

Statistika 1 - elementarni kurs

Odabrana poglavlja iz matematike

školska 2024/25

Literatura

- 1 Chihara LM, Hesterberg TC, Mathematical Statistics with Resampling and R. John Wiley & Sons Hoboken NJ. (2018) ISBN 978-1-119-41653-1
2. Ghilezan et. al., Zbirka rešenih zadataka iz Verovatnoće i statistike, CMS, NS (2009)
3. Annette J. Dobson, Adrian G. Barnett, An Introduction to Generalized Linear Models, 4th Edition, Chapman and Hall/CRC (2018)
4. Ivetić J. Folije Praktikuma iz statistike predavanja i vežbe

Statistika, osnovni pojmovi

Populacija je skup svih elemenata koje ispitujeemo.

Obeležje je numerička karakteristika elementa. Modeliramo ga slučajnom promenljivom.

Uzorak je odabrani deo populacije na kojem ispitujeemo realizovanu vrednost obeležja X .

Prost slučajni uzorak je n -dimenzionalna slučajna promenljiva čije komponente su nezavisne i imaju raspodelu posmatranog obeležja (X_1, X_2, \dots, X_n) .

Uzoračka funkcija raspodele $F(x_1, x_2, \dots, x_n) = F(x_1)F(x_2) \cdots F(x_n)$

Realizovane vrednosti slučajnih promenljivih obeležavamo malim slovima $X_i \rightarrow x_i$.

Realizovana vrednost prostog slučajnog uzorka $(X_1, X_2, \dots, X_n) \rightarrow (x_1, x_2, \dots, x_n)$

Statistika je funkcija uzorka, $Y = h(X_1, X_2, \dots, X_n)$.

Realizovana vrednost statistike je $y = h(x_1, x_2, \dots, x_n)$.

Kao transformacija slučajnih promenljivih, **statistika je slučajna promenljiva**.

Raspodela statistike obeležja koje modeliramo se koristi za statističko zaključivanje.

Deskriptivna statistika

Važne statistike uzorka (X_1, X_2, \dots, X_n) **obeležja** X

Aritmetička sredina uzorka

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$$

Uzoračka disperzija (varijansa)

$$\bar{S}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2,$$

Korigovana varijansa

$$\bar{S}_n^{2'} = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2, E(\bar{S}_n^{2'}) = D(X), \bar{S}_n' = \sqrt{\bar{S}_n^{2'}}$$

PRIMER 1 *Izračunati aritmetičku sredinu i uzoračku varijansu za $(14, 8, 8, 11, 6, 8, 4, 3, 12, 14)$.*

$$\bar{x}_n = (14 + 8 + 8 + 11 + 6 + 8 + 4 + 3 + 12 + 14) / 10 = 8.8$$

$$\bar{s}_n^2 = ((14 - 8.8)^2 + (8 - 8.8)^2 + \dots + (3 - 8.8)^2 + (12 - 8.8)^2 + (14 - 8.8)^2) / 10 = 13.56$$

Intervalni uzorak

Intervalni uzorak nastaje grupisanjem elemenata početnog uzorka u intervale I_i .

Ako imamo granice intervala I_i , odnosno deobene tačke m_i , $i = 0, 1, \dots, k$ i broj elemenata uzorka u intervalu i : **frekvencije** f_i , $i = 1, 2, \dots, k$, kažemo da je to **intervalni uzorak**.

Delimična rekonstrukcija početnog uzorka sredinama intervala: smatramo da imamo f_i komada elemenata jednakih $x_i = (m_i + m_{i-1})/2$, sredini i -tog intervala.

Ponekad se anketiranjem podaci prikupljaju u intervalni uzorak.

Formule za računanje aritmetičke sredine i varijanse intervalnog uzorka sa sredinama x_i , $i = 1, 2, \dots, k$ i frekvencijama f_i , $i = 1, 2, \dots, k$ su:

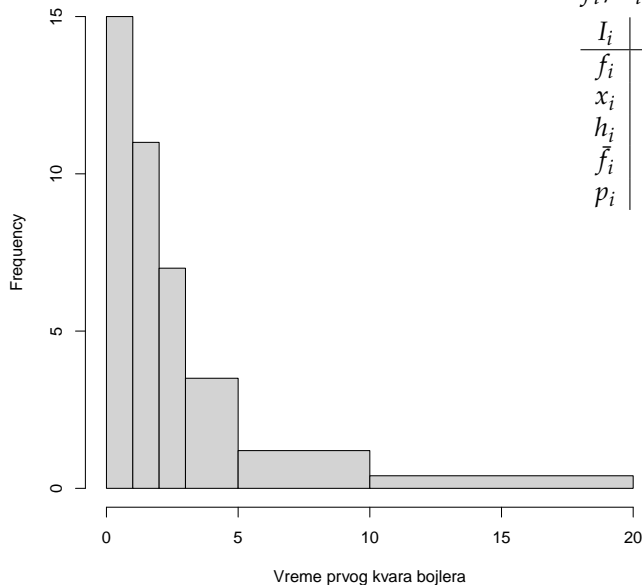
$$n = \sum_{i=1}^k f_i, \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^k x_i f_i, \quad \bar{s}_n^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_n)^2 f_i, \quad \bar{s}_n = \sqrt{\bar{s}_n^2}, \quad \bar{s}_n^{2'} = \frac{n}{n-1} \bar{s}_n^2.$$

PRIMER 2 Anketirani su kupci o vremenu u godinama do prvog kvara na bojleru

I_i	$[0,1]$	$(1,2]$	$(2,3]$	$(3,5]$	$(5,10]$	$(10,20]$
f_i	15	11	7	7	6	4

$$n = 15 + 11 + 7 + 7 + 6 + 4 = 50, \quad \bar{x}_n = (0.5 \cdot 15 + 1.5 \cdot 11 + \dots + 15 \cdot 4) / 50 = 3.49, \\ \bar{s}_n^2 = ((0.5 - 3.49)^2 \cdot 15 + \dots + (15 - 3.49)^2 \cdot 4) / 50 = 16.2549, \quad \bar{s}_n = \sqrt{16.2549} = 4.0317.$$

Histogram



Dopunili smo tabelu sa četiri nove vrste:

$(x_i = (m_i + m_{i-1})/2, h_i = m_i - m_{i-1}, \bar{f}_i = f_i/h_i, p_i = \bar{f}_i/n)$:

I_i	[0,1]	(1,2]	(2,3]	(3,5]	(5,10]	(10,20]
f_i	15	11	7	7	6	4
x_i	0.5	1.5	2.5	4	7.5	15
h_i	1	1	1	2	5	10
\bar{f}_i	15	11	7	3.5	1.2	0.4
p_i	0.30	0.22	0.14	0.07	0.024	0.008

Tabelarni uzorak

Za diskretno obeležje uzorak se može zadati tabelom u kojoj se navode vrednosti obeležja i **frekvencije** pojavljivanja u uzorku.

Formule za računanje aritmetičke sredine i standardne devijacije tabelarnog uzorka sa vrednostima $x_i, i = 1, 2, \dots, k$ i frekvencijama $f_i, i = 1, 2, \dots, k$ su:

$$n = \sum_{i=1}^k f_i, \quad \bar{x}_n = \frac{1}{n} \sum_{i=1}^k x_i f_i, \quad \bar{s}_n^2 = \frac{1}{n} \sum_{i=1}^k (x_i - \bar{x}_n)^2 f_i, \quad \bar{s}_n^{2'} = \frac{n}{n-1} \bar{s}_n^2.$$

PRIMER 3 Kockica je bačena 100 puta. Rezultati bacanja su u tabeli. Izračunati srednju vrednost i varijansu. Nacrtati stubičasti dijagram (Bar Chart).

x_i	1	2	3	4	5	6
f_i	15	17	18	21	15	14

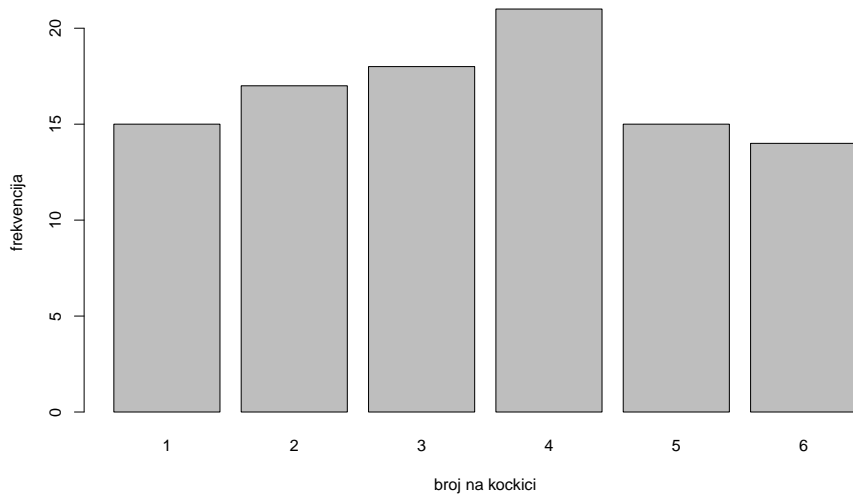
$$n = 15 + 17 + 18 + 21 + 15 + 14 = 100,$$

$$\bar{x}_n = (1 \cdot 15 + 2 \cdot 17 + \dots + 6 \cdot 14) / 100 = 3.46,$$

$$\bar{s}_n^2 = ((1 - 3.46)^2 \cdot 15 + (2 - 3.46)^2 \cdot 17 + \dots + (6 - 3.46)^2 \cdot 14) / 100 = 2.6284$$

x_i	1	2	3	4	5	6
f_i	15	17	18	21	15	14

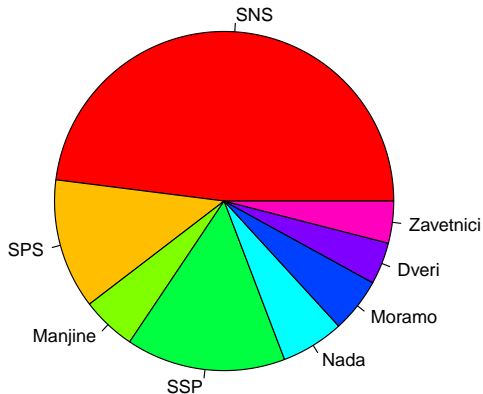
Stubicasti dijagram



PRIMER 4 *Na osnovu sadržaja parlamenta Srbije pre raspuštanja 2023. nacrtati pitu stranačke pripadnosti.*

SNS	SPS	Manjine	SSP	Nada	Moramo	Dveri	Zavetnici
120	31	13	38	15	13	10	10

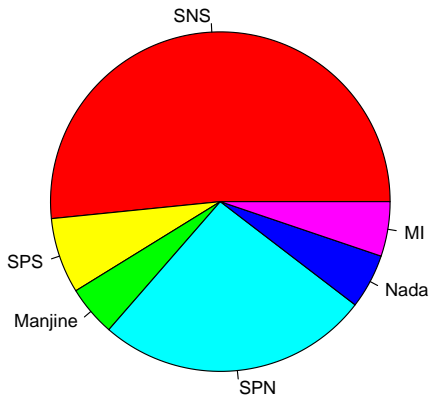
Parlament Srbije 2023



PRIMER 5 *Na osnovu sadržaja parlamenta Srbije posle izbora 2023. nacrtati pitu stranačke pripadnosti.*

SNS	SPS	Manjine	SPN	Nada	Mi
129	18	12	65	13	13

Parlament Srbije 2024



Modus uzorka

Modus je ona vrednost obeležja X kojoj odgovara najveća frekvencija. Ako je uzorak intervalni sa intervalima iste veličine, onda se modus nalazi na sledeći način $Mo = m_{s-1} + d \frac{r_1}{r_1+r_2}$ gde je $I_s = (m_{s-1}, m_s)$ interval sa najvećom frekvencijom (modalni interval), d je dužina intervala, $r_1 = f_s - f_{s-1}$ je razlika najveće frekvenije i frekvencije iz intervala koji prethodi modalnom, $r_2 = f_s - f_{s+1}$ je razlika najveće frekvencije i frekvencije iz intervala posle modalnog.

Medijana uzorka

Medijana Me je sredina uzorka, odnosno to je ona vrednost realizovanog uzorka za koju važi $P(X < Me) = P(X > Me)$. Ako je uzorak neopadajući medijana se izračunava: $Me = x_{\frac{n+1}{2}}$, za n neparno, odnosno $\frac{1}{2}(x_{\frac{n}{2}} + x_{(\frac{n}{2}+1)})$ za n parno.

Ako je uzorak intervalni veličine n onda se medijana računa $Me = m_{l-1} + h_l \frac{\frac{n}{2} - k_{l-1}}{f_l}$, gde je $I_l = (m_{l-1}, m_l)$ medijalni interval, $h_l = m_l - m_{l-1}$ širina medijalnog intervala, $k_{l-1} = \sum_{i=1}^{l-1} f_i$ kumulativna frekvencija intervala I_{l-1} koji prethodi medijalnom intervalu I_l , f_l frekvencija medijalnog intervala. Medijalni interval I_l je interval sa najmanjom kumulativnom frekvencijom većom od $\frac{n}{2}$.

Uzoračka funkcija raspodele

Uzoračka (empirijska) funkcija raspodele F_n^* obeležja X je funkcija definisana za svako x na sledeći način:

$$F_n^*(x) = \frac{N_x}{n}$$

gde je N_x broj elemenata uzorka koji su manji ili jednaki od x , a n je obim realizovanog uzorka. **Realizovana empirijska funkcija raspodele** f_n^* je data sa

$$f_n^*(x) = \frac{n_x}{n}$$

gde je n_x realizovana vrednost promenljive N_x na uzorku (x_1, x_2, \dots, x_n) .

Kvantili (percentili)

Za slučajnu promenljivu X

p -ti **kvantil** je vrednost x za koju je $F(x) = p$. (za percentil $p/100$)

Vrednost x za koju je $F(x) = P(X \leq x) = k/4$ zovemo k -ti **kvartil**, Q_k , $k = 1, 2, 3$.

Za $X : \mathcal{N}(0, 1)$ u R-u `qnorm(.25)` daje prvi kvartil. `pnorm(x)` daje funkciju $F(x) = \Phi(x)$.

Za uzorak obeležja X

Ako je (x_1, x_2, \dots, x_n) sortiran uzorak, $q = (n - 1)p + 1$ i $m = \lfloor q \rfloor$, p -ti **kvantil** je: $x_{(p)} = x_m + (q - m)(x_{m+1} - x_m)$. $x_{(k/4)}$ je k -ti **kvartil**, $k = 1, 2, 3$. $Me = x_{(1/2)}$.

Inter-kvartilni razmak (IQR)

Mera rasutosti uzorka $IQR = Q_3 - Q_1$, gde su Q_3 i Q_1 redom treći i prvi kvartil.

Q-Q plot

Crtaju se tačke u ravni. Apscise se uzimaju iz realizovane vrednosti uzorka, ordinate su kvantili iz pretpostavljene raspodele. Dobijeni skup tačaka treba da daje pravu liniju ako se raspodele slažu.

Pri crtanju se može povući linija kvantila raspodele. U R-u: `qqnorm` i `qqline`.

Box plot

Box plot je kutija (pravougaonik) sa telom od prvog do trećeg kvartila, linijom preko mediane i brkovima na $Q_1 - 1.5 IQR$ i $Q_3 + 1.5 IQR$ ili minimumu i maksimumu uzorka.

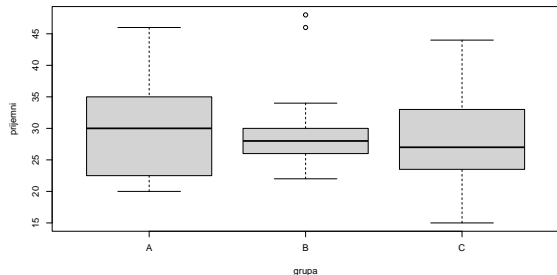
Five-number summary

Pet brojeva koji se često koriste da opišu uzorak su: $\min, Q_1, Q_2, Q_3, \max$.

PRIMER 6 *Naći Five-number summary i nacrtati boxplot za uspeh na prijemnom po grupama.*

A	20	40	36	23	23	31	22	33	21	25	22	36	46	30	34		
B	25	34	25	30	27	29	25	29	30	46	33	26	26	48	27	22	28
C	26	27	28	21	38	44	15										

	Min.	1st Qu.	Median	3rd Qu.	Max.
A	20.00	22.50	30.00	35.00	46.00
B	22.00	26.00	28.00	30.00	48.00
C	15.00	23.50	27.00	33.00	44.00



Six-number summary pored tih pet brojeva uključuje i srednju vrednost uzorka.

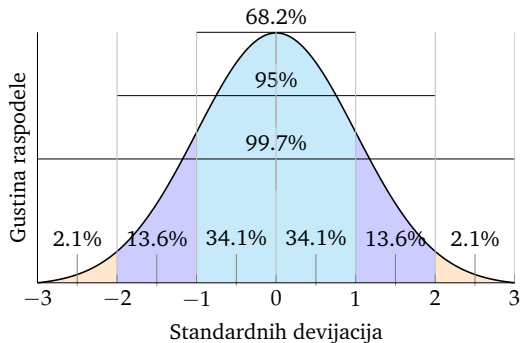
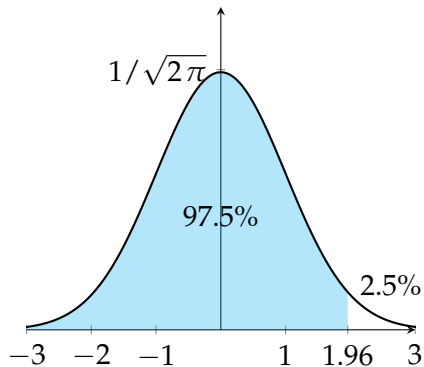
Normalna (Gausova) raspodela

$\mathcal{N}(\mu, \sigma)$, $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R};$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt.$$

Za $\mu = 0$ i $\sigma = 1$, $\mathcal{N}(0,1)$, standardna normalna raspodela $\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.



Slučajna promenljiva ima **Normalnu raspodelu**, $\mathcal{N}(\mu, \sigma)$, $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$ ako

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt, \quad E(X) = \mu, \quad D(X) = \sigma^2.$$

Centriranje normalne raspodele: $X : \mathcal{N}(\mu, \sigma) \Leftrightarrow X^* = \frac{X - \mu}{\sigma} : \mathcal{N}(0, 1).$

Teorema: Ako su X i Y nezavisne slučajne promenljive sa normalnim raspodelama $X : \mathcal{N}(\mu_1, \sigma_1)$, $Y : \mathcal{N}(\mu_2, \sigma_2)$, onda $X \pm Y : \mathcal{N}\left(\mu_1 \pm \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$

Posledica 1: Ako nezavisne slučajne promenljive imaju raspodelu $X_k : \mathcal{N}(\mu, \sigma)$, $k = 1, \dots, n$, onda slučajna promenljiva $X_1 + X_2 + \dots + X_n$ ima normalnu raspodelu $\mathcal{N}(n\mu, \sqrt{n}\sigma)$.

Posledica 2: Ako nezavisne slučajne promenljive imaju raspodelu $X_k : \mathcal{N}(\mu, \sigma)$, $k = 1, \dots, n$, onda slučajna promenljiva $Z = \frac{1}{n} (X_1 + X_2 + \dots + X_n)$ ima normalnu raspodelu $Z : \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, a slučajna promenljiva $Z^* = \frac{Z - \mu}{\frac{\sigma}{\sqrt{n}}} \sqrt{n} = \frac{\frac{1}{n} \sum_{k=1}^n X_k - \mu}{\frac{\sigma}{\sqrt{n}}} \sqrt{n}$ ima normalnu raspodelu $Z^* : \mathcal{N}(0, 1)$.

Centralna granična teorema - Central Limit Theorem (CLT)

Ako je X_1, X_2, \dots niz nezavisnih slučajnih promenljivih sa istom raspodelom čije su očekivanja i disprezija redom $E(X_k) = a$ i $D(X_k) = s^2$, $0 < s < \infty$, onda za svako x

$$\lim_{n \rightarrow \infty} P \left(\frac{\sum_{k=1}^n X_k - E \left(\sum_{k=1}^n X_k \right)}{\sqrt{D \left(\sum_{k=1}^n X_k \right)}} \leq x \right) = \lim_{n \rightarrow \infty} P \left(\frac{\sum_{k=1}^n X_k - n a}{s \sqrt{n}} \leq x \right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x).$$

Moavr-Laplasova teorema

$$X : \mathcal{B}(n, p) \Rightarrow \lim_{n \rightarrow \infty} P \left(\frac{X - n p}{\sqrt{n p (1 - p)}} \leq x \right) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = \Phi(x).$$

Poasonova aproksimacija

Za konačno k , ako je $\lim_{n \rightarrow \infty} n p = \lambda = \text{const}$, važi $\lim_{n \rightarrow \infty} \binom{n}{k} p^k (1 - p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$.

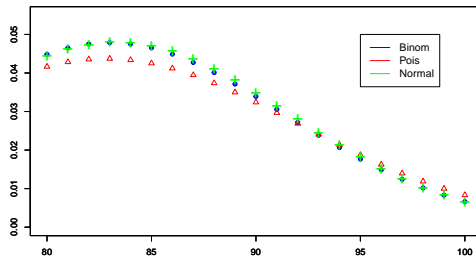
PRIMER 7 Kolika je verovatnoća da je broj šestica u 500 bacanja kocke između 80 i 100?

Označimo X broj šestica u 500 bacanja.

Onda $X : \mathcal{B}(500, \frac{1}{6})$.

$$P = \sum_{k=80}^{100} \binom{500}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{500-k} = \underline{0.6518}.$$

(R: sum(dbinom((80:100),500,1/6)))



Moavr-Laplasova aproksimacija

$$\begin{aligned} P(80 < X \leq 100) &= P\left(\frac{80-500\frac{1}{6}}{\sqrt{500\frac{1}{6}\frac{5}{6}}} < \frac{X-500\frac{1}{6}}{\sqrt{500\frac{1}{6}\frac{5}{6}}} \leq \frac{100-500\frac{1}{6}}{\sqrt{500\frac{1}{6}\frac{5}{6}}}\right) \approx \Phi\left(\frac{100-500\frac{1}{6}}{\sqrt{500\frac{1}{6}\frac{5}{6}}}\right) - \Phi\left(\frac{80-500\frac{1}{6}}{\sqrt{500\frac{1}{6}\frac{5}{6}}}\right) = \\ &= \Phi(2) - \Phi(-0.4) = \Phi(2) - 1 + \Phi(0.4) = \\ &= 0.9772 - 1 + 0.6554 = 0.6326 \end{aligned}$$

Poasonova aproksimacija, $\lambda = n \cdot p = 500 \cdot \frac{1}{6} = 83.33333$

$$P = \sum_{k=80}^{100} \binom{500}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{500-k} \approx \sum_{k=80}^{100} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=80}^{100} \frac{83.33333^k}{k!} \exp(-83.33333) = 0.6242$$

χ^2 raspodela (Pirsonova)

Ako X ima Normalnu raspodelu, $X : \mathcal{N}(0,1)$ i $Y = X^2$, onda je gustina raspodele Y :

$$\varphi_Y(y) = \begin{cases} 0, & y \leq 0, \\ \frac{1}{\sqrt{2\pi y}} e^{-y/2}, & y > 0, \end{cases}$$

kažemo da Y ima **hi-kvadrat raspodelu** sa jednim stepenom slobode, $Y : \chi_1^2$.

Slučajna promenljiva sa **hi-kvadrat raspodelom** sa n stepeni slobode, $X : \chi_n^2$, ima gustinu:

$$\varphi(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, \quad x > 0, \quad E(X) = n, \quad D(X) = 2n.$$

Teorema: Ako su $Y_1 : \chi_n^2$ i $Y_2 : \chi_m^2$ nezavisne slučajne promenljive, onda $Y = Y_1 + Y_2 : \chi_{n+m}^2$.

Posledica 1: Ako su X_1, X_2, \dots, X_n nezavisne slučajne promenljive sa normalnom $\mathcal{N}(0,1)$ raspodelom, onda $Y = X_1^2 + X_2^2 + \dots + X_n^2$ ima χ_n^2 raspodelu.

Posledica 2: Ako su X_1, X_2, \dots, X_n nezavisne slučajne promenljive sa normalnom $\mathcal{N}(\mu, \sigma)$ raspodelom, onda $Y = \left(\frac{X_1 - \mu}{\sigma}\right)^2 + \left(\frac{X_2 - \mu}{\sigma}\right)^2 + \dots + \left(\frac{X_n - \mu}{\sigma}\right)^2$ ima χ_n^2 raspodelu.

t raspodela (Studentova)

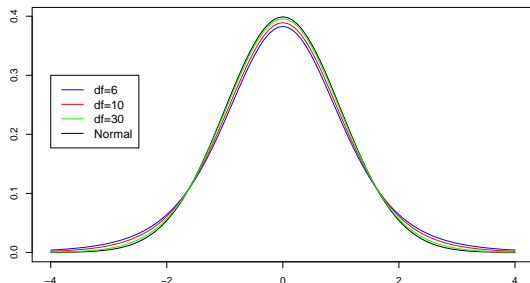
Slučajna promenljiva $T : t_n$ ima **Studentovu raspodelu** sa n stepeni slobode ako ima gustinu

$$\varphi(t) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \Gamma(n/2) (1+t^2/n)^{(n+1)/2}}.$$

Drugi zapis: $\varphi(t) = \left(\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right) \left(1 + \frac{t^2}{n}\right)^{(n+1)/2} \right)^{-1}$. $E(T) = 0, D(T) = \frac{n}{n-2}, n > 2$.

Teorema:

Ako su $X : \mathcal{N}(0,1)$ i $Y : \chi_n^2$ nezavisne sl. prom, onda $T = \frac{X}{\sqrt{\frac{Y}{n}}}$ ima t_n raspodelu.



Važne statistike uzorka (X_1, X_2, \dots, X_n) obeležja X

Aritmetička sredina uzorka

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k \quad \Rightarrow \quad \begin{cases} E(\bar{X}_n) = E(X), & D(\bar{X}_n) = \frac{1}{n} D(X) \\ X : \mathcal{N}(\mu, \sigma) \Rightarrow \bar{X}_n : \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} : \mathcal{N}(0, 1) \end{cases}$$

Uzoračka disperzija (varijansa)

$$\bar{S}_n^2 = \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X}_n)^2 = \frac{1}{n} \sum_{k=1}^n X_k^2 - \bar{X}_n^2, \quad E(\bar{S}_n^2) = \frac{n-1}{n} D(X), \quad \bar{S}_n = \sqrt{\bar{S}_n^2}.$$

$$X : \mathcal{N}(\mu, \sigma) \quad \Rightarrow \quad \begin{cases} Y = \frac{n\bar{S}_n^2}{\sigma^2} = \frac{(n-1)\bar{S}_n'^2}{\sigma^2} : \chi_{n-1}^2, \\ T = \frac{\bar{X}_n - \mu}{\bar{S}_n} \sqrt{n-1} = \frac{\bar{X}_n - \mu}{\bar{S}_n'} \sqrt{n} : t_{n-1} \end{cases}$$

Korigovana varijansa

$$\bar{S}_n'^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X}_n)^2, \quad E(\bar{S}_n'^2) = D(X), \quad \bar{S}_n' = \sqrt{\bar{S}_n'^2}$$

Tačkaste ocene parametara - Point Estimate

Raspodela obeležja zavisi od (nepoznatog) parametra θ , kojeg ocenjujemo pomoću (realizovane vrednosti) uzorka.

Ocenjivač neke funkcije parametra $\tau(\theta)$ je statistika $U = u(X_1, X_2, \dots, X_n)$ čija realizovana vrednost (**ocena**) $u(x_1, x_2, \dots, x_n)$ je bliska $\tau(\theta)$.

Ocenjivač U je **postojan** za $\tau(\theta)$ ako $\lim_{n \rightarrow \infty} P(|\tau(\theta) - u(X_1, X_2, \dots, X_n)| > \varepsilon) = 0$ za sve $\varepsilon > 0$.

Ocenjivač U je **centriran** za $\tau(\theta)$ ako $E(u(X_1, X_2, \dots, X_n)) = \tau(\theta)$, a **asimptotski centriran** ako $\lim_{n \rightarrow \infty} E(u(X_1, X_2, \dots, X_n)) = \tau(\theta)$.

PRIMER 8 Ispitati postojanost ocenjivača \bar{X}_n za μ , obeležja $X : \mathcal{N}(\mu, \sigma)$.

Aritmetička sredina uzorka \bar{X}_n je centriran i postojan ocenjivač parametra jednakog matematičkom očekivanju obeležja.

PRIMER 9 Naći centrirani ocenjivač parametra jednakog disperziji obeležja.

Srednja kvadratna greška ocenjivača U za $\tau(\theta)$ je $E((U - \tau(\theta))^2) = D(U) + ((E(U) - \tau(\theta))^2)$

Ako su U_1 i U_2 centrirani ocenjivači za $\tau(\theta)$ i $D(U_1) < D(U_2)$, kažemo da je ocenjivač U_1 **efikasniji** od U_2 . Za obeležje i parametar postoji **najbolja** disperzija σ_0^2 koja se može postići.

Metod momenata

Ocene parametara dobijamo iz jednačina u kojima izjednačavamo uzoračke momenta sa momentima obeležja.

PRIMER 10 Metodom momenata naći ocene parametara μ i σ obeležja $X : \mathcal{N}(\mu, \sigma)$.

Metod maksimalne verodostojnosti

Za ocenu parametra θ od koga zavisi gustina raspodele $\varphi(x, \theta)$ ili zakon raspodele $p_i = p(x_i, \theta)$ uzima se vrednost $\theta = \theta(x_1, x_2, \dots, x_n)$ za koju se ostvaruje maksimum (ako postoji) funkcije verodostojnosti koja se za realizovanu vrednost uzorka (x_1, x_2, \dots, x_n) računa:

$$L = L(x_1, x_2, \dots, x_n, \theta) = \begin{cases} \varphi(x_1, \theta) \varphi(x_2, \theta) \dots \varphi(x_n, \theta), & \text{neprekidno} \\ p(x_1, \theta) p(x_2, \theta) \dots p(x_n, \theta), & \text{diskretno obeležje} \end{cases}$$

PRIMER 11 Naći ocenu maksimalne verodostojnosti parametara μ i σ^2 za $X : \mathcal{N}(\mu, \sigma)$.

Rešenje primera 10 i 11: Metodom momenata i metodom maksimalne verodostojnosti se dobija ocenjivač za μ je $\mu := \bar{x}_n$, ocenjivač za σ^2 je $\sigma^2 := \bar{s}_n^2$.

Za svako obeležje X (slučajnu promenljivu) i prost slučajni uzorak obeležja (X_1, X_2, \dots, X_n) , centrirani ocenjivač za očekivanje je $E(X) := \bar{x}_n$, za varijansu obeležja $D(X) = \text{Var } X := \bar{s}_n^2$.

Intervali poverenja - Confidence Intervals (CI)

Za obeležje X raspodele $F(x, \theta)$, sa uzorkom (X_1, X_2, \dots, X_n) , ako su $U_1 = u_1(X_1, X_2, \dots, X_n)$ i $U_2 = u_2(X_1, X_2, \dots, X_n)$ statistike za koje važi $P(U_1 < \theta < U_2) = \beta$, gde je β unapred zadat **nivo poverenja**, onda je (U_1, U_2) **interval poverenja** širine β .

Za očekivanje μ obeležja $X : \mathcal{N}(\mu, \sigma)$, σ poznato

Ako $X : \mathcal{N}(\mu, \sigma)$ onda $\bar{X}_n : \mathcal{N}(\mu, \sigma / \sqrt{n})$, odnosno, onda $Z = \frac{\bar{X}_n - \mu}{\sigma} \sqrt{n} : \mathcal{N}(0, 1)$.

Označimo sa z_β vrednost za koju je $P(|Z| < z_\beta) = \beta$. ($z_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$ je $\frac{1+\beta}{2}$ kvantil).

Onda je $U_1 = \bar{X}_n - z_\beta \frac{\sigma}{\sqrt{n}}$, $U_2 = \bar{X}_n + z_\beta \frac{\sigma}{\sqrt{n}}$. Izraz $\frac{\sigma}{\sqrt{n}}$ nazivamo **standardna greška**.

Za očekivanje μ obeležja $X : \mathcal{N}(\mu, \sigma)$, σ nepoznato

Ako $X : \mathcal{N}(\mu, \sigma)$ onda $T = \frac{\bar{X}_n - \mu}{\bar{S}_n} \sqrt{n-1} = \frac{\bar{X}_n - \mu}{\bar{S}'_n} \sqrt{n} : t_{n-1}$.

Označimo sa t_β vrednost za koju je $P(|T| < t_\beta)$. (t_β je $(1 + \beta)/2$ kvantil raspodele t_{n-1} .)

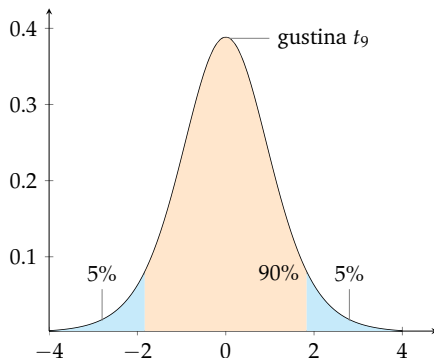
Onda je $U_1 = \bar{X}_n - t_\beta \frac{\bar{S}_n}{\sqrt{n-1}}$, $U_2 = \bar{X}_n + t_\beta \frac{\bar{S}_n}{\sqrt{n-1}}$. **Standardna greška** je $\frac{\bar{S}_n}{\sqrt{n-1}} = \frac{\bar{S}'_n}{\sqrt{n}}$.

PRIMER 12 Naći 90% interval poverenja za srednju vrednost μ obeležja sa normalnom $\mathcal{N}(\mu, \sigma)$ raspodelom

(a) Ako je poznato $\sigma = 3$, (b) ako je σ nepoznato,
za uzorak (17.3, 12.9, 10.4, 11.9, 9.9, 8.9, 9.9, 6.3, 12.9, 9.4).

($n = 10$, $\bar{x}_n = 10.98$, $\bar{s}'_n = 2.973139$, $z_{0.9} = 1.645$, $t_{0.9} = 1.833$)

```
> x<-c(17.3, 12.9, 10.4, 11.9, 9.9,
      8.9, 9.9, 6.3, 12.9, 9.4)
> n<-10; xn<-mean(x); sn<-sd(x)
> z<-qnorm(.95); t<-qt(.95,9)
> xn-z*3/sqrt(10)
[1] 9.419555
> xn+z*3/sqrt(10)
[1] 12.54045
> xn-t*sn/sqrt(10)
[1] 9.256527
> xn+t*sn/sqrt(10)
[1] 12.70347
```



PRIMER 13 Merena je visina muških studenata druge godine, rezultati su dati u tabeli:

173	174	174.5	175	176	177	177.5	178	178.5	179	179.5	180	180.5	181	181.5	182	182.5	183.5	184.5	185	186	186.5	189	189.5
1	2	1	2	1	1	1	1	5	3	1	3	2	2	3	3	2	2	2	3	3	2	2	2

Naći 95% interval visine studenata.

Rešenje:

$$1 - \alpha = 0.95, t_{1-\alpha;49} = 2.010, n = 50,$$

$$\bar{x}_n = \frac{1}{50}(173 \cdot 1 + 174 \cdot 2 + 174.5 \cdot 1 + 175 \cdot 2 + 176 \cdot 1 + 177 \cdot 1 + 177.5 \cdot 1 + 178 \cdot 1 + 178.5 \cdot 5 + 179 \cdot 3 + 179.5 \cdot 1 + 180 \cdot 3 + 180.5 \cdot 2 + 181 \cdot 2 + 181.5 \cdot 3 + 182 \cdot 3 + 182.5 \cdot 2 + 183.5 \cdot 2 + 184.5 \cdot 2 + 185 \cdot 3 + 186 \cdot 3 + 186.5 \cdot 2 + 189 \cdot 2 + 189.5 \cdot 2) = \underline{181.21},$$

$$\bar{s}_n^2 = \frac{1}{50}((173 - 181.21)^2 \cdot 1 + (174 - 181.21)^2 \cdot 2 + (174.5 - 181.21)^2 \cdot 1 + (175 - 181.21)^2 \cdot 2 + (176 - 181.21)^2 \cdot 1 + (177 - 181.21)^2 \cdot 1 + (177.5 - 181.21)^2 \cdot 1 + (178 - 181.21)^2 \cdot 1 + (178.5 - 181.21)^2 \cdot 5 + (179 - 181.21)^2 \cdot 3 + (179.5 - 181.21)^2 \cdot 1 + (180 - 181.21)^2 \cdot 3 + (180.5 - 181.21)^2 \cdot 2 + (181 - 181.21)^2 \cdot 2 + (181.5 - 181.21)^2 \cdot 3 + (182 - 181.21)^2 \cdot 3 + (182.5 - 181.21)^2 \cdot 2 + (183.5 - 181.21)^2 \cdot 2 + (184.5 - 181.21)^2 \cdot 2 + (185 - 181.21)^2 \cdot 3 + (186 - 181.21)^2 \cdot 3 + (186.5 - 181.21)^2 \cdot 2 + (189 - 181.21)^2 \cdot 2 + (189.5 - 181.21)^2 \cdot 2) = \underline{17.7309}$$

$$\bar{x}_n \pm t_{1-\alpha;49} \frac{\bar{s}_n}{\sqrt{n-1}} = 181.21 \pm 2.010 \cdot \frac{\sqrt{17.7309}}{\sqrt{49}} = \underline{(180.0009, 182.4191)}.$$

Studentove i Gausove tablice t i z vrednosti

Za $X : t_n$ raspodelu $P = P(X \leq t)$, za $n \rightarrow \infty$, $t_n \rightarrow \mathcal{N}$, $t \rightarrow z$

n^P	.75	.90	.95	.975	.990	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	1.638	2.353	3.182	4.541	5.841	12.924
...							
8	.706	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	4.587
...							
30	.683	1.310	1.697	2.042	2.457	2.750	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	3.551
49	.680	1.299	1.677	2.010	2.405	2.680	3.500
60	.679	1.296	1.671	2.000	2.390	2.660	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	3.373
...							
z	.674	1.282	1.645	1.960	2.326	2.576	3.291

Za disperziju σ^2 obeležja $X : \mathcal{N}(\mu, \sigma)$

Ako $X : \mathcal{N}(\mu, \sigma)$ onda $Y = \frac{n\bar{S}_n^2}{\sigma^2} : \chi_{n-1}^2$.

Neka su $y_{(1-\beta)/2}$ i $y_{(1+\beta)/2}$ redom $(1-\beta)/2$ i $(1+\beta)/2$ kvantili χ_{n-1}^2 raspodele, odnosno, $P(y_{(1-\beta)/2} < Y < y_{(1+\beta)/2}) = \beta$.

Onda $P\left(\frac{n\bar{S}_n^2}{y_{(1+\beta)/2}} < \sigma^2 < \frac{n\bar{S}_n^2}{y_{(1-\beta)/2}}\right) = \beta$, odnosno, $P\left(\sqrt{\frac{n\bar{S}_n^2}{y_{(1+\beta)/2}}} < \sigma < \sqrt{\frac{n\bar{S}_n^2}{y_{(1-\beta)/2}}}\right) = \beta$.

PRIMER 14 Naći 90% interval poverenja za nepoznatu varijansu obeležja iz primera 12.

```
> y0 <- -qchisq(.05,9); y1 <- -qchisq(.95,9); # nastavak primera 12  
> 9*sn ^ 2/y1  
[1] 4.702175  
> 9*sn ^ 2/y0  
[1] 23.9258
```

Za nepoznatu verovatnoću p

Ako obeležje ima Bernulijevu raspodelu: $X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$, naći interval poverenja za p .

Moavr-Laplasova teorema: za $K = \sum X_i$, $\frac{K - np}{\sqrt{np(1-p)}} \rightarrow Z : \mathcal{N}(0,1)$. Za $z_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$ važi

$$\begin{aligned}\beta &\approx P\left(\left|\frac{K - np}{\sqrt{np(1-p)}}\right| < z_\beta\right) = P\left(\left(\frac{K - np}{\sqrt{np(1-p)}}\right)^2 < z_\beta^2\right) = \\ &= P\left((n^2 + z_\beta^2 n)p^2 + (-2Kn - z_\beta^2 n)p + K^2 < 0\right) = P(U_1 < p < U_2),\end{aligned}$$

gde su $U_{1,2}$ rešenja kvadratne jednačine.

PRIMER 15 U filmu *I-origin* radi se test: 25 puta se postavlja pitanje sa istom verovatnoćom tačnog odgovora. Kandidat 11 puta odgovara tačno. Naći 90% interval poverenja za nepoznatu verovatnoću tačnog odgovora.

```
> n<-25; K<-11; z<-qnorm(.95);
> a<-n^2+z^2*n; b<-2*K*n-z^2*n; c<-K^2; d<-b^2-4*a*c;
> x1<-(-b-sqrt(d))/2/a; x2<-(-b+sqrt(d))/2/a;
> x1
[1] 0.2906299
> x2
[1] 0.6010886
```

Testiranje hipoteza

Statistički testovi

Hipoteza H_0 protiv H_1		Usvojena H_0	Usvojena H_1
	Tačna H_0	OK	Greška I vrste
	Tačna H_1	Greška II vrste	OK

Parametarske hipoteze

- Zadaje se prag značajnosti α (recimo $\alpha = 5\% = 0.05$)
- Bira se parametar raspodele obeležja (θ).
- Nalazi se statistika koja je ocenjivač parametra $\hat{\theta} = h(X_1, \dots, X_n)$.
- Nalazi se kritična oblast C (koja daje nedozvoljene vrednosti) parametra, takva da je $P_{H_0}(\hat{\theta} = h(X_1, \dots, X_n) \in C) = \alpha$.
- Računa se realizovana vrednost statistike uzorka $\theta = h(x_1, x_2, \dots, x_n)$ i ako $\theta \in C$, odbacujemo H_0 (i usvajamo H_1).

$H_0(\mu = \mu_0)$ **protiv** $H_1(\mu \neq \mu_0)$ **za** $X : \mathcal{N}(\mu, \sigma)$, σ **poznato**

Koristimo interval poverenja za nepoznato očekivanje μ obeležja sa normalnom raspodelom, σ poznato, širine $\beta = 1 - \alpha$. ($Z : \mathcal{N}(0, 1)$)

$$\mu_0 \in \mathbb{R} \setminus \left(\bar{x}_n \mp z_\beta \frac{\sigma}{\sqrt{n}} \right) \Leftrightarrow z := \frac{|\bar{x}_n - \mu_0|}{\sigma} \sqrt{n} > z_\beta \Leftrightarrow \alpha^* := P_{H_0} \left(|Z| > \frac{|\bar{x}_n - \mu_0|}{\sigma} \sqrt{n} \right) < \alpha,$$

($z_\beta = \Phi^{-1} \left(\frac{1+\beta}{2} \right)$ je $\frac{1+\beta}{2}$ kvantil)

$H_0(\mu = \mu_0)$ **protiv** $H_1(\mu \neq \mu_0)$ **za** $X : \mathcal{N}(\mu, \sigma)$, σ **nepoznato**

Koristimo interval poverenja za nepoznato očekivanje μ obeležja sa normalnom raspodelom, σ nepoznato, širine $\beta = 1 - \alpha$. ($T : t_{n-1}$)

$$\mu_0 \in \mathbb{R} \setminus \left(\bar{x}_n \mp t_\beta \frac{\bar{s}'_n}{\sqrt{n}} \right) \Leftrightarrow t := \frac{|\bar{x}_n - \mu_0|}{\bar{s}'_n} \sqrt{n} > t_\beta \Leftrightarrow \alpha^* := P_{H_0} \left(|T| > \frac{|\bar{x}_n - \mu_0|}{\bar{s}'_n} \sqrt{n} \right) < \alpha,$$

(t_β je $(1 + \beta)/2$ kvantil raspodele t_{n-1} .)

PRIMER 16 Testirati hipotezu $H_0(\mu = 13)$ za uzorak iz zadatka 12.

PRIMER 17 Testirati hipotezu $H_0(p = 1/3)$ za uzorak iz zadatka 15.

$H_0(\sigma^2 = \sigma_0^2)$ **protiv** $H_1(\sigma^2 \neq \sigma_0^2)$ **za** $X : \mathcal{N}(\mu, \sigma)$

Koristimo $\beta = 1 - \alpha$ interval poverenja za nepoznatu varijansu σ^2 obeležja sa normalnom raspodelom.

$$\frac{n \bar{s}_n^2}{y_{(1+\beta)/2}} < \sigma_0^2 < \frac{n \bar{s}_n^2}{y_{(1-\beta)/2}} \Leftrightarrow H_0 \text{ ne odbacujemo}$$

Jednostrani testovi

Alternativna hipoteza je $H_1(\mu < \mu_0)$ ili $H_1(\mu > \mu_0)$

Koristimo jednostrane intervale poverenja sa $z_1 = \Phi^{-1}(\beta)$ umesto $z_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$

Za varijansu $H_1(\sigma^2 > \sigma_0^2)$ koristimo jednostrani interval poverenja $\left(0, \frac{n \bar{s}_n^2}{y_{1-\beta}}\right)$, gde je $y_{1-\beta}$ kvantil koji odgovara $\alpha = 1 - \beta$ za χ_{n-1}^2 .

PRIMER 18 Za uzorak iz zadatka 12 testirati hipotezu $H_0(\sigma^2 = 25)$ protiv $H_1(\sigma^2 > 25)$.

$> 9 * s_n^2 / y_0$; # nastavak primera 12
[1] 23.9258 # odbacujemo hipotezu

Testiranje jednakosti srednjih vrednosti dva uzorka

$H_0(\mu_1 = \mu_2)$ **protiv** $H_1(\mu_1 \neq \mu_2)$, σ_1, σ_2 **poznato**, obeležja sa $\mathcal{N}(\mu_1, \sigma_1)$ i $\mathcal{N}(\mu_2, \sigma_2)$ **raspodelama**

Koristimo statistiku $Z := \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ koja ima $\mathcal{N}(0, 1)$ raspodelu

$H_0(\mu_1 = \mu_2)$ **protiv** $H_1(\mu_1 \neq \mu_2)$ **(T-test)**

Koristimo statistiku $T := \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\bar{S}_1'^2}{n_1} + \frac{\bar{S}_2'^2}{n_2}}}$, koja približno ima t_ν raspodelu,

gde se za ν uzima procena Welcha: $\nu = \frac{\left(\frac{\bar{s}_1'^2}{n_1} + \frac{\bar{s}_2'^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{\bar{s}_1'^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{\bar{s}_2'^2}{n_2}\right)^2}$.

PRIMER 19 Testirati jednakost srednjih vrednosti obeležja 9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95 i obeležja 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78 sa Normalnim raspodelama.

```
> t.test(c(9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95),  
         c( 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78))  
Welch Two Sample t-test  
t = 1.7536, df = 16.766, p-value = 0.09776  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval :  
  -0.1439427 1.5522760  
sample estimates:  
mean of x mean of y  
10.032500 9.328333
```

Razlika srednjih vrednosti je $10.032500 - 9.328333 = 0.704167$.

Vrednost 0 pripada intervalu poverenja širine 95% oko razlike $(-0.1439427, 1.5522760)$.

Welchova ocena je $\nu = 16.76$, p-vrednost je $\alpha^* = 0.09776 > 0.05$.

Realizovana vrednost statistike je $t = 1.7536 < t_{0.975;16} = 2.120$.

Zaključak: Sa pragom značajnosti $\alpha = 0.05$ ne odbacujemo Nultu hipotezu o jednakosti srednjih vrednosti.

Račun prethodnog zadatka:

	x_1	$x_1 - \bar{x}_{1,n}$	$(x_1 - \bar{x}_{1,n})^2$	x_2	$x_2 - \bar{x}_{2,n}$	$(x_2 - \bar{x}_{2,n})^2$
1	9.81	-0.22	0.05	9.85	0.52	0.27
2	9.83	-0.20	0.04	9.30	-0.03	0.00
3	10.43	0.40	0.16	9.08	-0.25	0.06
4	11.13	1.10	1.20	8.07	-1.26	1.58
5	9.70	-0.33	0.11	9.22	-0.11	0.01
6	9.59	-0.44	0.20	9.55	0.22	0.05
7	10.88	0.85	0.72	7.88	-1.45	2.10
8	10.97	0.94	0.88	7.84	-1.49	2.22
9	9.35	-0.68	0.47	8.50	-0.83	0.69
10	9.34	-0.69	0.48	11.95	2.62	6.87
11	9.41	-0.62	0.39	10.92	1.59	2.53
12	9.95	-0.08	0.01	9.78	0.45	0.20

Σ	120.39		4.696225	111.94		16.58837	
$\bar{x}_{1,n} =$	10.0325	$\bar{s}_{1,n}^2 =$	0.3913521	9.328333	$= \bar{x}_{2,n}$	1.382364	$= \bar{s}_{2,n}^2$

$$t := \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\bar{s}_1'^2}{n_1} + \frac{\bar{s}_2'^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\bar{s}_1^2}{n_1-1} + \frac{\bar{s}_2^2}{n_2-1}}} = \frac{10.0325 - 9.328333}{\sqrt{\frac{0.3913521}{11} + \frac{1.382364}{11}}} = 1.7536$$

Broj stepeni slobode, Weltchova procena:

$$\nu = \frac{\left(\frac{\bar{s}_1^2}{n_1-1} + \frac{\bar{s}_2^2}{n_2-1}\right)^2}{\frac{1}{n_1-1} \left(\frac{\bar{s}_1^2}{n_1-1}\right)^2 + \frac{1}{n_2-1} \left(\frac{\bar{s}_2^2}{n_2-1}\right)^2} = \frac{\left(\frac{0.3913521}{12-1} + \frac{1.382364}{12-1}\right)^2}{\frac{1}{12-1} \left(\frac{0.3913521}{12-1}\right)^2 + \frac{1}{12-1} \left(\frac{1.382364}{12-1}\right)^2} = 16.76614.$$

T-test parova

Koristi se kad imamo dva obeležja sa uparenim vrednostima ("pre" i "posle"): x_1, x_2, \dots, x_n i y_1, y_2, \dots, y_n .

Nalazimo $t_1 = x_1 - y_1, t_2 = x_2 - y_2, \dots, t_n = x_n - y_n$ i testiramo $H_0(\mu = 0)$ protiv $H_1(\mu \neq 0)$ ili $H_1(\mu < 0)$ ili $H_1(\mu > 0)$, sa σ nepoznato za uzorak t_1, t_2, \dots, t_n

PRIMER 20 *Testirati hipotezu o jednakosti srednjih vrednosti obeležja sa normalnom raspodelom pre i posle tretmana.*

pre	9.81	9.83	10.43	11.13	9.70	9.59	10.88	10.97	9.35	9.34	9.41	9.95
posle	9.85	9.30	9.08	8.07	9.22	9.55	7.88	7.84	8.50	11.95	10.92	9.78

Računamo: $9.81 - 9.84 = -0.04, 9.83 - 9.30 = 0.53, \dots, 9.95 - 9.78 = 0.17$.

Dobijamo uzorak razlika:

$(-0.04, 0.53, 1.35, 3.06, 0.48, 0.04, 3.00, 3.13, 0.85, -2.61, -1.51, 0.17)$, $n = 12$.

$\bar{x}_{12} = (-0.04 + 0.53 + 1.35 + 3.06 + 0.48 + 0.04 + 3.00 + 3.13 + 0.85 - 2.61 - 1.51 + 0.17) / 12 = 0.7042$,

$\bar{s}_{12}^2 = (0.04^2 + 0.53^2 + 1.35^2 + 3.06^2 + 0.48^2 + 0.04^2 + 3.00^2 + 3.13^2 + 0.85^2 + 2.61^2 + 1.51^2 + 0.17^2) / 12 - 0.7042^2 = 2.8659$,

$$t = \frac{|\bar{x}_n - m_0|}{\bar{s}_n} \sqrt{n-1} = \frac{|0.7042 - 0|}{\sqrt{2.8659}} \sqrt{11} = 1.380$$

Pošto je $t = 1.380 \leq 2.201$, ne odbacujemo H_0 .

Koristeći software R, vidimo i p-vrednost $0.1951 > 0.05$:

```
> t.test(c(9.81, 9.83, 10.43, 11.13, 9.70, 9.59, 10.88, 10.97, 9.35, 9.34, 9.41, 9.95),  
         c( 9.85, 9.30, 9.08, 8.07, 9.22, 9.55, 7.88, 7.84, 8.50, 11.95, 10.92, 9.78),  
         paired = T)  
Paired t-test  
t = 1.3796, df = 11, p-value = 0.1951  
alternative hypothesis: true mean difference is not equal to 0  
95 percent confidence interval: -0.4192785 1.8276118
```

Ne odbacujemo hipotezu o jednakosti srednjih vrednosti sa pragom značajnosti $\alpha = 0.05$.

n^P	.75	.90	.95	.975	.990	.995	.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	.816	1.886	2.920	4.303	6.965	9.925	31.599
3	.765	1.638	2.353	3.182	4.541	5.841	12.924
4	.741	1.533	2.132	2.776	3.747	4.604	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	6.869
6	.718	1.440	1.943	2.447	3.143	3.707	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	5.408
8	.706	1.397	1.860	2.306	2.896	3.355	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	4.587
11	.697	1.363	1.796	2.201	2.718	3.106	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.965
...							

Testiranje jednakosti proporcija dva uzorka

PRIMER 21 *Da li veruju u zagrobni život? Pitali su 684 žena, 550 odgovorilo sa DA i 563 muškarca, 425 odgovorilo sa DA. Testirati hipotezu da su proporcije jednake.*

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

Neka broj pozitivnih odgovora žena ima $X_1 : \mathcal{B}(n_1, p_1)$ a muškaraca $X_2 : \mathcal{B}(n_2, p_2)$.

Neka $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$, $\hat{p}_1 = \frac{X_1}{n_1}$, $\hat{p}_2 = \frac{X_2}{n_2}$, onda statistika $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}}$,
ako je tačna Nulta hipoteza, ima približno Normalnu $\mathcal{N}(0, 1)$ raspodelu $\Leftrightarrow Z^2 : \chi_1^2$.

```
> prop.test(c(550, 425), c(684, 563), correct=F)
2-sample test for equality of proportions without continuity correction
data:  c(550, 425) out of c(684, 563)
X-squared = 4.3848, df = 1, p-value = 0.03626
alternative hypothesis: two.sided
95 percent confidence interval:
 0.002870906 0.095547134
sample estimates:
   prop 1   prop 2 
0.8040936 0.7548845
```

Odbacujemo Nultu hipotezu o jednakosti proporcija sa pragom značajnosti $\alpha = 0.05$.

Neparametarski testovi $H_0(F(x) = F_0(x))$ protiv $H_1(F(x) \neq F_0(x))$

χ^2 test

Uzorak se grupiše u intervale I_i , sa deobenim tačkama m_i , $i = 0, 1, \dots, k$ i brojem elemenata uzorka u intervalu i jednak f_i , $i = 1, 2, \dots, k$. (Treba $f_i \geq 5$.)

Može se pokazati da se za dovoljno veliki obim uzorka n , raspodela statistike

$$Y = \sum_{i=1}^k \frac{(F_i - n p_i)^2}{n p_i}, \text{ gde je } p_i = P(m_{i-1} < X \leq m_i), f_i \text{ realizovana vrednost } F_i,$$

može aproksimirati χ^2_{k-1} raspodelom. Ako se ocenjuje s parametara, onda χ^2_{k-1-s} .

Ako realizovana vrednost statistike $y > y_{1-\alpha}$, gde je $y_{1-\alpha}$ kvantil χ^2 raspodele sa $k - 1 - s$ stepeni slobode, s = broj ocenjivanih parametara, odbacujemo nultu hipotezu H_0 .

PRIMER 22 U Mendelovim eksperimentima ukršteni pasulji su dali 315 okruglih žutih, 108 okruglih zelenih, 101 naboranih žutih i 32 naborana zelena zrna. Po njegovoj teoriji, njihov odnos bi trebao biti 9:3:3:1. Da li je njegova teorija ispravna? Kolika je p -vrednost?

```
> chisq.test(c(315,108,101,32),p=c(9,3,3,1)/16)
Chi-squared test for given probabilities
X-squared = 0.47002, df = 3, p-value = 0.9254
```

Ne odbacujemo nultu hipotezu. P -vrednost je velika!

Tabela kontigencije

χ^2 -test nezavisnosti obeležja. Obeležje X uzima m mogućih vrednosti, Y uzima n vrednosti. Formira se tabela $m \times n$ verovatnoća $p_{i,j}$ izračunatih preko marginalnih verovatnoća $p_{i\cdot} = p_{i\cdot} \cdot p_{\cdot j}$, koje se dobijaju koristeći marginalne frekvencije.

Statistika $Y = \sum_{i,j} \frac{(F_{i,j} - N p_{i\cdot} p_{\cdot j})^2}{N p_{i\cdot} p_{\cdot j}}$ ima približno χ^2 raspodelu sa $(m - 1)(n - 1)$ st. sl.

PRIMER 23 U tabeli su dati brojevi studenata koji su položili i pali kolokvijum kod tri asistenta. Testirati hipotezu da su procenti položenih nezavisni od asistenta.

	X	Y	Z	
pali	50	47	56	153
položili	5	14	8	27
ukupno	55	61	64	

```
> chisq.test(matrix(c(50,5,47,14,56,8), ncol = 3)))  
Pearson Chi-squared test  
X-squared = 4.8444, df = 2, p-value = 0.08873
```

Ne odbacujemo nultu hipotezu o nezavisnosti obeležja sa pragom značajnosti $\alpha = 0.05$.

Test Kolmogorov-Smirnov

Primenjujemo ga za poznatu neprekidnu raspodelu

Statistika koju koristimo je

$$D_n = \sup_x |F_n^*(x) - F(x)|, \text{ važi } P(\sqrt{n}D_n \leq \lambda) \rightarrow D(\lambda), \text{ za } n \rightarrow \infty, \text{ gde je}$$

$D(\lambda)$ funkcija raspodele Kolmogorov-Smirnov čiji kvantili su $\lambda_{0.95} = 1.36$ i $\lambda_{0.99} = 1.63$.

PRIMER 24 *Za 100 brojeva generisanih pseudo-slučajnim generatorom u intervalu (0,1) testirati da li su uniformno raspoređeni testom Kolmogorov-Smirnov sa pragom značajnosti $\alpha = 0.05$. Ponoviti testiranje 5000 puta. Proveriti u kojem procentu slučajeva hipoteza biva odbaćena.*

```
set.seed(12345); n<-5000; s<-numeric(n);  
for(k in 1:n){s[k]<-ks.test(runif(100),'punif')$p.value};  
sum(s<.05)/n
```

0.0436