Predavanje 6

NEPARAMETARSKI TESTOVI

Parametarski testovi

- Parametarski testovi se koriste kada se hipoteza odnosi na nepoznatu vrednost parametra populacije (srednja vrednost, proporcija, varijansa...)
- Većina parametarskih testova je ograničena uslovima koje obeležje mora da ispunjava da bi test bio validan.

Na primer:

Tri uslova za t-test dva uzorka:

1. Nasumično odabrani uzorci.
2. Nezavisni uzorci
3. Normalno distribuirane populacije.

Parametarski vs neparametarski testovi

Kada uslovi za primenu parametarskih testova nisu ispunjeni,

ili kada se hipoteza ne odnosi na parametar populacije, koristimo **neparametarske statističke testove.**

Postoje neparametarski analogoni parametarskim testovima, koji se koriste kao zamena kada nisu ispunjeni uslovi za parametarske.

- kada obeležje nije intervalno već **ordinalno.**
- kada obeležje jeste neprekidno, ali značajno **odstupa od normalne distribucije**.

Ovi testovi se zasnivaju na rangovima, a ne na aritmetičkoj sredini i standardnoj devijaciji.

Rang

- Rang elementa uzorka je pozicija elementa u sortiranom neopadajućem nizu.
- Npr:

element	4	20	6	2	15	10	18	22
rang	2	7	3	1	5	4	6	8

- Ukoliko se pojavljuju isti elementi, oni dele rang.

element	4	15	4	2	15	10	15	22
rang	2.5	6	2.5	1	6	4	6	8

- Zbir rangova je jednak zbiru brojeva od 1 do n (obim uzorka)

Testovi sa rangovima

- Neparametarski testovi koji se zasnivaju na rangovima su analogoni parametarskim testovima.

parametarski	neparametarski
T-test dva uzorka	Man-Whitney U test
ANOVA	Kruscal-Wallis ANOVA
Pirsonov koeficijent korelacije	Spirmanov koeficijent korelacije rangova

Spirmanov koeficijent korelacije

- Neparametarski metod za ocenu jačine zavisnosti dve varijable

Primenjuje se kada nisu ispunjeni uslovi za izračunavanje Pirsonovog koeficijenta korelacije, tj. kada:

- Bar jedna varijabla je data u vidu ordinalnih podataka
- bar jedna varijabla nije normalno distribuirana
- Zavisnost između varijabli nije linearna

Izračunavanje:

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

n - broj parova podataka

d – razlika u rangovima

Primer

- Odrediti Spirmanov koeficijent korelacije za sledeće podatke:

X	23	19	26	23	19	17	23	26	20	19
Y	18.5	14.5	15	17	16.5	14	15.5	18	16	16.5
rangX	7	3	9.5	7	3	1	7	9.5	5	3
rangY	10	2	3	8	6.5	1	4	9	5	6.5
d	3	1	6.5	1	3.5	0	3	0.5	0	3.5
d^2	9	1	42.25	1	12.25	0	9	0.25	0	12.25

$$s = 1 - \frac{6*87}{10(10^2 - 1)} = 0.473$$

Testiranje značajnosti

- Testiramo da li je Spirmanov koeficijent korelacije značajno različit od 0.
- Test-statistika je oblika: $t = s * \sqrt{\frac{n-2}{1-s^2}}$
- Kritična vrednost testa se određuje iz tablice Studentove raspodele, za zadati nivo značajnosti i n-2 stepena slobode
- Uslov za primenu testa je da je broj parova podataka n $\geq$ 10.

- **Primer:** Za podatke iz prethodnog primera testirati da li je Spirmanov koeficijent značajan na nivou 5%.

$$t = 0.473\sqrt{\frac{10-2}{1-0.473^2}} = 1.518$$

- Za $\alpha = 0.05$ i df=8, kritična vrednost je 2.306 (dvostrani test)
- Pošto test statistika ne upada u kritičnu oblast, zaključujemo da Spirmanov koeficijent korelacije nije statistički značajan na nivou 0.05, tj. da na datom nivou značajnosti varijable nisu statistički značajno povezane.

Hi-kvadrat testovi

- Drugu grupu neparametarskih testova čine testovi kod kojih se hipoteza odnosi na neko svojstvo obeležja koje ne predstavlja parametar populacije.
- Ovi testovi nemaju parametarske analogone
- Često se test statistika ponaša po Pirsonovoj χ^2- raspodeli
- U ovu grupu spadaju:

Pirsonov hi-kvadrat test saglasnosti sa raspodelom (Goodness-of-fit test)

Hi-kvadrat test nezavisnosti kategoričkih varijabli (tabele kontingencije)

χ^2 test saglasnosti

Hi-kvadrat test saglasnosti se koristi da bi proverili da li raspodela obeležja odgovara nekoj očekivanoj raspodeli, tj. da li su podaci saglasni sa nekom distribucijom.

Očekivana raspodela može biti:

1. teorijska (diskretna, neprekidna, neka od raspodela „sa imenom“)
2. empirijska (dobijena na osnovu nekog drugog uzorka)

Formulacija hipoteza

$H_0(F=F_0)$ – raspodela odgovara pretpostavljenoj

$H_1(F\neq F_0)$ – raspodela ne odgovara pretpostavljenoj

Opažene i očekivane frekvencije

Da bi primenili hi-kvadrat test, podaci treba da budu grupisani u klase (intervale ili klase sa pojedinačnim elementima)

Za svaku klasu određujemo opažene i očekivane frekvencije.

Opažena frekvencija O neke klase je broj elemenata realizovanog uzorka koji pripadaju toj klasi. Obeležava se i sa n_i.

Očekivana frekvencija E neke klase se računa na osnovu pretpostavljene raspodele F_0, i predstavlja proizvod obima uzorka n i verovatnoće da se neki element nađe u toj klasi p_i.

$$E_i = np_i$$

Uslovi za primenu testa

1. Opažene frekvencije treba da su dobijene na osnovu slučajnog uzorka obima bar 50.
2. Svaka opažena frekvencija treba da je bar 5.

Test-statistika

Ukoliko su prethodni uslovi zadovoljeni, test statistika se ponaša u skladu sa hi-kvadrat raspodelom sa k-1 stepenom slobode, gde je k broj klasa.

$$z = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Test-statistika predstavlja meru odstupanja očekivanih od opaženih frekvencija. Što je njena vrednost veća, veća je verovatnoća odbacivanja početne hipoteze o saglasnosti raspodela.

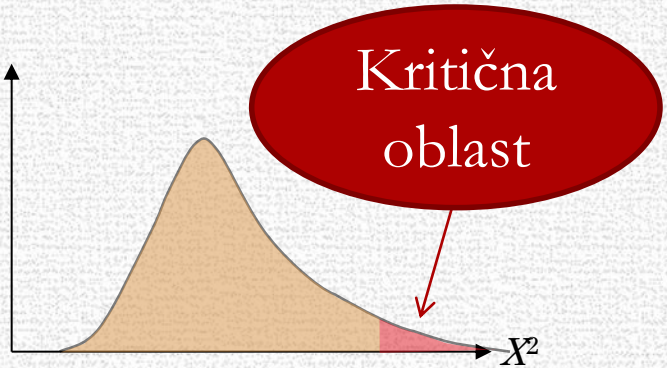
Zaključivanje

Odluka o (ne)odbacivanju nulte hipoteze se donosi na osnovu kritične oblasti, koja se za dati prag značajnosti α određuje iz χ^2- raspodele sa k-1 stepenom slobode.

Ukoliko je realizovana **test-statistika z u kritičnoj oblasti, odbacujemo početnu hipotezu** i zaključujemo da podaci nisu u saglasnosti sa pretpostavljenom raspodelom.

Ako je $z < \chi^2$, **ne odbacujemo H_0** sa pragom značajnosti α

Ako je $z > \chi^2$, **odbacujemo H_0** sa pragom značajnosti α



Napomena: ukoliko je pretpostavljena raspodela teorijska, i nisu poznati svi parametri raspodele, oni se procenjuju metodom momenata, a broj stepeni slobode se umanjuje za broj ocenjenih parametara.

Primer:

Ranije istraživanje je pokazalo da je raspodela omiljenih vrsta pica u nekoj piceriji kao u tabeli (Kolona 1). Rezultati nove ankete su dati u Koloni 2. Sa pragom značajnosti 0.01 proveriti da li se raspodela omiljenih pica promenila, tj. da li su rezultati nove ankete u saglasnosti sa pretpostavljenom raspodelom iz ranijeg istraživanja.

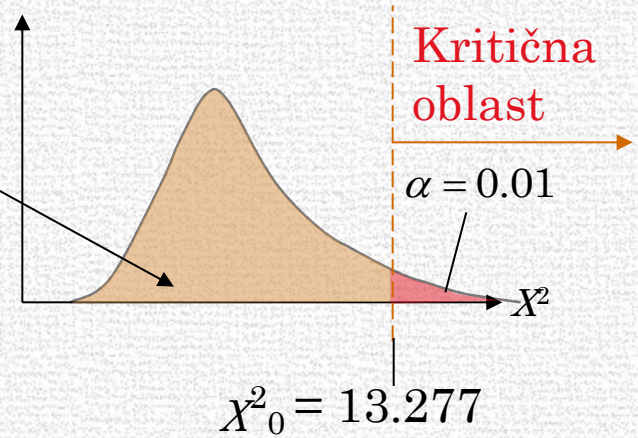
Vrsta pice	Kolona 1 (%)	Kolona 2 ($n = 200$)
Kapričoza	41%	78
Mađarica	25%	52
4 sira	15%	30
Vegetarijana	10%	25
Vojvođanka	9%	15

Nastavak primera:

Vrsta pice	O	E=np
Kapričoza	78	82
Mađarica	52	50
4 sira	30	30
Vegetarijana	25	20
Vojvođanka	15	18

$$X^2 = \Sigma \frac{(O-E)^2}{E} = \frac{(78-82)^2}{82} + \frac{(52-50)^2}{50} + \frac{(30-30)^2}{30} + \frac{(25-20)^2}{20} + \frac{(15-18)^2}{18} \approx 2.025$$

Test statistika nije u regionu odbacivanja nulte hipoteze ($2.025 < 13.277$)



Zaključak: Sa pragom značajnosti 0.01 zaključujemo da je raspodela u skladu sa pretpostavljenom, tj. da se nije promenila distribucija omiljenih pica između nove i stare ankete.

Test normalnosti

- Goodness-of-fit hi-kvadrat test se može koristiti i za proveru saglasnosti empirijskih podataka sa nekom od poznatih raspodela (binomna, normalna, eksponencijalna, Poasonova,...)
- Ukoliko je neki od parametara raspodele nepoznat, on se prethodno ocenjuje tačkastom ocenom, a broj stepeni slobode test statistike se umanjuje za broj ocenjenih parametara
- Specijalno, ukoliko ispitujemo saglasnost sa normalnom raspodelom, reč je o jednom od **testova normalnosti**

Ovi testovi služe da provere da li je ispunjen uslov za primenu parametarskih testova, ili treba koristiti neparametarske analogone.

- Drugi testovi normalnosti: Kolmogorov-Smirnov, Anderson-Darling, vizuelne metode (P-P plot, Q-Q plot, itd...)

Tabele kontingencije

***r*×*c* tabela kontingencije** prikazuje opažene frekvencije za dve kategoričke varijable.

Sadrži *r* vrsta, *c* kolona i *r*×*c* polja.

Sledeća *2*×*6* tabela kontingencije prikazuje distribuciju po polu i po uzrastu 321 vozača žrtve saobraćajnih nesreća u USA.

	uzrast					
pol	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 i više
M	32	51	52	43	28	10
Ž	13	22	33	21	10	6

Očekivane frekvencije

Ukoliko pretpostavimo da su dve varijable nezavisne, pomoću tabele kontingencije (koja sadrži opažene frekvencije) možemo izračunati i očekivane frekvencije.

Nalaženje očekivanih frekvencija

Očekivana frekvencija polja u preseku vrste r i kolone c $E_{r,c}$ u tabeli kontingencije je

$$\text{Expected frequency } E_{r,c} = \frac{(\text{Sum of row } r) \times (\text{Sum of column } c)}{\text{Sample size}}.$$

Primer:

Uz pretpostavku da su pol i starost vozača nezavisni, izračunati očekivane frekvencije za sva polja koja odgovaraju muškim vozačima.

	starost						
pol	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 i više	Total
M	32	51	52	43	28	10	216
Ž	13	22	33	21	10	6	105
Total	45	73	85	64	38	16	321

Nastavak primera:

	starost						
pol	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 i više	Total
M	32	51	52	43	28	10	216
Ž	13	22	33	21	10	6	105
Total	45	73	85	64	38	16	321

$$\text{Expected frequency } E_{r,c} = \frac{(\text{Sum of row } r) \times (\text{Sum of column } c)}{\text{Sample size}}$$

$$E_{1,1} = \frac{216 \cdot 45}{321} \approx 30.28 \qquad E_{1,2} = \frac{216 \cdot 73}{321} \approx 49.12 \qquad E_{1,3} = \frac{216 \cdot 85}{321} \approx 57.20$$

$$E_{1,4} = \frac{216 \cdot 64}{321} \approx 43.07 \qquad E_{1,5} = \frac{216 \cdot 38}{321} \approx 25.57 \qquad E_{1,6} = \frac{216 \cdot 16}{321} \approx 10.77$$

Hi-kvadrat test nezavisnosti

Hi-kvadrat test nezavisnosti se koristi za proveru nezavisnosti dve kategoričke varijable, tj. za utvrđivanje da li verovatnoća pojavljivanja jedne varijable utiče na verovatnoću javljanja druge.

H_0 – varijable su nezavisne.

Sledeći uslovi treba da budu ispunjeni:

1. Opažene frekvencije su dobijene na osnovu slučajno odabranog uzorka.
2. Svaka opažena frekvencija treba da je bar 5.

Hi-kvadrat test nezavisnosti

Ukoliko su navedeni uslovi ispunjeni, test statistika koja meri odstupanje opaženih i očekivanih frekvencija ima hi-kvadrat raspodelu sa

$$(r-1)(c-1)$$

stepenom slobode.

Test statistika se računa po formuli:

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Ukoliko je vrednost test statistike veća od kritične vrednosti za dati prag značajnosti, odbacujemo nultu hipotezu i zaključujemo da varijable nisu nezavisne.

Hi-kvadrat test nezavisnosti

Primer:

Sa pragom značajnosti $\alpha = 0.05$, proveriti na osnovu donje tabele da li su povezani pol i starost vozača koji su stradali u saobraćajnim nesrećama u USA?

	uzrast					
pol	16 – 20	21 – 30	31 – 40	41 – 50	51 – 60	61 i više
M	32	51	52	43	28	10
Ž	13	22	33	21	10	6

Nastavak primera:

Uslovi za hi-kvadrat test nezavisnosti su ispunjeni, pa možemo formulisati hipoteze.

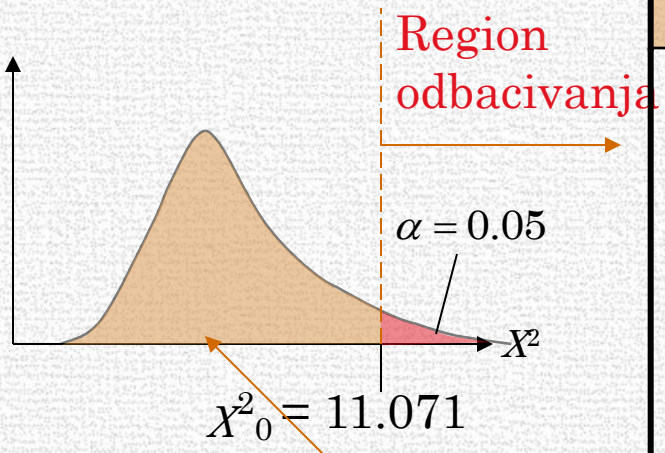
H_0: Starost i pol vozača su nezavisni

H_a: Starost i pol vozača nisu nezavisni (tvrdnja)

d.f. = $(r-1)(c-1) = (2-1)(6-1) = (1)(5) = 5$

Za 5 stepeni slobode i $\alpha = 0.05$, kritična vrednost $\chi^2_0 = 11.071$.

Nastavak primera:



$$X^2 = \Sigma \frac{(O-E)^2}{E} = 2.84$$

Ne odbacujemo H_0.

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
32	30.28	1.72	2.9584	0.0977
51	49.12	1.88	3.5344	0.072
52	57.20	–5.2	27.04	0.4727
43	43.07	–0.07	0.0049	0.0001
28	25.57	2.43	5.9049	0.2309
10	10.77	–0.77	0.5929	0.0551
13	14.72	–1.72	2.9584	0.201
22	23.88	–1.88	3.5344	0.148
33	27.80	5.2	27.04	0.9727
21	20.93	0.07	0.0049	0.0002
10	12.43	–2.43	5.9049	0.4751
6	5.23	0.77	0.5929	0.1134

Sa pragom značajnosti 0.05 zaključujemo da test govori u prilog nezavisnosti pola i uzrasta.