A november 28-i gyakorlat témája

Az előadáson szerepelt az az állítás, hogy a standard normális eloszlás sűrűségfüggvénynek nevezett $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ függvény valóban sűrűségfüggvény, tehát

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1.$$

Lássuk be (e tény felhasználásával),

1.) Egy standard normális eloszlású ξ valószínűségi változó, azaz egy $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ sűrűségfüggvénnyel rendelkező valószínűségi változó várható értéke nulla és szórásnégyzete 1.

Megoldás:

$$E\xi = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 0,$$

mert az $x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ integrandus páratlan függvény. Ezért $\operatorname{Var}\xi = E\xi^2$, és parciális integrálással $f(x) = x$ és $g(x) = x\frac{1}{\sqrt{2\pi}}e^{-x^2/2} = \frac{d}{dx}\left(-\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right)$ választással kapjuk, hogy

$$\begin{aligned} E\xi^2 &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = \int_{-\infty}^{\infty} -x\frac{d}{dx}\left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right) dx \\ &= \left[-x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1. \end{aligned}$$

2.) Egy szabályos dobókockát és egy szabályos érmét feldobunk 3300 alkalommal egymástól függetlenül. (Az érme és kockadobások eredményei is függetlenek egymástól.) Ha a kockadobás eredménye páros és az érme a fej oldalra esett, akkor annyi forintot nyerünk, amennyi a kockadobás eredménye. Ha az érme az írás oldalra esett vagy a kockadobás eredménye páratlan szám, akkor nem nyerünk, és nem is vesztünk semmit. Mi a valószínűsége annak, hogy az össznyereményünk 3190 és 3520 forint közé esik? Adjunk erre jó közelítő becslést a centrális határeloszlástétel és egy normális eloszlástáblázat segítségével.

Megoldás: Vezessük be a következő ξ_j, és η_j $1 \le j \le 3300$, valószínűségi változókat: $\xi_j = 2$, ha a j-ik kockadobás eredménye 2, $\xi_j = 4$, ha a j-ik kockadobás eredménye 4, $\xi_j = 6$, ha a j-ik kockadobás eredménye 6, $\xi_j = 0$, ha a j-ik kockadobás eredménye 1, 3 vagy 5. Legyen $\eta_j = 1$, ha a j-ik érmedobás eredménye fej, és $\eta_j = 0$, ha a j-ik érmedobás írás. Legyen $\zeta_j = \xi_j\eta_j$. Ekkor a j-ik dobásnál a nyereményünk ζ_j lesz, $1 \le j \le 3300$, és a $P\left(3190 \le \sum_{j=1}^{3300} \zeta_j \le 3520\right)$ valószínűséget kell jól megbecsülnünk. Ennek érdekében számoljuk ki a független ζ_j valószínűségi változók várható értékét és szórásnégyzetét. $E\zeta_j = E\xi_j\eta_j = E\xi_j E\eta_j = \frac{1}{6}(2+4+6)\frac{1}{2} = 1$,

$E\zeta_j^2 = E\xi_j^2 E\eta_j^2 = \frac{1}{6}(4+16+36)\cdot\frac{1}{2} = \frac{14}{3}$, $\operatorname{Var}\zeta_j^2 = E\zeta_j^2 - (E\zeta_j)^2 = \frac{11}{3}$. Innen a centrális határeloszlástétel alapján

$$P\left(3190 \le \sum_{j=1}^{3300}\zeta_j \le 3520\right) = P\left(\frac{-110}{\sqrt{3300\frac{11}{3}}} \le \frac{\sum_{j=1}^{3300}\zeta_j - \sum_{j=1}^{3300}E\zeta_j}{\sqrt{\sum_{j=1}^{3300}\operatorname{Var}\xi_j}} \le \frac{220}{\sqrt{3300\frac{11}{3}}}\right)$$
$$\sim \Phi(2) - \Phi(-1) = 0.9772 + 0.8413 - 1 = 0.9285.$$

3.) Egy pénzdarabról ellenőrizni akarjuk, hogy igaz-e az a hipotézis, amely szerint ez az érme legalább $\frac{3}{4}$ valószínűséggel esik a fej és legfeljebb $\frac{1}{4}$ valószínűséggel az írás oldalára. Ennek érdekében feldobjuk a pénzdarabot 30 000 alkalommal, és a következő döntési szabályt hozzuk. Választunk egy k számot, és akkor fogadjuk el a hipotézist helyesnek, ha legalább k fejdobás történt. Legalább mekkorának kell válassztanunk ezt a k számot, ha azt akarjuk, hogy egy a hipotézist teljesítő pénzdarab esetén legalább 0.9 valószínűséggel döntsünk úgy, hogy a hipotézis teljesül?

Megoldás: Vezessük be a következő valószínűségi változókat: $\xi_j = 1$, ha a j-ik dobás eredménye fej, $\xi_j = 0$, ha a j-ik dobás eredménye írás, $1 \le j \le 30\,000$, $S = S_{30000} = \sum_{j=1}^{30\,000}\xi_j$. Ha a fejdobás eredményének valószínűsége pontosan $\frac{3}{4}$, $E\xi_j = \frac{3}{4}$, $E\xi_j^2 = \frac{3}{4}$, $\operatorname{Var}\xi_j = E\xi_j^2 - (E\xi_j)^2 = \frac{3}{16}$, $ES = 30\,000E\xi_j = 22\,500$, $\operatorname{Var}S = 30\,000\operatorname{Var}\xi_j = 5625 = 75^2$. Innen és a centrális határeloszlástételből,

$$P(S > k) = P\left(\frac{S-ES}{\sqrt{\operatorname{Var}S}} > \frac{k-22\,500}{75}\right) = 1 - P\left(\frac{S-ES}{\sqrt{\operatorname{Var}S}} \le \frac{k-22\,500}{75}\right)$$
$$\sim 1 - \Phi\left(\frac{k-22\,500}{75}\right).$$

Válasszuk a k számot úgy, hogy a fenti valószínűség körülbelül 0.9 legyen. Ekkor a $\Phi\left(\frac{k-22\,500}{75}\right) = 0.1$ vagy ami ezzel ekvivalens, a $\Phi\left(\frac{22\,500-k}{75}\right) = 0.9$ egyenletet kell kielégítenünk. A normális eloszlás-táblázat alapján $\frac{22\,500-k}{75} \sim 1.28$, ami azt jelenti, hogy $k = 22\,500 - 75\times 1.28$ és $p = \frac{3}{4}$ esetén annak valószínűsége, hogy a fejdobások száma nagyobb mint $k = 22\,500 - 75\times 1.28 = 22\,212$ és $p = \frac{3}{4}$ esetében annak valószínűsége, hogy legalább ennyi fejdobás történik körülbelül 0.9. Ha $p \ge \frac{3}{4}$, akkor ez a valószínűség nagyobb. Ezért a $k = 22\,212$ helyes választás.

4.) Ledobunk egymástól függetlenül 24 000 pontot a $[0,2]$ intervallumra egyenletes eloszlással, (azaz annak a valószínűsége, hogy egy ledobott pont értéke x-nél kisebb $\frac{x}{2}$-vel egyenlő, ha $0 \le x \le 2$, eggyel egyenlő, ha $x \ge 2$, és nulla, ha $x \le 0$.) Őrizzük meg azokat a ledobott pontokat, melyek értéke 1-nél kisebb, és hagyjuk el azokat, melyek értéke, nagyobb mint egy. Mi annak a valószínűsége, hogy a megőrzött pontok értékeinek az összege 5900 és 6075 közé esik? Adjunk erre a valószínűségre jó közelítő becslést a mellékelt normális eloszlástáblázat segítségével.

Megoldás: Vezessük be a következő ξ_j, $1 \le j \le 24\ 000$, valószínűségi változókat: $\xi_j = x$, ha a j-ik ledobott pont értéke x, és $0 \le x \le 1$, és $\xi_j = 0$, ha a j-ik ledobott pont értéke az $(1, 2]$ intervallumba esik. Ekkor a megőrzött pontok összege $S = \sum_{j=1}^{24\ 000} \xi_j$, továbbá a ξ_j valószínűségi változók függetlenek és egyforma eloszlásúak. Ezért a centrális határeloszástétel segítségével jó becslést tudunk adni a minket érdeklő $P(5900 < S < 6075)$ valószínűségre. Ennek érdekében ki kell számolnunk a ξ_1 valószínűségi változó várható értékét és szórásnégyzetét.

A ξ_1 valószínűségi változó várható értékének és szórásnégyzetének a kiszámolása érdekében vezessük be az η_1 valószínűségi változót, amelyik megegyezik az első ledobott pont értékével, és a következő $h(x)$ függvényt a $[0, 2]$ intervallumon: Legyen $h(x) = x$, ha $0 \le x \le 1$, és $h(x) = 0$, ha $1 \le x \le 2$. Ekkor $\xi_1 = h(\eta_1)$, és η_1 sűrűségfüggvénye $f(x) = \frac{1}{2}$, ha $0 \le x \le 2$, $f(x) = 0$, ha $x < 0$, és $x > 2$. Innen $E\xi_1 = Eh(\eta_1) = \int h(x)\,dx = \int_0^1 x\frac{1}{2}\,dx = \frac{1}{4}$, $E\xi_1^2 = Eh(\eta_1)^2 = \int_0^1 x^2\frac{1}{2}\,dx = \frac{1}{6}$, és $\operatorname{Var}\xi_1 = E\xi_1^2 - (E\xi_1)^2 = \frac{1}{6} - \frac{1}{16} = \frac{5}{48}$. Ezért $ES = 6000$, $\operatorname{Var} S = 2500$. Innen

$$P(5900 < S < 6075) = P\left(-2 < \frac{S - ES}{\sqrt{\operatorname{Var} S}} < 1.5\right) \sim \Phi(1.5) - \Phi(-2) = \Phi(1.5) + \Phi(2) - 1.$$

A matematikai statisztika fontos feladata a paraméter becslés, a hipotézisvizsgálat és a konfidencia intervallum becslés. Az ilyen jellegű, a gyakorlatban előforduló feladatok szoros kapcsolatban vannak a centrális határeloszlástétellel. Ennek tárgyalása érdekében először értsük meg e fogalmakat.

Statisztikai becslés fogalma. *Adott eloszlásfüggvények egy valamely (ismeretlen λ paramétertől függő) $F(x, \lambda)$ családja. (Van amikor nem az $F(x, \lambda)$ eloszlás függvényt, hanem annak $f(x, \lambda)$ sűrűségfüggvényét adják meg. Az egyszerűség kedvéért egy λ paraméterről beszéltem. De vannak olyan fontos problémák, amikor több ismeretlen $\lambda_1, \dots, \lambda_k$ paramétertől függ az eloszlás. Ebben az esetben is beszélhetünk egy $\lambda = (\lambda_1, \dots, \lambda_k)$ vektor értékű paraméterről.) Legyen adva n darab $\xi_1, \dots, \xi_n$ $F(x, \lambda)$ eloszlású független valószínűségi változó, amelyeket általában n kísérlet eredményeként kapunk meg. A becslés feladata azt jelenti, hogy adjunk ezen valószínűségi változók (megfigyelések) segítségével egy jó közelítő értéket az ismeretlen λ paraméterre, azaz definiáljunk olyan $\hat{\lambda} = T(\xi_1, \dots, \xi_n)$ a $\xi_1, \dots, \xi_n$ valószínűségi változóktól függő valószínűségi változót, amely (nagy valószínűséggel) közel van a becsülendő λ paraméterhez, bármi is volt a λ értéke.*

Szokás bevezetni a következő terminológiát.

Minta fogalma. *Legyen adva eloszlások egy $F(x, \lambda)$, $\lambda \in \Lambda$, eloszláscsalád. Független $F(x, \lambda)$ (az ismeretlen becsülendő λ eloszlású valószínűségi paramétertől függő) $\xi_1, \dots, \xi_n$ valószínűségi változók sorozatát n elemű mintának nevezik.*

Valamilyen értelemben a becslésnek jónak kell lennie. Ilyen természetes a becslés jóságát megkövetelő tulajdonságot fejez ki a következő definició.

(Aszimptotikusan) torzítatlan becslés fogalma. *Legyen adva egy $F(x,\lambda)$ eloszlás függvény, és $\xi_1,\dots,\xi_n$ $F(x,\lambda)$ eloszlású n elemű minta. Egy $\hat\lambda = T(\xi_1,\dots,\xi_n)$ becslést torzítatlannak nevezünk, ha $\lambda = E\hat\lambda_n = E(T(\xi_1,\dots,\xi_n))$. Adjunk meg egy olyan az n paramétertől függő $T_n(x_1,\dots,x_n)$ függvényt minden $n=1,2,\dots$ számra, amelyre*

$$\lambda = \lim_{n\to\infty} E\hat\lambda_n = \lim_{n\to\infty} E(T_n(\xi_1,\dots,\xi_n))$$

egy n elemű $\xi_1,\dots,\xi_n$ mintára minden λ paraméterre. Az ilyen tulajdonságú $\hat\lambda_n = T_n(\xi_1,\dots,\xi_n)$ becsléseket aszimptotikusan torzítatlan becsléseknek nevezzük.

A becsléselmélet feladata a λ paraméter olyan torzítatlan (vagy ha az nem lehetséges aszimptotikusan torzítatlan) $T_n(\xi_1,\dots,\xi_n$ becslését adni, amelyre

$$\lim_{n\to\infty} \operatorname{Var} T_n(\xi_1,\dots,\xi_n) = 0.$$

5.) Legyen adva egy $F(x,\lambda)$ eloszláscsalád, és egy n-elemű $\xi_1,\dots,\xi_n$ mintától függő $T_n(\xi_1,\dots,\xi_n)$ becsléssorozat, $n=1,2,\dots$. Mutassuk meg, hogy

$$\lim_{n\to\infty} E(T_n(\xi_1,\dots,\xi_n)-\lambda)^2 = 0$$

akkor és csak akkor, ha a $T_n(\xi_1,\dots,\xi_n)$ becslés aszimptotikusan torzítatlan, és

$$\lim_{n\to\infty} \operatorname{Var} T_n(\xi_1,\dots,\xi_n) = 0.$$

Megoldás:

$$\begin{aligned} &E(T_n(\xi_1,\dots,\xi_n)-\lambda)^2 \\ &\quad = E\left([T_n(\xi_1,\dots,\xi_n) - E(T_n(\xi_1,\dots,\xi_n)]^2\right) + [\lambda - E(T_n(\xi_1,\dots,\xi_n)]^2 \\ &\quad = \operatorname{Var} T_n(\xi_1,\dots,\xi_n) + [\lambda - E(T_n(\xi_1,\dots,\xi_n)]^2. \end{aligned}$$

Ebből az azonosságból könnyen kiolvasható a feladat állítása.

A leggyakoribb becsléselméleti feladat egy valószínűségi változó várható értékének vagy szórásnégyzetének a becslése. Formálisan ez a feladat csak az előbb megfogalmazott becslési feladat egy általánosabb megfogalmazásaként tárgyalható, mert az összes lehetséges eloszlást tekintjük, és az összes eloszlásfüggvényt tartalmazó függvénycsalád nem paraméterezhető véges sok paraméter segítségével. Továbbá nem kívánjuk magát az eloszlást meghatározni, csak annak egy fontos paraméterét. Az ilyen feladatokat nem paraméteres becsléseknek nevezik az irodalomban. De ilyenkor is beszélhetünk a

paraméter torzítatlan vagy aszimptotikusan torzítatlan becsléséről, amit az előzőekhez hasonlóan definiálunk.

6.) Legyen adva független $F(x)$ eloszlású valószínűségi változók $\xi_1, \ldots, xi_n$ sorozata Mutassuk meg, hogy az $\bar{\xi} = \frac{1}{n}\sum_{k=1}^{n} \xi_k$ átlag a várható érték torzítatlan becslése. Ha ismerjük a ξ valószínűségi változók μ várható értékét, akkor az $\bar{S}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(\xi_j - \mu)^2$ kifejezés felírható, és ez a szórásnégyzet torzítatlan becslése.

Megoldás: A várható érték tulajdonsága alapján

$$E\bar{\xi} = E\left(\frac{1}{n}\sum_{k=1}^{n}\xi_k\right) = \frac{1}{n}\sum_{k=1}^{n} E\xi_k = E\xi_1,$$

és $E\bar{S}_n^2 = E\left(\frac{1}{n}\sum_{j=1}^{n}(\xi_j - \mu)^2\right) = \frac{1}{n}\sum_{j=1}^{n} E(\xi_j - \mu)^2 = E(\xi_1 - \mu)^2 = \operatorname{Var}\xi_1$. Ezt kellett belátnunk.

7.) Legyen adva független $F(x)$ eloszlású valószínűségi változók $\xi_1, \ldots, \xi_n$ sorozata Mutassuk meg, hogy az $S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2$ kifejezés, ahol $\bar{\xi} = \frac{1}{n}\sum_{k=1}^{n}\xi_k$, a szórásnégyzet torzítatlan becslése.

Megoldás: Írjuk át az S_n^2 kifejezést számunkra alkalmasabb alakban.

$$\begin{aligned} S_n^2 &= \frac{1}{n-1}\sum_{j=1}^{n}(\xi_j - \bar{\xi})^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\xi_j^2 + \bar{\xi}^2 - 2\bar{\xi}\xi_j) \\ &= \frac{1}{n-1}\left(\sum_{j=1}^{n}\xi_j^2 - n\bar{\xi}^2\right) = \frac{1}{n-1}\sum_{j=1}^{n}\xi_j^2 - \frac{1}{(n-1)n}\left(\sum_{j=1}^{n}\xi_j\right)^2. \end{aligned}$$

Továbbá, $E\sum_{j=1}^{n}\xi_j^2 = nE\xi_1^2$, $E\left(\sum_{j=1}^{n}\xi_j\right)^2 = \sum_{j=1}^{n}E\xi_j^2 + 2\sum_{1\le j<k\le n}E\xi_j\xi_k = nE\xi_1^2 + n(n-1)E\xi_1\xi_2 = nE\xi_1^2 + n(n-1)(E\xi_1)^2$. Innen $ES_n^2 = \frac{1}{n-1}nE\xi_1^2 - \frac{1}{n-1}E\xi_1^2 - (E\xi_1)^2 = E\xi_1^2 - (E\xi_1)^2 = \operatorname{Var}\xi_1$.

A hipotézisvizsgálat azzal a kérdéssel foglalkozik, hogy azt a hipotézist (feltételezést), mely szerint egy véletlen mennyiség valamilyen előírt eloszlású vagy vagy valamely előírt eloszlások családjába tartozik hogyan ellenőzízzük bizonyos megfigyelések alapján. E kérdés tárgyalásakor érdemes bevezetni a következő fogalmakat.

Nullhipotézis és ellenhipotézis, egyszerű és összetett hipotézis fogalma. *A hipotézisvizsgálat feladata a következő: Adott eloszlásfüggvények egy Λ_0 és Λ_1 családja. Egy véletlen mennyiség ismeretlen $F(x)$ eloszlása vagy a Λ_0 vagy a Λ_1 eloszlásfügggvény*

családba tartozik, (más lehetőség nincsen.) Azt akarjuk eldönteni független $F(x)$ eloszlású valószínűségi változók megfigyelése alaján, melyik eset áll fenn. Az a feltevésünk, hogy $F \in \Lambda_0$. Ezt a feltevést null-hipotézisnek, az $F \in \Lambda_1$ feltevést ellenhipotézisnak nevezik. Ha Λ_0 egyetlen eloszlásból áll, akkor a null-hipotézist egyszerűnek, ha több elemből áll, akkor összetettnek nevezik. Hasonló terminológiát lehet használni a Λ_1 ellenhipotézisre is.

A hipotézisvizsgálat alapfeladata, első és másodfajú hiba fogalma. *Legyen adva a null-hipotézisek Λ_0 és ellenhipotézisek Λ_1 családja, valamint ismert független $F(x)$ eloszlású valószínűségi változók $\xi_1, \dots, \xi_n$ sorozata. Definiáljuk az R^n Euklideszi tér valamely alkalmasan definiált $A \subset R^n$ részhalmazát, és hozzuk a következő döntést. Ha $(\xi_1, \dots, \xi_n) \in A$ akkor a null-hipotézist fogadjuk el, ha $(\xi_1, \dots, \xi_n) \notin A$ akkor az ellenhipotézist fogadjuk el. Elsőfajú hibának a*

$$\sup_{F \in \Lambda_0} P_F((\xi_1, \dots, \xi_n) \notin A)$$

másodfajú hibának a

$$\sup_{F \in \Lambda_1} P_F((\xi_1, \dots, \xi_n) \in A)$$

mennyiséget nevezzük, ahol $P_F((\xi_1, \dots, \xi_n) \in A)$ annak a valószínűségét jelöli, hogy $\xi_1, \dots, \xi_n$ független, F eloszlású valószínűségi változók sorozata az A halmazba esik. Az elsőfajú hiba a null-hipotézis elutasításának a valószínűségét jelenti abban a legkellemetlenebb esetben, amikor azt el kellene fogadni. A másodfajú hiba a null-hipotézis elfogadását jelenti a legkellemetlenebb esetben, amikor azt el kellene utasítani. A hipotézisvizsgálat feladata a következő. Rögzítsünk egy $\varepsilon > 0$ számot, és keressünk olyan döntési eljárást, amelyben az elsőfajú hiba kisebb vagy egyenlő, mint ε, és minimalizáljuk eme feltétel mellett a másodfajú hibát.

Röviden ismertetem a konfidencia intervallum (konfidencia=megbízhatóság) problémáját is, ami a becsléselmélet problémájának természetes folytatása.

Konfidencia intervallum konstrukciójának a feladata. *Legyen adva $F(x, \lambda)$ eloszlásfüggvények egy családja, független $F(x, \lambda_0)$ eloszlású $\xi_1, \dots, \xi_n$ valószínűségi változók egy sorozata. Legyen adva a λ_0 paraméter egy jó $\hat{\lambda}_0 = T(\xi_1, \dots, \xi_n)$ becslése. Rögzítsünk egy kis $\varepsilon > 0$ pozitív számot. A konfidencia intervallum konstrukciójá olyan a $\xi_1, \dots, \xi_n$ megfigyelésektől függő lehetőleg kicsi $\tilde{\lambda}_{0,1,n} = U_1(\xi_1, \dots, \xi_n)$ és $\tilde{\lambda}_{0,2,n} = U_2(\xi_1, \dots, \xi_n)$ végpontú véletlen $[\tilde{\lambda}_{0,1,n}, \tilde{\lambda}_{0,2,n}]$ intervallum konstrukcióját jelenti, amelyre*

$$P(\lambda_0 \in [\tilde{\lambda}_{0,1,n}, \tilde{\lambda}_{0,2,n}] \geq 1 - \varepsilon.$$

Azaz olyan intervallumot keresünk, amelybe az ismeretlen paraméter majdnem biztos, hogy beleesik, pontosabban ennek valószínűsége legalább $1 - \varepsilon$.