20.SE001.SW.Statistika

Softversko inženjerstvo i informacione tehnologije

školska 2022/23

Literatura

[1] Ghilezan et. al., Zbirka rešenih zadataka iz Verovatnoće i statistike, CMS, NS, 2009.

[2] Stojaković M., Matematička statistika, Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Novi Sad, 2000.

[3] Chihara L., Hesterberg T, Mathematical Statistics with Resampling and R by John Wiley & Sons, Ltd

[4] Grbić T., Nedović Lj., Zbirka odabranih rešenih ispitnih zadataka iz Verovatnoće, Statistike i Slučajnih procesa, Novi Sad, Fakultet tehničkih nauka, 2016.

Bodovi i datumi

	Kol. 1	Kol. 2	Test R1	Test R2	Test	Usm.	Σ
MAX	30	20	15	15	10	10	100
MIN	10	8	0	0	0	0	51
Datumi							

Prostor verovatnoće

σ-polje događaja

Definicija 1 *Ako je $\Omega \neq \emptyset$ (neprazan skup ishoda), $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ i važi:*

(i) $\Omega \in \mathcal{F}$,

(ii) $\forall A \in \mathcal{F}, \overline{A} := \Omega \backslash A \in \mathcal{F}$,

(iii) *za prebrojivu familiju* $A_1, A_2, \ldots \in \mathcal{F}, \bigcup\limits_{j=1}^{\infty} A_j \in \mathcal{F}$,

*onda za $\mathcal{F}$ kažemo da je **σ-polje događaja (nad Ω)**.*

Elemente $\mathcal{F}$ nazivamo **događaji**. Ω je **siguran događaj**.
Za događaj A, njegov komplement $\overline{A}$ je **suprotan događaj**.

Primer 1 *Partitivni skup $\mathcal{F}_1 = \mathcal{P}(\Omega)$ i $\mathcal{F}_2 = \{\emptyset, \Omega\}$ su σ-polja događaja nad Ω.*

Primer 2 *Za skup $\Omega = \{1,2,3,4,5,6\}$, $\mathcal{F} = \{\emptyset, \Omega, \{1,2\}, \{3,4,5,6\}\}$ je σ-polje događaja.*

Osobine σ-polja događaja

$\emptyset \in \mathcal{F}$. $A_1, A_2, \ldots, A_n \in \mathcal{F} \;\Rightarrow\; A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{j=1}^{n} A_j \in \mathcal{F}$.

$A, B \in \mathcal{F} \;\Rightarrow\; A \cup B \in \mathcal{F}, A\,B = A \cap B \in \mathcal{F}, A \backslash B \in \mathcal{F}, A \Delta B = (A \backslash B) \cup (B \backslash A) \in \mathcal{F}$.

$A_1, A_2, \ldots \in \mathcal{F} \;\Rightarrow\; \bigcap_{j=1}^{\infty} A_j \in \mathcal{F}$.

Ako je $A\,B = \emptyset$, kažemo da su događaji A i B **disjunktni**.

Ako za prebrojivu familiju $A_1, A_2, \ldots$ važi

$$\forall k, j \in \{1, 2, \ldots\}, k \neq j \Rightarrow A_k A_j = \emptyset,$$

kažemo da su događaji te familije **disjunktni po parovima**.

Verovatnoća

Definicija 2 *Za σ-polje $\mathcal{F}$ nad nepraznim skupom Ω,* ***verovatnoća*** *je funkcija $P : \mathcal{F} \to \mathbb{R}$ koja zadovoljava*

1. $\forall A \in \mathcal{F}, P(A) \geq 0.$
2. *Za prebrojivu familiju događaja disjunktnih po parovima $A_1, A_2, \ldots \in \mathcal{F}$,*
   $P\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} P(A_k).$
3. $P(\Omega) = 1.$

Uređenu trojku $(\Omega, \mathcal{F}, P)$ nazivamo ***prostor verovatnoće****.*

Primer 3 *Za skup $\Omega = \{1,2,3,4,5,6\}$ i $\mathcal{F} = \{\emptyset, \Omega, \{1,2\}, \{3,4,5,6\}\}$, funkcija*

$$P = \begin{pmatrix} \emptyset & \Omega & \{1,2\} & \{3,4,5,6\} \\ 0 & 1 & 1/3 & 2/3 \end{pmatrix}$$

je verovatnoća, odnosno, $(\Omega, \mathcal{F}, P)$ je prostor verovatnoće.

Osobine verovatnoće

Za proizvoljne događaje A i B:

4. $P(A) \leq 1$
5. $P(\emptyset) = 0$
6. $P(\overline{A}) = 1 - P(A)$
7. $P(A \cup B) = P(A) + P(B) - P(AB)$
8. $A \subseteq B \Rightarrow P(A) \leq P(B)$

Za prebrojivu familiju događaja $A_1, A_2, \ldots$ i događaj A

9. $A_k \uparrow A$ ili $A_k \downarrow A \Rightarrow P(A_k) \to P(A)$
10. $P\left(\bigcup_{k=1}^{\infty} A_k\right) \leq \sum_{k=1}^{\infty} P(A_k)$

PRIMER 4 *Neka je skup ishoda konačan:* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, *i neka je* σ*-polje skup svih podskupova* $\mathcal{F} = \mathcal{P}(\Omega)$.

Neka je $p_k = P(\{\omega_k\}) \geq 0$, $k = 1, 2, \ldots, n$ *i* $\sum_{k=1}^{n} p_k = 1$.

Verovatnoća je definisana

$$P(A) = \sum_{k:\omega_k \in A} p_k.$$

Ovaj prostor zovemo ***diskretni prostor verovatnoće.***

PRIMER 5 *Ako u primeru 4 važi i* $p_1 = p_2 = \cdots = p_n = 1/n$, *dobijamo* $P(A) = \#A/\#\Omega$, *to je* ***klasična definicija*** verovatnoće. *U klasičnoj definiciji verovatnoće se kaže da je verovatnoća* broj povoljnih podeljen sa brojem mogućih ishoda. *Klasična definicija verovatnoće se koristi kada imamo konačno mnogo jednako verovatnih ishoda, odnosno, kada se vrši slučajan izbor.*

PRIMER 6 *Neka je* $\Omega \subseteq \mathbb{R}^n$ *Euklidski prostor.*

Neka je $\mathcal{F}$ *skup podskupova od* Ω *koji su merljivi merom* m *i neka je* $m(\Omega) > 0$.

Geometrijsku verovatnoću *za* $A \subseteq \Omega$ *definišemo:* $P(A) = m(A)/m(\Omega)$.

Onda je $(\Omega, \mathcal{F}, P(\cdot))$ *prostor verovatnoće.*

Skup čija verovatnoća je 0 zovemo **nemoguć događaj**. Na primer, $\emptyset$ je nemoguć događaj. Ako podskupovi nemogućeg događaja pripadaju $\mathcal{F}$, kažemo da je prostor **kompletan**. Ako nije kompletan, prostor se može kompletirati proširivanjem.

Ako je $P(A) = 1$, kažemo da je A **skoro siguran skup**. Ako nešto važi na skupu verovatnoće 1, kažemo da **skoro sigurno važi**.

Primer 7 *Tri dečaka i tri devojčice sedaju na slučajan način u red sa 6 mesta. (Svi rasporedi sedenja su jednako verovatni.) Kolika je verovatnoća da nema dve osobe istog pola koje sede jedna do druge?*

Primer 8 *Iz špila od 52 karte na slučajan način se izvlači jedna karta. Kolika je verovatnoća da je izvučena karta dama ili herc?*

Ako u prostoru verovatnoće $(\Omega, \mathcal{F}, P(\cdot))$ za događaje $A, B \in \mathcal{F}$ važi $P(AB) = P(A)P(B)$, kažemo da su A i B **nezavisni**.

Familija događaja $A_1, A_2, \ldots$ je **nezavisna u ukupnosti** ako za proizvoljan skup indeksa $i_1, i_2, \ldots i_k$ važi $P(A_{i_1} A_{i_2} \cdots A_{i_k}) = P(A_{i_1}) P(A_{i_2}) \cdots P(A_{i_k})$.

Primer 9 *Novčić se baca tri puta. Bacanja su nazavisna. Izračunati verovatnoće p_k da će pasti k grbova za $k = 0, 1, 2, 3$.*

Primer 10 *Novčić se baca dok se ne dobije grb. Izračunati ver. da bude paran broj bacanja.*

Primer 11 *(Bernulijeva shema) Pozitivna realizacija eksperimenta u svim pokušajima ima istu verovatnoću $p \in (0,1)$. Eksperiment se vrši n puta. Kolika je verovatnoća da će biti k, za $0 \leq k \leq n$ pozitivnih realizacija?*

Primer 12 *U odeljenju od 30 đaka ima 12 dečaka. Na slučajan način se bira petočlana komisija. Kolika je verovatnoća da u komisiji ima (barem) 2 dečaka?*

Primer 13 *Oko kocke je opisana lopta. Na slučajan način se bira tačka u lopti. Kolika je verovatnoća da je izabrana tačka u kocki?*

Primer 14 *Na slučajan način se biraju brojevi a i b u intervalu $[0,1]$. Kolika je verovatnoća da će jednačina $x^2 + ax + b = 0$ imati realna rešenja?*

Primer 15 *Dve osobe dolaze na sastanak na dogovoreno mesto u slučajno odabranom momentu između 12 i 13 časova. Dogovor je da se čeka 20 minuta.*

Kolika je verovatnoća da će se sresti?

Primer 16 *(Bertranov paradoks) Izračunati verovatnoću da slučajno izabrana tetiva kružnice bude veća od stranice jednakostraničnog trougla upisanog u kružnicu.*

(a) Ako se jedan kraj tetive fiksira, a drugi se bira slučajno.

(b) Ako se fiksira pravac tetive.

(c) Ako se slučajno bira središte tetive (unutar kružnice).

Uslovna verovatnoća

Neka je $(\Omega, \mathcal{F}, P(\cdot))$ prostor verovatnoće i neka je $A \in \mathcal{F}$ i $P(A) > 0$.
Definišemo za $B \in \mathcal{F}$ **verovatnoću pod uslovom da se desio događaj** A:

$$P(B|A) = \frac{P(A\,B)}{P(A)}.$$

Onda je $(\Omega, \mathcal{F}, P(\cdot|A))$ prostor verovatnoće.
Prebrojiva familija događaja $H_1, H_2, \ldots$ čini **potpun sistem događaja** ako su događaji disjunktni po parovima i ako važi $\bigcup_{j=1}^{\infty} H_j = \Omega$.

Formula totalne verovatnoće za potpuni sistem događaja $H_1, H_2, \ldots$ i događaj A:

$$P(A) = \sum_{j=1}^{\infty} P(H_j)\, P(A|H_j).$$

Bejzova (Bayes) formula: $P(H_j|A) = \dfrac{P(H_j)\,P(A|H_j)}{\sum\limits_{j=1}^{\infty} P(H_j)\,P(A|H_j)}$.

PRIMER 17 *Simptom X se pojavljuje usled bolesti A, B i C. Poznato je da se bolest A, B i C pojavljuju kod redom 10%, 5%, 20% populacije. Bolesti A, B i C isključuju jedna drugu. Simptom X se u slučaju bolesti A razvija u 90% slučajeva, u slučaju bolesti B razvija se u 95% slučajeva, i u slučaju bolesti C razvija u 75% slučajeva.*

Kolika je verovatnoća da će se kod slučajno odabranog čoveka pojaviti simptom X?

Ako se pojavio simptom X, kolika je verovatnoća da ima bolest A, B, odnosno C?

PRIMER 18 *Od n novčića jedan je neispravan: ima grb sa obe strane. Na slučajan način se bira novčić i baca k puta. Kolika je verovatnoća da svih k puta padne grb?*

Ako je svih k puta pao grb, kolika je verovatnoća da je u pitanju neispravan novčić?

PRIMER 19 *Osobe A, B, C i D prenose informaciju koju dobiju u obliku iskaza DA ili NE u jednom od tri slučaja. Osoba A dobija informaciju, prenosi je osobi B, zatim ona osobi C, zatim ona osobi D i na kraju osoba D saopštava informaciju.*

Kolika je verovatnoća da je prva osoba prenela početnu informaciju ako se zna da je poslednja osoba prenela početnu informaciju?

Uopštena formula preseka:

$$P(A_1 A_2 \cdots A_n) = P(A_1)\,P(A_2|A_1)\,P(A_3|A_1 A_2) \cdots P(A_n|A_1 A_2 \cdots A_{n-1})$$

Primer 20 *Koliko treba da ima osoba u nekoj grupi pa da verovatnoća da barem dve osobe iz grupe imaju rođendan istog dana bude veća od $\frac{1}{2}$?*

Primer 21 *Koliko osoba treba da pitam za rođendan da bih sreo osobu koja ima rođendan istog dana kad i ja sa verovatnoćom većom od $\frac{1}{2}$?*

Uopštena formula unije:

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{1 \le i_1 \le n} P(A_{i_1}) - \sum_{1 \le i_1 < i_2 \le n} P(A_{i_1} A_{i_2}) + \sum_{1 \le i_1 < i_2 < i_3 \le n} P(A_{i_1} A_{i_2} A_{i_3}) -$$

$$\cdots + (-1)^{(n-1)} P(A_1 A_2 \cdots A_n)$$

Primer 22 *Nestalo je struje u pozorištu i svih n lica su u mraku (na slučajan način) uzeli kaput u garderobi. Kolika je verovatnoća da je barem jedno lice uzelo svoj kaput?*

Kojem broju teži dobijena verovatnoća kad $n \to \infty$?

Slučajne promenljive

Ako $X : \Omega \to \mathbb{R}$, za $S \subseteq \mathbb{R}$, **inverzna sliku** od S je

$$X^{-1}(S) = \{\omega \in \Omega : X(\omega) \in S\}\,.$$

Definicija

Ako je $(\Omega, \mathcal{F}, P)$ prostor verovatnoće i $X : \Omega \to \mathbb{R}$ zadovoljava:

$$\forall x \in \mathbb{R},\; X^{-1}((-\infty, x]) \in \mathcal{F}\,,$$

kažemo da je X **slučajna promenljiva**.

Primer 23 *Za prostor verovatnoće iz primera 3, možemo definisati slučajnu promenljivu koja registruje sa 1 da li je broj veći od 2 (inače je nula):*

$$X = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Vidimo da je $X^{-1}((-\infty, x]) = \begin{cases} \emptyset, & x < 0 \\ \{1,2\}, & 0 \le x < 1 \\ \{1,2,3,4,5,6\}, & x \ge 1 \end{cases}$

Funkcija raspodele

Za slučajnu promenljivu X nad $(\Omega, \mathcal{F}, P)$ definišemo **funkciju raspodele**:

$$F(x) = P\left(X^{-1}((-\infty, x])\right) = P(\{\omega : X(\omega) \leq x\}) = P(X \leq x)\,.$$

Primer 24 *Za slučajnu promenljivu X iz primera 23 funkcija raspodele je*

$$F(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{3}, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases}$$

Osobine funkcije raspodele

1. $\lim\limits_{x \to -\infty} F(x) = 0$
2. $\lim\limits_{x \to \infty} F(x) = 1$
3. $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$
4. $\forall a \in \mathbb{R}, \lim\limits_{x \to a^+} F(x) = F(a)$
5. $P(a < X \leq b) = F(b) - F(a)$

Diskretne slučajne promenljive

Neka je X slučajna promenljiva nad $(\Omega, \mathcal{F}, P)$. Sliku skupa ishoda Ω označavamo sa $\mathcal{R}_X$. Kažemo da je X **diskretna slučajna promenljiva** ako je slika $\mathcal{R}_X$ konačan ili prebrojiv skup. Ako je $\mathcal{R}_X = \{x_1, x_2, \ldots\}$ onda $\Omega = \sum\limits_{n=1}^{\infty} (X = x_n)$ sledi da je $1 = \sum\limits_{n=1}^{\infty} P(X = x_n)$. Možemo uvesti oznake $p_n = P(X = x_n)$. Funkciju $x_n \mapsto p_n$ zovemo **zakon raspodele** slučajne promenljive X i zapisujemo $X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$. Važi $F(x) = \sum\limits_{n: x_n \leq x} p_n$.

Bernulijeva raspodela sa parametrom $p \in (0,1)$ je $X : \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$.

Binomna raspodela sa parametrima $p \in (0,1)$ i $n \in \mathbb{N}$ u oznaci $\mathcal{B}(n,p)$:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k \in \{0, 1, \ldots, n\},.$$

Poasonova raspodela sa parametrom $\lambda \in (0, \infty)$ u oznaci $\mathcal{P}(\lambda)$:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, k = 0, 1, \ldots$$

Geometrijska raspodela sa parametrom $p \in (0,1)$ u oznaci $\mathcal{G}(p)$:

$$P(X = k) = p(1-p)^{k-1}, k = 1, 2, \ldots$$

Neprekidne slučajne promenljive

Kažemo da je slučajna promenljiva **apsolutno neprekidnog tipa** ako postoji funkcija $\varphi : \mathbb{R} \to [0, \infty)$ takva da je $\forall x \in \mathbb{R}$, $F(x) = \int\limits_{-\infty}^{x} \varphi(t)dt$.

Funkciju φ nazivamo **funkcija gustine raspodele** slučajne promenljive X.

Osobine gustine raspodele

1. Ako je φ neprekidno u x, onda $\varphi(x) = F'(x)$
2. $\int\limits_{-\infty}^{\infty} \varphi(t)dt = 1$
3. $P(a < X < b) = P(a \leq X < b) = P(a \leq X \leq b) = F(b) - F(a) = \int_a^b \varphi(x)dx$

Uniformna raspodela $\mathcal{U}(a,b)$, $a < b \in \mathbb{R}$ ima gustinu i funkciju raspodele

$$\varphi(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b), \\ 0, & x \notin (a,b); \end{cases} \qquad F(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x < b, \\ 1, & x \geq b. \end{cases}$$

Eksponencijalna raspodela $\mathcal{E}(\lambda)$, $\lambda \in (0,\infty)$

$$\varphi(x) = \begin{cases} 0, & x < 0, \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

Normalna (Gausova) raspodela $\mathcal{N}(m,\sigma)$, $m,\sigma \in \mathbb{R}$, $\sigma > 0$

$$\varphi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \, x \in \mathbb{R}; \qquad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt.$$

Specijalno, za $m = 0$ i $\sigma = 1$, vrednosti funkcije raspodele $\mathcal{N}(0,1)$ možemo očitati iz tablica sa kraja knjige.

Funkciju koja je tabelirana obeležavamo $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$.

Primer 25 *Ako $X : \mathcal{N}(0,1)$, iz tablica funkcije Φ očitati vrednosti $P(X \leq 0.55), P(X > 1), P(|X| < 2)$ i naći vrednost x za koju je $P(X \leq x) = F(x) = 0.975$.*

Transformacija slučajne promenljive

Neka je $X : \Omega \to \mathbb{R}$ slučajna promenljiva nad prostorom verovatnoće $(\Omega, \mathcal{F}, P)$.

Neka $f : \mathbb{R} \to \mathbb{R}$ i neka je $Y = f \circ X$ slučajna promenljiva. $(\omega \in \Omega, Y(\omega) = f(X(\omega)))$

Ako je f neprekidna funkcija, onda je Y slučajna promenljiva.

Obeležavamo $Y = f(X)$.

Zadatak je naći raspodelu slučajne promenljive Y kada je poznata raspodela slučajne promenljive X i funkcija f.

X je diskretna slučajna promenljiva

Aka je X diskretna slučajna promenljiva sa raspodelom $X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$, onda je Y diskretna slučajna promenljiva.

Neka je $\{y_1, y_2 \ldots\}$ skup slika za Y. Onda je zakon raspodele za Y:

$$Y : \begin{pmatrix} y_1 & y_2 & \cdots \\ q_1 & q_2 & \cdots \end{pmatrix}, \text{ gde je } q_i = \sum_{\substack{m \\ y_i = f(x_m)}} p_m.$$

Primer 26 *Neka je X slučajna promenljiva koja predstavlja broj koji padne na kockici za igru. Naći raspodelu slučajne promenljive $Y = (X-3)^2$.*

$$X : \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}, \; Y : \begin{pmatrix} 0 & 1 & 4 & 9 \\ \frac{1}{6} & \frac{2}{6} & \frac{2}{6} & \frac{1}{6} \end{pmatrix}.$$

X je neprekidna slučajna promenljiva

Neka je X slučajna promenljiva sa gustinom raspodele φ_X. Neka je f rastuća ili opadajuća neprekidna funkcija. Onda je inverzna funkcija f^{-1} rastuća, odnosno, opadajuća funkcija. Onda je gustina za Y: $\varphi_Y(y) = \varphi_X(f^{-1}(y))\,|(f^{-1}(y))'|$, $y \in Y(\Omega) = f(X(\Omega))$.

PRIMER 27 *Neka $X : \mathcal{U}(0,1)$ i neka je $Y = -\ln X$. Naći raspodelu za Y.*

PRIMER 28 *Neka $X : \mathcal{N}(0,1)$ i neka je $Y = aX + b$, $a \neq 0$. Naći raspodelu za Y.*

PRIMER 29 *Neka $X : \mathcal{N}(0,1)$ i neka je $Y = X^2$. Naći raspodelu za Y.*

PRIMER 30 *Neka $X : \mathcal{N}(0.5,2)$. Izračunati verovatnoće $P(X \leq 0.55), P(X > 1), P(|X| < 2)$ i naći vrednost x za koju je $P(X \leq x) = F(x) = 0.975$.*

Dvodimenzionalne slučajne promenljive

Kažemo da je (X,Y) **dvodimenzionalna slučajna promenljiva**, odnosno **dvodimenzional-ni slučajni vektor**, nad prostorom verovatnoće $(\Omega, \mathcal{F}, P)$, ako su X i Y slučajne promenljive nad $(\Omega, \mathcal{F}, P)$.

Funkcija raspodele slučajnog vektora (X,Y) je

$$F(x,y) = P(\{\omega : X(\omega) \le x\} \cap \{\omega : Y(\omega) \le y\}) = P(X \le x, Y \le y)\,.$$

Osobine funkcije raspodele dvodimenzionalnog vektora

1. $F(-\infty,y) = F(x,-\infty) = 0$
2. $F(\infty,\infty) = 1$
3. $F(x,y)$ je neprekidna s desna po obe promenljive.
4. $F(x,y)$ je neopadajuća po obe promenljive.
5. $P(a < X \le b, c < Y \le d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$.

Marginalne raspodele slučajnog vektora (X,Y) su raspodele slučajnih promenljivih X i Y čije funkcije raspodele dobijamo:

$$F_X(x) = F(x,\infty)\,,\ F_Y(y) = F(\infty,y)\,.$$

Diskretna dvodimenzionalna slučajna promenljiva

Neka je (X,Y) slučajni vektor nad $(\Omega,\mathcal{F},P)$.
Neka je $\mathcal{R}_{(X,Y)} = (X,Y)(\Omega) \subseteq \mathbb{R}^2$ prebrojiv skup.
Preslikavanje koje elementima slike $\mathcal{R}_{(X,Y)}$ pridružuje verovatnoće $(x_i,y_j) \mapsto p_{i,j}$ definisano:

$$p_{i,j} = P(\{\omega : X(\omega) = x_i \wedge Y(\omega) = y_j\})$$

zovemo **zakon raspodele vektora** (X,Y).
Često zakon raspodele zadajemo tabelom na čijim marginama možemo izračunati verovatnoće marginalnih raspodela. $p_{i\cdot} = \sum_j p_{i,j}, \quad p_{\cdot j} = \sum_i p_{i,j}, \quad i,j = 1,2,\dots$

$X\backslash Y$	y_1	y_2	$\cdots$	
x_1	$p_{1,1}$	$p_{1,2}$	$\cdots$	$p_{1\cdot}$
x_2	$p_{2,1}$	$p_{2,2}$	$\cdots$	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$p_{\cdot 1}$	$p_{\cdot 2}$	$\cdots$	

$$X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_{1\cdot} & p_{2\cdot} & \cdots \end{pmatrix}, \; Y : \begin{pmatrix} y_1 & y_2 & \cdots \\ p_{\cdot 1} & p_{\cdot 2} & \cdots \end{pmatrix}.$$

Primer 31 *Tri puta se baca novčić. Neka X predstavlja broj grbova, a Y broj promena. Naći zakon raspodele slučajnog vektora (X,Y) i marginalne zakone raspodele.*

Neprekidna dvodimenzionalna slučajna promenljiva

Neka je $(\Omega, \mathcal{F}, P)$ prostor verovatnoće i (X, Y) slučajni vektor sa funkcijom raspodele $F(x,y)$. Ako postoji integrabilna funkcija $\varphi(x,y) : \mathbb{R}^2 \to [0,\infty]$ takva da je $\forall (x,y) \in \mathbb{R}^2$

$$F(x,y) = \iint\limits_{D_{x,y}} \varphi(u,v)\,du\,dv, \text{ gde je } D_{x,y} = (-\infty, x] \times (-\infty, y]\,,$$

kažemo da je (X, Y) **neprekidna dvodimenzionalna slučajna promenljiva** i da je $\varphi(x,y)$ njena **gustina raspodele**.

Poslednji dvostruki integral se računa preko ponovljenih (nesvojstvenih) integrala:

$$F(x,y) = \int\limits_{-\infty}^{x} \left(\int\limits_{-\infty}^{y} \varphi(u,v)dv \right) du\,.$$

Osobine dvodimenzionalne gustine raspodele

1. Ako je $\varphi(x,y)$ neprekidna u (x,y), onda je $\varphi(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}$.

2. $\iint\limits_{\mathbb{R}^2} \varphi(x,y)\,dxdy = 1$

3. $P((X,Y) \in S) = \iint\limits_{S} \varphi(x,y)\,dxdy$

Gustine marginalnih raspodela slučajnog vektora (X,Y) su

$$\varphi_X(x) = \int_{-\infty}^{\infty} \varphi(x,y)\,dy\,, \qquad \varphi_Y(y) = \int_{-\infty}^{\infty} \varphi(x,y)\,dx\,.$$

Uslovne raspodele

Neka je (X,Y) slučajni vektor nad prostorom verovatnoće $(\Omega,\mathcal{F},P)$ i neka je $P(X \in S) > 0$. Uslovna funkcija raspodele slučajne promenljive $Y|X \in S$ u $(\Omega,\mathcal{F},P(\cdot|X \in S))$ je

$$F_{Y|X\in S}(y) = P(\{\omega : Y(\omega) \le y\}|X \in S) = \frac{P(Y \le y, X \in S)}{P(X \in S)}.$$

Diskretne uslovne raspodele

Neka je dat zakon raspodele diskretnog slučajnog vektora (X,Y). Ako je $p_{\cdot j} > 0$, **uslovni zakon raspodele** slučajne promenljive X ako je $Y = y_j$ je $p(x_i|y_j) = p_{i,j}/p_{\cdot j}$:

$$X|Y = y_j : \begin{pmatrix} x_1 & x_2 & \cdots \\ \frac{p_{1,j}}{p_{\cdot j}} & \frac{p_{2,j}}{p_{\cdot j}} & \cdots \end{pmatrix} \text{ slično } Y|X = x_i : \begin{pmatrix} y_1 & y_2 & \cdots \\ \frac{p_{i,1}}{p_{i\cdot}} & \frac{p_{i,2}}{p_{i\cdot}} & \cdots \end{pmatrix}, \; p(y_j|x_i) = \frac{p_{i,j}}{p_{i\cdot}}$$

Primer 32 *Za diskretnu slučajnu promenljivu iz prethodnog primera naći uslovne zakone raspodele $Y|X = 2$ i $X|Y = 1$.*

Neprekidne uslovne raspodele

Za neprekidni slučajni vektor (X,Y) sa gustinom $\varphi(x,y)$ funkciju raspodele slučajne promenljive $Y|X = x$ definišemo:

$$F_{Y|X=x}(y) = \lim_{h\to 0^+} P(Y \leq y | x \leq X < x+h)\,.$$

Može se dokazati da je

$$F_{Y|X=x}(y) = \int_{-\infty}^{y} \frac{\varphi(x,v)}{\varphi_X(x)}\, dv\,, \quad \text{za} \quad \varphi_X(x) > 0, \ \text{ odnosno,}$$

$\varphi(y|x) := \frac{\varphi(x,y)}{\varphi_X(x)}$ za $\varphi_X(x) > 0$ je **uslovna gustina raspodele** za $Y|X = x$.

Primer 33 *X se na slučajan način bira iz intervala $(0,1)$. Potom se Y bira na slučajan način iz intervala $(X,1)$. Naći gustinu raspodele za (X,Y) i marginalnu raspodelu za Y.*

Nezavisnost slučajnih promenljivih

Neka je (X,Y) slučajni vektor nad prostorom $(\Omega,\mathcal{F},P)$ sa funkcijom raspodele $F(x,y)$ i neka su $F_X(x)$ i $F_Y(y)$ marginalne funkcije raspodele.

Kažemo da su X i Y **nezavisne** slučajne promenljive ako $\forall x,y \in \mathbb{R}$, $F(x,y) = F_X(x)\, F_Y(y)$.

Diskretne slučajne promenljive X i Y su nezavisne ako i samo ako je $p_{i,j} = p_{i\cdot}\ p_{\cdot j}$ za sve i,j, gde je $p_{i,j}$ zajednički zakon raspodele za (X,Y), a $p_{i\cdot}$ i $p_{\cdot j}$ marginalni zakoni raspodele.

Neprekidne slučajne promenljive X i Y su nezavisne ako i samo ako je $\varphi(x,y) = \varphi_X(x)\, \varphi_Y(y)$ za sve x,y, gde je $\varphi(x,y)$ zajednička gustina, a $\varphi_X(x)$ i $\varphi_Y(y)$ marginalne gustine.

Ako za niz promenljivih $X_1, X_2, \ldots$ kažemo da su **nezavisne** ako $\forall i, j,\ i \neq j \Rightarrow X_i$ nezavisno od X_j.

Transformacija dvodimenzionalne slučajne promenljive

Posmatraćemo transformacije $\mathbb{R}^2 \to \mathbb{R}^2$ i $\mathbb{R}^2 \to \mathbb{R}$.

Diskretne slučajne promenljive

Ako je (X,Y) diskretni slučajni vektor sa zakonom raspodele $(x_i, y_j) \mapsto p_{i,j}$ i transformacija $(X,Y) \mapsto (U,Z) = (f_1(X,Y), f_2(X,Y))$, onda je zakon raspodele $(U,Z):\ (u_k, z_l) \mapsto q_{k,l}$,

$$q_{k,l} = P(U = u_k, Z = z_l) = \sum_{\substack{(i,j) \\ (u_k,z_l)=(f_1(x_i,y_j),f_2(x_i,y_j))}} p_{i,j},$$

za sve vrednosti (u_k, z_l) iz slike $\mathcal{R}_{(U,Z)}$.
Slično ako je $Z = f(X,Y)$.

Primer 34 *Nezavisne slučajne promenljive X i Y imaju istu Poasonovu raspodelu $\mathcal{P}(\lambda)$. Naći raspodelu slučajne promenljive $Z = X + Y$.*

Primer 35 *Neka su slučajne promenljive $X_1, X_2, \ldots, X_n$ nezavisne sa istom Bernulijevom raspodelom $\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$. Naći raspodelu slučajne promenljive $Y = X_1 + X_2 + \cdots + X_n$.*

Neprekidne slučajne promenljive

Nalaženje raspodele transformisane slučajne promenljive ćemo pokazati za $Z = X + Y$ (tzv. konvolucija slučajnih promenljivih X i Y).

Neka je (X, Y) dvodimenzionalni neprekidni slučajni vektor sa gustinom raspodele $\varphi(x, y)$.

Gustina raspodele slučajne promenljive $Z = X + Y$ je $\varphi_Z(z) = \int\limits_{-\infty}^{\infty} \varphi(z - y, y)\, dy$.

Numeričke karakteristike slučajnih promenljivih

Matematičko očekivanje

Za diskretnu slučajnu promenljivu $X : \begin{pmatrix} x_1 & x_2 & \cdots \\ p_1 & p_2 & \cdots \end{pmatrix}$, definišemo $E(X) = \sum\limits_{i=1}^{\infty} x_i p_i$.

Za neprekidnu slučajnu promenljivu $X : \varphi(x)$, definišemo $E(X) = \int\limits_{-\infty}^{\infty} x\,\varphi(x)\,dx$.

(Ako suma, odnosno integral, apsolutno konvergira.)

Osobine matematičkog očekivanja

1. $X = c,\ c = const \;\Rightarrow\; E(X) = c$
2. $Y = f(X),\ E(Y) = E(f(X)) = \begin{cases} \sum\limits_{i=1}^{\infty} f(x_i)\,p_i, & X \text{ diskretna} \\ \int\limits_{-\infty}^{\infty} f(x)\,\varphi(x)\,dx, & X \text{ neprekidna} \end{cases}$

3. $Z = f(X,Y)$, $E(Z) = \begin{cases} \sum\limits_{i=1}^{\infty}\sum\limits_{j=1}^{\infty} f(x_i,y_j)\, p_{i,j}, & (X,Y) \text{ diskretna} \\ \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} f(x,y)\, \varphi(x,y)\, dy \right) dx, & (X,Y) \text{ neprekidna} \end{cases}$

4. $E(X+Y) = E(X) + E(Y)$

5. $E(cX) = c\,E(X)$

6. X i Y nezavisne $\Rightarrow$ $E(XY) = E(X)\,E(Y)$

7. $a \leq X \leq b \Rightarrow a \leq E(X) \leq b$

Disperzija (Varijansa)

$$D(X) = E\left((X - E(X))^2\right) = E\left(X^2\right) - (E(X))^2$$

(Ako postoji.)

Osobine disperzije

1. $X = c,\ c = const \ \Leftrightarrow \ D(X) = 0$

2. $D(X) \geq 0$

3. $c = const \ \Rightarrow \ D(cX) = c^2 D(X),\ D(X + c) = D(X)$

4. X i Y nezavisne $\ \Rightarrow \ D(X + Y) = D(X) + D(Y)$

Standardna devijacija

$$\sigma(X) = \sqrt{D(X)}$$

Standardizacija slučajne promenljive

$$X^* = \frac{X - E(X)}{\sigma(X)} \ \Rightarrow \ E(X^*) = 0,\ D(X^*) = 1$$

Medijana

$$P(X < Me) = P(X > Me)$$

Modus

X diskretna $\Rightarrow$ Mo je vrednost sa najvećom verovatnoćom.
X neprekidna $\Rightarrow$ Mo je vrednost za koju gustina dostiže maksimum.

Momenti

Obični: $m_k(X) = E(X^k) = \begin{cases} \sum\limits_{i=1}^{\infty} x_i^k p_i, & X \text{ diskretna} \\ \int\limits_{-\infty}^{\infty} x^k \varphi(x)\, dx, & X \text{ neprekidna} \end{cases}$

Centralni: $\mu_k(X) = m_k(X - E(X)) = E\left((X - E(X))^k\right)$.

(Ako postoji.)

Numeričke karakteristike dvodimenzionalne slučajne promenljive

Očekivanje i disperzija: $E(X,Y) = (E(X), E(Y))$, $D(X,Y) = (D(X), D(Y))$

Mešoviti momenti:

$$m_{k,n}(X,Y) = E(X^k Y^n) = \begin{cases} \sum\limits_{i=1}^{\infty} \sum\limits_{j=1}^{\infty} x_i^k y_j^n \, p_{i,j}, & (X,Y) \text{ diskretna} \\ \int\limits_{-\infty}^{\infty} \left(\int\limits_{-\infty}^{\infty} x^k y^n \, \varphi(x,y) \, dy \right) dx, & (X,Y) \text{ neprekidna} \end{cases}$$

Kovarijansa: $\operatorname{cov}(X,Y) = E\left((X - E(X))\,(Y - E(Y))\right) = E(XY) - E(X)\,E(Y)$

Koeficijent korelacije:

$$\rho_{X,Y} = \operatorname{cov}(X^*, Y^*) = \operatorname{cov}\left(\frac{X - E(X)}{\sqrt{D(X)}}, \frac{Y - E(Y)}{\sqrt{D(Y)}} \right) = \frac{E(XY) - E(X)\,E(Y)}{\sqrt{D(X)\,D(Y)}}$$

Osobine:

1. X i Y nezavisne $\Rightarrow \rho_{X,Y} = 0$
2. $|\rho_{X,Y}| \leq 1$
3. $|\rho_{X,Y}| = 1 \Leftrightarrow Y = aX + b$, $a, b \in \mathbb{R}$, $a \neq 0$

Regresija

Za diskretnu dvodimenzionalnu slučajnu promenljivu (X,Y) definišemo **uslovno matematičko očekivanje** za X ako je $Y = y_j$: $E(X|Y = y_j) = \sum_{i=1}^{n} x_i\, p(x_i|y_j) = \frac{1}{p_{\cdot j}} \sum_{i=1}^{n} x_i\, p_{i,j}$

Funkciju $y_j \mapsto E(X|Y = y_j)$ nazivamo **regresija** X po Y, obeležavamo $r_1(y)$. Regresija definiše novu slučajnu promenljivu koju obeležavamo $E(X|Y) = r_1(Y)$, čija je raspodela:

$E(X|Y) : \begin{pmatrix} E(X|Y = y_1) & E(X|Y = y_2) & \cdots \\ p_{\cdot 1} & p_{\cdot 2} & \cdots \end{pmatrix}$. Važi $E(E(X|Y)) = E(X)$.

Za neprekidnu dvodimenzionalnu slučajnu promenljivu (X,Y) definišemo **uslovno matematičko očekivanje** za X ako je $Y = y$: $E(X|Y = y) = \int\limits_{-\infty}^{\infty} x\, \varphi(x|y)\, dx = \frac{1}{\varphi_Y(y)} \int\limits_{-\infty}^{\infty} x\, \varphi(x,y)\, dx$

Funkciju $y \mapsto E(X|Y = y)$ nazivamo **regresija** X po Y, obeležavamo $r_1(y)$.
Regresija definiše novu slučajnu promenljivu koju obeležavamo $E(X|Y) = r_1(Y)$.
Važi $E(E(X|Y)) = E(X)$.

Primer 36 *Naći regresiju X po Y za primer 31.*

Granične teoreme

Nejednakost Čebiševa

Neka za slučajnu promenljivu X postoji $E(X^2)$ i neka je $\varepsilon > 0$. Onda $P(|X| \geq \varepsilon) \leq \dfrac{E(X^2)}{\varepsilon^2}$.

Ako za slučajnu promenljivu X postoji $D(X)$, onda za $\varepsilon > 0$, $P(|X - E(X)| \geq \varepsilon) \leq \dfrac{D(X)}{\varepsilon^2}$.

Primena: Ako $X : \mathcal{B}(n, p)$, $\varepsilon > 0$, onda $P(|X - np| \geq \varepsilon) \leq \dfrac{np(1-p)}{\varepsilon^2}$.

Zakoni velikih brojeva

Ako je $X_1, X_2, \ldots$ niz nezavisnih slučajnih promenljivih sa istom Bernulijevom raspodelom $\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$ i $\varepsilon > 0$, onda

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{k=1}^{n} X_k - p\right| \geq \varepsilon\right) = 0. \qquad \text{(Bernulijev slabi zakon velikih brojeva)}$$

Za niz slučajnih promenljivih $X_1, X_2, \ldots$ kažemo da važi

- **slabi zakon velikih brojeva** ako $\forall \varepsilon > 0,\ \lim\limits_{n\to\infty} P\left(\left|\frac{1}{n}\sum\limits_{k=1}^{n}(X_k - E(X_k))\right| > \varepsilon\right) = 0,$

- **jaki zakon velikih brojeva** ako $P\left(\lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{k=1}^{n}(X_k - E(X_k)) = 0\right) = 1.$

Za niz nezavisnih slučajnih promenljivih sa istom raspodelom i konačnim očekivanjem važi slabi zakon velikih brojeva. (**Hinčin**)

Ako postoji konstanta $C > 0$ tako da za niz nezavisnih slučajnih promenljivih $X_1, X_2, \ldots$ važi $D(X_k) < C, k = 1, 2, \ldots$, onda za taj niz važi slabi zakon velikih brojeva. (**Čebišev**)

Za niz nezavisnih slučajnih promenljivih $X_1, X_2, \ldots$ sa Bernulijevom raspodelom $\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$ važi jaki zakon velikih brojeva: $P\left(\left\{\omega : \lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{k=1}^{n} X_k(\omega) = p\right\}\right) = 1.$ (**Bernuli, Borel**)

Za niz nezavisnih, jednako raspoređenih, slučajnih promenljivih sa konačnim očekivanjem važi jaki zakon velikih brojeva. (**Kolmogorov**)

Normalna raspodela

Primer 37 *Naći očekivanje i varijansu za normalnu raspodelu* $\mathcal{N}(m,\sigma)$.

$X:\mathcal{N}(m,\sigma) \Leftrightarrow \frac{X-m}{\sigma}:\mathcal{N}(0,1)$

X i Y nezavisne i $X:\mathcal{N}(m_1,\sigma_1)$, $Y:\mathcal{N}(m_2,\sigma_2) \Rightarrow X\pm Y:\mathcal{N}\left(m_1\pm m_2,\sqrt{\sigma_1^2+\sigma_2^2}\right)$

Primer 38 *U jednoj školi težina dečaka [kg] ima raspodelu:* $X:\mathcal{N}(50,2.5)$, *a devojčice:* $Y:\mathcal{N}(45,3)$. *Na slučajan način je odabran dečak i, nezavisno, devojčica. Kolika je verovatnoća da će dečak imati barem 3 kg više od devojčice?*

Ako nezavisne slučajne promenljive imaju raspodelu $X_k:\mathcal{N}(m_k,\sigma_k), k=1,2,\ldots,n$, onda slučajna promenljive $Y=a_1X_1+a_2X_2+\cdots+a_nX_n$ ima normalnu raspodelu $\mathcal{N}(m,\sigma)$, gde je $m=a_1m_1+a_2m_2+\cdots+a_nm_n$ i $\sigma^2=a_1^2\sigma_1^2+a_2^2\sigma_2^2+\cdots+a_n^2\sigma_n^2$.

Centralne granične teoreme

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

$$\lim_{n\to\infty} P\left(\frac{\sum\limits_{k=1}^{n} X_k - E\left(\sum\limits_{k=1}^{n} X_k\right)}{\sqrt{D\left(\sum\limits_{k=1}^{n} X_k\right)}} \le x\right) = \lim_{n\to\infty} P\left(\frac{\sum\limits_{k=1}^{n} X_k - n\,a}{s\sqrt{n}} \le x\right) = \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt,$$

kažemo da za $X_1, X_2, \ldots$ važi **centralna granična teorema**.

Ako su $X_1, X_2, \ldots$ nezavisne, $E(X_k) = a_k$ i $D(x_k) = s_k^2$, i $\lim\limits_{n\to\infty} \frac{\max_k s_k^2}{\sum_{k=1}^{n} s_k^2} = 0$, onda važi CGT.

Posledica: $S_n : \mathcal{B}(n,p) \;\Rightarrow\; \lim\limits_{n\to\infty} P\left(\frac{S_n - n\,p}{\sqrt{n\,p\,(1-p)}} \le x\right) = \int\limits_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$. (**Moavr-Laplas**)

PRIMER 39 *Kolika je verovatnoća da je broj grbova u 100 bacanja novčića između 40 i 60?*

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

Statistika

Važne raspodele

χ_n^2 raspodela (hi kvadrat sa n stepeni slobode)

$X : \chi_n^2$ ima gustinu $\varphi(x) = \dfrac{x^{n/2-1} e^{-x/2}}{2^{n/2}\Gamma(n/2)}$, $x > 0$, $E(X) = n$, $D(X) = 2n$

Osobina: $X : \chi_n^2$ i $Y : \chi_m^2$ nezavisne, onda $X + Y : \chi_{n+m}^2$

$$\int_{-\infty}^{\infty} \varphi(z-y,y)\,dy = \int_{-\infty}^{0} + \int_{0}^{z} + \int_{z}^{\infty} = \int_0^z \frac{(z-y)^{n/2-1} e^{-(z-y)/2}}{2^{n/2}\Gamma(n/2)} \frac{y^{m/2-1} e^{-y/2}}{2^{m/2}\Gamma(m/2)}\,dy =$$

$$= \frac{z^{(n+m)/2-1} e^{-z/2}}{2^{(n+m)/2}\Gamma(n/2)\Gamma(m/2)} \int_0^z (1-\frac{y}{z})^{n/2-1} \left(\frac{y}{z}\right)^{m/2-1} \frac{1}{z}\,dy = \frac{z^{(n+m)/2-1} e^{-z/2}}{2^{(n+m)/2}\Gamma((n+m)/2)},\ z > 0$$

jer je za $x > 0$, $y > 0$, $B(x,y) = \displaystyle\int_0^1 (1-t)^{x-1} t^{y-1}\,dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$.

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

Napomena: Za $n = 1$ smo radili kao transformaciju, za $n = 2$ imamo $\mathcal{E}(\frac{1}{2})$.

t_n **raspodela (Studentova sa n stepeni slobode)**

$T : t_n$ ima gustinu $\varphi(t) = \dfrac{\Gamma((n+1)/2)}{\sqrt{n\pi}\;\Gamma(n/2)\,(1+t^2/n)^{(n+1)/2}}$

Drugi zapis $\varphi(t) = \left(\sqrt{n}\,B\left(\frac{1}{2},\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{(n+1)/2}\right)^{-1}$

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

$$E(T) = 0, \qquad D(T) = \frac{n}{n-2}, n > 2$$

$F_{m,n}$ **raspodela (Fišerova sa m,n stepeni slobode)**

$X : F_{m,n}$ ima gustinu $\varphi(x) = \dfrac{\Gamma((m+n)/2)\,m^{m/2}\,n^{n/2}}{\Gamma(m/2)\,\Gamma(n/2)} \cdot \dfrac{x^{m/2-1}}{(mx+n)^{(m+n)/2}}$, za $x > 0$.

$$X : F_{m,n} \;\Rightarrow\; E(X) = \frac{m}{n-2}, n > 2, \quad D(X) = \frac{2n^2\,(m+n-2)}{m\,(n-2)^2\,(n-4)}, n > 4$$

PRIMER 40 *Neka $T : t_{10}$ i $Y : \chi^2_4$. Naći vrednost za koju je: $P(Y < y_1) = 0.9$, $P(Y > y_2) = 0.95$, $P(T < t_1) = 0.95$, $P(T > t_2) = 0.25$, $P(|T| < t_3) = 0.975$.*

PRIMER 41 *Neka $F : F_{9,15}$. Naći vrednost za koju je $P(F > f_1) = 0.05$ i $P(F < f_2) = 0.99$.*

Statistika, osnovni pojmovi

Populacija je skup svih elemenata koje ispitujemo.

Obeležje je numerička karakteristika elementa. Modeliramo ga slučajnom promenljivom.

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

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

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

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

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

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

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

Važne statistike

Aritmetička sredina uzorka

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

$$E(\bar{X}_n) = E(X) \qquad D(\bar{X}_n) = \frac{1}{n}D(X)$$

$$X : \mathcal{N}(m,\sigma) \;\Rightarrow\; \bar{X}_n : \mathcal{N}\left(m, \frac{\sigma}{\sqrt{n}}\right)$$

Uzoračka disperzija (varijansa)

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

$$E(\bar{S}_n^2) = \frac{n-1}{n}D(X)$$

$$X : \mathcal{N}(m,\sigma) \;\Rightarrow\; \frac{n\bar{S}_n^2}{\sigma^2} : \chi^2_{n-1}$$

Za obeležje sa Normalnom raspodelom $X : \mathcal{N}(m, \sigma)$

Statistika $T = \frac{\bar{X}_n - m}{\bar{S}_n}\sqrt{n-1}$ ima Studentovu t_{n-1} raspodelu.

Za X **i** Y **nezavisne slučajne promenljive sa raspodelom** $X : \chi^2_m$**,** $Y : \chi^2_n$

Slučajna promenljiva $F = \frac{\frac{X}{m}}{\frac{Y}{n}}$ ima Fišerovu $F_{m,n}$ raspodelu

Uzorački momenti

Za uzorak $(X_1, X_2, \ldots, X_n)$ definišemo **momenat reda** r kao $M_r = \frac{1}{n}\sum_{k=1}^{n} X_k^r$,

centralni momenat reda r: $\mu_r = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X}_n)^r$

Intervalni uzorak

Intervalni uzorak nastaje grupisanjem elemenata početnog uzorka u intervale I_i.

Ako imamo granice intervala I_i, odnosno deobene tačke m_i, $i = 0,1,\ldots k$ i broj elemenata uzorka u intervalu i: **frekvencije** f_i, $i = 1,2,\ldots k$, 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 standardne devijacije intervalnog uzorka sa sredinama x_i, $i = 1,2,\ldots,k$ i frekvencijama f_i, $i = 1,2,\ldots,k$ su

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

Primer: 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

U tabelu dodajemo: sredine, širine intervala, kumulativne i korigovane frekvencije.

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
$\sum f_i$	15	26	33	40	46	50
$\bar{f}_i$	15	11	7	3.5	1.2	0.4

Histogram i poligon

Neka interval $(a,b) \subseteq \mathbb{R}$ sadrži sve vrednosti obeležja X. Taj interval delimo tačkama $a := m_0 < m_1 < \cdots < m_k =: b$ na k podintervala: $I_1 = [m_0, m_1]$, $I_2 = (m_1, m_2], \ldots, I_k = (m_{k-1}, m_k]$. **Širina** intervala I_i je $h_i := m_i - m_{i-1}$, a **frekvencija** f_i je broj elemenata u intervalu I_i

Nad svakim od podintervala I_i, $i \in \{1,2,\ldots,k\}$ nacrtamo pravougaonik visine $\bar{f}_i = \frac{f_i}{h_i}$, gde je f_i frekvencija, a h_i širina i-tog intervala. Dobili smo **histogram** realizovanog uzorka.

Neka je x_i sredina intervala I_i, $i \in \{1,2,\ldots,k\}$. Neka je x_0 tačka na x-osi koja je od a manja za onoliko koliko je x_1 veća od a i neka je x_{k+1} tačka na x-osi koja je od b veća za onoliko koliko je x_k manja od b. Izlomljenu linija koja polazi od x_0, prolazi kroz tačke $(x_i, \frac{f_i}{h_i})$ i završava u tački x_{k+1} nazivamo **poligonom** realizovanog uzorka.

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 $\mathsf{P}(X < Me) = \mathsf{P}(X > Me)$. 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 $(x_1, x_2, \ldots, x_n)$.

Kvantili (percentili)

Za slučajnu pormenljivu 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 $(x_1, x_2, \ldots, x_n)$ 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 medijane i brkovima na $Q_1 - 1.5\,IQR$ i $Q_3 + 1.5\,IQR$ ili minimumu i maksimumu uzorka.

Tačkaste ocene parametara

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

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

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

Ocenjivač U je **centriran** za $\tau(\theta)$ ako $E(u(X_1, X_2, \ldots, X_n)) = \tau(\theta)$, a **asimptotski centriran** ako $\lim\limits_{n\to\infty} E(u(X_1, X_2, \ldots, X_n)) = \tau(\theta)$.

Primer 42 *Ispitati postojanost ocenjivača $\bar{X}_n$ za m, obeležja $X : \mathcal{N}(m, \sigma)$.*

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

Primer 43 *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 44 *Metodom momenata naći ocene parametara m i σ obeležja $X : \mathcal{N}(m,\sigma)$.*

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 $\theta = \theta(x_1, x_2, \ldots, x_n)$ za koju se ostvaruje maksimum (ako postoji) funkcije verodostojnosti koja se za realizovanu vrednost uzorka $(x_1, x_2, \ldots, x_n)$ računa:

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

Primer 45 *Naći ocenu maksimalne verodostojnosti parametara m i σ^2 za $X : \mathcal{N}(m,\sigma)$.*

Primer 46 *Naći ocenu maksimalne verodostojnosti parametra λ obeležja $X : \mathcal{P}(\lambda)$, ispitati njenu centriranost i postojanost.*

Intervali poverenja

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 m obeležja $X : \mathcal{N}(m,\sigma)$, σ poznato

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

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

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

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

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

Označimo sa t_β vrednost za koju je $P\left(|T| < t_\beta\right)$. ($t_\beta$ je $(1+\beta)/2$ kvantil raspodele t_{n-1}.)
Onda je $U_1 = \bar{X}_n - t_\beta \dfrac{\bar{S}_n}{\sqrt{n-1}}$, $U_2 = \bar{X}_n + t_\beta \dfrac{\bar{S}_n}{\sqrt{n-1}}$. **Standardna greška** je $\dfrac{\bar{S}_n}{\sqrt{n-1}} = \dfrac{\bar{S}'_n}{\sqrt{n}}$.

PRIMER 47 *Naći 90% interval poverenja za srednju vrednost m obeležja sa normalnom $\mathcal{N}(m,\sigma)$ 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
```

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

Ako $X : \mathcal{N}(m,\sigma)$ onda $Y = \dfrac{n\bar{S}_n^2}{\sigma^2} : \chi^2_{n-1}$.

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

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

PRIMER 48 *Naći 90% interval poverenja za nepoznatu varijansu obeležja iz primera 47.*

```
> y0<-qchisq(.05,9); y1<-qchisq(.95,9); # nastavak iz primera 48
> 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 $K = \sum X_i$, $\dfrac{K - np}{\sqrt{np(1-p)}} \to Z : \mathcal{N}(0,1)$. Za $z_\beta = \Phi^{-1}\left(\frac{1+\beta}{2}\right)$ važi

$$\beta \approx P\left(\left|\frac{K - np}{\sqrt{np(1-p)}}\right| < z_\beta\right) = P\left(\left(\frac{K - np}{\sqrt{np(1-p)}}\right)^2 < z_\beta^2\right) =$$

$$= P\left((n^2 + z_\beta^2 n)p^2 + (-2Kn - z_\beta^2 n)p + K^2 < 0\right) = P(U_1 < p < U_2),$$

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

Primer 49 *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*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] 0.2811693
> x2
[1] 0.646047
```

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 ocena parametra $\theta = h(x_1, \ldots, x_n)$.
- Nalazi se kritična oblast C (koja daje nedozvoljene vrednosti) parametra, takva da je $P_{H_0}(\hat{\theta} = h(X_1, \ldots, X_n) \in C) = \alpha$.
- Računa se statistika uzorka $\theta = h(x_1, x_2, \ldots, x_n)$ i ako $\theta \in C$, odbacujemo H_0 (i usvajamo H_1)

$H_0(m = m_0)$ **protiv** $H_1(m \neq m_0)$ **za** $X : \mathcal{N}(m, \sigma)$, σ **poznato**

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

$$m_0 \in \mathbb{R} \backslash \left(\bar{x}_n \mp z_\beta \frac{\sigma}{\sqrt{n}} \right) \Leftrightarrow z := \frac{|\bar{x}_n - m_0|}{\sigma}\sqrt{n} > z_\beta \Leftrightarrow \alpha^* := P_{H_0}\left(|X^*| > \frac{|\bar{x}_n - m_0|}{\sigma}\sqrt{n} \right) < \alpha$$

$H_0(m = m_0)$ **protiv** $H_1(m \neq m_0)$ **za** $X : \mathcal{N}(m, \sigma)$, σ **nepoznato**

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

$$m_0 \in \mathbb{R} \backslash \left(\bar{x}_n \mp t_\beta \frac{\bar{s}'_n}{\sqrt{n}} \right) \Leftrightarrow t := \frac{|\bar{x}_n - m_0|}{\bar{s}'_n}\sqrt{n} > t_\beta \Leftrightarrow \alpha^* := P_{H_0}\left(|T| > \frac{|\bar{x}_n - m_0|}{\bar{s}'_n}\sqrt{n} \right) < \alpha$$

PRIMER 50 *Testirati hipotezu $H_0(m = 13)$ za uzorak iz zadatka 47.*

PRIMER 51 *Testirati hipotezu $H_0(p = 1/3)$ za uzorak iz zadatka 49.*

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

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 $H_1(m < m_0)$ ili $H_1(m > m_0)$

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 $\left(0, \dfrac{n\bar{s}_n^2}{y_{1-\beta}}\right)$, gde je $y_{1-\beta}$ kvantil koji odgovara $\alpha = 1 - \beta$ za χ^2_{n-1}.

PRIMER 52 *Za uzorak iz zadatka 47 testirati hipotezu* $H_0(\sigma^2 = 25)$ *protiv* $H_1(\sigma^2 > 25)$.

```
> 9*sn^2/y0; # nastavak iz primera 48 i 49
[1] 23.9258     # odbacujemo hipotezu
```

Testiranje jednakosti srednjih vrednosti dva uzorka

$H_0(m_1 = m_2)$ **protiv** $H_1(m_1 \neq m_2)$**,** σ_1**,** σ_2 **poznato, obeležja sa** $\mathcal{N}(m_1,\sigma_1)$ **i** $\mathcal{N}(m_2,\sigma_2)$ **raspodelama**

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

$H_0(m_1 = m_2)$ **protiv** $H_1(m_1 \neq m_2)$ **(T-test)**

Koristimo statistiku $T := \dfrac{\bar{X}_1 - \bar{X}_2 - (m_1 - m_2)}{\sqrt{\frac{\bar{S}_1^{2\prime}}{n_1} + \frac{\bar{S}_2^{2\prime}}{n_2}}}$, koja približno ima t_ν raspodelu,

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

```
> 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
```

T-test parova

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

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

```
> 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
```

Testiranje dva uzorka, nastavak

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 $H_1(p_1 \neq p_2)$

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

Neka $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$, onda statistika $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\,(1-\hat{p})\,(1/n_1 + 1/n_2)}}$,

ako je tačna Nulta hipoteza, ima približno Normalnu $\mathcal{N}(0,1)$ raspodelu.

```
> 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
```

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, $i = 0, 1, \ldots k$ i brojem elemenata uzorka u intervalu i jednak f_i, $i = 1, 2, \ldots k$. (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 hoipotezu H_0.

Primer 53 *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?*

Tabela kontigencije

χ^2-test nezavisnosti obeležja. Obeležje X uzima m mogućih vrednosti, Y uzima n mogućih vrednosti.

Formira se tabela $m \times n$ verovatnoća izračunatih preko marginalnih verovatnoća $p_{i,j} = p_{i\cdot}\, p_{\cdot j}$, 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 $(m-1)(n-1)$ stepeni slobode.

PRIMER 54 *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))  # p-value = 0.08873
```

Test Kolmogorov-Smirnov

Primenjujemo ga za poznatu neprekidnu raspodelu

Statistika koju koristimo je

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

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

Primer 55 *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 odbačena.*

```
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
```

Regresija

Za slučajne promenljive X i Y definišemo **kovarijansu**:

$$\operatorname{cov}(X,Y) = E((X-E(X))(Y-E(Y))) = E(XY) - E(X)E(Y)$$

i **koeficijent korelacije**:

$$\rho_{X,Y} = \operatorname{cov}(X^*,Y^*) = \operatorname{cov}\left(\frac{X-E(X)}{\sqrt{D(X)}}, \frac{Y-E(Y)}{\sqrt{D(Y)}}\right) = \frac{E(XY)-E(X)E(Y)}{\sqrt{D(X)D(Y)}}$$

Osobine:

1. $\operatorname{cov}(X,X) = D(X) =: \operatorname{var}(X)$
2. X i Y nezavisne $\Rightarrow \rho_{X,Y} = \operatorname{cov}(X,Y) = 0$
3. $|\rho_{X,Y}| \leq 1, \quad |\rho_{X,Y}| = 1 \Leftrightarrow Y = a\mathrm{X} + b,\ a,b \in \mathbb{R},\ a \neq 0$
4. $\operatorname{cov}(\sum_{i=1}^{m} X_i, \sum_{j=1}^{n} Y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} \operatorname{cov}(X_i,Y_j)$
5. $\operatorname{var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{cov}(X_i,X_j) = \sum_{i=1}^{n} \operatorname{var}(X_i) + 2\sum_{1\leq i<j\leq n} \operatorname{cov}(X_i,X_j)$

6. $\rho_{X,Y} = \rho_{X_1,Y_1}$, gde su $X_1 = a + bX$ i $Y_1 = c + dY$, za pozitivne konstante a,b,c,d.

Ako posmatramo dvodimenzionalni uzorak $(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)$, odnosno, za realizovanu vrednost $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, definišemo **uzorački koeficijent korelacije**:

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

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

U R-u cov $(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n)$, cor $(x,y) =$ cov (x,y) / sd (x) / sd (y)

Linearna regresija najmanjih kvadrata

Za n parova tačaka $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$, tražimo vezu između x i y u obliku prave linije $y = a + bx$.

Tražimo vrednosti a i b za koje funkcija $g(a,b) = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$ ostvaruje minimum.

Funkcija g je konveksna i minimum je stacionarna tačka:

$$\frac{\partial g}{\partial a} = 2\sum_{i=1}^{n}(y_i - (a + bx_i))\,(-1) = 0, \qquad \frac{\partial g}{\partial b} = 2\sum_{i=1}^{n}(y_i - (a + bx_i))\,(-x_i) = 0.$$

Rešavanje ovog sistema po a i b daje: $b = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x}_n)\,(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} = r\dfrac{s_y}{s_x} = r\dfrac{s'_y}{s'_x}$, $a = \bar{y}_n - b\bar{x}_n$,

gde su s_x i s_y standardne devijacije uzorka x i y, a s'_x i s'_y korigovane standardne devijacije.

Za tako izračunate a i b funkciju $\hat{y} = a + bx$ zovemo **prava najmanjih kvadrata**. Vrednosti $\hat{y}_i = a + bx_i$ su **predikcije**. Važi $\bar{\hat{y}}_n = \bar{y}_n$. Prava najmanjih kvadrata prolazi kroz $(\bar{x}_n, \bar{y}_n)$.

$$ss_x = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2,\ ss_y = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2,\ s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x}_n)\,(y_i - \bar{y}_n),\ \Rightarrow\ r = \frac{s_{xy}}{\sqrt{ss_x\, ss_y}},\ b = \frac{s_{xy}}{ss_x}.$$

Sa tim oznakama imamo: $s_{xy} = r\sqrt{ss_x}\sqrt{ss_y}$, takođe: $b = r\frac{\sqrt{ss_x}\sqrt{ss_y}}{ss_x} = r\frac{\sqrt{ss_y}}{\sqrt{ss_x}}$.

Može se pokazati: $\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}_n))^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2$.

Takođe: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2 = \sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}_n))^2 = b^2\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 = b^2 ss_x = r^2 ss_y$.

Odatle: $r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_n)^2} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_n)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2} = \frac{\text{varijansa predikcija } y}{\text{varijansa realizovanih } y}$, odnosno, $r^2 \cdot 100\%$ je procenat varijanse objašnjene pravom linijom najmanjih kvadrata.

Vrednosti $\epsilon_i = y_i - \hat{y}_i$ zovemo **reziduali**. **Rezidual plot** je skup tačaka (x_i, ϵ_i).

Primer 56 *Napraviti scatter plot, linearni model, naći koeficijent korelacije, napraviti rezidual plot i naći procenat varijacije koji se objašnjava linearnim modelom za uzorak:*

s	15	13	20	23	11	21.5	12	23	23	19	19	19	21.8	17	20	17	20	16	11	12
f	8	8	16	14	7	15.5	8	21	15.5	16	4.5	13	13.5	12	16	8	15	8	7	6

Analiza varijanse ANOVA (jednofaktorska)

$H_0(m_1 = m_2 = \cdots = m_G)$, $H_1(\exists i, j, m_i \neq m_j)$

Poljoprivredni proizvođač želi da testira kvalitet četiri vrste semena soje: A, B, C, D i u tom cilju je odabrao 30 parcela iste površine koje imaju sličan kvalitet zemljišta, drenažu i izloženost suncu. Dobijeni su sledeći prinosi:

Seme	Prinos
A	46,43,43,46,44,42
B	51,58,62,49,53,51,50,59
C	37,39,41,38,39,37,42,36,40
D	42,43,42,45,47,50,48

Metodom analize varijanse ispitati da li postoje razlike u prosečnim prinosima soje kod semena A, B, C, D sa nivoom značajnosti $\alpha = 0.05$.

```
> read.csv("prinosi.csv")->prinosi
> boxplot(prinos ~ seme, data = prinosi)
> summary(prinosi)
     prinos          seme
 Min.   :36.00   A:6
 1st Qu.:41.25   B:8
 Median :43.50   C:9
 Mean   :45.43   D:7
 3rd Qu.:49.75
 Max.   :62.00
> prinostab<-lm(prinos~seme,data=prinosi)
```

Analiza varijanse Fišerovom statistikom

Grupa	Merenje					Grupna sredina
1	Y_{11}	Y_{12}	$\cdots$	Y_{1n_1}		$\bar{Y}_{1\cdot}$
2	Y_{21}	Y_{22}	$\cdots$		Y_{2n_2}	$\bar{Y}_{2\cdot}$
$\vdots$			$\ddots$			
G	Y_{G1}	Y_{G2}	$\cdots$	Y_{Gn_G}		$\bar{Y}_{G\cdot}$

$$\left(\bar{Y}_{g\cdot} = \frac{1}{n_g}\sum_{k=1}^{n_g} Y_{gk}\right)$$

$$Y_{gk} = m_g + \epsilon_{gk}, \text{ gde je } m_g = E(Y_{gk}), \quad \bar{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{g=1}^{G}\sum_{k=1}^{n_g} Y_{gk} = \frac{1}{n}\sum_{g=1}^{G} n_g \bar{Y}_{g\cdot}, \quad n = \sum_{k=1}^{G} n_g$$

$$\text{Treatment: } SSTR = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(\bar{Y}_{g\cdot} - \bar{Y}_{\cdot\cdot})^2 = \sum_{g=1}^{G} n_g(\bar{Y}_{g\cdot} - \bar{Y}_{\cdot\cdot})^2$$

$$\text{Error: } SSE = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g\cdot})^2, \qquad \text{Total: } SST = \sum_{g=1}^{G}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{\cdot\cdot})^2$$

$$SST = SSTR + SSE, \quad \bar{S}_g^{2'} = \frac{1}{n_g - 1}\sum_{k=1}^{n_g}(Y_{gk} - \bar{Y}_{g\cdot})^2, \quad SSE = \sum_{g=1}^{G}(n_g - 1)\bar{S}_g^{2'}$$

$$\frac{(n_1-1)\,\bar{S}_1^{2'} + (n_2-1)\,\bar{S}_2^{2'} + \cdots + (n_G-1)\,\bar{S}_G^{2'}}{(n_1-1) + (n_2-1) + \cdots + (n_G-1)} = \frac{\sum_{g=1}^{G}(n_g-1)\,\bar{S}_g^{2'}}{n-G} = \frac{SSE}{n-G}$$

Neka su Y_{gk}, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$ nezavisne slučajne promenljive sa istim očekivanjem u grupi: $E(Y_{gk}) = m_g$ i istom varijansom $D(Y_{gk}) = \sigma^2$ i neka $m = \sum_{g=1}^{G} n_g m_g \,/\, n$.

Može se dokazati da je $E\left(\frac{SSTR}{G-1}\right) = \sigma^2 + \frac{1}{G-1}\sum_{g=1}^{G} n_g\,(m_g - m)^2$ i $E\left(\frac{SSE}{n-G}\right) = \sigma^2$.

Takođe važi: Ako su Y_{gk}, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$ nezavisne slučajne promenljive sa normalnom raspodelom $Y_{gk} : \mathcal{N}(m_g,\sigma)$, $k = 1,2,\ldots,n_g$, $g = 1,2,\ldots,G$, onda

1. SSE i $SSTR$ su nezavisne
2. SSE/σ^2 ima χ^2_{n-G} raspodelu
3. Ako $m_1 = m_2 = \cdots = m_G$, onda $SSTR/\sigma^2$ ima χ^2_{G-1} raspodelu

Odatle sledi: ako $MSTR = \frac{SSTR}{G-1}$ i $MSE = \frac{SSE}{n-G}$, statistika $F = \frac{MSTR}{MSE}$ ima $F_{G-1,n-G}$ raspodelu

```
> anova(prinostab)
Analysis of Variance Table
Response: prinos
          Df  Sum Sq Mean Sq F value    Pr(>F)
seme       3 1015.51  338.50  32.614 5.781e-09 ***
Residuals 26  269.86   10.38
```

Multipla regresija

U linearnoj regresiji najmanjih kvadrata smo za parove tačaka $(x_i, y_i), i = 1, \ldots, n$ tražili a i b koje za linearnu zavisnost $y = a + bx$ daju minimalnu sumu kvadrata reziduala.

Moguće je pronaći i koeficijente zavisnosti y od parametara $x_1, \ldots, x_p$ u formuli

$$y = a + b_1 x_1 + \cdots + b_p x_p.$$

Može i za više zavisnih promenljivih $y_i = a_i + b_{i,1} x_1 + \cdots + b_{i,p} x_p$, $i = 1, \ldots, n$. Koeficijenti se najlakše nalaze matričnim računom. Mi koristimo funkcije ugrađene u R.

Primer 6

Posmatramo Case study `Spruce.csv`, 72 zasađena stabla praćena 5 godina.

Variable	Description
Tree	Tree number
Competition	C (competition), CR (competition removed)
Fertilizer	F (fertilized), NF (not fertilized)
...	...
Ht.change	Change (cm) in height
Di.change	Change (cm) in diameter

Kodiraćemo Competition i Fertilizer brojevima:
C $\rightarrow 1 =$ ima konkurenciju, CR $\rightarrow 0 =$ nema konkurenciju i
F $\rightarrow 1 =$ jeste đubreno, NF $\rightarrow 0 =$ nije đubreno.

Tree	Competition	Fertilizer		Ht.change	Di.change
1	0	1	$\cdots$	45	5.415625
2	0	1		36.2	4.009375
			$\ddots$		
72	1	0		19	2.11875

```
read.csv("Spruce.csv") -> Spruce
lm(Di.change ~ Ht.change + Fertilizer + Competition, data = Spruce)
```

```
Coefficients:
(Intercept)     Ht.change    Fertilizer   Competition
     0.5116        0.1040        1.0266       -0.4895
```

Formula za zavisnost porasta prečnika stabla u zavisnosti od promene visine, đubrenja i uništavanja konkurencije :

```
Di.change = 0.5116 + 0.1040 Ht.change + 1.0266 Fertilizer -
    0.4895 Competition
```

Jednostavni linearni model

Za uzorak $(x_1, y_1), \ldots, (x_n, y_n)$ linearna regresija najmanjih kvadrata je $y = a + bx$.
Neka su vrednosti y_i realizovane vrednosti Y_i, posmatramo parove $(x_1, Y_1), \ldots, (x_n, Y_n)$.
Jednostavni linearni model (JLM) pretpostavlja:

- Za neke α i β važi $E(Y_i | x_i) = \alpha + \beta x_i =: \mu_i$, $i = 1, \ldots, n$.
- Reziduali $\epsilon_i := Y_i - \mu_i$ su nezavisne slučajne promenljive sa raspodelom $\mathcal{N}(0, \sigma)$.

Posledica je da su $Y_i : \mathcal{N}(\mu_i, \sigma)$, $i = 1, \ldots, n$ nezavisne slučajne promenljive.
Ocena metodom maksimalne verodostojnosti za β, α, σ^2 daje redom ocenjivače:

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \quad \hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \hat{Y}_i)^2, \text{ gde je } \hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i,\ \bar{Y} = \frac{1}{n}\sum Y_i.$$

Vidimo da realizovane vrednosti ocena $\hat{\beta}$ i $\hat{\alpha}$ odgovaraju formulama najmanjih kvadrata.
Ako $(x_1, Y_1), \ldots, (x_n, Y_n)$ zadovoljavaju pretpostavke JLM, onda

- $\hat{\beta}$, $\hat{\alpha}$, $\bar{Y}$ su nezavisne,
- $n\hat{\sigma}^2/\sigma^2$ ima χ^2 raspodelu sa $n - 2$ stepeni slobode,

(Koristimo centriranu ocenu $S^2 = \frac{n}{n-2}\hat{\sigma}^2 = \frac{1}{n-2}\sum(Y_i - \hat{Y}_i)^2$, tzv. **varijansa reziduala**.)

- $\hat{\beta}$ i $\hat{\alpha}$ imaju normalnu raspodelu,
- $E(\hat{\beta}) = \beta$, $E(\hat{\alpha}) = \alpha$,
- Var $(\hat{\beta}) = \sigma^2 / ss_x$,
- Var $(\hat{\alpha}) = \sigma^2 \left(1/n + \bar{x}^2 / ss_x\right)$.

Posledice: $Z = \dfrac{\hat{\beta} - \beta}{\sigma / \sqrt{ss_x}} : \mathcal{N}(0,1), \quad T = \dfrac{\hat{\beta} - \beta}{S / \sqrt{ss_x}} : t_{n-2}$, gde je $S = \sqrt{S^2}$.

Uvodimo i oznaku $\hat{SE}(\hat{\beta}) = S / \sqrt{ss_x}$, tzv. **standardna greška** ocenjivača $\hat{\beta}$.

Uobičajeno je testiranje hipoteze $H_0(\beta = 0) - H_1(\beta \neq 0)$ koristeći $T = \dfrac{\hat{\beta}}{\hat{SE}(\hat{\beta})} : t_{n-2}$.

Interval poverenja za β širine $(1 - \alpha) \cdot 100\%$ je $\left(\hat{\beta} \pm q\, \hat{SE}(\hat{\beta})\right), \quad P(|T| > q) = \alpha$.

Primer 7

Na takmičenju u klizanju izvodi se dvominutni obavezni program i četvorominutni slobodni program. U fajlu `Skating2010.csv` su bodovi za 24 klizača sa Olimpijade 2010.

Da li postoji i kako glasi zavisnost između dobijenih ocena?

```
read.csv("Skating2010.csv") -> Skating2010
skate.lm <- lm(Free ~ Short, data=Skating2010)
summary(skate.lm)
Residuals:
    Min      1Q  Median      3Q     Max
-23.314  -6.780   0.710   6.407  21.205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.9691    18.1175   0.440    0.664
Short         1.7347     0.2424   7.157 3.56e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.36 on 22 degrees of freedom
Multiple R-squared:  0.6995,Adjusted R-squared:  0.6859
F-statistic: 51.22 on 1 and 22 DF,  p-value: 3.562e-07
```

Postoji linearna veza između Free i Short programa ($\rho^2 = 0.6995$, $\rho = 0.8364$). Odbačena je $H_0(\beta = 0)$, p-value $= \alpha^* =$ `3.562e-07`. Formula: `Free = 7.9691 + 1.7347 Short`.

Statističko zaključivanje za predikcije

Očekivanje predikcije

Neka $(x_1, Y_1), \ldots, (x_n, Y_n)$ zadovoljavaju pretpostavke Jednostavnog linearnog modela. Slučajnu promenljivu $Y_s = \alpha + \beta x_s$ nazivamo predikcija za x_s. Ocenjivač očekivane vrednosti predikcije $E(Y_s)$ za dato x_s je $\hat{Y}_s = \hat{\alpha} + \hat{\beta} x_s$. Onda

1. $\bar{Y} = \frac{1}{n} \sum Y_i$ je slučajna promenljiva sa Normalnom raspodelom.
2. $E(\bar{Y}) = \alpha + \beta \bar{x}$.
3. Var $(\bar{Y}) = \sigma^2 / n$.
4. Y_s ima normalnu raspodelu.
5. $\hat{Y}_s$ je centrirani ocenjivač za očekivanu vrednost predikcije: $E(\hat{Y}_s) = E(Y_s)$.
6. Var $(\hat{Y}_s) = \sigma^2 \left[1/n + (x_s - \bar{x})^2 / ss_x) \right]$.

Posledica: $T = \dfrac{\hat{Y}_s - E(\hat{Y}_s)}{S \sqrt{1/n + (x_s - \bar{x})^2 / ss_x)}}$, gde je S standardna greška reziduala, ima Studentovu raspodelu sa $n - 2$ stepeni slobode.

Za dato x_s, interval poverenja širine β za $E(Y_s)$ je

$$\left(\hat{Y}_s \pm t_{(1+\beta)/2;n-2}\, \hat{SE}(\hat{Y}_s)\right),$$

gde je $t_{(1+\beta)/2;n-2}$ kvantil $(1+\beta)/2$ Stundetove raspodele sa $n-2$ stepeni slobode i gde je $\hat{SE}(\hat{Y}_s) = S\sqrt{1/n + (x_s - \bar{x})^2/ss_x}$ standardna greška ocenjivača $\hat{Y}_s$.

Pojedinačna predikcija

Varijansa greške pojedinačne predikcije $Y = \alpha + \beta x$ je

$$\text{Var}\,(Y - \hat{Y}) = \text{Var}\,(Y) + \text{Var}\,(\hat{Y}) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x_s - \bar{x})^2}{ss_x}\right].$$

Za dato $x = x_s$ interval poverenja širine β predikcije $Y_s = \alpha + \beta x_s$ je

$$\left(\hat{Y}_s \pm t_{(1+\beta)/2;n-2}\, \hat{SE}(Y_s)\right),$$

gde je $t_{(1+\beta)/2;n-2}$ kvantil $(1+\beta)/2$ Stundetove raspodele sa $n-2$ stepeni slobode i gde je $\hat{SE}(Y_s) = S\sqrt{1 + 1/n + (x_s - \bar{x})^2/ss_x}$ standardna greška predikcije Y_s.

Permutacioni testovi

Primer 1

Mereno je u sekundama vreme potrebno mišu da izađe iz lavirinta.
Pod uticajem leka: 30, 25, 20 i bez uticaja leka: 18, 21, 22 (kontrolna grupa).
Ostvarena je razlika srednjih vrednosti:

$$\bar{x}_d - \bar{x}_c = (30+25+20)/3 - (18+21+22)/3 = 4.667s. \qquad \binom{6}{3} = 20$$

Ako se posmatraju kao jednako verovatni svih 20 ishoda izbora 3 od 6 merenja, kolika je verovatnoća da je razlika srednjih vrednosti veća ili jednaka od ostvarene?
Testiramo H_0: "lek nema uticaja" protiv H_1: "lek usporava", $H_0(\mu_d = \mu_c) - H_1(\mu_d > \mu_c)$.

```
x<-c(30, 25, 20, 18, 21, 22)
ind<-t(matrix(c(1,2,3,1,2,4,1,2,5,1,2,6,1,3,4,1,3,5,...),nrow=3))
index<-ind[1,]; observed<-mean(x[index])-mean(x[-index])
result<-numeric(20)
for(i in 1:20)
{ index<-ind[i,]
  result[i]<-mean(x[index])-mean(x[-index])}
sum(result >= observed)/20
```

Dobijena verovatnoća $3/20 = 0.15$ ne protivreči H_0 sa pragom značajnosti $\alpha = 0.05$.

Primer 2

Osoba A tvrdi da uvek ispadne grb kada baci novčić.
Da bi dokazala, bacila je novčić 3 puta i sva tri puta je ispao grb.

Kolika je verovatnoća da ispadne grb u 3 bacanja novčića?

Testiramo H_0: "bacanje novčića osobe A ima uobičajenu verovatnoću"
protiv H_1: "osoba A može da baci grb svaki put".

Ne odbacujemo nultu hipotezu jer $1/8 = 0.125 = 12.5\%$ nije manja od $\alpha = 0.05 = 5\%$.

Primer 3

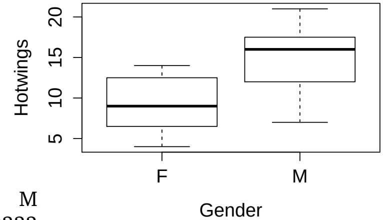
Posmatramo Case study Beerwings.

Variable	Description
Gender	Male or female
Beer	Ounces of beer consumed
Hotwings	Number of hot wings eaten

ID	Hotwings	Beer	Gender
1	4	24	F
2	5	0	F
	...		
30	21	42	M

Posmatramo Hotwings u odnosu na faktor Gender. Testiramo hipotezu H_1 da M pojede više Hotwingsa od F. H_0 je da nema razlike u Hotwings u zavisnosti od Gender.

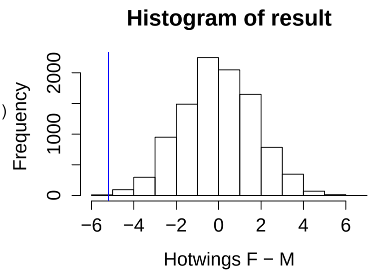
```
       ID              Hotwings           Beer        Gender
 Min.   : 1.00    Min.   : 4.00    Min.   : 0.0    F:15
 1st Qu.: 8.25    1st Qu.: 8.00    1st Qu.:24.0    M:15
 Median :15.50    Median :12.50    Median :30.0
 Mean   :15.50    Mean   :11.93    Mean   :26.2
 3rd Qu.:22.75    3rd Qu.:15.50    3rd Qu.:36.0
 Max.   :30.00    Max.   :21.00    Max.   :48.0
```




```
        F         M
 9.333333 14.533333
```

Ostvarena vrednost razlike srednjih vrednosti je $9.333333 - 14.533333 = -5.2$. Imamo $\binom{30}{15} = 155117520$ mogućih izbora $\rightarrow$ preobimno. Vršimo reuzorkovanje (resampling).

```
read.csv("Beerwings.csv")->krilca; summary(krilca)
plot(Hotwings~Gender,data=krilca)
tapply(krilca$Hotwings,krilca$Gender,mean)
observed <- -5.2
Hotwings <- krilca$Hotwings; N <- 9999; result <- numeric(N)
for (i in 1:N)
{ index<-sample(30,size=15,replace=FALSE)
  result[i]<-mean(Hotwings[index])-mean(Hotwings[-index])}
hist(result,xlab="Hotwings F - M")
abline(v=observed,col="blue")
(sum(result <= observed) + 1)/(N + 1)
```



p-value = 4e-4 $\Leftrightarrow \alpha^* = 4 \cdot 10^{-4}$

Vidimo da je verovatnoća ovako velike razlike daleko manja od $\alpha = 0.05$, odbacujemo H_0.

procedure TWO-SAMPLE PERMUTATION TEST(x, m, n, dx)
    repeat
        Izaberi poduzorak m od $m + n$ vrednosti x (bez vraćanja)
        Uporedi izabranu statistiku za izabranih m i preostalih n vrednosti
    until ima dovoljno uzoraka
    Izračunaj p-value kao procenat slučajeva u kojima je poređenje statistika $\geq dx$
    Pomnoži p-value sa 2 ako je u pitanju dvostrani test
    Nacrtaj histogram i označi p-vrednost (Opciono)
end procedure

Najčešće se za statistiku koristi aritmetička sredina, ali mogu i druge statistike.

Ekvivalentni rezultati se dobijaju primenom rastuće funkcije na statistiku.

Dodajemo 1 na brojilac i imenilac da bismo izbegli p-value $= 0$.

Ovaj test ne zahteva da uzorak ima normalnu raspodelu.

Ovaj test je manje osetljiv na poduzorke nejednakih obima od t-testa ($m \gg n$). Pažnja!!!

Uzastopne primene ovog testa ne daju istu p-value.

Jednostrani test se primenjuje ako je suprotna alternativa očigledno nemoguća. Odlučivanje za primenu jednostranog testa se ne sme vršiti posle testiranja.

Za veliko N, uzimanje n uzoraka sa vraćanjem i bez vraćanja daje približno iste verovatnoće.

Tabela kontigencije

Primer 4

Tabela odgovora za i protiv smrtne kazne u odnosu na najviši nivo obrazovanja iz `GSS2002.csv`

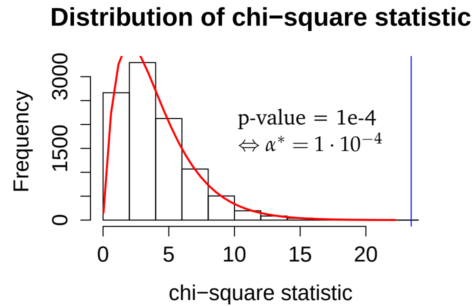
$$\chi^2 = \sum_{sve\ \acute{c}elije} \frac{(ostvareno - o\check{c}ekivano)^2}{o\check{c}ekivano}$$

Education	Favor	Oppose	rowsum	%
Bachelors	135	71	206	15.8
Graduate	64	50	114	8.7
HS	511	200	711	54.4
JrColl	71	16	87	6.7
Left HS	117	72	189	14.5
colsum	898	409	1307	
%	68.7	31.3		

procedure PERMUTATION TEST FOR INDEPENDENCE OF TWO VARIABLES($t_{n\times 2}$)
  Izračunaj ostvarenu $\chi^2(t)$
  repeat
    Permutuj na slučajan način vrste jedne kolone
    Izračunaj $\chi^2(t)$ i upamti rezultat
  until ima dovoljno uzoraka
  Izračunaj p-value kao procenat slučajeva u kojima je upamćena χ^2 veća od ostvarene
  Nacrtaj histogram i označi p-vrednost (Opciono)
end procedure

Testira se H_0: parametri su nezavisni.

```
read.csv("GSS2002.csv")->GSS2002
Education <- GSS2002$Education
DeathPenalty <- GSS2002$DeathPenalty
observed <- chisq.test(table(Education,
    DeathPenalty))$statistic
N <- 10^4-1; result <- numeric(N)
for (i in 1:N)
{ DP.permuted <- sample(DeathPenalty)
GSS.table <- table(Education, DP.permuted)
result[i] <- chisq.test(GSS.table)$statistic}
hist(result, xlab = "chi-square statistic",
    main = "Distribution");
abline(v = observed, col = "blue")
(sum(result >= observed) + 1)/(N+1) # p-value
```



Kako je p-value = 1e-4 daleko manje od $\alpha = 0.05$, odbacujemo H_0.

Ako je H_0 tačna, verovatnoće pojedinačne ćelije (i,j) su $p_{i,j} = p_{i\cdot}\, p_{\cdot j} = \frac{\text{rowsum}}{n}\, \frac{\text{colsum}}{n}$, a očekivane vrednosti su $n\, p_{i,j} = n \frac{\text{rowsum}}{n}\, \frac{\text{colsum}}{n} = \frac{\text{rowsum}\cdot\text{colsum}}{n}$.

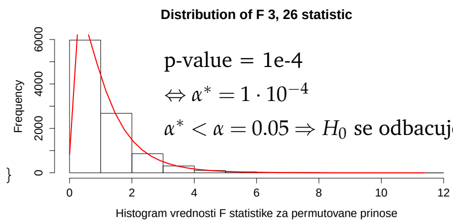
Ostvarena vrednost χ^2 statistike je bila `observed = 23.45`, p-value za χ^2 test nezavisnosti je `1-pchisq(observed,4) = 1.029e-4`, što daje isti rezultat kao permutacioni test.

Permutacioni test za jednakost srednjih vrednosti u više grupa (ANOVA)

Primer 5

Slično kao u prethodnom primeru, možemo uraditi resampling od (recimo) $N = 10^4$ permutacija vrednosti prinosa sa istim brojem po grupama. Onda umesto p-vrednosti za ostvareni kvantil u Fišerovoj raspodeli, možemo koristiti proporciju broja vrednosti Fišerove statistike koje prelaze preko ostvarene vrednosti zadate raspodelom po grupama.

```
observed <- anova(prinostab)$F[1]
prinos <- prinosi$prinos
n <- length(prinos)
N <- 10^4 - 1
results <- numeric(N)
for (i in 1:N)
{ index <- sample(n)
  prinosi$prinos <- prinos[index]
  results[i] <- anova(lm(prinos~seme,data=prinosi))$F[1]}
(sum(results> observed) + 1) / (N + 1) # p-value
```



Setimo se da je `observed = 32.61` i `1-pf(observed,3,26) = 5.781e-9`, dobili smo isto kao sa kvantilom Fišerove statistike, H_0 se odbacuje.