Statistika

Statistički eksperiment podrazumijeva mjerenja obilježja na nekoliko izabranih članova populacije. Mi ćemo se uglvnom baviti nezavisnim mjerenjima. Rezultati mjerenja se tretiraju kao realizacije nezavisnih slučajnih promjenljivih $X_1, X_2, ..., X_n$ i svaka slučajna promjenljiva je raspodijeljena kao obilježje X.

DEFINICIJA 0.1 *Neka obilježje X ima funkciju raspodjele F. Prost slučajni uzorak obima n iz raspodjele F je slučajni vektor $(X_1, X_2, ..., X_n)$ gdje su $X_1, X_2, ..., X_n$ nezavisne slučajne promjenljive sa istom funkcijom raspodjele F.*

Prost slučajni uzorak je matematički model n nezavisnih mjerenja obilježja. Često se na osnovu predznanja o izučavanom fenomenu može pretpostaviti da raspodjela obilježja pripada nekoj familiji $\mathfrak{R}$ koju nazivamo familijom dopustivih raspodjela. Ista oznaka $\mathfrak{R}$ se koristi i za familiju dopustivih funkcija raspodjela, motivisani smo činjenicom da raspodjela slučajne promjenljive jednoznačno određuje njenu funkciju raspodjele i obrnuto, funkcija raspodjele slučajne promjenljive jednoznačno određuje njenu raspodjelu. Zadatak matematičke statistike je da se na osnovu registrovanih podataka zaključi što je moguće više o nepoznatoj funkciji raspodjele obilježja, a što je ekvivalentno sa tim da se sazna što je moguće više o nepoznatoj raspodjeli obilježja. Ubuduće ćemo umjesto prost slučajni zorak govoriti uzorak.

Primjer 1 Registrujemo α čestice koje emituje grumen radioaktivne materije, intenzitet potoka čestica je λ, λ je nepoznato. Slučajna promjenljiva X_1 predstavlja broj čestica koje se emituju u toku prvog minuta, X_2 predstavlja broj čestica koje se emituju u toku drugog minuta,...,X_n predstavlja broj čestica koje se emituju u toku n-tog minuta. Obilježje X je broj čestica koje se emituju u toku jednog minuta i X ima $\mathcal{P}(\lambda)$ raspodjelu. Slučajne promjenljive $X_1, ..., X_n$ su nezavisne i svaka ima $\mathcal{P}(\lambda)$ raspodjelu. $(X_1, ..., X_n)$ je prost slučajan uzorak. Koristili smo činjenice 1. slučajne promjenljive koje prebrojavaju čestice na disjunktnim vremenskim intervalima su nezavisne 2. raspodjela slučajne promjenljive koja prebrojava čestice na vremenskom intervalu dužine t je $\mathcal{P}(\lambda t)$. U ovom primjeru familiju dopustivih raspodjela čine Puasonove raspodjele sa parametrom $\lambda, \lambda > 0$. Skup $\Lambda = (0, \infty)$ mogućih vrijednosti za λ nazivamo parametarskim skupom. Ako imamo neku dodatnu informaciju o λ, recimo da je $1 < \lambda < 2$ tada je $\Lambda = (1, 2)$. U ovom modelu označimo sa W_1 vrijeme do emitovanja prve čestice, sa W_2 vrijeme od emitovanja prve do emitovanja druge, sa W_3 vrijeme od emitovanja deruge do emitovanja treće, sa W_n vrijeme od emitovanja $n-1$-e do emitovanja n-te. Pomenuta vremena su slučajna te su $W_1, ..., W_n$ slučajne promjenljive i one su nezavisne, $(W_1, ..., W_n)$ je prost slučajni uzorak, a odgovarajuća familija raspodjela je $\mathcal{E}(\lambda), \lambda > 0$. Teorijsko objašnjenje zbog čega se u modelu pojavljuju Puasonova i eksponencijalna raspodjela te zbog čega su promjenljive $X_1, ..., X_n$ odnosno $W_1, ..., W_n$ nezavisne biće dato u okviru predmeta Slučajni procesi.

Primjer 2. Obilježje X je rezultat mjerenja fizičke veličine, dok rezultati n mjerenja predstavljaju prost slučajni uzorak (uzorak). Gaus je tvorac teorije grešaka kod mjerenja fizičkih veličina. Primijetio ja da rezultati mjerenja udaljenosti konkretnog nebeskog tijela od Zemlje variraju kao posljedica postojanja greške mjerenja. Greška mjerenja ε se dobija sumiranjem malih grešaka uzrokovanim faktorima: temperaturna kolebanja koja utiču na mjerni instrument, intenzitet sunčevih zraka koji može dovesti do savijanja instrumenta, vazdušna strujanja, vibracije i mnogim drugim. Sumarna greška ε na osnovu CGT (ako ne postoji sistematska greška) ima $\mathcal{N}(0,\sigma^2), m > 0$ raspodjelu. Ako je stvarna udaljenost d tada obilježje X koje predstavlja rezultat mjerenja ima $\mathcal{N}(d,\sigma^2)$ raspodjelu. Dakle, $\mathfrak{R} = \{\mathcal{N}(d,\sigma^2), d > 0, \sigma^2 > 0\}$.

Primjer 3. Obilježje X je rezultat mjerenja koeficijenta inteligencije osobe, dok rezultati mjerenja koeficijenta inteligencije n osoba predstavljaju prost slučajni uzorak (uzorak). Familija dopustivih raspodjela je normalna.

Primjer 4. Obilježje je kilometraža koju obezbjeđuje jedna baterija automobilu na električni pogon. Familija dopustivih raspodjela je eksponencijalna.

Primjer 5. Obilježje X je 1 ako osoba koja ima pravo glasa, glasa za kandidata A, a 0 ako ne glasa za kandidata A. Raspodjela obilježja X je

$$X : \begin{array}{cc} 0 & 1 \\ 1-p=q & p \end{array}$$

i p je relativna učestalost onih koji glasaju za kandidata A. Drugim riječima p je vjerovatnoća da slučajno izbrana osoba sa pravom glasa, glasa za A. Ako osobe sa pravom glasa u uzorak biramo metodom bez vraćanjanja, tada su promjenljive $X_1, ..., X_n$ zavisne i raspodijeljene kao i obilježje.

Primjer 6. $\mathfrak{R} = \{\mathcal{U}(-\theta;\theta),\ \theta \in (1,2)\}$. $\mathfrak{R} = \{\mathcal{U}(a;b),\ a < 0;\ b > 3\}$.

DEFINICIJA 0.2 *Neka je $f : \mathrm{R}^n \to \mathrm{R}$ Borelova funkcija i $(X_1, ..., X_n)$ uzorak. Slučajna promjenljiva $V = f(X_1, ..., X_n)$ se naziva statistika.*

Neki autori statistiku izjednačavaju sa funkcijom f. Realizaciju uzorka $(X_1, ..., X_n)$ označavamo sa $(x_1, ..., x_n)$ i nazivamo realizovanim uzorkom. Realizovana statistika je $v = f(x_1, ..., x_n)$.

Primjeri statistika.

Uzoračka sredina. $f(x_1, ..., x_n) = \frac{X_1+...+X_n}{n}, \overline{X}_n = \frac{X_1+...+X_n}{n}$. Neka je $EX = a, DX = \sigma^2$. Tada je $E\overline{X}_n = a, D\overline{X}_n = \frac{\sigma^2}{n}$.

Uzoračka disperzija. $\overline{S}_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n}\sum\limits_{i=1}^{n} X_i^2 - \overline{X}_n^2; E\overline{S}_n^2 = EX^2 - E\overline{X}_n^2 = EX^2 - \frac{1}{n}EX^2 - \frac{n^2-n}{n^2}(EX)^2 = \frac{n-1}{n}\sigma^2$.

Popravljena uzoračka disperzija. $\hat{S}_n^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{1}{n-1}\sum\limits_{i=1}^{n} X_i^2 - \frac{n}{n-1}\overline{X}_n^2; E\hat{S}_n^2 = \sigma^2.$

Uzorački minimum. $Y_1 = \min\limits_{1\leqslant i\leqslant n} X_i$. **Uzorački maksimum.** $Y_n = \max\limits_{1\leqslant i\leqslant n} X_i$.

χ^2 **raspodjela sa n stepeni slobode.** Neka su $X_1, ..., X_n$ nezavisne slučajne promjenljive, svaka promjenljiva ima $\mathcal{N}(0,1)$ raspodjelu. Raspodjela slučajne promjenljive $\chi_n^2 = \sum\limits_{i=1}^{n} X_i^2$ se naziva χ^2 sa n stepeni slobode. Potražimo odgovarajuću gustinu raspodjele, označimo je sa g_n. Označimo sa $V_n(x)$ zapreminu lopte $x_1^2 + ... + x_n^2 < x, x > 0$. Iz analize znamo da je $V_n(x) = x^{\frac{n}{2}} V_n(1)$. Znamo da je $(t+h)^{\frac{n}{2}} - t^{\frac{n}{2}} = \frac{n}{2} t^{\frac{n}{2}-1} h + o(h)$.

$$P\{t \leqslant \chi_n^2 < t+h\} = \int\limits_{t\leqslant\sum\limits_{i=1}^{n} x_i^2 < t+h}\cdots\int (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum\limits_{i=1}^{n} x_i^2} dx_1...dx_n = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}(t+\theta h)} [V_n(t+h) - V_n(t)],$$

$$\Rightarrow g_n(t) = \frac{n}{2}(2\pi)^{-\frac{n}{2}} e^{-\frac{t}{2}} t^{\frac{n}{2}-1} V_n(1), \text{ te iz } \int\limits_0^{\infty} g_n(t)dt = 1 \Rightarrow g_n(t) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} t^{\frac{n}{2}-1} e^{-\frac{t}{2}}, t > 0.$$

Konstatujmo da je $E\chi_n^2 = n; D\chi_n^2 = 2n, f_n(t) = (1-2it)^{-\frac{n}{2}}$.

Studentova raspodjela sa n stepeni slobode. Neka su $X_0, X_1, X_2, ..., X_n$ nezavisne slučajne promjenljive, svaka ima $\mathcal{N}(0,1)$ raspodjelu. Za slučajnu promjenljivu $t_n = X_0 \left(\frac{\sum\limits_{i=1}^{n} X_i^2}{n}\right)^{-\frac{1}{2}}$ kažemo da ima Studentovu raspodjelu sa n stepeni slobode. Gustina $\psi_n(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}, t \in \mathbb{R}$.
Konstatujmo, $Et_n = 0, n > 1; Dt_n =: \begin{cases} \infty, & n \leqslant 2 \\ \frac{n}{n-2}, & n > 2. \end{cases}$

Teorema 0.1 Fišerova teorema. *Neka obilježje X ima $\mathcal{N}(m, \sigma^2)$ raspodjelu. Tada važi:*

a) $\frac{\overline{X}_n - m}{\sigma}\sqrt{n} : \mathcal{N}(0,1)$.

b) Statistike $\overline{X}_n$ i $\overline{S}_n^2$ su nezavisne.

c) $\frac{n\overline{S}_n^2}{\sigma^2} : \chi_{n-1}^2$ *i* $\frac{(n-1)\hat{S}_n^2}{\sigma^2} : \chi_{n-1}^2$

d) $\frac{\overline{X}_n - m}{\hat{S}_n}\sqrt{n} : t_{n-1}$.

Glivenko-Kantelijeva centralna teorema matematičke statistike.

DEFINICIJA 0.3 *Neka je $(X_1, ..., X_n)$ uzorak iz raspodjele F. Funkcija*

$$F_n(x, \omega) = F_n(x) = \frac{1}{n}\sum_{k=1}^{n} I\{X_k(\omega) < x\}, \; x \in \mathbb{R}$$

se naziva empirijska funkcija raspodjele.

Ako se fiksira ω dobija se funkcija po x i ona je jednaka relativnoj učestalosti izmjerenih podataka manjih od x. Pretpostavimo da su izmjereni podaci $x_1, x_2, ..., x_n$ (tj. realizovani uzorak je je $(x_1, ..., x_n)$) i da su različiti. Imamo

$$F_n(x) = \begin{cases} 0, & \text{ako je } x \leqslant x_{(1)}, \\ \frac{k}{n}, & \text{ako je } x_{(k)} < x \leqslant x_{(k+1)},\ k = 1, 2, ..., n-1, \\ 1, & \text{ako je } x > x_{(n)}. \end{cases}$$

Primijetimo da je funkcija $F_n(x)$ funkcija raspodjele slučajne promjenljive koja ima ravnomjernu diskretnu raspodjelu na skupu $\{x_1, x_2, ..., x_n\}$.

Ako fiksiramo x tada je $F_n(x,\omega)$ slučajna promjenljiva. Primijetimo, $\sum\limits_{k=1}^{n} I\{X_k(\omega) < x\}$ je slučajna promjenljiva sa $\mathcal{B}(n, F(x))$ raspodjelom te je

$$P\Big\{F_n(x,\omega) = \frac{k}{n}\Big\} = \binom{n}{k} F^k(x)(1 - F(x))^{n-k},\ k = 0, 1, ..., n.$$

$F_n(x,\omega)$ je slučajni proces. Funkcija F se naziva teorijska funkcija raspodjele obilježja X.

Empirijska funkcija raspodjele u slučaju kada je obim uzorka veliki aproksimira teorijsku funkciju raspodjele. Preciznije, važe sljedeće dvije teoreme.

Teorema 0.2 *Za svaki realni broj x važi*

$$P\{\lim_{n\to\infty} F_n(x) = F(x)\} = 1.$$

Tvrđenje teoreme je neposredna posljedica jakog zakona velikih brojeva.

Teorema 0.3 Glivenko-Kantelijeva centralna teorema matematičke statistike. *Neka je F teorijska funkcija raspodjele obilježja X. Važi*

$$P\Big\{\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \mid F_n(x) - F(x) \mid = 0\Big\} = 1.$$

Drugim riječima, skoro izvjesno $F_n(x) \rightrightarrows F(x)$. Neka je $A = \{\omega : \lim\limits_{n\to\infty} \sup\limits_{x\in\mathbb{R}} \mid F_n(x,\omega) - F(x) \mid = 0\}$. Glivenko-Kantelijeva teorema tvrdi da je vjerovatnoća-mjera skupa A jednaka 1. Neka je $\omega_0 \in A$. Tada za $\forall\varepsilon > 0$, $\exists n_0 = n(\omega_0, \varepsilon)$ takav da za $\forall n > n_0$ i $\forall x \in \mathbb{R}$ važi $\mid F_n(x,\omega_0) - F(x) \mid < \varepsilon$.

Tačkasta ocjena nepoznatog parametra

Neka teorijska tj. stvarna funkcija raspodjele obilježja X pripada familiji funkcija raspodjela $\mathfrak{R}$ i neka je $\mathbb{X} = (X_1, ..., X_n)$ odgovarajući uzorak. Neka je τ parametar koji je jednoznačno određen funkcijama raspodjela iz $\mathcal{R}$. Formalno, $\tau = \tau(F)$, $F \in \mathfrak{R}$, tj. $\tau(F)$ je funkcional. Zadatak ocjenjivanja je da se nađe statistika $T(\mathbb{X})$ čija realizacija $t = T(x_1, ..., x_n)$ može poslužiti kao dobra-razumna aproksimacija za $\tau_0 = \tau(F_X)$ tj. $t \approx \tau_0$. Statistika $T(\mathbb{X})$ se naziva ocjena parametra τ, a $t = T(x_1, ..., x_n)$ je ocjena koju dobijamo nakon prikupljanja podataka tj. nakon statističkog eksperimenta. Kako se nepoznati parametar ocjenjuje brojem odnosno **tačkom** to se govori o tačkastoj ocjeni parametra. Mi ćemo najčešće raditi sa parametarskim familijama

$$\mathfrak{R} = \{F(x, \theta),\, \theta \in \Theta\}.$$

Da bi ocjena bila "dobra", poželjno je da ima neke osobine. Ocjena T_n, n je obim uzorka, kojom ocjenjujemo nepoznati parametar τ, je **centrirana-nepristrasna** ako je $ET_n = \tau$ za $\forall \tau$. Ocjena T_n je **asimptotski centrirana-nepristrasna** ako je $\lim\limits_{n\to\infty} ET_n = \tau$ za $\forall\tau$. Ocjena T_n je postojana ako $T_n \xrightarrow{p} \tau$ za $\forall\tau$.

Neka je T_n ocjena nepoznatog parametra τ. Tačnost ocjene se mjeri parametrom $\mathcal{K}(T_n) := E(T_n - \tau)^2 = DT_n + (ET_n - \tau)^2$. Ocjena je tim bolja što je parametar $\mathcal{K}$ manji. Ako je ocjena T_n centrirana tada je $\mathcal{K}T_n = DT_n$. Ako su U_n i V_n centrirane ocjene nepoznatog parametra τ i ako je $DU_n \leqslant DV_n$ tada kažemo da ocjena U_n nije gora od ocjene V_n.

Primjer. $X : \mathcal{U}(0, \theta), \theta > 0$, θ je nepoznati parametar. Statistike $T_1 = 2\overline{X}_n$ i $T_2 = \frac{n+1}{n} Y_n$ su centrirane, $DT_1 = \frac{\theta^2}{3n}$; $DT_2 = \frac{\theta^2}{n(n+2)}$. Za $n \geqslant 2$ ocjena T_2 je bolja od ocjene T_1.

Primjer. $X : \mathcal{U}(0, 2\theta), \theta > 0$, θ je nepoznati parametar. Za ocjenjivanje parametra θ koristimo statistike $T_1 = \frac{X_1+X_2+3X_3+2X_4+X_5}{10}$ i $T_2 = X_1 - X_2 + X_3 - X_4 + X_5$. Imamo: $ET_1 = \frac{9}{10}\theta$, $DT_1 = \frac{19}{300}$, $\mathcal{K}T_1 = 0,073\theta^2$; $ET_2 = \theta$, $DT_2 = \frac{5}{3}\theta^2$, $\mathcal{K}T_2 = 1,67\theta^2$.

Primjer. U kutiji se nalazi N listica i na njima su zapisani brojevi $a_1, a_2, ..., a_N$. Obilježje je broj na listici. Ocjenjujemo nepoznato $a = \frac{a_1+...+a_N}{N}$.

a) model sa vraćanjem, vadimo n listica.

$P\{X_i = a_j\} = \frac{1}{N}, i = 1, 2, ..., n; j = 1, 2, ..., N$, $\frac{a_1+...+a_N}{N} = a = EX_i$, $DX_i = \sigma^2 = \frac{a_1^2+...+a_N^2}{N} - a^2$, $E\overline{X}_n = a$, $D\overline{X}_n = \frac{\sigma^2}{n}$. Familja dopustivih funkcija raspodjele nije parametarska.

b) model bez vraćanja, $n \leqslant N$.

$P\{X_i' = a_j\} = \frac{1}{N}, i = 1, 2, ..., n; j = 1, 2, ..., N$, $a = EX_i'$, $DX_i' = \sigma^2$, $\rho_{X_k', X_l'} = -\frac{1}{N-1}$, $E\overline{X}_n' = a$, $D\overline{X}_n' = \frac{\sigma^2}{n}\frac{N-n}{N-1}$, $D\overline{X}_n' < D\overline{X}_n' n \geqslant 2$.

Primjer. Neka funkcija raspodjele obilježja X pripada familiji $\mathfrak{R} = \{F(x, \theta),\, \theta \in \Theta\}$. Označimo sa $a = a(\theta) = \int\limits_{-\infty}^{\infty} x dF(x, \theta)$ i sa $\sigma^2 = \sigma^2(\theta) = \int\limits_{-\infty}^{\infty} (x - a(\theta)^2 dF(x, \theta)$ očekivanja i disperzije slučajnih promjenljivih čije funkcije raspodjele pripadaju $\mathfrak{R}$. Znamo, $E\overline{X}_n = a$, $D\hat{S}_n^2 = \sigma^2$, $\forall\theta \in \Theta$ te nepoznato $EX = a_0$ ocjenjujemo sa $\overline{X}_n$, a nepoznato $DX = \sigma_0^2$ sa $\hat{S}_n^2$. Kada sakupimo podatke $x_1, ..., x_n$ tada nepoznato a_0 aproksimiramo sa $\overline{x}_n$, a nepoznato σ_0^2 sa $\hat{s}_n^2$.

Teorema 0.4 *Neka je T_n asimptotski centrirana ocjena nepoznatog parametra τ i neka $DT_n \to 0$, $n \to \infty$. Tada je ocjena T_n postojana.*

$$P\{| T_n - \tau |> \varepsilon\} \leqslant P\{| T_n - ET_n |> \varepsilon\} + P\{| ET_n - \tau |> \varepsilon\} \leqslant \frac{DT_n}{\varepsilon^2} + P\{| ET_n - \tau |> \varepsilon\} \to 0.$$

Rao-Kramerova nejednakost.

Ocjenjujemo nepoznati parametar τ, kolekciju centriranih ocjena ćemo označiti sa $\mathcal{C}$. Za ocjenu $V \in \mathcal{C}$ kažemo da je najbolja ako je $DV' \leqslant DV$ za svako $V' \in \mathcal{C}$. Najbolja ocjena ne mora postojati. Naime, postoji model u kome je $W = \{DV,\, V \in \mathcal{C}\} = (1, 2)$.

Teorema 0.5 *Ako postoji najbolja ocjena ona je skoro izvjesno jedinstvena tj. ako je V najbolja ocjena i $V' \in \mathcal{C}$ takva da je $DV = DV' \Rightarrow V \overset{si}{=} V'$.*

Neka je U statistika za koju važi $EU = 0$. Očigledno je $V + \alpha U \in \mathcal{C}$ za $\forall\alpha \in \mathbb{R}$. Dalje, $D(V + \alpha U) = DV + 2\alpha cov(U, V) + \alpha^2 DV \geqslant DV \Rightarrow cov^2(U, V) \leqslant 0 \Rightarrow cov(U, V) = 0$. U posljednju jednakost uvrstimo $U = V - V'$. Kako statistike V i V' imaju isto očekivanje i disperziju, dobijamo $EV^2 = EVV' = EV'^2$. I na kraju $E(V - V')^2 = EV^2 - EVV' - EVV' + EV'^2 = 0 \Rightarrow V - V' \overset{si}{=} 0$.

Radićemo sa familijama raspodjela diskretnog i apsolutno neprekidnog tipa.

A) Za familiju raspodjela diskretnog tipa ćemo koristiti zapis $\{p(w_k, \theta),\, \theta \in \Theta\}$ gdje je $p(w_k, \theta) = P_\theta\{X = w_k\}$.

B) Za familiju raspodjela apsolutno neprekidnog tipa ćemo koristiti zapis $\{\varphi(x, \theta),\, \theta \in \Theta\}$ gdje je $\varphi(x, \theta)$ gustina obilježja.

Definicija 0.4 *Za familiju raspodjela diskretnog tipa kažemo da je regularna ako važi*

$$\sum_k \frac{\partial}{\partial\theta} p(w_k, \theta) = 0, \ \textit{za}\ \forall\theta \in \Theta.$$

Za familiju raspodjela apsolutno neprekidnog tipa kažemo da je regularna ako važi

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} \varphi(x, \theta) dx = 0, \ \textit{za}\ \forall\theta \in \Theta.$$

Teorema 0.6 Rao-Kramerova nejednakost. *Za svaku centriranu ocjenu* $V = f(X_1, ..., X_n)$ *parametra* θ *regularne familije* $\{p(w_k, \theta),\ \theta \in \Theta\}$ *odnosno* $\{\varphi(x, \theta),\ \theta \in \Theta\}$ *važi:*

$$DV \geqslant \frac{1}{nE\left(\frac{\partial}{\partial\theta}\ln p(X,\theta)\right)^2} \text{ odnosno } DV \geqslant \frac{1}{nE\left(\frac{\partial}{\partial\theta}\ln \varphi(X,\theta)\right)^2}.$$

$$\int\limits_{-\infty}^{\infty} \tfrac{\partial}{\partial\theta}\varphi(x_i,\theta)dx_i = \int\limits_{-\infty}^{\infty} \left[\tfrac{\partial}{\partial\theta}\ln\varphi(x_i,\theta)\right]\varphi(x_i,\theta)dx_i = 0, i = 1, 2, ...n.$$

$$EV = \int\limits_{\mathbb{R}^n}\!\!\ldots\!\int f(x_1,...,x_n)\varphi(x_1,\theta)...\varphi(x_n,\theta)dx_1...dx_n = \theta \Rightarrow \text{ (diferenciranje)}$$

$$1 = \int\limits_{\mathbb{R}^n}\!\!\ldots\!\int f(x_1,...,x_n)\Big[\sum_{i=1}^{n}\tfrac{1}{\varphi(x_i,\theta)}\tfrac{\partial}{\partial\theta}\varphi(x_i,\theta)\Big]\varphi(x_1,\theta)...\varphi(x_n,\theta)dx_1...dx_n =$$

$$\int\limits_{\mathbb{R}^n}\!\!\ldots\!\int f(x_1,...,x_n)\Big[\sum_{i=1}^{n}\tfrac{\partial}{\partial\theta}\ln\varphi(x_i,\theta)\Big]\varphi(x_1,\theta)...\varphi(x_n,\theta)dx_1...dx_n.$$

Neka je statistika $W = \sum\limits_{i=1}^{n}\frac{\partial}{\partial\theta}\ln\varphi(X_i,\theta)$. Imamo

$$EW = \sum_{i=1}^{n}\int\limits_{-\infty}^{\infty}\Big(\frac{\partial}{\partial\theta}\ln\varphi(x_i,\theta)\Big)\varphi(x_i,\theta)dx_i = 0, i = 1, ..., n;\ DW = nE\Big(\frac{\partial}{\partial\theta}\ln\varphi(X,\theta)\Big)^2.$$

Kako je $EVW = 1$ imamo $\varrho_{V,W} = \frac{1}{\sqrt{DVDW}} \Rightarrow DVDW \geqslant 1 \Rightarrow DV \geqslant \frac{1}{DW}$ Q.E.D.

Primjer. Ocjenjujemo nepoznatu vjerovatnoću padanja pisma. Novčić bacamo n puta. Obilježje X je indikator padanja pisma, X_i je indikator padanja pisma u $i$tom bacanju.

$$X_i : \begin{matrix} 0 & 1 \\ 1-p & p \end{matrix}, i = 1, 2, ..., n,\ \sum_{i=1}^{n} X_i : \mathcal{B}(n,p),\ E\overline{X}_n = p,\ D\overline{X}_n = \frac{p(1-p)}{n}.$$

Isti model imamo u slučaju kada ocjenjujemo vjerovatnoću p realizacije događaja A. Obilježje X je indikator događaja A, a X_i indikator događaja A u $i$tom ponavljanju opita.

Ocjenjujemo nepoznatu učestalost osoba čiji je koeficijent inteligencije između 105 i 115. $P\{105 < X < 115\} = p$, p je nepoznato. Zbog nezavisnosti i jednake raspodijeljenosti promjenljivih $X_1, .., X_n$ događaji $A_1 = \{105 < X_1 < 115\}, ..., A_n = \{105 < X_n < 115\}$ su nezavisni. Formiranjem slučajnih promjenljivih $X_1 = I_{A_1}, ..., X_n = I_{A_n}$ problem ocjene nepoznatog p (kao i formiranja odgovarajućeg intervala povjerenja) se rješava korišćenjem statistike $\overline{X}_n$.

Familija Benulijevih raspodjela je regularna. $X : \begin{matrix} 0 & 1 \\ 1-p & p \end{matrix}, 0 < p < 1,\ \frac{\partial}{\partial p}p(0,p) = -1,\ \frac{\partial}{\partial p}p(1,p) =$

1. Statistika $\overline{X}_n$ je najbolja centrirana ocjena parametra p. Naime, $D\overline{X}_n = \frac{p(1-p)}{n}$,

$$p(X,p) = (1-p)I\{X=0\} + pI\{X=1\},\ \frac{\partial}{\partial p}\ln p(X,p) = \frac{I\{X=1\}-I\{X=0\}}{(1-p)I\{X=0\}+pI\{X=1\}}$$

$$E\left(\frac{I\{X=1\}-I\{X=0\}}{(1-p)I\{X=0\}+pI\{X=1\}}\right)^2 = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)},$$

prvo smo integralili na skupu $\{\omega : X(\omega)=1\}$, a zatim na skupu $\{\omega : X(\omega)=0\}$. Dakle, disperzija uzoračke sredine je dostigla donju (Rao-Kramerovu) granicu.

Primjer. Familija Puasonovih raspodjela je regularna. Imamo: $p(k,\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}, k = 0,1,2,...,$ $\frac{\partial}{\partial\lambda}p(0.\lambda) = -e^{-\lambda}$, $\frac{\partial}{\partial\lambda}p(k.\lambda) = \left(\frac{l^{k-1}}{(k-1)!} - \frac{l^k}{k!}\right)e^{-\lambda}$, $k = 1,2,...,$ $\sum\limits_{k=0}^{\infty}\frac{\partial}{\partial\lambda}p(k,\lambda) = 0$. Statistika $\overline{X}_n$ je najbolja centrirana ocjena parametra λ.

Primjer. Familija $\mathcal{N}(m,\sigma_0^2), m \in \mathbb{R}, \sigma_0^2$ je poznato, je regularna. Statistika $\overline{X}_n$ je najbolja centrirana ocjena parametra m.

Primjer. Familija eksponencijalnih raspodjela, gustina $g(x,\theta) = \frac{1}{\theta}e^{-\frac{x}{\theta}}, x > 0,\ \theta > 0$ je regularna. Statistika $\overline{X}_n$ je najbolja centrirana ocjena parametra θ. Kako je $EX = \theta$, $DX = \theta^2$ to je $E\overline{X}_n = \theta$, $D\overline{X}_n = \frac{\theta^2}{n}$. Imamo $\frac{\partial}{\partial\theta}\ln g(x,\theta) = \frac{x}{\theta^2} - \frac{1}{\theta}$. I na kraju, $E\left(\frac{\partial}{\partial\theta}\ln g(X,\theta)\right)^2 = E\left(\frac{X-\theta}{\theta^2}\right)^2 = \frac{DX}{\theta^4} = \frac{1}{\theta^2}$ odakle slijedi da je dostignuta Rao-Kramerova granica.

Primjer. Familija $\mathcal{U}(0,\theta), \theta > 0$, nije regularna. Kako je $\varphi(x,\theta) = \frac{1}{\theta}, 0 < x < \theta$, imamo $\int\limits_0^\theta \frac{\partial}{\partial\theta}\varphi(x,\theta)dx = -\frac{1}{\theta}$. Statistika $T = \frac{n+1}{n}Y_n$ je centrirana i $DT = \frac{\theta^2}{n(n+2)}$. Zbog neregularnosti ne možemo primijeniti Rao-Kramerovu nejednakost. Primijetimo da je Rao-Kramerova donja granica disperzije ocjene reda $\frac{1}{n}$ dok je disperzija ocjene T u našem modelu reda $\frac{1}{n^2}$, dakle znatno je manja.

Metod maksimalne vjerodostojnost.

Primjer. Ocjenjujemo nepoznati broj N riba u ribnjaku. Specijalnom mrežom izvadimo n riba, markiramo ih i vratimo u ribnjak. Sačekamo neko vrijeme da bi se markirane pomiješale sa nemarkiranim ribama, a zatim ponovo vadimo n riba. Označimo sa X slučajnu promjenljivu koja je jednaka broju markiranih među izvađenim. Imamo

$$P_N\{X=r\} = \frac{\binom{n}{r}\binom{N-n}{n-r}}{\binom{N}{n}},\ r = 0,1,...,n.$$

Pretpostavimo da je broj registrovanih markiranih riba r_0. Primijetimo da je $P_N\{X=r_0\}$ funkcija od N. Nepoznato N ocjenjujemo onim brojem za koji funkcija $P_N\{X=r_0\}$ ima maksimum, a on se dostiže za cio broj najbliži broju $\frac{n^2}{r_0}$. Ako ocjenu označimo sa $\hat{N}$ imamo $\hat{N} \approx \frac{n^2}{r_0} \Rightarrow \frac{n}{\hat{N}} \approx \frac{r_0}{n}$ što je u saglasju sa intuicijom. Ocjenjujemo sa onom vrijednošću za N koja generiše model u kome je najveća vjerovatnoća da se registruje upravo r_0 markiranih riba. U ovakav način zaključivanja je ugrađena logika: ako se nešto desi, najrealnije je da se desilo u okviru modela u kome je vjerovatnoća dešavanja najveća.

Neka je $\mathbb{X} = (X_1, ..., X_n)$ uzorak obilježja X čija funkcija raspodjele pripada familiji

$$\mathfrak{R} = \{F(x, \theta),\, \theta \in \Theta\}.$$

Definisaśimo funkciju vjerodostojnosti u slučaju kada raspodjela obilježja pripada familiji raspodjela diskretnog tipa $\mathfrak{R} = \{p(w_k, \theta),\, \theta \in \Theta\}$, $p(w_k, \theta) = P_\theta\{X = w_k\}$.

Definicija 0.5 *Funkcija vjerodostojnost* $L(\theta, x_1, ..., x_n) := p(x_1, \theta) \cdots p(x_n, \theta) = P_\theta(X_1 = x_1) \cdots P_\theta(X_n = x_n) = P_\theta\{(X_1, ..., X_n) = (x_1, ..., x_n)\}$, $\theta \in \Theta$, *je funkcija čiji je argument* θ, *a* $(x_1, ..., x_n)$ *je neka realizacija uzorka obilježja* X, *raspodjela obilježja je određena sa* θ; $x_1, ..., x_n$ *tretiramo kao parametre.*

Iz definicije se vidi da je $L(\theta, x_1, ..., x_n)$ vjerovatnoća sa kojom se u modelu u kome je raspodjela obilježja određena sa θ dobija realizovani uzorak $(x_1, ..., x_n)$.

Definisaśimo sada funkciju vjerodostojnosti u slučaju kada raspodjela obilježja pripada familiji raspodjela apsolutno neprekidnog tipa $\mathfrak{R} = \{g(x, \theta),\, \theta \in \Theta\}$, g je gustina.

Definicija 0.6 *Funkcija vjerodostojnost* $L(\theta, x_1, ..., x_n) := g(x_1, \theta) \cdots g(x_n, \theta)$, $\theta \in \Theta$, *je funkcija čiji je argument* θ, *a* $(x_1, ..., x_n)$ *je neka realizacija uzorka obilježja* X, *raspodjela obilježja je određena sa* θ; $x_1, ..., x_n$ *tretiramo kao parametre.*

Na $L(\theta, x_1, ..., x_n)\Delta_1 \cdots \Delta_n$ možemo gledati kao približnu vjerovatnoću da uzorak $(X_1, ..., X_n)$ iz raspodjele koja je određena sa θ, "upadne" u "mali" paralelopiped $[x_1, x_1 + \Delta_1) \times ... \times [x_n, x_n + \Delta_n)$. Ovo slijedi iz

$$P_\theta\{\mathbb{X} \in [x_1, x_1 + \Delta_1) \times ... \times [x_n, x_n + \Delta_n)\} = L(\theta, x_1, ..., x_n)\Delta_1 \cdots \Delta_n + o(\Delta_1 \cdots \Delta_n),\; \Delta_1, ..., \Delta_n \to 0.$$

Dakle, $L(\theta, x_1, ..., x_n)$ "mjeri" šansu da se dobije realizovani uzorak u gore pomenutom malom paralelopipedu. Slobodnije, funkcija vjerodostojnosti govori o izgledima sa kojima u konkretnim modelima tj. u zavisnosti od parametra θ, realizovani uzorak "upada" u pojedine oblasti iz $\mathbb{R}^n$.

Pretpostavimo da je $\Theta = \{\theta_1, \theta_2\}$ i da je recimo $L(\theta_1, x_1, ..., x_n) < L(\theta_2, x_1, ..., x_n)$. Smatrajući da se realizuje ono što ima veću vjerovatnoću, kao ocjenu ćemo uzeti θ_2.

Za ocjenu $\hat{\theta} = h(X_1, ..., X_n)$, gdje je $\hat{\theta} = h(x_1, ..., x_n)$ vrijednost parametra θ za koju funkcija $L(\theta, x_1, ..., x_n)$, $\theta \in \Theta$ dostiže maksimum, kažemo da je **ocjena maksimalne vjerodostojnosti**.

Kako su kod funkcije $L(\theta, x_1, ..., x_n)$, $\theta \in \Theta$, $x_1, ..., x_n$ parametri, $\hat{\theta}$ za koje funkcija dostiže maksimum se dobija kao funkcija od $(x_1, ..., x_n)$. Kada u raspodjeli obilježja figuriše eksponencijalna funkcija, umjesto traženja tačke maksimuma za $L(\theta, x_1, ..., x_n)$ jednostavnije je tražiti tačku maksimuma za $\ln L(\theta, x_1, ..., x_n)$. Dobija se isti rezultat jer zbog monotonosti funkcije ln funkcije $L(\theta, x_1, ..., x_n)$ i $\ln L(\theta, x_1, ..., x_n)$ maksimum dostižu u istoj tački.

Primjer. $X : \mathcal{U}(0, \theta)$, $\theta > 0$. Naći ocjenu maksimalne vjerodostojnosti.

Znamo, $g(x, \theta) = \frac{1}{\theta} I(0 \leqslant x \leqslant \theta)$. Imamo,

$$L(\theta, x_1, ..., x_n) = \frac{1}{\theta} I(\theta \geqslant x_1) \cdots \frac{1}{\theta} I(\theta \geqslant x_n) = \frac{1}{\theta^n} I(\theta \geqslant y_n),\ y_n = \max_{1 \leqslant k \leqslant n} x_k.$$

Dakle, $\hat{\theta} = Y_n$. U prethodnoj analizi smo koristili činjenicu da je $(x_1, ..., x_n) \in (0, \theta)^n$. Da je u funkciji $L(\theta, x_1, ..., x_n)$ vektor $(x_1, ..., x_n)$ tretiran kao vektor iz $\mathbb{R}^n$, tada bi imali

$$L(\theta, x_1, ..., x_n) = \frac{1}{\theta} I(\theta \geqslant x_1) I(x_1 \geqslant 0) \cdots \frac{1}{\theta} I(\theta \geqslant x_n) I(x_n \geqslant 0) = \frac{1}{\theta^n} I(\theta \geqslant y_n) I(y_1 \geqslant 0).$$

Primjer. a) $X : \mathcal{N}(m, 1), m \in \mathbb{R}$. b) $X : \mathcal{N}(m, 1), m \geqslant m_0$.

a) $L(m, x_1, ..., x_n) = (2\pi)^{-\frac{n}{2}} e^{-\frac{\sum_{k=1}^{n} (x_k - m)^2}{2}}$, $\frac{\partial}{\partial m} \ln L(m, x_1, ..., x_n) = \sum_{k=1}^{n} (x_k - m)$. Iz $\frac{\partial}{\partial m} \ln L = 0$ slijedi $\hat{m} = \overline{x}_n$. Dakle, $\hat{m} = \overline{X}_n$.

b) $\ln L(m, x_1, ..., x_n) = -\frac{n}{2} m^2 + m \sum_{k=1}^{n} x_k - \sum_{k=1}^{n} x_k^2 - \frac{n}{2} \ln(2\pi)$, $m \geqslant m_0$. Nakon traženja maksimuma upravo dobijene kvadratne funkcije, dobijamo $\hat{m} = \max\{m_0, \overline{X}_n\}$.

Intervali povjerenja

Neka je $\mathbb{X} = (X_1, ..., X_n)$ uzorak obilježja X čija funkcija raspodjele pripada familiji

$$\mathfrak{R} = \{F(x, \theta),\ \theta \in \Theta\}.$$

Neka su $\theta_1 = \theta_1(X_1, ..., X_n)$ i $\theta_2 = \theta_2(X_1, ..., X_n)$ statistike takve da za $\forall \theta \in \Theta$ važi:

a) $P_\theta\{\theta_1 < \theta_2\} = 1$,

b) $P_\theta\{\theta_1 < \theta < \theta_2\} = \gamma$.

Tada se interval (θ_1, θ_2) naziva interval povjerenja za nepoznati parametar θ, a broj γ se naziva nivo povjerenja. Na osnovu zakona velikih brojeva, približno $100\gamma\%$ realizovanih intervala sadrži nepoznati parametar θ.

Primjer. $X : \mathcal{N}(m, \sigma_0^2)$, $m \in \mathbb{R}$. Znamo, $\frac{\overline{X}_n - m}{\sigma_0}\sqrt{n} : \mathcal{N}(0,1)$, $\forall m \in \mathbb{R}$. Imamo

$$P_m\Big\{-z_\gamma < \frac{\overline{X}_n - m}{\sigma_0}\sqrt{n} < z_\gamma\Big\} = \gamma \Leftrightarrow P_m\Big\{\overline{X}_n - \frac{\sigma_0}{\sqrt{n}}z_\gamma < m < \overline{X}_n - \frac{\sigma_0}{\sqrt{n}}z_\gamma\Big\} = \gamma, \int_0^{z_\gamma} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}dt = \frac{\gamma}{2}.$$

Zaključujemo da je $I = \Big(\overline{X}_n - \frac{\sigma_0}{\sqrt{n}}z_\gamma, \overline{X}_n + \frac{\sigma_0}{\sqrt{n}}z_\gamma\Big)$. Dužina intervala je $d(I) = 2\frac{\sigma_0}{\sqrt{n}}z_\gamma$. Pokazuje se da

$$\frac{Y_{[\frac{n}{2}]} - m}{\sigma_0}\sqrt{\frac{2n}{\pi}} \xrightarrow{R} X^*$$

odakle slijedi da je $I = \Big(Y_{[\frac{n}{2}]} - \sqrt{\frac{\pi}{2}}\frac{\sigma_0}{\sqrt{n}}z_\gamma, Y_{[\frac{n}{2}]} + \sqrt{\frac{\pi}{2}}\frac{\sigma_0}{\sqrt{n}}z_\gamma\Big)$ i $d(I) = 2\sqrt{\frac{\pi}{2}}\frac{\sigma_0}{\sqrt{n}}z_\gamma$.

Primjer. $X : \mathcal{N}(m, \sigma^2)$, $m \in \mathbb{R}, \sigma^2 > 0$. Znamo, $\frac{\overline{X}_n - m}{\hat{S}_n}\sqrt{n} : t_{n-1}$, $\forall m \in \mathbb{R}$. Imamo

$$P_{m,\sigma^2}\Big\{-t_\gamma < \frac{\overline{X}_n - m}{\hat{S}_n}\sqrt{n} < t_\gamma\Big\} = \gamma \Leftrightarrow P_{m,\sigma^2}\Big\{\overline{X}_n - \frac{\hat{S}_n}{\sqrt{n}}t_\gamma < m < \overline{X}_n - \frac{\hat{S}_n}{\sqrt{n}}t_\gamma\Big\} = \gamma,$$

za t_γ važi $P\{0 < t_{n-1} < t_\gamma\} = \frac{\gamma}{2}$ i nalazimo ga u tablici Studentove raspodjele. Zaključujemo da je $I = \Big(\overline{X}_n - \frac{\hat{S}_n}{\sqrt{n}}t_\gamma, \overline{X}_n - \frac{\hat{S}_n}{\sqrt{n}}t_\gamma\Big)$.

Primjer. $X : \mathcal{N}(m, \sigma^2)$, $m \in \mathbb{R}, \sigma^2 > 0$. Znamo da $\frac{n\overline{S}_n^2}{\sigma^2} : \chi^2_{n-1}$. Neka je v_γ broj takav da je $P_{m,\sigma^2}\{\chi^2_{n-1} < v_\gamma\} = \gamma$.

$$a)\ P_{m,\sigma^2}\Big\{v_{n-1,\frac{1-\gamma}{2}} < \frac{n\overline{S}_n^2}{\sigma^2} < v_{n-1,\frac{1+\gamma}{2}}\Big\} = \gamma = P_{m,\sigma^2}\left\{\frac{n\overline{S}_n^2}{v_{n-1,\frac{1+\gamma}{2}}} < \sigma^2 < \frac{n\overline{S}_n^2}{v_{n-1,\frac{1-\gamma}{2}}}\right\}.$$

Dobijamo $I = \left(\frac{n\overline{S}_n^2}{v_{n-1,\frac{1+\gamma}{2}}}, \frac{n\overline{S}_n^2}{v_{n-1,\frac{1-\gamma}{2}}}\right)$.

b) $P_{m,\sigma^2}\Big\{\frac{n\overline{S}_n^2}{\sigma^2} > v_{n-1,1-\gamma}\Big\} = \gamma = P_{m,\sigma^2}\Big\{0 < \sigma^2 < \frac{n\overline{S}_n^2}{v_{n-1,1-\gamma}}\Big\} \Rightarrow I = \Big(0, \frac{n\overline{S}_n^2}{v_{n-1,1-\gamma}}\Big)$.

c) $P_{m,\sigma^2}\Big\{\frac{n\overline{S}_n^2}{\sigma^2} < v_{n-1,\gamma}\Big\} = \gamma = P_{m,\sigma^2}\Big\{\sigma^2 > \frac{n\overline{S}_n^2}{v_{n-1,\gamma}}\Big\} \Rightarrow I = \Big(\frac{n\overline{S}_n^2}{v_{n-1,\gamma}}, \infty\Big)$.

Trebaće nam dva tvrđenja iz teorije. Neka za obilježje X važi $EX = a$, $DX = \sigma^2$. Znamo $\frac{\overline{X}_n - a}{\sigma}\sqrt{n} \longrightarrow X^* : \mathcal{N}(0,1)$. Ako je σ_n^2 postojana ocjena σ^2, tada $\frac{\overline{X}_n - a}{\sigma_n}\sqrt{n} \longrightarrow X^*$. Takođe, ako $X_n \xrightarrow{p} X$ i ako je f neprekidna funkcija, tada $f(X_n) \xrightarrow{p} f(X)$.

Primjer. Obilježje $X : \mathcal{P}(\lambda)$, $\lambda > 0$. Znamo da je $EX = DX = \lambda$ i da je $\overline{X}_n$ postojana ocjena λ. Iz $\frac{\overline{X}_n - \lambda}{\sqrt{\overline{X}_n}}\sqrt{n} \longrightarrow X^* \Rightarrow I = \left(\overline{X}_n - z_\gamma\sqrt{\frac{\overline{X}_n}{n}}, \overline{X}_n + z_\gamma\sqrt{\frac{\overline{X}_n}{n}}\right)$.

Primjer. Obilježje $X : \mathcal{B}(1,p), 0 < p < 1$, tj. Bernulijevu raspodjelu. Znamo, $S_n = \sum\limits_{k=1}^{n} X_k$:

$\mathcal{B}(n,p)$, $EX_i = p$, $DX_i = p(1-p)$, $\overline{X}_n = \frac{S_n}{n}$, $\frac{S_n}{n}\Big(1-\frac{S_n}{n}\Big) \xrightarrow{p} p(1-p)$. Imamo,

$$\frac{\frac{S_n}{n}-p}{\sqrt{\frac{S_n}{n}\big(1-\frac{S_n}{n}\big)}}\sqrt{n} \xrightarrow{p} X^* \Rightarrow I = \left(\frac{S_n}{n} - \frac{z_\gamma}{\sqrt{n}}\sqrt{\frac{S_n}{n}\Big(1-\frac{S_n}{n}\Big)}, \frac{S_n}{n} + \frac{z_\gamma}{\sqrt{n}}\sqrt{\frac{S_n}{n}\Big(1-\frac{S_n}{n}\Big)}\right).$$

Primjer. U kutiji se nalazi M bijelih i $N-M$ crnih kuglica. Ocjenjujemo relativnu učestalost bijelih kuglica to jest $\frac{M}{N}$. Ono što ćemo uraditi istovremeno obuhvata slučajeve a) M nepoznato, N poznato i b) M nepoznato i N nepoznato.

1. Model sa vraćanjem, vadimo n kuglica.

$$X_i : \begin{matrix} 0 & 1 \\ 1-\frac{M}{N} & \frac{M}{N} \end{matrix}, i = 1,2,...,n, \sum_{i=1}^{n} X_i : \mathcal{B}(n, \frac{M}{N}), E\overline{X}_n = \frac{M}{N}, D\overline{X}_n = \frac{M}{nN}(1-\frac{M}{N}).$$

$(X_1, ..., X_n)$ je prost slučajni uzorak.

2. Model bez vraćanja, vadimo n kuglica, $n \leqslant N$.

$$X_i' : \begin{matrix} 0 & 1 \\ 1-\frac{M}{N} & \frac{M}{N} \end{matrix}, i = 1,2,...,n, \rho_{X_k', X_l'} = \frac{\frac{n-1}{N-1}-\frac{n}{N}}{1-\frac{n}{N}}, E\overline{X}_n' = \frac{M}{N}, D\overline{X}_n' = \frac{M}{nN}(1-\frac{M}{N})\frac{N-n}{N-1}.$$

Konstatujmo: ako je $\lim\limits_{n\to\infty} \frac{n}{N} = \gamma$, $0 < \gamma \leqslant 1$, tada je $\lim\limits_{n\to\infty} \frac{D\overline{X}_n'}{D\overline{X}_n} = 1-\gamma$; ako je $\lim\limits_{n\to\infty} \frac{n}{N} = 0$, tada je $\lim\limits_{n\to\infty} \frac{D\overline{X}_n'}{D\overline{X}_n} = 1$. Takođe, iz $D\overline{X}_n' < D\overline{X}_n$, $n = 2,3,...$ slijedi da je ocjena $\overline{X}_n'$ bolja. U uzorku $(X_1', ..., X_n')$ komponente su jednako raspodijeljene ali su zavisne.

Veza modela sa vraćanjem i bez vraćanja Statistika koja prebrojava bijele kuglice u uzorku obima n (model bez vraćanja) ima hipergeometrijsku raspodjelu. Označimo tu statistiku sa W. Razmotrimo granični slučaj u kome $\frac{M}{N} \to p$. Lako se provjerava

$$P\{W = k\} \to \binom{n}{k} p^k (1-p)^{n-k}, k = 0,1,...,n.$$

Dakle, u graničnom slučaju, statistika W ima $\mathcal{B}(n,p)$ raspodjelu. Budući da statistika $\sum_{i=1}^{n} X_i$ iz modela 1 ima $\mathcal{B}(n, \frac{M}{N})$ raspodjelu, dobijeni granični rezultat u praksi koristimo na sljedeći način: Ako je broj članova populacije N velik, uzoračke sredine tj. srednji broj izvučenih bijelih kuglica u oba modela imaju istu raspodjelu (preciznije približno istu, ali razlika je zanemarljiva). Ovo zapažanje je značajno kod formiranja intervala povjerenja nepoznate učestalosti. Naime, podatke biramo po modelu bez vraćanja. a kad formiramo interval povjerenja, koristimo rezultat dobijen za model sa vraćanjem. Na primjer, formiramo interval povjerenja za nepoznatu učestalost pušača.

1 Testiranje statističkih hipoteza

Pojam statističke hipoteze i ideju testiranja približimo kroz jedan primjer.

Primjer 1. U kutiji se nalazi 10 kuglica i znamo da je kutija napunjena po jednoj od dvije strategije:

0) Sa 9 kuglica na kojima je broj 1 i jednom kuglicom na kojoj je broj 2.

1) Sa 9 kuglica na kojima je broj 2 i jednom kuglicom na kojoj je broj 1.

Dozvoljeno nam je da iz kutije izvučemo 4 kuglice po modelu sa vraćanjem. Na osnovu izvučenih kuglica tj. brojeva na njima, treba da se odlučimo za jednu od dvije hipoteze (pretpostavke, mogućnosti):

1^0 Kutija je napunjena po strategiji 0 – govorićemo o hipotezi H_0.

2^0 Kutija je napunjena po strategiji 1 – govorićemo o hipotezi H_1.

Postupak presuđivanja u korist jedne od hipoteza, tj. postupak prihvatanja jedne od hipoteza, zvaćemo **testom**. Ako je kutija napunjena po strategiji 0, tada je zbog velike vjerovatnoće pojave broja 1, vjerovatnoća da se registruje mala suma brojeva, velika. Ako je kutija napunjena po strategiji 1, tada je zbog velike vjerovatnoće pojave broja 2, vjerovatnoća da se registruje velika suma brojeva, velika. Jasno, suma brojeva je u rasponu od 4 do 8 i kada govorimo o maloj odnosno velikoj sumi imamo u vidu male odnosno velike vrijednosti u odnosu na interval omeđen brojevima 4 i 8. Nakon ove analize, nameće se kao razuman sljedeći postupak odlučivanja: Ako je zbir izvučenih brojeva $\geqslant 7$ prihvatićemo H_1, a H_0 odbaciti. U suprotnom tj. ako je zbir izvučenih brojeva < 7 prihvatićemo H_0, a H_1 odbaciti.

Označimo sa X obilježje koje predstavlja broj na izvučenoj kuglici. U slučaju kada je važeća strategija 0, obilježje X ima raspodjelu

$$X : \begin{array}{cc} 1 & 2 \\ \frac{9}{10} & \frac{1}{10} \end{array},$$

a u slučaju kada je važeća strategija 1, obilježje X ima raspodjelu

$$X : \begin{array}{cc} 1 & 2 \\ \frac{1}{10} & \frac{9}{10} \end{array}.$$

Ove dvije raspodjele možemo tretirati kao familiju raspodjela obilježja X. U uzorku (X_1, X_2, X_3, X_4) komponente predstavljaju redom prvi, drugi, treći i četvrti izvučeni broj. U statistici je uobičajeno da se hipoteze izražavaju u terminima raspodjela. U našem primjeru H_0 je hipoteza da obilježje X ima gornju, a H_1 donju raspodjelu.

Ako je (x_1, x_2, x_3, x_4) realizovani uzorak, tada se uslov "zbir izvučenih brojeva je $\geqslant 7$" zapisuje sa $x_1 + x_2 + x_3 + x_4 \geqslant 7$. Ovaj uslov generiše oblast

$$C = \{(x_1, x_2, x_3, x_4) : x_1 + x_2 + x_3 + x_4 \geqslant 7\},\ C \subset \mathbb{R}^4.$$

Nakon uvođenja oblasti C, pravilo odlučivanja možemo ovako formulisati: Ako realizovani uzorak (x_1, x_2, x_3, x_4) "upadne" u oblast C, tada prihvatamo H_1, a ako "upadne" u C^c, tada prihvatamo H_0. Oblast C određuje postupak – pravilo odlučivanja tj. test. "Upadanje" uzorka u oblast C ne odgovara hipotezi H_0 (tada imamo onu situaciju kada je zbir izvučenih brojeva veliki tj. među brojevima dominira 2; ostvaruje se događaj čija je vjerovatnoća realizacije mala ako je tačna hipoteza H_0). Zbog toga oblast C nazivamo kritična oblast za H_0.

U postupku odlučivanja nema izričitosti. Mi samo konstatujemo da na osnovu registrovanih brojeva prednost dajemo jednoj hipotezi (prihvatamo jednu hipotezu). Naravno, postoji mogućnost greške. Grešku pravimo kada povodeći se za pravilom odlučivanja odbacimo hipotezu koja je faktički tačna. Preciznije, moguće je da odbacimo H_0 koja je tačna i samim tim prihvatimo H_1 koja je netačna. Jasno, moguća je i situacija u kojoj H_0 i H_1 imaju zamijenjene uloge.

Izračunajmo vjerovatnoću α da odbacimo tačnu hipotezu H_0.

$$\alpha = P_{H_0}\{(X_1, X_2, X_3, X_4) \in C\} = \frac{1}{10^4} + 4 \cdot \frac{9}{10^4} = 0,0037.$$

Izračunajmo vjerovatnoću β da odbacimo tačnu hipotezu H_1.

$$\beta = P_{H_1}\{(X_1, X_2, X_3, X_4) \in C^c\} = \frac{1}{10^4} + 4 \cdot \frac{9}{10^4} + 6 \cdot \frac{9^2}{10^4} = 0,0523.$$

Interesantno je vidjeti šta se dešava ako promijenimo test na taj način što za kritičnu oblast za H_0 sada uzmemo

$$D = \{(x_1, x_2.x_3, x_4) : x_1 + x_2 + x_3 + x_4 \geqslant 8\},\ D \subset \mathbb{R}^4.$$

Lako se dobija $\alpha = 0,0001$, $\beta = 0,3439$. Primijetimo, vjerovatnoća da se odbaci faktički tačna H_0 je znatno smanjena, ali je vjerovatnoća odbacivanja faktički tačne H_1 postala enormno velika. Praktično, u jednom od tri slučaja ćemo odbaciti tačnu hipotezu H_1. U ovakvoj situaciji je bolje testiranje obaviti prvim postupkom.◀

Motivisani prethodnim primjerom, možemo preći na izlaganje teorije.

Neka je X obilježje čija funkcija raspodjele vjerovatnoća pripada familiji

$$\mathfrak{R} = \{F(x, \theta),\ \theta \in \Theta\}.$$

Pretpostavka oblika $H_0(\theta \in \Theta_0)$, $\Theta_0 \subset \Theta$, zove se **statistička hipoteza**. Malo drugačije rečeno, pretpostavka se sastoji u tome da obilježje X ima raspodjelu koja pripada užoj familiji određenoj parametarskim skupom Θ_0. Ako je Θ_0 jednočlani skup, tada kažemo da je hipoteza H_0 prosta. U protivnom govorimo o složenoj hipotezi. Hipotezi H_0 ćemo suprotstaviti hipotezu $H_1(\theta \in \Theta \backslash \Theta_0)$. Hipoteza H_o se naziva nulta, a hipoteza H_1 alternativna. Postupak odlučivanja u korist jedne hipoteze, tj. postupak prihvatanja jedne od hipoteza, na osnovu realizovanog uzorka se naziva **statistički test**. Taj postupak je određen zadavanjem kritične oblasti $C \subset \mathbb{R}^n$ i sprovodi se na sljedeći način: Ako $(x_1, ..., x_n) \in C$ tada H_0 odbacujemo u korist H_1, a ako $(x_1, ..., x_n) \in C^c$ tada H_1 odbacujemo u korist H_0. Zbog upravo izloženog se kaže da kritična oblast zadaje test.

Posvetimo se slučaju kada je $\Theta = \{\theta_0, \theta_1\}$, $H_0(\theta = \theta_0), H_1(\theta = \theta_1)$, dakle obje hipoteze su proste. Pretpostavimo da je $C \subset \mathbb{R}^n$ kritična oblast koja zadaje test.

Prilikom odlučivanja, postoji mogućnost da se napravi greška.

1^o **Grešku prve vrste** pravimo kada odbacimo faktički tačnu hipotezu H_0.

2^o **Grešku druge vrste** pravimo kada odbacimo faktički tačnu hipotezu H_1.

Vjerovatnoća greške prve vrste se označava sa α i za nju, na osnovu rečenog, važi

$$\alpha = P_{H_0}\{(X_1, \ldots, X_n) \in C\}.$$

α se naziva **pragom značajnosti testa**, a C se naziva **kritična oblast veličine** α.

Vjerovatnoća greške druge vrste se označava sa β i za nju, na osnovu rečenog, važi

$$\beta = P_{H_1}\{(X_1, \ldots, X_n) \in C^c\}.$$

Prirodna je potreba da se pronađe test u kome su brojevi α i β mali. Kod rješavanja ovog zadatka poteškoće izviru iz činjenice da smanjivanje jednog od ova dva parametra povlači uvećavanje drugog. Postupamo na sljedeći način. Među svim skupovima $S \subset \mathbb{R}^n$ za koje je

$$P_{\theta_0}\{(X_1, ..., X_n) \in S\} = \alpha$$

tražimo skup C za koji je vjerovatnoća $P_{\theta_1}\{(X_1, ..., X_n) \in C^c\}$ najmanja. Ako skup C postoji nazivamo ga **najbolja kritična ooblast veličine** α, a odgovarajući test **najbolji test sa pragom značajnosti** α. U nekim modelima postoji efektivni postupak za dobijanje najbolje kritične oblasti. Postupak dobijanja najbolje kritične oblasti generiše Nejman-Pirsonova lema. Mi ćemo Nejman-Pirsonovu lemu formulisati u slučaju kada obilježje ima apsolutno neprekidnu raspodjelu. Gustina koja odgovara raspodjeli zadatoj parametrom θ_0 je $g_0(x)$, gustina koja odgovara raspodjeli zadatoj parametrom θ_1 je $g_1(x)$.

U tekstu koji slijedi ćemo koristiti jedan detalj koji smo obrazložili u okviru predmeta Teorija vjerovatnoće. Ako je $h : \mathbb{R}^n \to \mathbb{R}$ Borelova funkcija, $(X_1, ..., X_n)$ slučajni vektor sa funkcijom gustine $g(x_1, ..., x_n)$, tada je

$$p\{h(X_1, ..., X_n) \in W\} = \int \underset{S}{\dots} \int g(x_1, ..., x_n) dx_1 ... dx_n, S = \{(x_1, ..., x_n) : h(x_1, ..., x_n) \in W\}.$$

Definišimo jednu statistiku i jednu funkciju.

$$K(\mathbf{X}) = K(X_1, ..., X_n) := \frac{L(\theta_1, X_1, ..., X_n)}{L(\theta_0, X_1, ..., X_n)} = \frac{L(\theta_1, \mathbf{X})}{L(\theta_0, \mathbf{X})} = \frac{g_1(X_1) ... g_1(X_n)}{g_0(X_1) ... g_0(X_n)},$$
$$\mathbf{X} = (X_1, ..., X_n),\ \mathbf{x} = (x_1, ..., x_n), h(c) := P_{\theta_0}\{K(X_1, ..., X_n) \geqslant c\},\ c \geqslant 0.$$

Primijetimo, $h(c) \downarrow$ i $h(0) = 1$. Neka je $A_c = \{(x_1, ..., x_n) : K(x_1, ..., x_n) \geqslant c\}$. Sada je

$$1 \geqslant P_{\theta_1}\{K(X_1, ..., X_n) \geqslant c\} = \int_{A_c} L(\theta_1, \mathbf{x}) d\mathbf{x} = \int_{A_c} g_1(x_1) ... g_1(x_n) d\mathbf{x} \geqslant c \int_{A_c} g_0(x_1) ... g_0(x_n) d\mathbf{x}$$
$$= c \int_{A_c} L(\theta_0, \mathbf{x}) d\mathbf{x} = c P_{\theta_0}\{K(X_1, ..., X_n) \geqslant c\} = c h(c) \Rightarrow h(c) \leqslant \frac{1}{c} \Rightarrow \lim_{c \to \infty} h(c) = 0.$$

Ako je $h(c)$ neprekidna funkcija, tada za proizvoljni broj $\alpha \in (0, 1)$ postoji $c \in \mathbb{R}^+$ takav da je $h(c) = \alpha$.

Nejman Pirsonova lema. Neka za dato $\alpha \in (0, 1)$ postoji $c > 0$ takav da je $h(c) = \alpha$. Tada postoji najbolja kritična oblast veličine α za testiranje $H_0(\theta = \theta_0)$ protiv $H_1(\theta = \theta_1)$ i data je sa $W_0 = \{\mathbf{x} : K(\mathbf{x}) \geqslant c\}$.

Dokaz. Primijetimo, $\alpha = h(c) = P_{\theta_0}\{K(\mathbf{X}) \geqslant c\} = P_{\theta_0}\{\mathbf{X} \in W_0\}$. Neka je W proizvoljna kritična oblast veličine α tj. $\alpha = P_{\theta_0}\{\mathbf{X} \in W\}$.

Treba da pokažemo $P_{\theta_1}\{\mathbf{X} \in W^c\} \geqslant P_{\theta_1}\{\mathbf{X} \in W_0^c\}$ što je ekvivalentno sa $P_{\theta_1}\{\mathbf{X} \in W\} \leqslant P_{\theta_1}\{\mathbf{X} \in W_0\}$.

$$P_{\theta_1}\{\mathbf{X} \in W\} - P_{\theta_1}\{\mathbf{X} \in W_0\}$$
$$= \int_{W \cap W_0^c} L(\theta_1, \mathbf{x}) d\mathbf{x} + \int_{W \cap W_0} L(\theta_1, \mathbf{x}) d\mathbf{x} - \int_{W \cap W_0} L(\theta_1, \mathbf{x}) d\mathbf{x} - \int_{W^c \cap W_0} L(\theta_1, \mathbf{x}) d\mathbf{x}$$
$$\leqslant c\Big(\int_{W \cap W_0^c} L(\theta_0, \mathbf{x}) d\mathbf{x} - \int_{W^c \cap W_0} L(\theta_0, \mathbf{x}) d\mathbf{x}\Big) = c\Big(\int_{W \cap W_0^c} + \int_{W \cap W_0} - \int_{W \cap W_0} - \int_{W^c \cap W_0}\Big)$$
$$= c\Big(\int_W L(\theta_0, \mathbf{x}) d\mathbf{x} - \int_{W_0} L(\theta_0, \mathbf{x}) d\mathbf{x}\Big) = c(\alpha - \alpha) = 0.$$

Primjer. Obilježje X ima $\mathcal{N}(m, \sigma_0^2)$, σ_0^2 poznato, raspodjelu (gustina $g(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(x-m)^2}{2\sigma_0^2}}$).

Tražimo najbolju kritičnu oblast veličine α (najbolji test sa pragom značajnosti α) za testiranje $H_0(m = m_0)$ protiv $H_1(m = m_1)$, $m_1 > m_0$.

$$h(c) = P_{m_0}\left\{ \frac{\frac{1}{\sqrt{2\pi\sigma_0^2}^n} e^{-\sum\limits_{k=1}^{n} \frac{(X_k - m_1)^2}{2\sigma_0^2}}}{\frac{1}{\sqrt{2\pi\sigma_0^2}^n} e^{-\sum\limits_{k=1}^{n} \frac{(X_k - m_0)^2}{2\sigma_0^2}}} \geqslant c \right\} = P_{m_0}\left\{ e^{\frac{(m_1-m_0)}{\sigma_0^2}\sum\limits_{k=1}^{n} X_k + \frac{n(m_0^2 - m_1^2)}{2\sigma_0^2}} \geqslant c \right\} =$$

$$P_{m_0}\left\{ \overline{X}_n \geqslant \frac{\sigma_0^2 \ln c}{(m_1-m_0)n} + \frac{m_1+m_0}{2} \right\} = P_{m_0}\left\{ \frac{\overline{X}_n - m_0}{\sigma_0}\sqrt{n} \geqslant \frac{\sigma_0 \ln c}{(m_1-m_0)\sqrt{n}} + \frac{(m_1-m_0)\sqrt{n}}{2\sigma_0} \right\}.$$

Neka je $w(c) := \frac{\sigma_0 \ln c}{(m_1-m_0)\sqrt{n}} + \frac{(m_1-m_0)\sqrt{n}}{2\sigma_0}$, $c > 0$. Funkcija $w(c)$ je neprekidna i strogo monotono rastuća. Budući da statistika $\frac{\overline{X}_n - m_0}{\sigma_0}\sqrt{n} : \mathcal{N}(0,1)$ u modelu u kome $X : \mathcal{N}(m_0, \sigma_0^2)$, zaključujemo da je $h(c) = P\{X^* \geqslant w(c)\}$. Funkcija $h(c)$ je neprekidna i strogo monotono opadajuća, te postoji jedinstveno c takvo da je $h(c) = \alpha$ i to c je rješenje jednačine $w(c) = z_{0,5-\alpha} = w_\alpha$. Dakle, $W_0 = \left\{(x_1, ..., x_n) : \frac{\overline{x}_n - m_0}{\sigma_0}\sqrt{n} \geqslant w_\alpha\right\} = \left\{(x_1, ..., x_n) : \overline{x}_n \geqslant m_0 + \frac{\sigma_0 w_\alpha}{\sqrt{n}}\right\}$. U slučaju testiranja $H_0(m = m_0)$ protiv $H_1(m = m_1)$, $m_1 < m_0$, primjenom istog postupka se dobija:

$$W_0 = \left\{(x_1, ..., x_n) : \frac{\overline{x}_n - m_0}{\sigma_0}\sqrt{n} \leqslant -w_\alpha\right\} = \left\{(x_1, ..., x_n) : \overline{x}_n \leqslant m_0 - \frac{\sigma_0 w_\alpha}{\sqrt{n}}\right\}.$$

U slučaju $n = 9$, $\alpha = 0,05$, $m_0 = 0$, $m_1 = 1$, $\sigma_0^2 = 1$ imamo: $W_0 = \{(x_1, ..., x_9) : \overline{x}_9 \geqslant \frac{1{,}65}{3}\} = \{(x_1, ..., x_9) : \overline{x}_9 \geqslant 0,55\}$, dok je $\beta = P_1\{\overline{x}_9 < 0,55\} = P_1\{(\overline{x}_9 - 1)3 < -1,35\} = 0,09$.

Vratimo se na opšti slučaj i nađimo najmanje n takvo da je uz zadato α vjerovatnoća greške druge vrste $\leqslant \beta$. Iz

$$P_{m_1}\left\{\overline{X}_n < m_0 + \frac{\sigma_0 w_\alpha}{\sqrt{n}}\right\} = P\left\{X^* < \frac{m_0 - m_1}{\sigma_0}\sqrt{n} + w_\alpha\right\} \leqslant \beta$$

Iz

$$\frac{m_0 - m_1}{\sigma_0}\sqrt{n} + w_\alpha < -w_\beta \Rightarrow n \geqslant \left(\frac{\sigma_0(w_\alpha + w_\beta)}{m_1 - m_0}\right)^2.$$

Primjer. $X : \mathcal{E}(\theta)$ (tj. $g(x) = \theta e^{-\theta x}$, $x > 0$, $\theta > 0$). Znamo $EX = \frac{1}{\theta}$; $DX = \frac{1}{\theta^2}$ i u slučaju kad je

n veliko $\left(\overline{X}_n - \frac{1}{\theta}\right)\theta\sqrt{n} : \mathcal{N}(0,1)$. Testirajmo $H_0(\theta = \theta_0)$ protiv $H_1(\theta = \theta_1)$, $\theta_1 > \theta_0$, n je veliko.

$$h(c) = P_{\theta_0}\left\{\frac{\theta_1^n e^{-\theta_1 \sum\limits_{k=1}^{n} x_k}}{\theta_0^n e^{-\theta_0 \sum\limits_{k=1}^{n} x_k}} \geqslant c\right\} = P_{\theta_0}\left\{\overline{X}_n \leqslant \frac{\ln c(\frac{\theta_0}{\theta_1})^n}{n(\theta_0-\theta_1)}\right\}$$
$$= P_{\theta_0}\left\{\left(\overline{X}_n - \frac{1}{\theta_0}\right)\theta_0\sqrt{n} \leqslant \left(\frac{\ln c(\frac{\theta_0}{\theta_1})^n}{n(\theta_0-\theta_1)} - \frac{1}{\theta_0}\right)\theta_0\sqrt{n}\right\} = \alpha.$$

Očigledno, funkcija $h(c)$ je neprekidna (kompozicija dvije neprekidne funkcije) te postoji najbolja kritična oblast veličine α. Uzimajući u obzir gore pomenutu aproksimaciju raspodjele statistike $\left(\overline{X}_n - \frac{1}{\theta_0}\right)\theta_0\sqrt{n} : \mathcal{N}(0,1)$ u modelu koji generiše H_0 dobijamo

$$W_0 = \left\{(x_1, ..., x_n) : \left(\overline{x}_n - \frac{1}{\theta_0}\right)\theta_0\sqrt{n} \leqslant -w_\alpha\right\}.$$

Primijetimo, W_0 ne zavisi od θ_1. I na kraju, kad je $n = 100$, $\alpha = 0,05$, $\theta_0 = 1$, $\theta_1 = 1,5$ dobijamo $W_0 = \{(x_1, ..., x_{100}) : \overline{x}_{100} \leqslant 0,835\}$, $\beta = P_{1,5}\{\overline{X} > 0,835\} = P\{X^* > (0,835 - 1,5^{-1})15\} = P\{X^* > 2,520\} = 0,0059$.

Najbolja kritična oblast iz formulacije Nejman-Pirsonove leme se može zapisati i u ekvivalentnom obliku $W_0 = \left\{\mathbf{x} : \frac{L(\theta_0,\mathbf{x})}{L(\theta_1,\mathbf{x})} \leqslant c\right\}$. Oblik kritične oblasti W_0 je očekivan (razuman, saglasan sa intuicijom). Ovu konstataciju objasnimo na neformalnom nivou. Prisjetimo se da funkcija vjerodostojnosti u zavisnosti od parametra θ daje informaciju o izgledima da realizovani uzorak "upadne" u proizvoljnu malu oblasti iz $\mathbb{R}^n$. Stoga su u kritičnoj oblasti W_0 vrijednosti $L(\theta_0, x_1, ..., x_n)$ male, a $L(\theta_1, x_1, ..., x_n)$ velike. Odatle slijedi da je količnik $\frac{L(\theta_0,\mathbf{x})}{L(\theta_1,\mathbf{x})}$ mali, što se prevodi na nejednakost $\frac{L(\theta_0,\mathbf{x})}{L(\theta_1,\mathbf{x})} \leqslant c$. Izložena analiza je motiv za

Test količnika vjerodostojnosti

Neka je parametarski skup Θ, $\Theta_0 + \Theta_1 = \Theta$, $H_0(\theta \in \Theta_0)$, $H_1(\theta \in \Theta_1)$. Neka je θ^* vrijednost parametra θ za koju funkcija $L(\theta, \mathbf{x})$, $\theta \in \Theta_0$ dostiže maksimum. Neka je θ^{**} vrijednost parametra θ za koju funkcija $L(\theta, \mathbf{x})$, $\theta \in \Theta$ dostiže maksimum. U testu količnika vjerodostojnosi kritična oblast je

$$\mathcal{C} = \left\{\mathbf{x} : \frac{L(\theta^*, \mathbf{x})}{L(\theta^{**}, \mathbf{x})} \leqslant c\right\}, 0 < c < 1,$$

i

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta\{\mathbf{X} \in \mathcal{C}\} = \sup_{\theta \in \Theta_0} P_\theta\left\{\frac{L(\theta^*, \mathbf{X})}{L(\theta^{**}, \mathbf{X})} \leqslant c\right\}.$$

Slučaj c smo isključili zbog toga što bi implicirao $\mathcal{C} = R^n$, $\alpha = 1$.

Neka je C kritična oblast koja određuje neki test za testiranje $H_o(\theta \in \Theta_o)$ protiv $H_1(\theta \in \Theta \backslash \Theta_0)$.

Funkcija moći testa se zadaje sa

$$W(\theta) = P_\theta\{(X_1, \ldots, X_n) \in C\},\ \theta \in \Theta.$$

Na jeziku funkcije moći testa, Nejman-Pirsonova lema daje konstrukciju najmoćnijeg testa, tj. testa za koji je $W(\theta_1)$ najveći pri datom $W(\theta_0)$. Vratimo se za trenutak na primjer koji smo uradili neposredno nakon dokaza Nejman-Pirsonove leme. Primijetimo, W_0 koje smo dobili primjenom Nejman Pirsonove leme, ne zavisi od m_1 te isti test (tj. istu kritičnu oblast) dobijamo bez obzira na to koje $m_1 > m_0$ zadaje alternativnu hipotezu. Zbog ove činjenice, za naš test kažemo da je **ravnomjerno najmoćniji** za testiranje $H_0(m = m_0)$ protiv $H_1(m = m_1), m_1 > m_0$.

Primjer. $X : \mathcal{N}(m, \sigma_0^2)$, σ_0^2 je poznato, $H_0(m \geqslant m_0)$, $H_1(m < m_0)$.

Budući da je $L(m, x_1, ..., x_n) = (2\pi\sigma_0^2)^{-\frac{n}{2}} e^{-\frac{\sum_{k=1}^{n}(x_k-m)^2}{2\sigma_0^2}}$, maksimum funkcije vjerodostojnosti se dobija za ono m za koje kvadratna funkcija $g(m) = \sum_{k=1}^{n}(x_k - m)^2 = nm^2 - 2m\sum_{k=1}^{n} x_k + \sum_{k=1}^{n} x_k^2$ ima minimum. Jednostavno se dobija $m^* = \max\{\overline{x}_n, m_0\}$, $m^{**} = \overline{x}_n$.

$$\frac{L(m^*, \mathbf{x})}{L(m^{**}, \mathbf{x})} = \begin{cases} 1, & \overline{x}_n \geqslant m_0 \\ e^{-\frac{1}{2\sigma_0^2}[\sum_{k=1}^{n}(x_k-m_0)^2 - \sum_{k=1}^{n}(x_k-\overline{x}_n)^2]}, & \overline{x}_n < m_0 \end{cases}$$

Riješimo nejednačinu

$$e^{-\frac{1}{2\sigma_0^2}[\sum_{k=1}^{n}(x_k-m_0)^2 - \sum_{k=1}^{n}(x_k-\overline{x}_n)^2]} \leqslant c$$

Naša nejednačina je ekvivalentna sa

$$\sum_{k=1}^{n}(x_k - m_0)^2 - \sum_{k=1}^{n}(x_k - \overline{x}_n)^2 \geqslant c,\ c > 0 \Leftrightarrow \overline{x}_n^2 - 2m_0\overline{x}_n + m_0^2 \geqslant c,\ c > 0 \Leftrightarrow (m_0 - \overline{x}_n)^2 \geqslant c,\ c > 0$$
$$\Leftrightarrow m_0 - \overline{x}_n \geqslant c,\ c > 0 \Leftrightarrow \overline{x}_n \leqslant m_0 - c,\ c > 0,$$

gdje zbog ekonomičnosti zapisa koristimo samo jednan simbol za konstantu iako je riječ o različitim konstantama.

$$\alpha = \sup_{m \geqslant m_0} P_m\Big\{\overline{X}_n \leqslant m_0 - c\Big\} = \sup_{m \geqslant m_0} P_m\left\{\frac{\overline{X}_n - m_0}{\sigma_0}\sqrt{n} \leqslant \frac{m_0 - c - m}{\sigma_0}\sqrt{n}\right\} =$$
$$\sup_{m \geqslant m_0} P\left\{X^* \leqslant \frac{m_0 - c - m}{\sigma_0}\sqrt{n}\right\},$$

Očigledno, supremum se dostiže za $m = m_0$ i nakon rješavanje jednačine $-\frac{c\sqrt{n}}{\sigma_0} = -w_\alpha$ dobijamo

$c=\frac{\sigma_0 w_\alpha}{\sqrt{n}}$. Dakle, $\mathcal{C}=\Big\{(x_1,...,x_n):\overline{x}_n\leqslant m_0-\frac{\sigma_0 w_\alpha}{\sqrt{n}}\Big\}$.

Potražimo funkciju moći testa.

$$W(m)=P_m\Big\{\overline{X}_n<m_0-\frac{w_\alpha\sigma_0}{\sqrt{n}}\Big\}=P\Big\{X^*<\frac{m_0-m-\frac{w_\alpha\sigma_0}{\sqrt{n}}}{\sigma_0}\sqrt{n}\Big\}=\Phi\Big(\frac{m_0-m}{\sigma_0}\sqrt{n}-w_\alpha\Big),$$

gdje sa Φ označavamo funkciju raspodjele slučajne promjenljive X^*. Funkcija $W(m)$ je neprekidna, monotono opada, teži ka 1 kad $m\to-\infty$ i teži ka 0 kad $m\to\infty$.

U slučaju $H_0(m\leqslant m_0)$, $H_1(m>m_0)$ se dobija $\mathcal{C}=\Big\{(x_1,...,x_n):\overline{x}_n\geqslant m_0+\frac{\sigma_0 w_\alpha}{\sqrt{n}}\Big\}$.

U slučaju $H_0(m=m_0)$, $H_1(m\neq m_0)$ imamo $m^*=m_0$, $m^{**}=\overline{x}_n$ te je

$$\mathcal{C}=\Big\{(x_1,...,x_n):\frac{|\,\overline{x}_n-m_0\,|}{\sigma_0}\sqrt{n}\geqslant w_{\frac{\alpha}{2}}\Big\}.$$

Primjer. $X:\mathcal{U}[0,\theta]$, $\theta>0$, $H_0(\theta=\theta_0)$, $H_1(\theta\neq\theta_0)$.

$\theta^*=\theta_0$, $L(\theta^*,x_1,...,x_n)=\frac{1}{\theta_0^n}$, $y_n\leqslant\theta_0$, $\theta^{**}=y_n$, $L(\theta^{**},x_1,...,x_n)=\frac{1}{y_n^n}$ te je

$$\frac{L(\theta^*,\mathbf{x})}{L(\theta^{**},\mathbf{x})}=\begin{cases}\left(\frac{y_n}{\theta_0}\right)^n, & y_n<\theta_0,\\ 1, & y_n=\theta_0,\\ 0, & y_n>\theta_0.\end{cases}\quad C=\Big\{(x_1,...,x_n):\left(\tfrac{y_n}{\theta_0}\right)^n\leqslant c \text{ ili } y_n>\theta_0\Big\},$$

$$\alpha=P_{\theta_0}\left\{\left(\frac{Y_n}{\theta_0}\right)^n\leqslant c\text{ ili }Y_n>\theta_0\right\}=P_{\theta_0}\{Y_n\leqslant\sqrt[n]{c}\theta_0\}=c;\ C=\{\mathbf{x}:\ y_n\leqslant\sqrt[n]{\alpha}\theta_0\text{ ili }y_n>\theta_0\}.$$

1^0 $0<\theta\leqslant\sqrt[n]{\alpha}\theta_0$, $W(\theta)=1$.

2^0 $\sqrt[n]{\alpha}\theta_0<\theta\leqslant\theta_0$, $W(\theta)=\alpha\Big(\frac{\theta_0}{\theta}\Big)^n$.

3^0 $\theta>\theta_0$, $W(\theta)=\alpha\Big(\frac{\theta_0}{\theta}\Big)^n+1-\Big(\frac{\theta_0}{\theta}\Big)^n=1-(1-\alpha)\Big(\frac{\theta_0}{\theta}\Big)^n$.

Primjer. $X:\mathcal{U}[0,\theta]$, $\theta>0$, $H_0(\theta\in[1,2])$.

Potražimo θ^*.

a) $y_n<1$, $L(\theta,\mathbf{x})=\frac{1}{\theta^n}$, $1\leqslant\theta\leqslant2$. Maksimum se dostiže za $\theta=1$.

b) $1\leqslant y_n\leqslant2$, $L(\theta,\mathbf{x})=\frac{1}{\theta^n}$, $y_n\leqslant\theta\leqslant2$ te se maksimum dostiže za $\theta=y_n$.

Dakle, $\theta^*=\max\{1,y_n\}$, a $\theta^{**}=y_n$.

$$\frac{L(\theta^*,\mathbf{x})}{L(\theta^{**},\mathbf{x})}=\begin{cases} y_n^n, & y_n<1,\\ 1, & 1\leqslant y_n\leqslant 2,\\ 0, & y_n>2.\end{cases}$$

$C=\{\mathbf{x}: y_n\leqslant \sqrt[n]{c} \text{ ili } y_n>2\}$, $\alpha=\sup\limits_{\theta\in[1,2]} P\{Y_n\leqslant \sqrt[n]{c} \text{ ili } Y_n>2\}=c\Rightarrow$
$C=\{\mathbf{x}: y_n\leqslant \sqrt[n]{\alpha} \text{ ili } y_n>2\}$.

$$W(\theta)=\begin{cases} 1, & 0<\theta\leqslant\sqrt[n]{\alpha},\\ \frac{\alpha}{\theta^n}, & \sqrt[n]{\alpha}<\theta\leqslant 2,\\ 1+\frac{\alpha}{\theta^n}-\frac{2^n}{\theta^n}, & \theta>2.\end{cases}$$

Primjer. $X:\mathcal{N}(m,\sigma^2)$ oba parametra su nepoznata, $H_0(m\leqslant m_0)$, $H_1(m>m_0)$. Metodom maksimalne vjerodostojnosti ćemo ocijeniti nepoznati vektor (m,σ^2). Zbog toga što je funkcija ln strogo monotono rastuća, traženje maksimuma funkcije $L(m,\sigma^2,\mathbf{x})$ ćemo zbog jednostavnije analize konvertovati u traženje maksimuma funkcije $\ln L(m,\sigma^2,\mathbf{x})$.

$\frac{\partial}{\partial m}\ln L=0\Rightarrow m=\overline{x}_n$. Zamijenimo dobijeno m u L i rješavamo $\frac{\partial}{\partial\sigma^2}\ln L=0\Rightarrow\sigma^2=\frac{1}{n}\sum\limits_{k=1}^{n}(x_k-\overline{x}_n)^2=\overline{s}_n^2$. Dakle $\theta^{**}=(\overline{x}_n,\overline{s}_n^2)$. Potražimo θ^*. $\frac{\partial}{\partial\sigma^2}\ln L=0\Rightarrow\sigma^2=\frac{1}{n}\sum\limits_{k=1}^{n}(x_k-m)^2$. Zamijenimo upravo dobijeno σ^2 u $L(m,\sigma^2,\mathbf{x})$, a zatim tražimo maksimum funkcije $\ln L(m,\frac{1}{n}\sum\limits_{k=1}^{n}(x_k-m)^2,\mathbf{x})$, $m\leqslant m_0$. Maksimum se dostiže u tački $\min\{\overline{x}_n,m_0\}$. Dakle,

$$\theta^*=\begin{cases} (\overline{x}_n,\overline{s}_n^2), & x_n\leqslant m_0,\\ (m_0,\frac{1}{n}\sum\limits_{k=1}^{n}(x_k-m_0)^2), & x_n>m_0.\end{cases}$$

$$\frac{L(\theta^*,\mathbf{x})}{L(\theta^{**},\mathbf{x})}=\begin{cases} 1, & \overline{x}_n\leqslant m_0,\\ \frac{\sum\limits_{k=1}^{n}(x_k-\overline{x}_n)^2}{\sum\limits_{k=1}^{n}(x_k-m_0)^2}, & \overline{x}_n>m_0.\end{cases}$$

$\frac{\sum\limits_{k=1}^{n}(x_k-\overline{x}_n)^2}{\sum\limits_{k=1}^{n}(x_k-m_0)^2}<c\Leftrightarrow\frac{\sum\limits_{k=1}^{n}(x_k-\overline{x}_n)^2}{\sum\limits_{k=1}^{n}(x_k-\overline{x}_n+\overline{x}_n-m_0)^2}<c\Leftrightarrow\frac{n\overline{s}_n^2}{n\overline{s}_n^2+n(\overline{x}_n-m_0)^2}<c\Leftrightarrow\frac{\overline{x}_n-m_0}{\hat{s}_n}\sqrt{n}\geqslant c_\alpha, \overline{s}_n^2=\frac{n-1}{n}\hat{s}_n^2.$

Znamo da statistika $\frac{\overline{X}_n-m_0}{\hat{s}_n}\sqrt{n}$ ima Studentovu raspodjelu sa $n-1$ stepena slobode. Slučajnu promjenljivu koja ima t_{n-1} raspodjelu ćemo oznčiti sa T.

$$\alpha=\sup_{m\leqslant m_0}P_m\left\{\frac{\overline{X}_n-m_0}{\hat{S}_n}\sqrt{n}\geqslant c_\alpha\right\}=\sup_{m\leqslant m_0}P_m\left\{\frac{\overline{X}_n-m}{\hat{S}_n}\sqrt{n}+\frac{m-m_0}{\hat{S}_n}\sqrt{n}\geqslant c_\alpha\right\}=$$

$$=P\{T\geqslant v_{n-1;0{,}5-\alpha}\}.$$

Naime, supremum se dostiže za $m = m_0$, a $v_{n-1;0,5-\alpha}$ se čita iz tablica za Studentovu raspodjelu. I na kraju

$$C = \left\{ \mathbf{x} : \overline{x}_n > m_0 + \frac{v_{n-1;0,5-\alpha}\hat{s}_n}{\sqrt{n}} \right\}.$$