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

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

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

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

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 daje Nejman-Pirsonova lema. Mi ćemo Nejman-Pirsonovu lemu formulisati u slučaju kada obilježje ima apsolutno neprekidnu raspodjelu. Gustinu obilježja u slučaju $\theta = \theta_0$ ćemo označavati sa g_0, a u slučaju $\theta = \theta_1$ ćemo označavati sa g_1.

Definišimo jednu statistiku i jednu funkciju.

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

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 $W_0 = \{\mathbf{x} : K(\mathbf{x}) \geqslant c\}$.

♦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 $P_{\theta_1}\{\mathbf{X} \in W\} \leqslant P_{\theta_1}\{\mathbf{X} \in W_0\}$.

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

Primjer 1.1 *Obilježje X ima $\mathcal{N}(m, \sigma_0^2), m \in \{m_0, m_1\}, m_1 > m_0$, σ_0^2 poznato. Naći najbolju*

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

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

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

Neka je $w(c) = \frac{\sigma_0 \ln c}{(m_1 - m_0)\sqrt{n}} + \frac{(m_1 - m_0)\sqrt{n}}{2\sigma_0}$, $c > 0$. 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 $h(c) = P\{X^* \geqslant w(c)\} = 1 - \Phi(w(c))$, gdje je Φ uobičajena oznaka za funkciju raspodjele slučajne promjenljive $X^* : \mathcal{N}(0,1)$. Funkcija $w(c)$ je strogo monotono rastuća i njen kodomen je $(-\infty, \infty)$. Funkcija $h(c)$ je strogo monotono opadajuća i njen kodomen je $(0,1)$ te postoji jedinstveno $c > 0$ takvo da je $h(c) = \alpha$ i to c je rješenje jednačine $w(c) = z_{1-2\alpha}$. Primijetimo, nema potreba za računanjem broja c. Dakle,

$$W_0 = \left\{ (x_1, ..., x_n) : \frac{\overline{x}_n - m_0}{\sigma_0}\sqrt{n} \geqslant z_{1-2\alpha} \right\} = \left\{ (x_1, ..., x_n) : \overline{x}_n \geqslant m_0 + \frac{\sigma_0 z_{1-2\alpha}}{\sqrt{n}} \right\}.$$

Primijetimo, kritična oblast ne zavisi od konkretne vrijednosti za m_1.

U slučaju testiranja $H_0(m = m_0)$ protiv $H_1(m = m_1)$, $m_1 < m_0$, primjenom istog postupka se dobija:

$$W_0 = \left\{ (x_1, ..., x_n) : \frac{\overline{x}_n - m_0}{\sigma_0}\sqrt{n} \leqslant -z_{1-2\alpha} \right\} = \left\{ (x_1, ..., x_n) : \overline{x}_n \leqslant m_0 - \frac{\sigma_0 z_{1-2\alpha}}{\sqrt{n}} \right\}.$$

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

$$P_{m_1}\left\{ \overline{X}_n < m_0 + \frac{\sigma_0 z_{1-2\alpha}}{\sqrt{n}} \right\} = P\left\{ X^* < \frac{m_0 - m_1}{\sigma_0}\sqrt{n} + z_{1-2\alpha} \right\} \leqslant \beta$$

Iz

$$\frac{m_0-m_1}{\sigma_0}\sqrt{n}+z_{1-2\alpha}<-z_{1-2\beta}\Rightarrow n\geqslant\left(\frac{\sigma_0(z_{1-2\alpha}+z_{1-2\beta})}{m_1-m_0}\right)^2.\blacktriangleleft$$

Primjer 1.2 $X:\mathcal{E}(\theta^{-1}),\theta\in\{\theta_0,\theta_1\},0<\theta_1<\theta_0$. *Testirati* $H_0(\theta=\theta_0)$ *protiv* $H_1(\theta=\theta_1)$, n *je veliko.*

►Znamo, ako $X:\mathcal{E}(\theta^{-1})$ tada statistika $\left(\overline{X}_n-\theta\right)\frac{\sqrt{n}}{\theta}$ ima asimptotski $\mathcal{N}(0,1)$.

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

Neka je $w(c)=\left(\frac{\theta_0\theta_1}{\theta_0-\theta_1}\ln\frac{\theta_0}{\sqrt[n]{c}\theta_1}-\theta_0\right)\frac{\sqrt{n}}{\theta_0}$. Uzimajući u obzir gore pomenutu aproksimaciju raspodjele statistike $\left(\overline{X}_n-\theta_0\right)\frac{\sqrt{n}}{\theta_0}:\mathcal{N}(0,1)$ u modelu koji generiše H_0, zaključujemo da je $h(c)=P\{X^*\leqslant w(c)\}$. Funkcija $w(c)$ je strogo monotono opadajuća i njen kodomen je $(-\infty,\infty)$. Funkcija $h(c)$ je strogo monotono opadajuća i njen kodomen je $(0,1)$ te postoji jedinstveno $c>0$ takvo da je $h(c)=\alpha$ i to c je rješenje jednačine $w(c)=-z_{1-2\alpha}$.

Dakle,

$$W_0=\left\{(x_1,...,x_n):\left(\overline{x}_n-\theta_0\right)\frac{\sqrt{n}}{\theta_0}\leqslant-z_{1-2\alpha}\right\}.$$

Primijetimo, W_0 ne zavisi od θ_1. I na kraju, kad je $n=100$, $\alpha=0,05$, $\theta_0=1$, $\theta_1=\frac{2}{3}$ dobijamo $W_0=\{(x_1,...,x_{100}):\overline{x}_{100}\leqslant 0,835\}$, $\beta=P_{\frac{2}{3}}\{\overline{X}_{100}>0,835\}=P\{X^*>(0,835-\frac{2}{3})15\}=P\{X^*>2,520\}=0,0059.\blacktriangleleft$

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

email adresa nastavnika: sstamatovic@ucg.ac.me