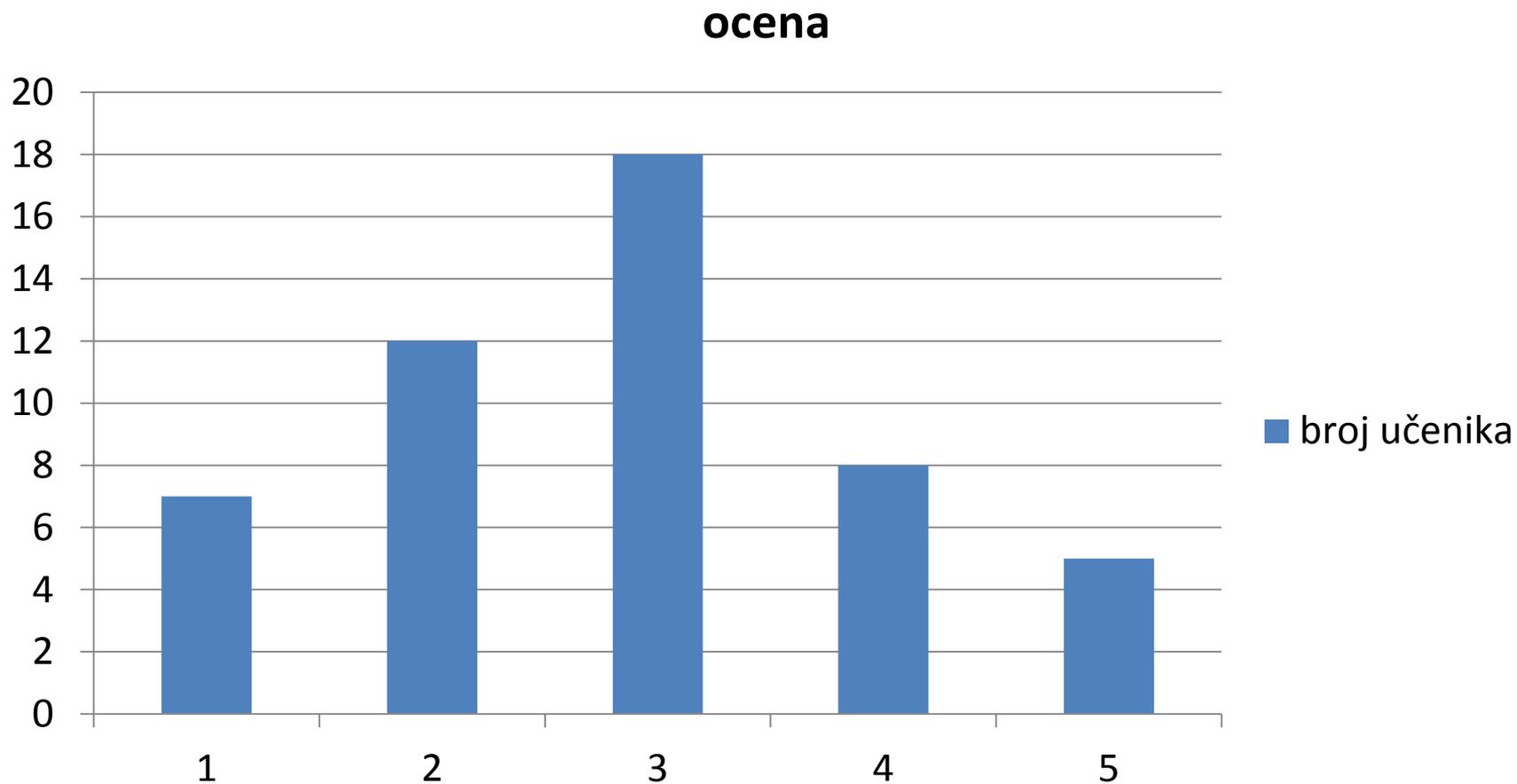
- 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5

- Distribucija frekvencija

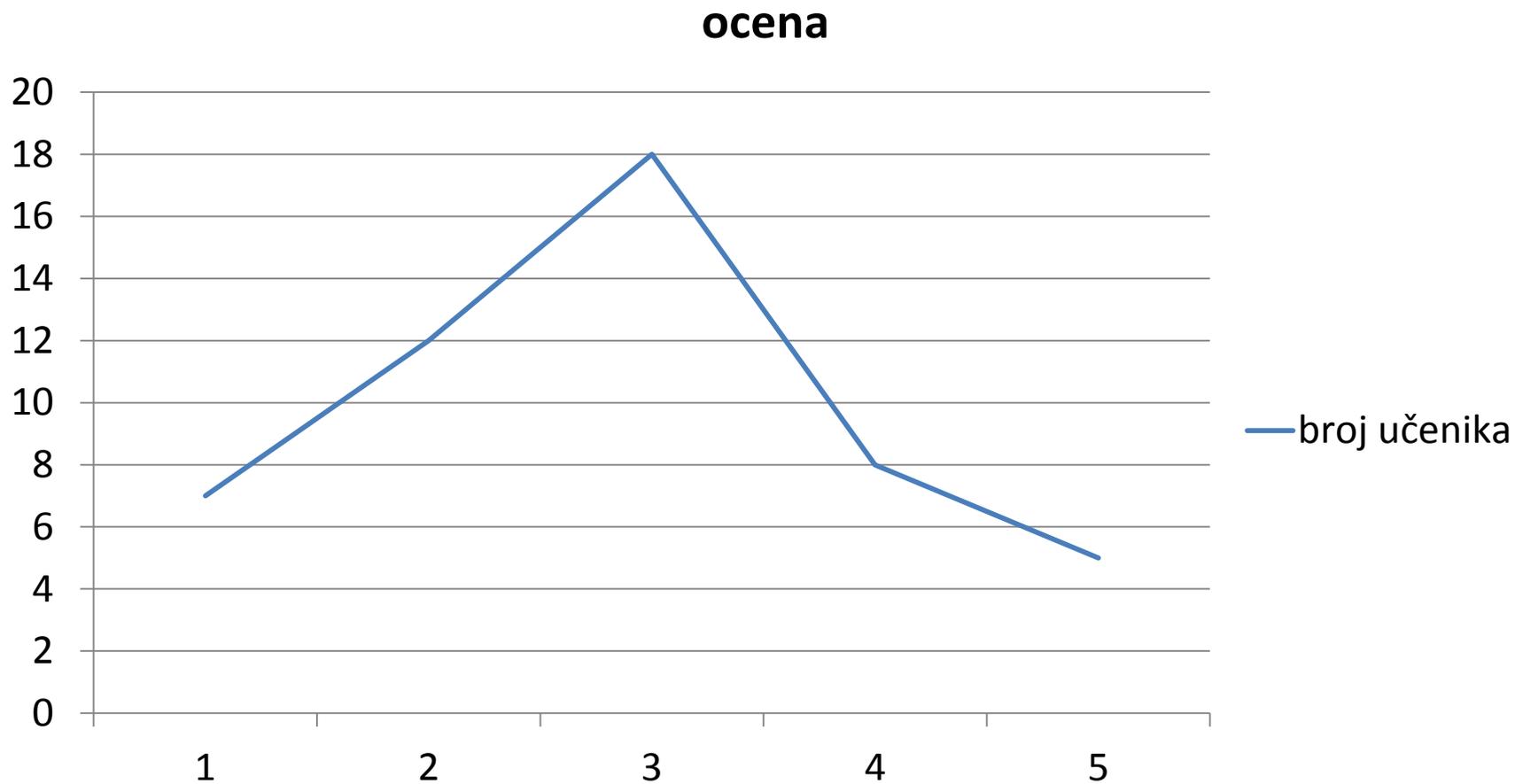
Ocena x	Broj učenika f
1	7
2	12
3	18
4	8
5	5

Grafičko prikazivanje podataka

štapičasti dijagram



Grafičko prikazivanje podataka izlomljena linija



Srednje vrednosti

- Prosečna ili srednja vrednost jednog skupa podataka predstavlja meru centralne tendencije
- Mere centralne tendencije imaju za cilj da odrede centar populacije osnovnog skupa tj. da utvrde ono što je tipično, zajedničko za sve elemente jedne populacije. Iz ovoga proizlazi da se elementi skupa najviše grupišu oko tog proseka
- Vrste: modus, medijana, aritmetička sredina

Srednje vrednosti: modus

- Modus je ona vrednost obeležja x koja se najčešće javlja

Student (x)	1	2	3	4	5	6	7
Broj pohađanih časova	62	81	72	62	72	79	62

- Gruba najčešće samo prva informacija o prosečnoj vrednosti
- Serija može biti i **bimodalna** tj. da ima dva modusa, a može biti i bez modusa ako se sve vrednosti pojavljuju jednako

Srednje vrednosti medijana

- Srednja vrednost obeležja x pošto je serija prethodno uređena od minimalne do maksimalne vrednosti. Iz ovog proizlazi da onoliko članova serije koliko je ispod medijane toliko je i iznad

Student (x)	1	2	3	4	5	6	7
Broj pohađanih časova po studentu	62	62	72	72	74	79	81

- Za razliku od modusa svaka serija ima medijanu

Nedostatak modusa i medijane

- Ne uzimaju u obzir vrednosti svih članova serije nego samo jednu ili dve vrednosti. Dosledno izvedena logika ocene centralne tendencije, onog što je zajedničko svim elementima jednog skupa, mora uzeti u obzir sve elemente populacije. Aritmetička sredina rešava ovaj problem

Aritmetička sredina

- Suma vrednosti podataka podeljena brojem podataka

$$\bar{x} = \frac{\sum x}{n} = \frac{x_1 + x_2 + x_3 \dots x_n}{n}$$

\bar{x} = aritmetička sredina (x – bar)

$x_{1=}$ vrednost i-tog podatka

n = broj podataka ili veličina uzorka

Aritmetička sredina populacije

- \bar{x} = aritmetička sredina uzorka
- μ (mi)=aritmetička sredina populacije
- n = broj elemenata uzorka
- N = broj elemenata populacije
- $\mu = \frac{\sum x}{N} = \frac{x_1 + x_2 + x_3 \dots x_n}{N}$
- Merljive karakteristike uzorka često se nazivaju statistikom
- Merljive karakteristike populacije nazivaju se parametrom
- Uobičajeno se statistika izračunava u cilju ocene nepoznatog parametra populacije. Zato se aritmetička osobina uzorka može smatrati ocenom aritmetičke sredine populacije iz koje je uzorak izvučen.

Osobine aritmetičke sredine

- Može se izračunati za bilo koji niz intervalnih podataka što znači da uvek postoji,
- Bilo koji niz podataka ima samo jednu aritmetičku sredinu, što znači da je ona jedinstvena vrednost bilo kog niza,
- Vrlo je pouzdana
- Za njeno računanje uzimaju se u obzir svi podaci

Izbor mere centralne tendencije

- Aritmetička sredina je najčešće korišćena mera centralne tendencije. Smatra se najpouzdanijom i najpreciznijom merom zato što se aritmetičke sredine uzoraka uzetih iz iste populacije neće toliko razlikovati kao modusi i medijane istih podataka
- Izbor mere delimično zavisi i od tipa podataka. Aritmetička sredina zahteva intervalni nivo podataka. Aritmetička sredina ordinalnih podataka je besmislena kao i aritmetička sredina nominalnih podataka

Izbor mere centralne tendencije u odnosu na tip podataka

A Intervalni podaci	B Ordinalni podaci		C Nominalni podaci	
težine šestorice studenata (kg)	rang	broj	fakultet	Broj studenata u 000
71,5	Poručnik	70	Medicinski	4,2
73,0	Kapetan	120	Mašinski	2,6
68,2	Major	60	Pravni	3,7
71,6	Pukovnik	25	Filozofski	3,9
75,3	General	10	Ekonomski	4,1
50,8				

Ne možemo dobiti aritmetičku sredinu činova u vojsci. Takođe nije moguće dobiti ni aritmetičku sredinu fakulteta jer se oni ne mogu tretirati ni intervalno ni hijerarhijski

Aritmetička sredina grupisanih podataka

- Podaci o starosti, obrazovanju, dohotku i sl. prikazuju se u obliku distribucije frekvencija. Da bi se izračunala aritmetička sredina distribucije frekvencija pretpostavlja se da su vrednosti jednako raspoređene u svakoj klasi. Logično, aritmetička sredina vrednosti u jednoj klasi je centar te klase. Zato se centar klase koristi kao reprezent klase
- $\bar{x} = \frac{\sum fx}{\sum f} \quad \sum f = n$
- \bar{x} = aritmetička težina
- f = frekvencija date klase
- x = centar date klase
- n = broj jedinica u uzorku

Aritmetička sredina grupisanih podataka - primer

Raspored beba po težini			
Klasni intervali (težina u kg)	Frekvencija f	Centar klase x	fx
0,0-0,1	4	0,5	2,0
1,1-2,0	9	1,5	13,5
2,1-3,0	42	2,5	105,0
3,1-4,0	56	3,5	196,0
4,1-5,0	17	4,5	76,5
5,1-6,0	9	5,5	49,5
	137		442,5

$$\Sigma f = n = 137$$

Četiri bebe u klasi 0,0-0,1 imaju prosečnu težinu od 0,5 (centar klase).
Ukupna težina sve četiri bebe ove klase iznosi 2,0 kg ...

Prosečna težina svih beba je:

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{442,5}{137} = \mathbf{3,2 \text{ kg}}$$

Mere disperzije

- Direktno poređenje prosečnih vrednosti dva ili više rasporeda može dovesti do pogrešnih zaključaka.
- Da bismo mogli da poredimo dve i više serija pored informacije o proseku, moramo imati i informaciju o odstupanjima od proseka
- Dve vrste mera s obzirom na način merenja disperzije
 - Mere intervala disperzije – interval varijacije, interkvartilna razlika
 - Mere srednje disperzije – srednje apsolutno odstupanje, varijansa i standardna devijacija

Varijansa i standardna devijacija

- Varijansa je mera disperzije svih elemenata serije od aritmetičke sredine te serije. Srednje kvadratno odstupanje članova serije od centra serije (njene aritmetičke sredine)
 - $\delta^2 = \frac{\sum (x-\bar{X})^2}{N}$ - za populaciju
 - $s^2 = \frac{\sum (x-\bar{X})^2}{n-1}$ - za uzorak
- Standardna devijacija je kvadratni koren srednjeg kvadratnog odstupanja
 - $\delta = \sqrt{\delta^2}$ - za populaciju
 - $s = \sqrt{s^2}$ - za uzorak

Testiranje statističkih hipoteza

- Statističke hipoteze su pretpostavke o parametrima tj. karakteristikama osnovnog skupa
- Testiranje hipoteze vrši se na osnovu uzorka
- Konstrukcija testiranja statističkih hipoteza bazira se na verifikaciji suprotne hipoteze od one koju isputujemo
 - Ako ispitujemo da li je tačno da je $\lambda < \lambda_H$ mi ćemo stvarno ispitivati da li je obrnuta hipoteza tačna $\lambda \geq \lambda_H$. Ako odbacimo $\lambda \geq \lambda_H$ onda ćemo prihvatiti da je tačno $\lambda < \lambda_H$. Hipoteza $\lambda \geq \lambda_H$ naziva se nultom hipotezom i ima sledeći izraz:
 - $H_0: \lambda \geq \lambda_H$
 - λ – bilo koji parametar osnovnog skupa
 - λ_H – hipotetička vrednost datog parametra osnovnog skupa
 - H_0 – nulta hipoteza
 - Njoj alternativna hipoteza za nas početna jeste:
 - $H_1: \lambda < \lambda_H$
 - H_1 – alternativna hipoteza
 - Preko ispitivanja verodostojnosti nulte hipoteze verifikujemo našu originalnu tj. alternativnu hipotezu

Koraci u testiranju statističkih hipoteza

1. Formira se hulta hipoteza; formira se i alternativna hipoteza; alternativna se odbacuje ako je prihvaćena hulta i obrnuto
2. Određuje se prag značajnosti; najčešće je to 0,05 ili 0,01
3. Bira se statistički test na osnovu koga se vrši testiranje; fiksira se promenljiva – parametar odgovarajući izabranom testu
4. Određuje se kriterijum na osnovu koga se donosi odluka o prihvatanju ili odbacivanju nulte, odnosno alternativne hipoteze. Kriterijum se bazira na pragu značajnosti i izabranom testu
5. Obradi se jedan ili više uzoraka čije rezultate koristimo za izračunavanje teorijske promenljive definisane tačkom 3. na kraju na osnovu kriterijuma iz četvrtog koraka donosimo odluku o verodostojnosti nulte hipoteze

Osnovna razlika u tretmanu ispitivanja hipoteza proizlazi iz veličine uzorka: uzorci veći od 30 jedinica smatraju se dovoljno velikim za primenu normalnog rasporeda; uzorci manji od 30 elemenata statistički su mali i za njih se koristi Studentov raspored – t- test.

Testiranje hipoteze o jednakosti aritmetičkih sredina dva osnovna skupa

- $Z = \frac{\bar{X}_1 - \bar{X}_2}{\delta_d}$ $\delta_d = \sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}$

- δ -standardna devijacija osnovnog skupa
- \bar{x} - aritmetička sredina
- n – broj elemenata uzorka
- δ_d – standardna devijacija distribucije razlika parova aritmetičkih sredina uzorka

Testiranje hipoteze o jednakosti aritmetičkih sredina dva osnovna skupa -primer

Brzina odgovora hitne pomoći u odnosu na starost pacijenata u minutama

pacijenti	Srednja vrednost odziva \bar{x}	Standardna devijacija uzorka s	Broj jedinica u uzorku n
stariji	25	5,9	100
mlađi	22	4,1	120

H_0 : nema razlike u vremenu odziva između ove dve grupe

H_1 : vreme odziva se razlikuje između ove dve grupe

Nivo značajnosti 0,01

Oblast prihvatanja hulte hipoteze definisana je izrazom $z \leq 2,33$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\delta_d}$$

$$\delta_d = \sqrt{\frac{\delta_1^2}{n_1} + \frac{\delta_2^2}{n_2}}$$

$$Z = \frac{25 - 22}{0,70} = 4,29$$

$$\delta_d = \sqrt{\frac{5,9^2}{100} + \frac{4,1^2}{120}}$$

$Z > 2,33$ - sa pouzdanošću od 99% odbacujemo nultu hipotezu da je vreme odziva za starije pacijente isto kao i za mlađe tj. prihvatamo alternativnu hipotezu da hitna pomoć sistematski sporije odgovara na pozive starijih građana

Hi kvadrat test – neparametrijski test

- Pored podele s obzirom na veličinu uzorka testovi se mogu podeliti na parametrijske i neparametrijske
 - Parametrijski se bave testiranjem vrednosti parametara osnovnog skupa, a neparametrijski vrednostima neparametara tj. frekvencijama osnovnog skupa
 - Pretpostavke neparametrijskih testova:
 - Ne mora se znati raspored osnovnog skupa
 - Obeležja su nominalna ili deskriptivna

Hi kvadrat (χ^2) test

- χ^2 test se bazira na teorijskom χ^2 - rasporedu koji ima sledeće karakteristike:
- Raspored je definisan u oblasti od 0 do $+\infty$
- Raspored je nesimetričan; međutim s povećanjem broja modaliteta obeležja tj. stepeni slobode raste
- Za svaki stepen slobode postoji i određen χ^2 - raspored
- Suština problema koji rešava χ^2 – test je u kom se stepenu neka empirijska distribucija frekvencija približava nekoj teorijskoj ili očekivanoj distribuciji frekvencija

Hi kvadrat test - primer

- $$\chi^2 = \sum \frac{(f_e - f_t)^2}{f_t}$$

tipprestupa * starost						
		starost				
			ispod 25	25-49	50 i vise	Total
tipprestupa	nasilnicki	Count	15	30	10	55
		Expected Count	11.0	33.0	11.0	55.0
	nenasilnicki	Count	5	30	10	45
		Expected Count	9.0	27.0	9.0	45.0
	Total	Count	20	60	20	100
		Expected Count	20.0	60.0	20.0	100.0

- f_t je teorijska (očekivna) frekvencija
 - za nasilnički tip prestupa do 25 godina $f_t = 0,55 \cdot 20 = 11$
 - Za nenasilnički tip prestupa do 25 godina $f_t = 0,45 \cdot 20 = 9$

Hi kvadrat test - primer

- H_0 : između tipa nasilja i godina prestupnika ne postoji veza
- H_1 : između tipa nasilja i godina prestupnika ne postoji veza
- Nivo značajnosti je 5%

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.040 ^a	2	.133
Likelihood Ratio	4.231	2	.121
Linear-by-Linear Association	2.500	1	.114
N of Valid Cases	100		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 9.00.

- Ako je Asymp.Sig. (2-sided) > 0,05 Prihvata se H_0 tj. da nema povezanosti između varijabli
- Vrednost hi kvadrat testa daje nam statistički kriterijum stepena različitosti dva rasporeda, empirijskog i teorijskog. U slučaju kada prihvatimo alternativnu hipotezu ne znamo koliki je stepen intenziteta veze među varijablama. Ta informacija dobija se na osnovu koeficijenta kontigencije C

Hi kvadrat test - primer

- Vrednost hi kvadrat testa daje nam statistički kriterijum stepena različitosti dva rasporeda, empirijskog i teorijskog. U slučaju kada prihvatimo alternativnu hipotezu ne znamo koliki je stepen intenziteta veze među varijablama. Ta informacija dobija se na osnovu koeficijenta kontigencije C

$$- C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

		Value	Approx. Sig.
Nominal by Nominal	Phi	.201	.133
	Cramer's V	.201	.133
	Contingency Coefficient	.197	.133
	N of Valid Cases	100	

- Kod tabela većih od 2x2 očekivane frekvencije manje od 5 mogu biti zastupljene u manje od 20% ćelija. U suprotnom se kategorije sažimaju.

Korelaciona analiza

- Korelaciona analiza bavi se merenjem intenziteta promene između dve varijable
- Da li su promene u jednoj varijabli praćene promenama u drugoj varijabli (npr. da li su promene u starosti radnika praćene promenama u dužini radnog staža)
- Koliki je intenzitet povezanosti između promena dve varijable
- Metode korelacione analize razlikuju se po tipu varijabli na koje se primenjuju.
- Pirsonov koeficijent korelacije primenjuje se na intervalne podatke tj. meri stepen povezanosti varijabli dveju intervalnih varijabli u linearnom odnosu
- Spirmanov koeficijent korelacije primenjuje se na ordinalne podatke

Pirsonov koeficijent korelacije

Radnici po godinama starosti i radnom stažu

radnik	1	2	3	4	5	6	7	8
Starost x	26	31	36	41	46	51	51	56
Radni staž y	7	10	13	25	30	27	22	40

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}} \quad r_{xy} = \frac{8 \cdot 8104 - 338 \cdot 174}{\sqrt{8 \cdot 15068 - (338)^2} \sqrt{8 \cdot 4656 - (174)^2}} \quad r_{xy} = 0,91$$

- Oblast definisanosti Pirsonovog koeficijenta korelacije je između -1 i +1
 - 1= jaka negativna korelacija
 - 0,8 = srednja korelacija
 - 0,6= slaba korelacija
 - 0,3 =zanemarljiva korelacija
 - Pirsonov koeficijent korelacije daje samo osnovnu informaciju: da li je povezanost u varijacijama između pojava slaba, srednja ili jaka. Ako hoćemo da znamo koliko je to umerena ili jaka korelacija koristimo **koeficijent determinacije**. Koeficijent je kvadrat Pirsonovog koeficijenta korelacije. Ova determinisanost izražava se koeficijentom ili procentom
 - $R^2 = 0,91^2 = 0,83$ ili 83%
- Dužina radnog staža određena je sa 83% starošću radnika

Spirmanov koeficijent korelacije – korelacija ranga

- Ako je merenje vršeno na ordinalnoj skali koristimo koeficijent korelacije ranga
- Najčešće se koristi Spirmanov koeficijent korelacije
 - Oblast definisanosti je između -1 i $+1$; što mu je vrednost bliža 0 korelacija je slabija

Spirmanov koeficijent korelacije - primer

Vrednovanje televizijskih emisija

emisija	mladi	stari	rang		d	d ²
			mladi	stari		
1	Loša	Veoma dobra	5	2	3	9
2	odlična	Izuzetno loša	1	7	-6	36
3	Veoma loša	Dobra	6	3	3	9
4	podnošljiva	Veoma loša	4	6	-2	4
5	Veoma dobra	Podnošljiva	2	4	-2	4
6	Izuzetno loša	Veoma dobra	7	2	5	25
7	odlična	loša	1	3	-4	16

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

d – razlika između parova
n – broj parova opservacija

$$\rho = 1 - \frac{6 * 103}{7(7^2 - 1)} = 1 - 1,83 = -0,83$$

Između strukture interesovanja mlađih i starijih građana postoji jača inverzna korelacija. Starijima se više sviđaju emisije koje se manje sviđaju mladima i obrnuto.