Matematika i statistika

Doktorski studij Strojarskog fakulteta

Sveučilište u Slavonskom Brodu

Predavanje 3

Predavač:
izv.prof.dr.sc. Nenad Šuvak, Odjel za matematiku, Sveučilište u Osijeku

RAZLIKE U DISTRIBUCIJI IZMEĐU DVIJU VARIJABLI

- dolazi li u nekim drugim uvjetima do promjene varijable, tj. obilježja koje proučavamo?
- **nevezani uzorci** - mjerenje istog obilježja vrši se na dva uzorka iz različitih populacija
- uspoređivanje varijabli u nezavisnim uzorcima - npr. želimo uspoređivati rezultate dijagnostičkog postupka za bolesnike u dva različita medicinska centra
- **vezani uzorci** - mjerenje istog obilježja vrši se na istom uzorku prije i poslije nekog tretmana
- uspoređivanje varijabli u zavisnim uzorcima - npr. želimo uspoređivati rezultate dijagnostičkog postupka za iste bolesnike prije i nakon liječenja
- cilj je utvrditi mogu li se razlike koje uočavamo među uzorcima pripisati samo slučajnosti ili ima razloga vjerovati da su izazvane postojanjem razlika između stvarnih distribucija promatranih obilježja u populaciji (razlike su **statistički značajne**)

USPOREDBA OČEKIVANJA

- pretpostavimo da imamo dva nevezana uzorka:
  - $x_1^{(1)}, \ldots, x_{n_1}^{(1)}$ iz distribucije varijable $X^{(1)}$ s očekivanjem μ_1 i standardnom devijacijom σ_1
  - $x_1^{(2)}, \ldots, x_{n_2}^{(2)}$ iz distribucije varijable $X^{(2)}$ s očekivanjem μ_2 i standardnom devijacijom σ_2

- želimo usporediti ta očekivanja, tj. **testirati jednakost očekivanja** promatranog obilježja (varijable) u dvije nezavisne populacije

USPOREDBA OČEKIVANJA - WELCHOV t-TEST

- hipoteze:

$$\mathcal{H}_0 : \ \mu_1 = \mu_2$$

$$\mathcal{H}_1 : \ \mu_1 > \mu_2 \quad \text{ili} \quad \mathcal{H}_1 : \ \mu_1 < \mu_2$$

- pretpostavke testa:
  - $X^{(1)} \sim \mathcal{N}(\mu_1, \sigma_1^2)$
  - $X^{(2)} \sim \mathcal{N}(\mu_2, \sigma_2^2)$
  - σ_1^2 i σ_2^2 nepoznate - procjenjujemo ih korigiranim varijancama uzoraka, $s_{n_1}^2$ i $s_{n_2}^2$

USPOREDBA OČEKIVANJA - WELCHOV t-TEST

- vrijednost test-statistike:

$$\hat{t} = \frac{\overline{x}_{n_1} - \overline{x}_{n_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

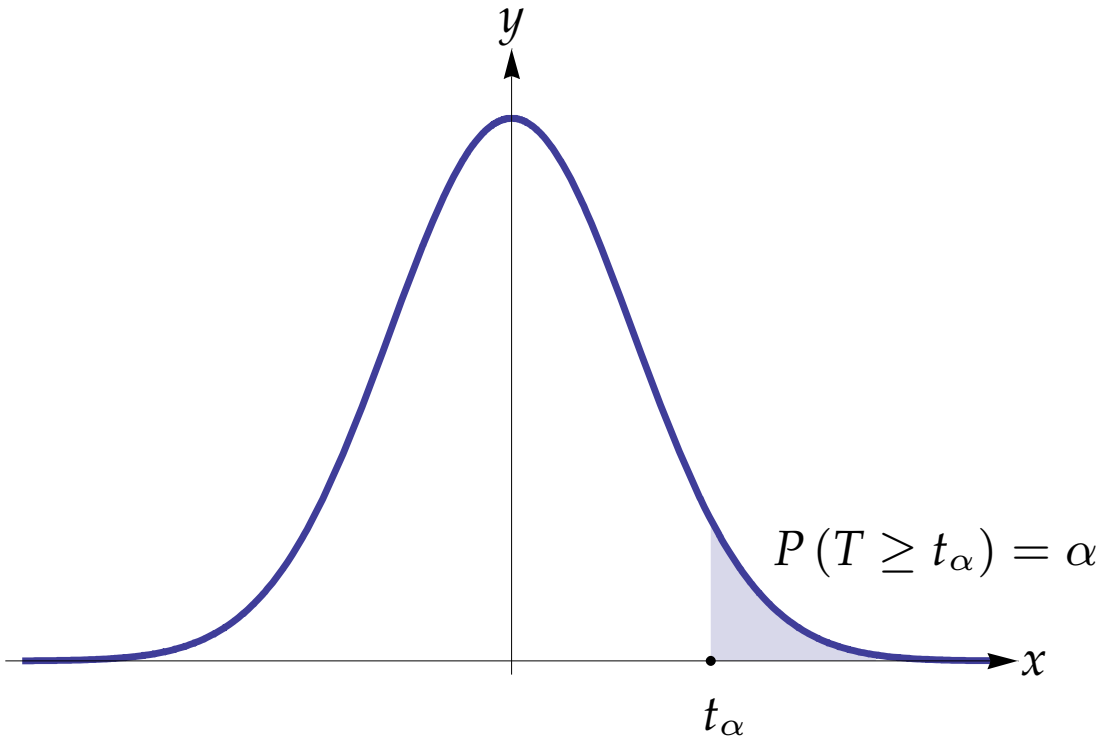
gdje su n_1 i n_2 veličine uzoraka, $\overline{x}_{n_1}$ i $\overline{x}_{n_2}$ su aritmetičke sredine uzoraka i

$$s_p^2 = \frac{(n_1 - 1)s_{n_1}^2 + (n_2 - 1)s_{n_2}^2}{n_1 + n_2 - 2}$$

- p-vrijednost - vjerojatnost da u uvjetima istinitosti $\mathcal{H}_0$ test-statistika bude jednaka ili ekstremnija od $\hat{t}$:
  - $p = P(T \geq \hat{t})$ ako je $\mathcal{H}_1 : \mu_1 > \mu_2$
  - $p = P(T \leq \hat{t})$ ako je $\mathcal{H}_1 : \mu_1 < \mu_2$

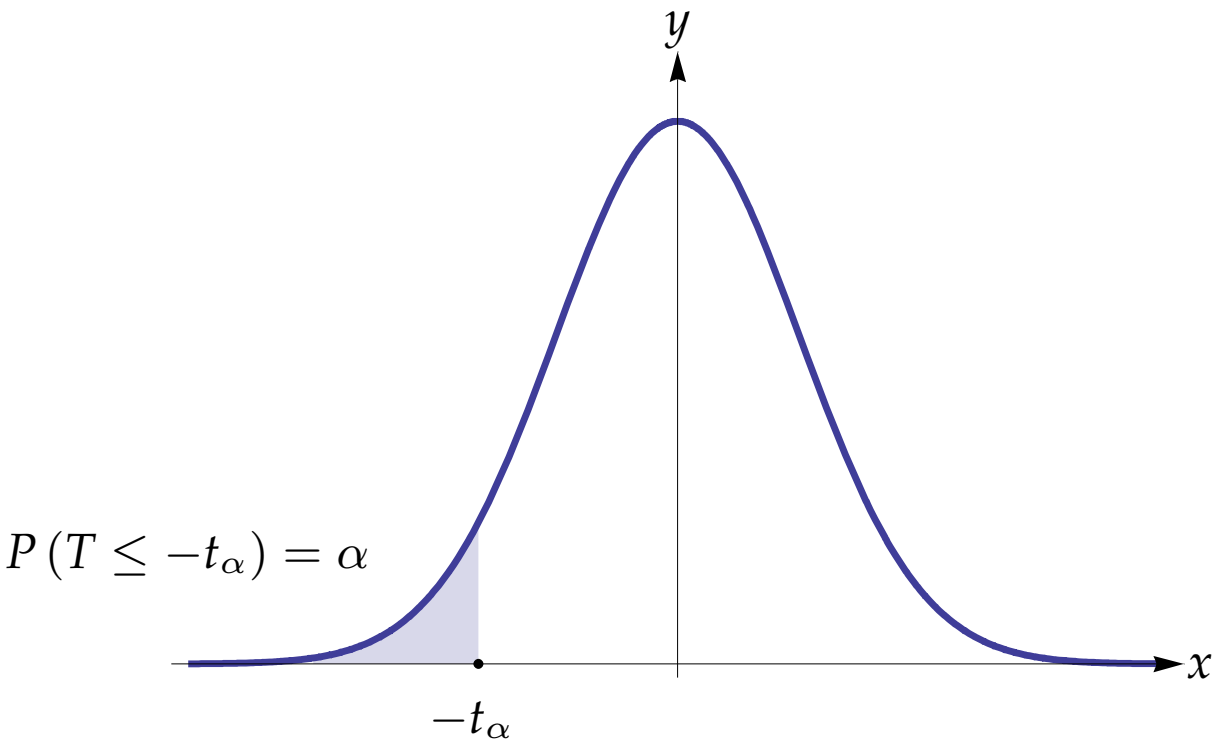
gdje test-statistika T ima Studentova $t(n_1 + n_2 - 2)$ distribuciju

USPOREDBA OČEKIVANJA - WELCHOV t-TEST



zaključivanje u slučaju alternativne hipoteze $\mathcal{H}_1 : \ \mu_1 > \mu_2$

USPOREDBA OČEKIVANJA - WELCHOV t-TEST



zaključivanje u slučaju alternativne hipoteze $\mathcal{H}_1 : \ \mu_1 < \mu_2$

USPOREDBA OČEKIVANJA - WELCHOV t-TEST

- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α:
  - ako je $p < \alpha$ odbacujemo $\mathcal{H}_0$ na razini značajnosti α
  - ako je $p > \alpha$ ne odbacujemo $\mathcal{H}_0$ na razini značajnosti α

- **u slučaju velikih uzoraka iz nepoznatih distribucija također možemo koristiti Welchov t-test**
- R - primjeri 1 i 2

USPOREDBA PROPORCIJA

- želimo testirati jednakost proporcija (ili vjerojatnosti događaja, tj. "uspjeha") u dvije nezavisne populacije
- $\left(x_1^{(1)}, \ldots, x_{n_1}^{(1)}\right)$ uzorak 0 (neuspjeha) i 1 (uspjeha) iz Bernoullijeve distribucije s parametrom p_1
- $\left(x_1^{(2)}, \ldots, x_{n_2}^{(2)}\right)$ uzorak 0 (neuspjeha) i 1 (uspjeha) iz Bernoullijeve distribucije s parametrom p_2
- hipoteze:

$$\mathcal{H}_0 : \ p_1 = p_2$$

$$\mathcal{H}_1 : \ p_1 > p_2 \quad \text{ili} \quad \mathcal{H}_1 : \ p_1 < p_2$$

USPOREDBA PROPORCIJA

- vrijednost test-statistike:

$$\hat{h} = \frac{\left(\overline{X}_{n_1} - \overline{Y}_{n_2}\right)^2}{\frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\left(1 - \frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  gdje je

$$\frac{n_1\overline{X}_{n_1} + n_2\overline{Y}_{n_2}}{n_1+n_2}$$

  relativna frekvencija "uspjeha" u oba uzorka zajedno
- p-vrijednost - vjerojatnost da u uvjetima istinitosti $\mathcal{H}_0$ test-statistika bude jednaka ili ekstremnija od $\hat{h}$
- tako izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i odluku donosimo na klasičan način
- R - primjer 3

USPOREDBA OČEKIVANJA

- želimo testirati jednakost očekivanja obilježja u istoj populaciji prije i poslije tretmana - mjerenja se promatraju u paru
- zaključci se donose na temelju razlika u vrijednostima varijabli izmjerenih na istom uzorku prije i poslije tretmana

ispitanik	prije tretmana	poslije tretmana	razlika
1	x_1	y_1	$d_1 = x_1 - y_1$
2	x_2	y_2	$d_2 = x_1 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	$d_n = x_1 - y_n$

USPOREDBA OČEKIVANJA

- $(x_1, \ldots, x_n)$ uzorak prije tretmana iz varijable s očekivanjem μ_1
- $(y_1, \ldots, y_n)$ uzorak poslije tretmana iz varijable s očekivanjem μ_2
- hipoteze:

$$\mathcal{H}_0 : \ \mu_1 = \mu_2$$

$$\mathcal{H}_1 : \ \mu_1 > \mu_2 \quad \text{ili} \quad \mathcal{H}_1 : \ \mu_1 < \mu_2$$

- testiranje hipoteze $\mathcal{H}_0$ može se svesti na testiranje ekvivalentne hipoteze

$$\mathcal{H}_0 : \mu_D = 0$$

koja se odnosi na očekivanje razlika

$$\mu_D = \mu_1 - \mu_2$$

pa se test svodi na T-test za testiranje hipoteze o očekivanju na jednom uzorku

- test se može koristiti za velike uzorke, a za male uzorke ako su razlike normalno distribuirane
- R - primjer 4

USPOREDBA PROPORCIJA

- $\left(x_1^{(1)}, \ldots, x_{n_1}^{(1)}\right)$ uzorak 0 (neuspjeha) i 1 (uspjeha) iz Bernoullijeve distribucije s parametrom p_1
- $\left(x_1^{(2)}, \ldots, x_{n_2}^{(2)}\right)$ uzorak 0 (neuspjeha) i 1 (uspjeha) iz Bernoullijeve distribucije s parametrom p_2
- ako promatramo uzorak razlika važni su samo parovi $(0,1)$ (daju razliku -1) i $(1,0)$ (daju razliku 1)
- razlike čine uzorak iz Bernoullijeve distribucije s parametrom p
- ekvivalentne hipoteze:

$$\mathcal{H}_0 : \ p = 1/2$$

$$\mathcal{H}_1 : \ p_1 > 1/2 \quad \text{ili} \quad \mathcal{H}_1 : \ p < 1/2$$

- test se svodi na binomi test za testiranje hipoteze o proporciji na jednom uzorku
- R - primjer 5

Distribucija dvodimenzionanog diskretnog slučajnog vektora

- jedna realizacija dvodimenzionalnog slučajnog vektora - uređeni par brojeva
- **diskretan slučajni vektor** - realizacija može biti samo konačno ili prebrojivo mnogo
- promatramo samo slučajne vektore s konačnim skupom vrijednosti $\mathcal{R}(X, Y)$
- distribucija slučajnog vektora - poznata ako znamo vjerojatnosti svih njegovih mogućih realizacija

$$P(X = x_i, Y = y_j), \text{ za sve } x_i \in \mathcal{R}(X), y_j \in \mathcal{R}(Y)$$

- $P(X = x_i, Y = y_j)$ - vjerojatnost da je *istovremeno* $X = x_i$ i $Y = y_j$
- te brojeve organiziramo u tablicu distribucije

Distribucija dvodimenzionanog diskretnog slučajnog vektora

- neka je (X, Y) slučajni vektor takav da je $\mathcal{R}(X) = \{x_1, \ldots, x_m\}$ i $\mathcal{R}(Y) = \{y_1, \ldots, y_n\}$
- **tablica distribucije** od (X, Y)

		Y			
		y_1	y_2	$\ldots$	y_n
X	x_1	$P(X = x_1, Y = y_1)$	$P(X = x_1, Y = y_2)$	$\ldots$	$P(X = x_1, Y = y_n)$
	x_2	$P(X = x_2, Y = y_1)$	$P(X = x_2, Y = y_2)$	$\ldots$	$P(X = x_2, Y = y_n)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	x_m	$P(X = x_m, Y = y_1)$	$P(X = x_m, Y = y_2)$	$\ldots$	$P(X = x_m, Y = y_n)$

Distribucija dvodimenzionanog diskretnog slučajnog vektora

- **marginalne distribucije** - distribucije od X i Y dobivaju se zbrajanjem vjerojatnosti u odgovarajućim redovima (za Y), odnosno stupcima (za X)

		Y			
		y_1	$\dots$	y_n	zbroj
X	x_1	$P(X=x_1, Y=y_1)$	$\dots$	$P(X=x_1, Y=y_n)$	$P(X=x_1)$
	x_2	$P(X=x_2, Y=y_1)$	$\dots$	$P(X=x_2, Y=y_n)$	$P(X=x_2)$
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_m	$P(X=x_m, Y=y_1)$	$\dots$	$P(X=x_m, Y=y_n)$	$P(X=x_m)$
	zbroj	$P(Y=y_1)$	$\dots$	$P(Y=y_n)$	1

PROCJENA TABLICE DISTRIBUCIJE

- **tablicu distribucije procjenjujemo** na temelju uzorka

$$(x_1, y_1), \ldots, (x_n, y_n)$$

- primjer - broj grešaka u proizvodnji na dvije proizvodne linije

sat	prva linija - broj grešaka	druga linija - broj grešaka
1	0	0
2	1	0
3	2	2
⋮	⋮	⋮
400	3	1

PROCJENA TABLICE DISTRIBUCIJE

- **zajednička tablica frekvencija** broja grešaka na prvoj i drugoj liniji

		druga linija - Y					
		0	1	2	3	4	zbroj
prva linija - X	0	22	12	13	12	7	66
	1	20	24	14	30	10	98
	2	15	20	30	10	7	82
	3	6	5	10	32	20	73
	4	5	7	13	31	25	81
zbroj		68	68	80	115	69	400

Procjena tablice distribucije

- **Empirijska distribucija para obilježja** (X, Y) - zajednička tablica relativnih frekvencija

		Y				
		0	1	2	3	4
	0	0.0550	0.0300	0.0325	0.0300	0.0175
	1	0.0500	0.0600	0.0350	0.0750	0.0250
X	2	0.0375	0.0500	0.0750	0.0250	0.0175
	3	0.0150	0.0125	0.0250	0.0800	0.0500
	4	0.0125	0.0175	0.0325	0.0775	0.0625

PROCJENE MARGINALNIH DISTRIBUCIJA

- **empirijske distribucije obilježja X i Y**

vrijednost od X	0	1	2	3	4
relativna frekvencija	0.165	0.245	0.205	0.1825	0.2025

vrijednost od Y	0	1	2	3	4
relativna frekvencija	0.17	0.17	0.2	0.2875	0.1725

UVJETNE DISTRIBUCIJE

- **uvjetna distribucija** - distribucija jednog obilježja uz uvjet fiksne vrijednosti drugog obilježja
- **uvjetna distribucija od X uz uvjet da se Y realizira s y_j** (oznaka $X|_{Y=y_j}$)

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \quad i = 1, \ldots, m$$

- **uvjetna distribucija od Y uz uvjet da se X realizira s x_i** (oznaka $Y|_{X=x_i}$)

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}, \quad j = 1, \ldots, n$$

UVJETNE DISTRIBUCIJE I NEZAVISNOST

- za obilježja X i Y kažemo da su **nezavisna** ako za sve $i = 1, \ldots, m$ i sve $j = 1, \ldots, n$ vrijedi da je

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

- u suprotnom kažemo da su X i Y **zavisna** obilježja
- ako su X i Y nezavisna obilježja, tada vrijedi:
  - distribucija od Y i uvjetna distribucija $Y|_{X=x_i}$ su jednake za sve $x_i \in \mathcal{R}(X)$
  - distribucija od X i uvjetna distribucija $X|_{Y=y_j}$ su jednake za sve $y_j \in \mathcal{R}(Y)$
  - tj. marginalne i odgovarajuće uvjetne distribucije slučajnog vektora (X, Y) su jednake

ANALIZA ZAVISNOSTI

- **procjene** - na temelju podataka možemo odrediti empirijsku distribuciju slučajnog vektora (X, Y), empirijske marginalne distribucije i empirijske uvjetne distribucije
- zavisnost slučajnih varijabli X i Y definirana je na temelju pravih, a ne empirijskih distribucija
- procjene odstupaju od stvarnih distribucija - **kako testirati nezavisnost?**
- χ^2 **test** - statistički test kojim možemo testirati hipotezu o nezavisnosti **diskretnih slučajnih varijabli (obilježja)**
- hipoteze:

$$\mathcal{H}_0 : \text{obilježja } X \text{ i } Y \text{ su nezavisna}$$
$$\mathcal{H}_1 : \text{obilježja } X \text{ i } Y \text{ su zavisna}$$

TESTIRANJE HIPOTEZE O NEZAVISNOSTI

- hipotezu $\mathcal{H}_0$ možemo zapisati na sljedeći način:

$$H_0 : P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

$$\text{za sve } i = 1, \ldots, m,\ j = 1, \ldots, n$$

- test-statistika χ^2-testa temelji se na usporedbi očekivanih frekvencija s empirijskim frekvencijama uzimajući u obzir sva polja tablice distribucije
- za velike uzorke, u uvjetima istinitosti nul-hipoteze, test-statistika ima približno $\chi^2\left((m-1)(n-1)\right)$ distribuciju
- usporedbom dobivene p-vrijednosti s razinom značajnosti α donosimo odluku:
  - ako je $p < \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da podaci potvrđuju postojanje zavisnosti između varijabli X i Y na razini značajnosti α
  - ako je $p > \alpha$, nemamo dovoljno argumenata za odbacivanje nul-hipoteze, tj. kažemo da podaci ne daju potvrdu o postojanju zavisnosti među varijablama X i Y

TESTIRANJE HIPOTEZE O NEZAVISNOSTI

- **zavisnost obilježja, tj. varijabli, ne znači da među njima postoji uzročna veza**
- za varijable koje nisu uzročno povezane može postojati neka treća varijabla koja je s objema u uzročnoj vezi
- R - primjeri 6 i 7

KOEFICIJENT KORELACIJE

- koeficijent korelacije - broj koji predstavlja mjeru jakosti linearne veze među varijablama X i Y:

$$\rho = \frac{E(X-\mu)(Y-\nu)}{\sigma_X \sigma_Y}$$

$$\mu = EX, \nu = EY, \sigma_X = \sqrt{Var(X)}, \sigma_Y = \sqrt{Var(Y)}$$

- važna svojstva koeficijenta korelacije:
  - $\rho \in [-1, 1]$
  - ako su X i Y nezavisne, tada je $\rho = 0$
  - $Y = aX + b, a > 0$, onda i samo onda ako je $\rho = 1$
  - $Y = aX + b, a < 0$, onda i samo onda ako je $\rho = -1$
  - ako je $\rho = 0$ varijable X i Y su **nekorelirane**
  - ako je $\rho \neq 0$ varijable X i Y su **korelirane**

Procjena koeficijenta korelacije

- **zavisnost varijabli X i Y** - potvrđena ako pokažemo da je njihov koeficijent korelacije različit od 0
- koeficijent korelacije 1 ili -1 - znamo da je veza između X i Y linearna
- **Pearsonov koeficijent korelacije** - procjena koeficijenta korelacije za neprekidne varijable
- ako su $(x_1, y_1), \ldots, (x_n, y_n)$ parovi nezavisnih realizacija slučajnog vektora (X, Y), onda se Pearsonov koeficijent korelacije računa na sljedeći način:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y}_n)^2}}$$

TESTIRANJE HIPOTEZE O NEKORELIRANOSTI

- statističke hipoteze:

$$\mathcal{H}_0 : \rho = 0$$

$$\mathcal{H}_1 : \rho \neq 0$$

- da bismo na temelju procjene koeficijenta korelacije r potvrdili zavisnost slučajnih varijabli na razini značajnosti α, potrebno je odbaciti hipotezu $\mathcal{H}_0$
- kod testiranja hipoteze o nekoreliranosti vrijednost test-statistike računamo po formuli

$$\hat{t} = \frac{\sqrt{n-2}\, r}{\sqrt{1-r^2}}$$

- ako je nul-hipoteza istinita, za normalno distribuirane varijable X i Y vrijednost test-statistike $\hat{t}$ je realizacija Studentove $t(n-2)$ distribucije

TESTIRANJE HIPOTEZE O NEKORELIRANOSTI

- p-vrijednost:

$$p = P(T \geq |\hat{t}|),$$

  gdje T ima Studentovu $t(n-2)$ distribuciju
- izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku:
  - ako je $p \leq \alpha$, odbacujemo nul-hipotezu i na razini značajnosti α prihvaćamo alternativnu hipotezu, tj. kažemo da su varijable X i Y korelirane, pa onda i zavisne
  - ako je $p > \alpha$, nemamo razloga odbaciti nul-hipotezu, tj. kažemo da nemamo dovoljno argumenata tvrditi da su X i Y korelirane

SPEARMANOV KOEFICIJENT KORELACIJE

- **Spearmanov koeficijent korelacije** ρ_S - daje informaciju o tome u kojoj se mjeri veza među varijablama X i Y može opisati monotonom funkcijom
- važna svojstva:
  - $\rho_S \in [-1, 1]$
  - ako $\rho_S = 0$ ne **postoji monotona veza** između X i Y
  - ako $\rho_S > 0$ veza između X i Y je **rastuća**
  - ako $\rho_S < 0$ veza između X i Y je **padajuća**

SPEARMANOV KOEFICIJENT KORELACIJE

procjena Spearmanovog koeficijenta korelacije za uzorak sparenih mjerenja $(x_1, y_1), \ldots, (x_n, y_n)$:

- sortiramo niz podataka $x_1, \ldots, x_n$ i niz podataka $y_1, \ldots, y_n$ svaki posebno u rastućem poretku
- odredimo r_{x_i} - redni broj podatka x_i u sortiranom nizu (ako u nizu ima jednakih podataka, svi jednakim podacima pridružuje se isti "redni broj"), tj. **rang podatka**
- odredimo r_{y_i} - rang podatka y_i u sortiranom nizu
- procjena za Spearmanov koeficijent računamo na sljedeći način:

$$r_S = 1 - \frac{6 \sum_{i=1}^{n} (r_{x_i} - r_{y_i})^2}{n(n^2 - 1)}$$

TESTIRANJE HIPOTEZE O NEPOSTOJANJU MONOTONE VEZE

- statističke hipoteze:

$$\mathcal{H}_0 : \rho_S = 0$$

$$\mathcal{H}_1 : \rho_S \neq 0$$

- da bismo na temelju procjene Spearmanovog koeficijenta korelacije r_S potvrdili postojanje monotone veze među varijablama na razini značajnosti α, potrebno je odbaciti hipotezu $\mathcal{H}_0$
- kod testiranja ove hipoteze vrijednost test-statistike računamo po formuli

$$\hat{t} = \frac{\sqrt{n-2}r_S}{\sqrt{1-r_S^2}}$$

TESTIRANJE HIPOTEZE O NEPOSTOJANJU MONOTONE VEZE

- ako je nul-hipoteza istinita, za velike uzorke vrijednost test-statistike $\hat{t}$ je realizacija Studentove $t(n-2)$ distribucije
- p-vrijednost:

$$p = P(T \geq |\hat{t}|),$$

  gdje T ima približno Studentovu $t(n-2)$ distribuciju
- izračunatu p-vrijednost uspoređujemo s razinom značajnosti α i donosimo odluku na standardan način
- R - primjeri 8 i 9

LITERATURA

- Benšić, M. i Šuvak, N., *Primijenjena statistika*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2013.
- Benšić, M. i Šuvak, N., *Uvod u vjerojatnost i statistiku*, Odjel za matematiku, Sveučilište J.J. Strossmayera, Osijek, 2014.