

3. del

1. Vzorcenje in cenilke

- Definicija enostavnega slučajnega vzorca
 - Naj bo X slučajna spremenljivka. **Enostavni slučajni vzorec** je slučajni vektor (X_1, X_2, \dots, X_n) z vrednostmi meritev (x_1, \dots, x_n) (n = velikost vzorca) za katerega velja:
 - vsi členi vektorja X_i imajo **enako porazdelitev** kot spremenljivka X
 - členi X_i so med seboj **neodvisni**

- Vzorcna statistika
 - poljubna simetrična funkcija vzorca** (njena vrednost je neodvisna od permutacij argumentov)

$$Y = g(X_1, X_2, \dots, X_n)$$

- Vzorcna statistika je **slučajna spremenljivka**. Značilni vrednosti:
 - pricakovana vrednost $E(Y)$, za katero uporabljamo vzorcno povprečje
 - standardni odklon σ_Y (pravimo tudi standardna napaka statistike $SE(Y)$), za katerega upoštevamo vzorčni odklon
- Vzorcne sredinske mere (modus, mediana, povprečje)
 - vzorcni modus** je najpogostejša vrednost
 - vzorcna mediana** je srednja vrednost glede na urejenost

$$M_e = \begin{cases} Y_{n+1}/2 & , n - \text{liho} \\ \frac{Y_{n/2} + Y_{n/2+1}}{2} & , n - \text{sodo} \end{cases}$$

- vzorcno povprečje** je povprečna vrednost

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

- Vzorcne mere razpršenosti (razmik, varianca, standardni odklon)
 - vzorcni razmah/razmik**: $\max_i x_i - \min_i x_i$
 - Vzorcna disperzija**: $s_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
 - Popravljen vzorcna disperzija** $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ uporabimo jo ko je vzorec majhen
 - vzorcna odklona**: s_0 in s
- Vzorcne mere oblike porazdelitve (koeficienta asimetrije in sploscenosti)
 - koeficient asimetrije** (s centralnimi momenti): $g_1 = \frac{m_3}{m_2^{3/2}}$
 - koeficient sploscenosti** (s centralnimi momenti): $K = g_2 = \frac{m_4}{m_2^2} - 3$
 - $K = 0$ ~ normalna porazdelitev zvonaste oblike
 - $K < 0$ ~ bolj kopasta kot normalna porazdelitev, s krajšimi repi

- $$\begin{aligned} &|| \text{Pricakovana vrednost } \mu \mid \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \mid \overline{x} = \frac{1}{n} \sum_{i=1}^n x_i || \text{Standardni odklon } \sigma \mid \\ S &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X})^2} \mid s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2} \mid \text{Verjetnost } p \mid \overline{X} = \frac{1}{n} \sum_{i=1}^n X_i \mid \\ \overline{x} &= \frac{1}{n} \sum_{i=1}^n x_i \mid \end{aligned}$$

- CLI
 - Naj bo (X_1, X_2, \dots, X_n) normlano porazdeljen enostavni slučajni vzorec.
 - $X_i \sim N(\mu, \sigma)$

$$\sum_{i=1}^n X_i \sim N(n\mu, \sigma\sqrt{n})$$

- teorija

- $$E(\bar{X}) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} \sum E(X_i) = \frac{\sum E(X_i)}{n} = \frac{\sum \mu}{n} = \mu$$

$$\text{SE}(\bar{X}) = D(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^n D(Xi) = \frac{D(x)}{n} = \frac{\sigma}{\sqrt{n}}$$

- o **Kaj pa ce porazdelitev X ni normalna?** Pri vecjih vzorcih ($n>30$), lahko uporabimo centralni limitni izrek, ki zagotavlja, da je spremenljivka \bar{X} porazdeljena standardizirano normalno. Vzorcno povprecie ima tedaj porazdelitev priblizno

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- primer
 - Kolikšna je verjetnost, da bo pri 36 metih igralne kocke povprečno število pik večje ali enako 4? X je slučajna spremenljivka z vrednostmi 1,2,3,4,5,6 in verjetnostmi $1/6$. Zanj je $\mu = 3.5$ in $\sigma = 1.7$. Vseh 36 ponovitev meta lahko obravnavamo kot slučajni vzorec velikosti 36.

$$P(\bar{X} \geq 4) = 1 - \phi\left(\frac{E(\bar{X}) - \mu}{\frac{\sigma}{\sqrt{n}}}\right) = 1 - \phi\left(\frac{4 - 3.5}{\frac{1.7}{6}}\right) \approx 0.04$$

3. CLI za delez

- izrek
 - Denimo, da **zelimo na populaciji oceniti delez enot π z določeno lastnostjo**. V ta namen poiscemo vzorčni delez p . Pokazati se da, da se za dovolj velike slučajne vzorce s ponavljanjem (za deleze okoli 0.5 je dovolj 20 enot ali več), vzorčni delezi porazdeljujejo približno normalno s

$$E(\hat{P}) = \pi$$

$$SE(\hat{P}) = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

$$\hat{P} \sim N\left(\pi, \sqrt{\frac{\pi(1 - \pi)}{n}}\right)$$

- Za manjše vzorce ($n < 20$) se vzorčni delez porazdeljuje binomsko. Mimogrede, cenilka populacijskega deleza je nepristranska ker velja $E(\hat{P}) = \pi$

$$\hat{P} \sim B\left(\pi, \sqrt{\frac{\pi(1 - \pi)}{n}}\right)$$

- cenilka za delez π

$$\hat{P} = \frac{\sum X_i}{n} = \bar{X}$$

- primer
 - V izbrani populaciji je polovica zensk $\pi = 0.5$. Če tvorimo vzorce po $n = 25$ enot, nas zanima, kolikšna je verjetnost, da je v vzorcu več kot 55% zensk? To pomeni da iscemo verjetnost $P(p > 0.55)$. Uporabimo dejstvo da se vzorčni delezi p porazdeljujejo približno normalno

$$\hat{P} \approx N\left(0.5, \sqrt{\frac{0.5 \cdot 0.5}{25}}\right) = N(0.5, 0.1)$$

- Zato je

$$P(\hat{P} > 0.55) = 1 - \phi\left(\frac{0.55 - 0.5}{0.1}\right) \approx 0.31$$

- Torej pri približno 31% vzorcih zensk bo delež zensk večji od 55%.

4. CLI za S^2

- Naj bo slučajna spremenljivka X na neki populaciji porazdeljena normalno $N(\mu, \sigma)$. Kako bi določili porazdelitev za vzorčno disperzijo ali popravljeno vzorčno disperzijo tj.:

$$S_0^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \text{ oziroma } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Dobimo ju iz vzorčne statistike χ^2 :

$$\chi^2 = \frac{nS_0^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2$$

- Ker vemo, da je $E(\chi^2(n)) = n$ in $D(\chi^2(n)) = 2n$ lahko takoj izračunamo:

$$\circ E(S_0^2) = E\left(\frac{\sigma^2 \chi^2}{n}\right) = \frac{(n-1)\sigma^2}{n}, E(S^2) = E\left(\frac{\sigma^2 \chi^2}{n-1}\right) = \sigma^2$$

$$\circ D(S_0^2) = D\left(\frac{\sigma^2 \chi^2}{n}\right) = \frac{2(n-1)\sigma^4}{n^2}, D(S^2) = D\left(\frac{\sigma^2 \chi^2}{n-1}\right) = \frac{2\sigma^4}{n-1}$$

- Torej za dovolj velike n je:

$$\chi^2 \approx N(n-1, \sqrt{2(n-1)})$$

$$S_0^2 \approx N\left(\frac{(n-1)\sigma^2}{n}, \frac{\sigma^2 \sqrt{2(n-1)}}{n}\right)$$

$$S^2 \approx N\left(\sigma^2, \sigma^2 \sqrt{\frac{2}{n-1}}\right)$$

5. CLI za razliko vzorčnih povprečij

- definicija

- Denimo da imamo dve populaciji velikosti N_1 in N_2 in se spremenljivka X na prvi populaciji porazdeljuje normalno $N(\mu_1, \sigma)$ na drugi populaciji pa $N(\mu_2, \sigma)$ (standardna odklona sta na obeh populacijah enaka). V vsaki od obeh populacij tvorimo neodvisno slučajne vzorce velikosti n_1 in n_2 . Na vsakem vzorcu (s ponavljanjem) prve populacije izračunamo vzorčno povprečje \bar{X}_1 in podobno na vsakem vzorcu druge populacije \bar{X}_2 . Po reprodukcijski lastnosti normalne porazdelitve **je porazdelitev velikih vzorčnih povprečij normalna** kjer je:

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$D(\bar{X}_1 - \bar{X}_2) = D(\bar{X}_1) + D(\bar{X}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}$$

$$\overline{X}_1 - \overline{X}_2 \approx N \left(\mu_1 - \mu_2, \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}} \right)$$

- Primer

- Populacijama studentov na neki univerzi (tehtnikom in družboslovcem) so izmerili neko sposobnost s pričakovanimi vrednostima $\mu_t = 70$ ter $\mu_d = 80$ točk in standardnim odklonom, ki je na obeh populacijah enak $\sigma = 7$ točk.
- Kolikšna je verjetnost, da je pri naključnih vzorcih vzorčno povprečje družboslovcev ($n_d = 36$) večje za več kot 12 točk od vzorčnega povprečja tehnikov ($n_t = 64$)? Zanima nas torej verjetnost:

$$P(\overline{X}_d - \overline{X}_t > 12) = 1 - \phi \left(\frac{12 - 10}{7 \cdot \sqrt{\frac{36+64}{36 \cdot 64}}} \right) = 1 - \phi(1.37) = 0.085$$

6. CLI za razliko deležev

- definicija

- Podobno kot pri porazdelitvi razlik vzorčnih povprečij naj bosta dani dve populaciji velikosti N_1 in N_2 z deležema enot z neko lastnostjo π_1 in π_2 . Iz prve populacije tvorimo slučajne vzorce velikosti n_1 in na vsakem izračunamo delež enot s to lastnostjo p_1 . Podobno naredimo tudi na drugi populaciji; tvorimo slučajne vzorce velikosti n_2 in na njih določimo deleže p_2 .
- Pokazati se da, **da se za dovolj velike vzorce razlike vzorčnih deležev porazdeljujejo približno normalno s**

$$E(\hat{P}_1 - \hat{P}_2) = E(\hat{P}_1) - E(\hat{P}_2) = \pi_1 - \pi_2$$

$$D(\hat{P}_1 - \hat{P}_2) = D(\hat{P}_1) + D(\hat{P}_2) = \frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}$$

$$\hat{P}_1 - \hat{P}_2 \approx N \left(\pi_1 - \pi_2, \sqrt{\frac{\pi_1(1 - \pi_1)}{n_1} + \frac{\pi_2(1 - \pi_2)}{n_2}} \right)$$

7. CLI za kvocient S_1^2 / S_2^2

- uvod

- Želimo primerjati varianci teže prebivalcev dveh različnih populacij. Naj bo X teža odraslih moskih iz prve populacije, ter Y teža odraslih moskih iz druge populacije. Nimamo možnosti da izmerimo tezo celotne prve in druge populacije, zato bomo izbrali enostavni slučajni vzorec iz vsake od populacij in jim izmerimo tezo. Slučajni spremenljivki X in Y sta neodvisni. Obe sta porazdeljeni normalno $X \sim N(\mu_1, \sigma_1)$, $Y \sim N(\mu_2, \sigma_2)$. Na njih tvorimo slučajne vzorce (X_1, \dots, X_n) in (Y_1, \dots, Y_m) ter izračunamo vzorčni povprečji in popravljeni vzorčni varianci za obe spremenljivke. Nastavimo vzorčno statistiko za kvocient obe popravljeni vzorčni varianci:

$$F = \frac{s_1^2}{s_2^2}$$

- fisherjeva porazdelitev

- porazdelitev $F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$ je **fisherjeva** $F(n-1, m-1)$

- kjer sta parametra prostostne stopnje obeh vzorcev porazdeljeni po χ^2 z m-1 oziroma n-1 prostostnimi stopnjami in sta tudi neodvisni.
- $\chi^2(m-1) = (m-1)s_X^2/\sigma_X^2$

- parameter μ te normalne porazdelitve

- $\mu = \frac{d_2}{d_2-2}$

- parameter σ te normalne porazdelitve

- $\sigma = \frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$

- uporaba

- Uporabljamo pri intervalu zaupanja in statističnemu testu za kvocient populacijskih varianc $\frac{\sigma_1^2}{\sigma_2^2}$

- primer

Zelimo prejeti varianci teze prebivalcev dveh mest. X je teza odraslih moskih enega mesta in Y je teza odraslih moskih drugega mesta. Ker ne moramo izmeriti teze celotni populaciji si izberemo slucajni vzorec iz vsake od populacij in izmerimo njuno tezo. Slucai ni spremenljivki sta neodvisni, obe porazdeljeni normalno. Na njih tvorimo slucajane vzorce (X_1, \dots, X_m) in (Y_1, \dots, Y_m) ter izracunamo vzorčni povpreci in popravljeni vzorčni varianci za obe spremenljivki. Kvoceiwnt varianc ocenujemo s kvocientom popravljenih vzorčnih varianc.

8. Intervali zaupanja

- interval zaupanja σ je znan
 - Za konstrukcijo intervala zaupanja uporabljamo dejstvo

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

- Slucai na spremenljivka X je normlano porazdeljena ali imamo dovolj veliki vzorec (za uporabo CLI)
- Z verjetnostjo $1 - \alpha$ se μ nahaja na intervalu $[\bar{X} - \epsilon, \bar{X} + \epsilon]$
- Dobi se $\epsilon = c \frac{\sigma}{\sqrt{n}}$
- $I_\mu = \left[\bar{X} - c \frac{\sigma}{\sqrt{n}}, \bar{X} + c \frac{\sigma}{\sqrt{n}} \right]$
- $c = F^{-1}\left(1 - \frac{\alpha}{2}\right)$
- $1 - \alpha$ je **stopnja zaupanja**, α je **stopnja tveganja**
- **Sirina (dolžina)** intervala zaupanja je $l = 2c \frac{\sigma}{\sqrt{n}}$
- **Primer:** Signal intenzitete μ je poslan z lokacije A. Na lokaciji B se belezi sprejet signal. Zaradi sumenja signal zaznamo z naključno napako. Intenziteta signala na lokaciji B je normalno porazdeljena slucajna

spremenljivka s povprečjem μ in standardnim odklonom 3. Da bi zmanjšali napako, isti signal neodvisno beležimo 10-krat. Dobili smo naslednje vrednosti intenzitete signala na lokaciji:

$$B : 17, 21, 20, 18, 19, 22, 20, 21, 16, 19$$

- Doloci 95% interval zaupanja za μ
 - $n = 10, \sigma = 3, \alpha = 0.05, c = F^{-1}(1 - \frac{\alpha}{2}), \bar{x} = 19.3$
 - c pogledamo v tabeli za $c = F^{-1}(0.975) = 1.9$
 - Interval zaupanja $I_\mu = [17.5, 21.1]$
- interval zaupanja σ ni znan
 - Za konstrukcijo intervala zaupanja uporabljamo dejstvo

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

- Interval zaupanja za μ s stopnjo zaupanja $1 - \alpha$ je enak

$$I_\mu = \left[\bar{X} - c \frac{S}{\sqrt{n}}, \bar{X} + c \frac{S}{\sqrt{n}} \right]$$

- kjer je $c = t_{n-1; 1-\frac{\alpha}{2}}$ kvantil **Studentove porazdelitve** z $n - 1$ prostnostnimi stopnjami (stevilo vzorca)
- **Primer:** Na vzorcu 30 zensk so dobili povprečje 6 in popravljeni standardni odklon 5 količine PCB-jev. Doloci 99% interval zaupanja za povprečno količino PCB-jev.
 - $\bar{x} = 6, s = 5, \alpha = 0.01, n = 30, c = t_{29; 0.995}$
 - $I_\mu = \left[\bar{x} - c \frac{s}{\sqrt{n}}, \bar{x} + c \frac{s}{\sqrt{n}} \right] = [3.5, 8.5]$
- interval zaupanja za delez p
 - p je delez populacije z neko lastnostjo
 - naj bo (X_1, X_2, \dots, X_n) enostavni slučajni vzorec, kjer je $X_i \sim \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, $i = 1, \dots, n$
 - neznani delez p ocenjujemo z vzorcnim delezom $\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
 - za konstrukcijo intervala uporabimo dejstvo $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$
 - Interval zaupanja $I_p = \left[\hat{p} - c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$
 - kjer je $c = F^{-1}(1 - \frac{\alpha}{2})$ kvantil standardne normalne porazdelitve
- definicija tockovne ocene za parameter, primeri (vsaj 3)
 - Za konkreten vzorec naredimo **numericno oceno neznage parametara oz. "vrednost v točki"**
 - **Vzorcno povprečje, vzorcna varianca, vzorcni delez**
- kdaj uporabimo Studentovo porazdelitev
 - ko nepoznamo standardnega odklona populacije σ ter imamo dokaj majhen vzorec $n < 30$
- kaj je drugace, ko imamo majhen vzorec
 - nemoremo uporabiti izreka za CLI

- izpeljava formule za interval zaupanja

9. Preverjanje domnev

- uvod
 - Statisticna domneva** je vsaka domneva o porazdelitvi slučajne spremenljivke X na populaciji
 - Ce poznamo vrsto(obliko) porazdelitve in razkisujemo domnevo o parametru θ govorimo o **parametricni domnevi**
 - Ce pa je vpraskljiva vrsta porazdelitve govorimo o **neparametricni domnevi**
 - Domneva je:
 - enostavna**: ce natancno doloca porazdelitev (vrsto in točno vrednost parametra)
 - sestavljena**: sicer
 - Primer**: Naj bo $X \sim N(\mu, \sigma)$, ce poznamo σ je domneva $H : \mu = 0$ enostavna; ce pa parametra σ nepoznamo pa je sestavljena
 - primer sestavljene je tudi $H : \mu > 0$
 - Domneva je lahko:
 - pravilna** (podatki domnevo podpirajo)
 - napacna** (podatki prevec odstobajo od domneve)
- Nicelna in alternativna domneva
 - Nicelna domneva** (H_0)
 - je trditev o lastnosti populacije za katero predpostavimo da drzi (verjamemo da je resnicna)
 - je trditev ki jo test zeli ovreci
 - Alternativna (nasprotna) domneva** H_a ali H_1
 - trditev, ki ni zdruzljiva z nicelno domnevo
 - trditev, ki jo s testiranjem skusamo dokazati
 - Primer**: Ameriski sodni sistem
 - H_0 : obtozenec je nedolzen (nicelna domneva)
 - H_a : obtozenec je kriv (alternativna domneva)

Elementi preverjanja domneve



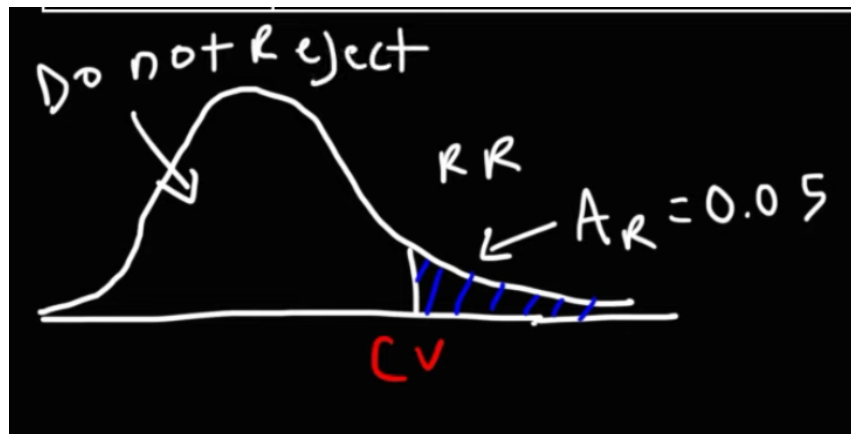
		odločitev	
		nedolžen	kriv
dejansko stanje	nedolžen	pravilna odločitev	napaka 1. vrste (α)
	kriv	napaka 2. vrste (β)	moč testa ($1 - \beta$)

- Napaka 1. vrste, 2. vrste
 - Napaka 1. vrste** je zavrnitev nicelne domneve, ce je le ta pravilna. Verjetnost da naredimo napako 1. vrste
 - Napaka 2. vrste** je ko ne zavrnemo nicelne domneve v primeru da je ta napacna. Verjetnost te napake je β
- P-vrednost
 - P-vrednost** oziroma **stopnja znacilnosti/signifikantnosti testa** je največja vrednost parametra α ki jo je vodja eksperimenta pripravljen sprejeti glede na dan vzorec.

- Moc statističnega testa
 - **Moc statističnega testa** ($1-\beta$) je verjetnost zavrnitve ničelne domneve v primeru, ko je ta v resnici napcna.
 - Preverjanje z P-testom
 - Preverjanje z Hi-kvadrat testom
- Ravnatelj bi rad izvedel odsotnost studentov na posamezen dan. Naredi vzorec z 100 naključnimi profesorji in jih vprasa katere dni so studenti največ manjkali. Rezultate je zbral v tabelo. **Ali se dnevi in odsotnosti povezani z enako povprečno frekvenco?** Uporabi $\alpha = 0.05$ (stopnja tveganja)

	Ponedeljek	Torek	Sreda	Četrtek	Petek
Izmerjene frekvence	23	16	14	19	28
Priackovane frekvence	20	20	20	20	20

- H_0 : enake frekvence (neodvisno od dneva)
- H_a : neneakovredne frekvence
- χ^2 je nesimetrična porazdelitev

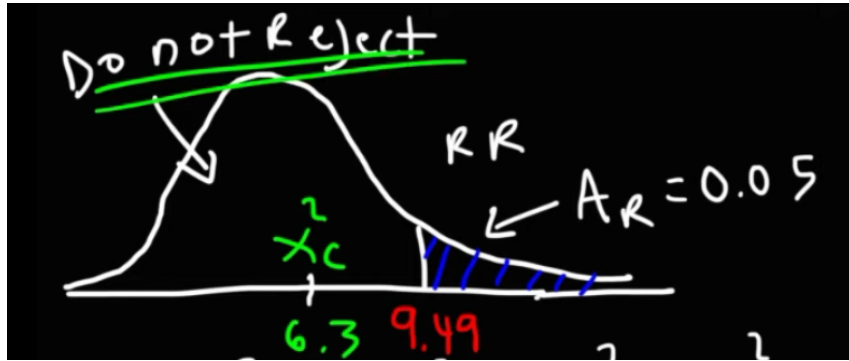


- Uporabimo tabelo za χ^2
 - stopnje prostosti = $n-1 = 4$, $\alpha = 0.05$

Area to the Right of the Critical Value									
df	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.320	2.710	3.840	6.630
2	0.020	0.103	0.211	0.575	1.386	2.770	4.610	5.990	9.210
3	0.115	0.352	0.584	1.212	2.366	4.110	6.250	7.810	11.34
4	0.297	0.711	1.064	1.923	3.357	5.390	7.780	9.490	13.28
5	0.554	1.145	1.610	2.675	4.351	6.630	9.240	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.840	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.040	12.02	14.07	18.48
8	1.646	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.954	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

- Dobimo kritično vrednost $c = 9.49$

- izračunamo se $\chi_c^2 = \sum \frac{(x_i - \bar{X})^2}{\bar{X}} = \frac{3^2}{20} + \frac{(-4)^2}{20} + \frac{(-6)^2}{20} + \frac{(-1)^2}{20} + \frac{8^2}{20} = 6.3$



- sprejmemo domnevo (relativno enake frekvence)

10. Bivariatna analiza in regresija

- Gledamo odvisnost oziroma povezanost spremenljivk
 - $X \leftrightarrow Y$ povezanost
 - $X \rightarrow Y$ odvisnost
- Tipi spremenljivk in testi za povezanost
 - **Imenski/nominalni**: χ^2 , kontingencni koeficienti, koeficient asociacije
 - **Ordinalni**: koeficient korelacije rangov
 - **Stevilski**: koeficient korelacije
- Povezanost dveh imenskih spremenljivk
 - teorija
 - Za preverjanje domneve o povezanosti med dvema **imenskima** spremenljivkama lahko uporabimo χ^2 test.
 - Testna statistika χ^2 , ki primerja dejanske in teoretične frekvence

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - f'_i)^2}{f'_i}$$

- k je število razredov (celic) v kontingencni tabeli
- testna statistika se porazdeljuje po χ^2 porazdelitvi s $(s - 1)(v - 1)$ prostostnimi stopnjami, kjer je s število vrstic v kontingencni tabeli in s število stolpcev. Nična in osnovna domneva sta v primeru tega testa:
 - $H_0 : \chi^2 = 0$ (spremenljivki nista povezani)
 - $H_1 : \chi^2 > 0$ (spremenljivki sta povezani)
- primer
 - Zanima nas ali sta spol in stanovanje v času studija povezana. Izmerimo podatke za naključni vzorec.

	starsi	st. dom	zasebno	skupaj
moski	16	40	24	80
zensek	48	36	36	120
skupaj	64	76	60	200

- Naredimo **kontingencno tabelo** (relativne frekvence po stolpcih)

	starsi	st. dom	zasebno	skupaj
moski	20	50	30	100
zenske	40	30	30	100
skupaj	32	38	30	100

- npr koliko moskih zivi pristarsih

$$P(M) = \frac{80}{200}, P(S) = \frac{64}{200}, P(MS) = P(M)P(S) = \frac{80}{200} \cdot \frac{64}{200} = 0.128$$

$$f'(MS) = n \cdot P(MS) = 200 \cdot \frac{80}{200} \cdot \frac{64}{200} = 25.6$$

- podobno izracunamo se ostale teoreticne frekvence

$$\chi^2_{1-\alpha}[(s-1)(v-1)] = \chi^2_{0.95}(2) = 5.99$$

$$\chi^2 = \frac{(16-26)^2}{26} + \frac{(40-30)^2}{30} + \dots = 12$$

- Izracunana vrednost je vecja od kriticne, zavrzemo osnovno domnevo
- povezanost dveh ordinalnih spremenljivk
 - Merimo s **koeficientom korelacije rangov** r_s (Spearman)

$$r_s := 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n \cdot (n^2 - 1)}$$

- d_i je razlika med **rangoma** v i-ti enoti
 - povezanost dveh številskih spremenljivk
- Uporabimo (**Pearsonov**) **koeficient korelacije**

$$r_{X,Y} = \frac{k(X,Y)}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Statisticno sklepanje: - $H_0: r = 0$ (spremenljivki nista linearno povezani) - $H_1: r \neq 0$ (spremenljivki sta linearno povezani)

Izkaze se da se **testna statistika**:

$$\frac{R_{X,Y} \cdot \sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} \sim t_{n-2}$$

Primer :

Preverimo domnevo, da sta izobrazba (stevilo priznanih let sole) in stevilo ur branja dnevnih časopisov na teden povezana med seboj na 5% stopnji značilnosti. Najprej izračunajmo vzorčni koeficient korelacije $r = r_{X,Y}$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
10	3	2	0	4	0	0
8	4	0	1	0	1	0
16	7	8	4	64	16	32
8	3	0	0	0	0	0
6	1	-2	-2	4	4	4
4	2	-4	-1	16	1	4
8	3	0	0	0	0	0
4	1	-4	-2	16	4	8
64	24	0	0	104	26	48

$$r = \frac{48}{\sqrt{104 \times 26}} = 0.92.$$

Vrednost testne statistike je:

$$\frac{0.92\sqrt{8-2}}{\sqrt{1-0.92^2}} = 2.66.$$

Kritično območje je določeno z kritičnima vrednostima $\pm t_{\alpha/2}(n-2) = \pm t_{0.025}(6) = \pm 2.447$

- linearna regresija (definicije, predpostavke, metoda najmanjših kvadratov)
 - Regresija prikazuje kakšen vpliv ima spremenljivka X na Y , če razen spremenljivke X ni drugih vplivov na Y
 - Slučajno spremenljivko Y zapisemo kot $Y_i = E(Y_i) + e_i = a + bx_i + e_i$, kjer je $1 \leq i \leq \text{todo}$
 $y = a + bx$ je enačba regresijske premice, kjer je a neznan odsek, b neznan naklon premice in e_i naključna napaka odvisna od X
 - Metodo najmanjših kvadratov** uporabimo ko želimo razdalje točk do regresijske premice čim bolj zmanjšati.
- časovne vrste in definicija trenda
 - Casovna vrsta je niz istovrstnih podatkov, ki se nanašajo na zaporedne časovne razmike ali trenutke
 - osnovni namen analize časovnih vrst:
 - opazovati časovni razvoj pojavitev
 - iskati njihove zakonitosti
 - predvideti nadaljni razvoj
 - Casovne vrste analiziramo tako, da opazujemo spreminjanje vrednosti členov v časovnih vrstah in iscemo zakonitosti tega spreminjanja
 - Naloga enostavne analize časovnih vrst je primerjava med deli v isti časovni vrsti
 - Z metodami, ki so specializirane za analizo časovnih vrst, analiziramo zakonitosti dinamike ene same vrste, s korelacijsko analizo pa zakonitosti odvisnosti v dinamiki več pojavov, ki so med seboj v zvezi
 - Trendi ali dolgoročno gibanje - X_T podaja dolgoročno smer razvoja. Običajno ga je mogoče izraziti s preprostimi rahlo ukrivljenimi krivuljami.
- Staticni test linearnosti modela
 - Validnost linearnega regresijskega modela lahko preverimo s tem, da narisemo graf ostankov v odvisnosti od X vrednosti. Ali pa od predvidenih vrednosti $\hat{y} = a\hat{x} + \hat{b}$ in preverimo obstoj kaksnega vzorca.
 - Ce so točke enakomerno raztresene nad in pod premico in ne vidimo nobene oblike, je linearni model validen. Če pa na grafu opazimo nekaksen vzorec, nam oblika vzorca daje informacijo, da v modelu manjka neka funkcija X .