A december 5-i gyakorlat témája

Tekintsük a következő statisztikai feladatot. Adott egy dobókocka, és el akarjuk dönteni, hogy az szabályos-e. Ennek érdekében feldobjuk a kockát n alkalommal, és megjegyezzük, hány alkalommal volt a dobás eredménye j, $1 \leq j \leq 6$. Jelölje $\nu_n(j)$, $1 \leq j \leq 6$, ezeket a (véletlen) számokat. Ha a $\nu_n(j) \sim \dfrac{n}{6}$ reláció teljesül minden $1 \leq j \leq 6$ számra akkor a kockát szabályosnak tekinthetjük, ellenkező esetben pedig szabálytalannak. Meg kell pontosabban fogalmazni, mit jelentenek a fenti aszimptotikus relációk. Egy jó eljárást akarunk kidolgozni annak mérésére, hogy a $\nu_n(j) - \frac{n}{6}$, $1 \leq j \leq 6$, számhatossal jellemzett ingadozás mikor tekinthető túlságosan nagynak és mikor tekinthető viszonylag kicsinek. Az eljárás kidolgozásába beletartozik annak a kérdésnek a tisztázása is, hogy hogyan érdemes mérni a $\nu_n(j) - \frac{n}{6}$, $1 \leq j \leq 6$, számhatossal jellemzett ingadozás nagyságát.

E feladat egy egyszerűbb változatával már találkoztunk, amikor azt a feltevés próbáltuk ellenőrizni sok kisérlet eredményének a segítségével, hogy egy pénzdarab szabályos-e. Megállapítottuk, hogy ezt az ellenőrzést a centrális határeloszlástétel segítségével lehet jól végrehajtani. Jelen esetben a centrális határeloszlástétel egy több-dimenziós általánosítása segítségével lehet jó módszert kidolgozni. Érdemes a dobókocka szabályosságának ellenőrzése helyett e probléma egy természetes általánosítását vizsgálni, és annak kapcsán tárgyalni a több-dimenziós centrális határeloszlástételt, illetve annak alkalmazását.

Az általánosított probléma a következő: Legyen adva k urna, és ellenőrizni akarjuk azt a feltételezést, mely szerint ha egy golyót véletlenül bedobunk ezen urnák valamelyikébe, akkor az p_j, $p_j > 0$, valószínűséggel esik a j-ik urnába, $\sum\limits_{j=1}^{k} p_j = 1$. Ennek a feltételezésnek az ellenőrzése érdekében dobjunk egymástól függetlenül n golyót ezekbe az urnákba, és jelölje $\nu_n(j)$ a j-ik urnába eső golyók számát. A $\nu_n = (\nu_n(1), \ldots, \nu_n(k))$ véletlen vektor viselkedésének megértése érdekében — mely vektor megadja, hogy az egyes urnákba hány golyó esett — érdemes bevezetni az alábbi $X_j = (X_j(1), \ldots, X_j(k))$, $1 \leq j \leq n$, véletlen vektorokat. Mindegyik X_j vektornak $k-1$ koordinátája 0-val és 1 (véletlenül választott) koordinátája 1-gyel egyenlő. Ha a j-ik dobás eredményeként az l-ik urnába esett a j-ik golyó, akkor legyen az l-ik koordináta értéke 1, azaz $X_j(l) = 1$ (és a többi koordináta értéke 0). Ekkor $\nu_n = \sum\limits_{j=1}^{n} X_j$, és ha bevezetjük az

$$\bar{X}_j = (X_j(1) - EX_j(1), \ldots, X_j(k) - EX_j(k)) = (X_j(1) - p_1, \ldots, X_j(k) - p_k)$$

valamint

$$\bar{\nu}_n = (\nu_n(1) - E\nu_n(1), \ldots, \nu_n(k) - E\nu_n(k)) = (\nu_n(1) - np_1, \ldots, \nu_n(k) - np_k)$$

normalizált valószínűségi változókat, akkor felírhatjuk a $\sum\limits_{j=1}^{n} \bar{X}_j = \bar{\nu}_n$ azonosságot is. Tekintsük független, egyforma eloszlású vektor értékű $X_j = (X_j(1), \ldots, X_j(k))$, $1 \leq j \leq n$, valószínűségi változók sorozatát (az $X_1, \ldots, X_n$ véletlen vektorok függetlenek

egymástól, de az egyes X_j vektorok $X_j(1), \ldots, X_j(k)$) koordinátái függhetnek egymástól). A több-dimenziós centrális határeloszlástétel azt állítja, hogy a véletlen $S_n = \frac{1}{\sqrt{n}} \sum_{j=1} \bar{X}_j$ összegek eloszlásban konvergálnak egy alkalmas eloszláshoz a k-dimenziós téren, ha $n \to \infty$, és megadja explicit módon ezt az eloszlást. Jelen tárgyalásban nem adom meg a határeloszlás pontos leírását, mert arra nem lesz szükségünk, hanem megelégszem az eredmény azon következményének a leírásával, amelyet a továbbiakban használni fogunk. (Az ilyen határeloszlástételben megjelenő határeloszlásokat nevezik több-dimenziós normális eloszlásnak.) Vegyük észre, hogy a több-dimenziós centrális határeloszlástétel leírja $\frac{1}{\sqrt{n}}\bar{\nu}_n$ eloszlását nagy n számokra jó közelítéssel. Ez van az alább ismertetendő χ-négyzet (ejtsd: khi-négyzet) próba, illetve az azt megalapozó határeloszlástétel hátterében. Ennek megfogalmazása érdekében vezessük be először a χ-négyzet eloszlás fogalmát.

A k szabadságfokú χ-négyzet eloszlás definiciója. *Legyenek $X_1, \ldots, X_k$ független standard normális eloszlású valószínűségi változók. Ekkor a $Z_k = X_1^2 + X_2^2 + \cdots + X_k^2$ valószínűségi változó, illetve minden a Z_k valószínűségi változóval megegyező eloszlású valószínűségi változó eloszlása a k szabadságfokú χ-négyzet eloszlás.*

Tétel. *Legyen adva k urna, és dobjunk be ezek valamelyikébe véletlenül n golyót egymástól függetlenül úgy, hogy mindegyik golyó p_j, $p_j > 0$, valószínűséggel esik a j-ik urnába, $\sum_{j=1}^{k} p_j = 1$. Jelölje $\nu_n(j)$ a j-ik urnába eső golyók számát, $1 \le j \le k$. A $Z_n = \sum_{j=1}^{k} \frac{(\nu_n(j) - np_j)^2}{np_j}$ valószínűségi változóknak létezik határeloszlása $n \to \infty$ esetén, és ez a $k-1$ szabadságfokú χ-négyzet eloszlás, ahol k az urnák száma.*

A következő feladat, amelyet a mobiDIÁK könyvtár Feladatok a hipotézisvizsgálat témaköréből jegyzetből vettem (2.19 Példa) példát mutat a fenti tétel alkalmazására.

1.) Egy újonnan kifejlesztett müzli ötféle magot (A, B, C, D és E) tartalmaz, melyek százalékos megoszlása a terméken levő tájékoztató szerint 35%, 25%, 20%, 10% illetve 10%. Egy véletlenül kiválasztott zacskóban az alábbi mennyiségi megoszlást találtuk:

A tipusú mag 184, B tipusú mag 145, C tipusú mag 100, D tipusú mag 68, E tipusú mag 63.

Döntsünk 90%-os szinten arról, hogy a minta összetétele megfelel-e a csomagoláson feltüntetetteknek.

Megoldás: Jelen feladatban a H_0 null-hipotézis azt jelenti, hogy az összetétel megfelel a csomagoláson feltüntetetteknek a H_1 ellenhipotézis pedig azt, hogy nem felel meg. Olyan döntési eljárást keresünk, amelynek elsőfajú hibája $\alpha = 0.1$. A kisérletben egy $n = 560 = 184 + 145 + 100 + 68 + 63$ elemű mintát vettünk, amelynek a null-hipotézis szerint minden eleme 0.35 valószínűséggel az A osztályba, 0.25 valószínűséggel a B osztályba, 0.2 valószínűséggel a C osztályba, 0.1 valószínűséggel a D osztályba és 0.1 valószínűséggel az E osztályba esik. A megfigyelt értékek

$k_1 = 184$, $k_2 = 145$, $k_3 = 100$, $k_4 = 68$ és $k_5 = 63$. A null-hipotézis teljesülése esetén a $\chi^2 = \sum_{l=1}^{5} \frac{(k_l - np_l)^2}{np_l}$ valószínűségi változó eloszlása a 4 szabadságfokú χ-négyzet eloszlás, amely a 0.9 értéket a 7.779 pontban veszi fel. Mivel a jelen mért értékek esetén a statisztika értéke 5.6454, ezért elfogadjuk a null-hipotézist, azaz azt, hogy a terméken megadott tájékoztató helyes.

Tétel független valószínűségi változók összegének a sűrűségfüggvényéről. *Legyen ξ és η két független valószínűségi változó $f(\cdot)$ és $g(\cdot)$ sűrűségfüggvénnyel. Ekkor a $\xi + \eta$ összegnek is létezik sűrűségfüggvénye, és az az*

$$f * g(x) = \int_{-\infty}^{\infty} f(u)g(x-u)\,du, \quad -\infty < x < \infty$$

függvény. (A fenti képletben definiált $f * g$ kifejezést az f és g függvény konvolúciójának nevezik az irodalomban.)

2.) Legyenek ξ_1 és ξ_2 független exponenciális eloszlású valószínűségi változók λ paraméterrel, azaz legyen sűrűségfüggvényük $f(x) = \lambda e^{-\lambda x}$ ha $x \geq 0$, és $f(x) = 0$, ha $x > 0$. Számítsuk ki $\xi_1 + \xi_2$ sűrűségfüggvényét.

Általánosabban, legyenek $\xi_1, \dots \xi_m$ független exponenciális eloszlású valószínűségi változók $\lambda > 0$ paraméterrel. Számítsuk ki $\xi_1 + \cdots + \xi_m$ sűrűségfüggvényét.

Megoldás: Ki kell számolnunk az $f * f(x)$ illetve $\underbrace{f * \cdots * f}_{m-\text{szer}}(x)$ konvolúciókat a fenti $f(x)$ sűrűségfüggvénnyel. Mivel $f(x) = 0$, ha $x \leq 0$, a konvolúciót meghatározó integrálban szereplő $f(y)f(x-y)$ integrandus nulla, ha $y \leq 0$ vagy $x - y \leq 0$. Innen a konvolúciót definiáló integrál csak $x \geq 0$ esetén lehet nulla, az $x \leq 0$ esetben $f(y)f(x-y) > 0$ minden y-ra nulla, és $x \geq 0$ esetén az $f(y)f(x-y) > 0$ integrandus csak $0 \leq y \leq x$ esetén nem nulla. Innen a $\xi_1 + \xi_2$ valószínűségi változó sűrűségfüggvénye $f_2(x) = f * f(x)$ $x < 0$-ra $f_2(x) = 0$, és

$$f_2(x) = f * f(x) = \int_{-\infty}^{\infty} f(y)f(x-y)\,dy = \int_0^x \lambda e^{-\lambda y}\lambda e^{-\lambda(x-y)}\,dy$$
$$= \int_0^x \lambda^2 e^{-\lambda x}\,dy = \lambda^2 x e^{-\lambda x}, \quad \text{ha } x \geq 0.$$

Hasonlóan, ha $f_m(x) = \underbrace{f * \cdots * f}_{m-\text{szer}}(x)$ jelöli $\xi_1 + \cdots \xi_m$ sűrűségfüggvényét, akkor $f_m(x) = 0$ minden $m \geq 1$ számra, ha $x < 0$. Azt állítjuk, hogy $f_m(x) = \frac{\lambda^m x^{m-1}}{(m-1)!} e^{-\lambda x}$, ha $x \geq 0$. Ezen állítás bizonyításához elég belátni teljes indukcióval azt, hogy $f_{m-1} * f(x) = f_m(x)$ a fent definiált f_m függvényekkel. Viszont

$$f_{m-1} * f(x) = \int_{-\infty}^{\infty} f_{m-1}(y)f(x-y)\,dy = \int_0^x \lambda^{m-1}\frac{y^{m-2}}{(m-2)!}\lambda e^{-\lambda y}e^{-\lambda(x-y)}\,dy$$
$$= \lambda^m e^{-\lambda x}\int_0^x \frac{y^{m-2}}{(m-2)!}\,dy = e^{-\lambda x}\frac{\lambda^m x^{m-1}}{(m-1)!}, \quad \text{ha } x \geq 0.$$

Másrészt $f_m(x) = 0$, ha $x \le 0$.

3.) Legyen ξ és η két független valószínűségi változó, mind a kettő $f(x) = \frac{1}{2}e^{-|x|}$, $-\infty < x < \infty$, sűrűségfüggvénnyel. Lássuk be először, hogy $f(x)$ valóban sűrűségfüggvény. Számítsuk ki a $\xi + \eta$ valószínűségi változó $g(x)$ sűrűségfüggvényét.

Megoldás: Az $f(x)$ függvény minden pontban nem negatív. Annak ellenőrzéséhez, hogy $f(x)$ sűrűségfüggvény azt kell megmutatnunk, hogy $\int_{-\infty}^{\infty} f(x)\,dx = 1$. Viszont $\int_{-\infty}^{\infty} f(x)\,dx = \frac{1}{2}\int_{-\infty}^{0} e^x\,dx + \frac{1}{2}\int_0^{\infty} e^{-x}\,dx = \frac{1}{2}\left[e^x\right]_{-\infty}^{0} + \frac{1}{2}\left[-e^{-x}\right]_0^{\infty} = 1$.

A $\xi+\eta$ valószínűségi változó $g(x)$ sűrűségfüggvényét a $g(x) = \int_{-\infty}^{\infty} f(y)f(x-y)\,dy$ formula segítségével számíthatjuk ki. Számítsuk ki ezt az integrált. Tekintsük először azt az esetet, amikor $x \ge 0$. Az integrált számítsuk ki úgy, hogy nézzük mind a négy (elvileg) lehetséges esetet, amikor

a) $y \ge 0$ és $x - y \ge 0$,

b) $y \ge 0$ és $x - y < 0$,

c) $y < 0$, $x - y \ge 0$,

d) $y < 0$, $x - y < 0$.

Számítsuk ki mind a négy esetben azt, hogy milyen tartományban veszi fel értékét az y változó, és mi az integrandus illetve az integrál értéke ebben a tartományban.

Az a) esetben $0 \le y \le x$, az integrandus $f(y)f(x-y) = \frac{1}{4}e^{-y}e^{-(x-y)} = \frac{e^{-x}}{4}$, az integrál pedig $\frac{xe^{-x}}{4}$.

A b) esetben $y > x$ és $f(y)f(x-y) = \frac{e^{-y}e^{x-y}}{4} = \frac{e^{x-2y}}{4}$ az integrál pedig $\frac{1}{4}\int_x^{\infty} e^{x-2y}\,dy = \frac{e^{-x}}{8}$.

A c) esetben $y < 0$ és $f(y)f(x-y) = \frac{1}{4}e^{y}e^{-(x-y)} = \frac{e^{2y-x}}{4}$, az integrál pedig $\frac{1}{4}\int_{-\infty}^{0} e^{2y-x}\,dy = \frac{e^{-x}}{8}$.

A d) eset nem lehetséges, mert ekkor egyrészt az $y < 0$ másrészt az $y > x \ge 0$ feltételeknek kellene teljesülniük.

Innen azt kapjuk, hogy $g(x) = \frac{(x+1)e^{-x}}{4}$, ha $x > 0$. Mivel f szimmetrikus függvény, ezért mint nem nehéz megmutatni, $f(x)$ is az. Tehát $g(-x) = g(x)$, és $g(x) = \frac{(|x|+1)e^{-|x|}}{4}$.

4.) Legyenek ξ és η független valószínűségi változók $f(x)$ és $g(x)$ sűrűségfüggvénnyel. Mutassuk meg, hogy $\xi-\eta$ sűrűségfüggvénye a $h(x) = \int_{-\infty}^{\infty} f(x+y)g(y)\,dy$ függvény. Értsük meg e formula szemléletes tartalmát is.

Megoldás Legyen $\bar{\eta} = -\eta$. Ekkor $\bar{\eta}$ sűrűségfüggvénye $g(-x)$, ξ és $\bar{\eta}$ függetlenek és $\xi - \eta = \xi + \bar{\eta}$. Innen $\xi - \eta$ sűrűségfüggvénye $h(x) = \int_{-\infty}^{\infty} f(x-y)g(-y)\,dy$, és elvégezve az $\bar{y} = -y$ helyettesítést az integrálban megkapjuk a kívánt állítást.

Szemléletes (nem precíz) magyarázat: Annak valószínűsége, hogy a $\xi - \eta$ valószínűségi változó az $[x, x+dx]$ intervallumba esik $h(x)\,dx$. Ez úgy következhet be, hogy az a ξ valószínűségi változó valamely az $[x+y, x+y+dx]$ intervallumbeli értéket vesz fel, az η pedig az $[y, y+dy]$ intervallumba esik. Ennek valószínűsége

rögzített y számra $f(x+y)g(y)\,dx\,dy$. Ezt a kifejezést az összes lehetséges y értékre összegezni, illetve mivel y kontinum sok értéket vehet fel, integrálni kell. Ez azt jelenti, hogy $h(x)\,dx = \left(\int f(x+y)g(y)\,dy\right)\,dx$. Hasonló módon megmagyarázható a konvolució formula is.)

Azon képlet segítségével, amely lehetővé teszi, hogy kiszámoljuk két független valószínűségi változó összegének a sűrűségfüggvényét be lehet látni a következő lemmát.

Lemma. *Legyen η_1 és η_2 két független normális eloszlású valószínűségi változó m_1 illetve m_2 várható értékkel, σ_1^2 és σ_2^2 szórásnégyzettel. Az $\eta_1+\eta_2$ összeg m_1+m_2 várható értékű és $\sigma_1^2+\sigma_2^2$ szórásnégyzetű normális eloszlású valószínűségi változó.*

A lemma igazolásának érdekében érdemes megjegyezni, hogy

$$\int_{-\infty}^{\infty} e^{-(x-A)^2/B}\,dx = \sqrt{B\pi}.$$

Ezt láthatjuk például az $y = \sqrt{2}\dfrac{x-A}{\sqrt{B}}$ helyettesítéssel, és abból a tényből, hogy a $\varphi(y) = \frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ függvény sűrűségfüggvény. Ez az észrevétel azért hasznos, mert ez lehetővé teszi, hogy amennyiben olyan integrált kell kiszámolni, amelyben az integrandus exponensében egy kvadratikus alak szerepel, akkor az inegrandusban szereplő kifejezést teljes négyzetté alakítva ki tudjuk számolni az integrált. Ez a gondolata a lemma igazolásának is.

A lemma igazolása: Az $\eta_1+\eta_2$ valószínűségi változó sűrűségfüggvénye

$$\begin{aligned}
f(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} e^{-(u-m_1)^2/2\sigma_1^2} e^{-(x-u-m_2)^2/2\sigma_2^2}\,du \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -u^2\left(\frac{1}{2\sigma_1^2}+\frac{1}{2\sigma_2^2}\right) + u\left(\frac{m_1}{\sigma_1^2}+\frac{x-m_2}{\sigma_2^2}\right) \right. \\
&\qquad\qquad \left. - \frac{m_1^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_2^2} \right\} du \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}\left(u - \frac{m_1\sigma_2^2+(x-m_2)\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2 \right. \\
&\qquad\qquad \left. + \frac{(m_1\sigma_2^2+(x-m_2)\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)} - \frac{m_1^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_2^2} \right\} du \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{ -\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1^2\sigma_2^2}u^2 \right\} du \\
&\qquad\qquad \exp\left\{ \frac{(m_1\sigma_2^2+(x-m_2)\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)} - \frac{m_1^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_2^2} \right\} \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left\{ \frac{(m_1\sigma_2^2+(x-m_2)\sigma_1^2)^2}{2\sigma_1^2\sigma_2^2(\sigma_1^2+\sigma_2^2)} - \frac{m_1^2}{2\sigma_1^2} - \frac{(x-m_2)^2}{2\sigma_2^2} \right\} \\
&= \frac{1}{\sqrt{2\pi(\sigma_1^2+\sigma_2^2)}} \exp\left\{ -\frac{(x-m_1-m_2)^2}{2(\sigma_1^2+\sigma_2^2)} \right\}.
\end{aligned}$$

5.) Legyen ξ egyenletes eloszlású valószínűségi változó a $[-1,1]$ intervallumon, azaz legyen sűrűségfüggvénye $f(x)=\frac{1}{2}$, ha $-1\le x\le 1$, és $f(x)=0$, ha $x>1$ vagy $x<-1$. Számoljuk ki a ξ^4 valószínűségi változó várható értékét és szórásnégyzetét.

Megoldás: $E\xi^4=\int_{-1}^{2}\frac{1}{2}x^4\,dx=\frac{1}{10}\left[x^5\right]_{-1}^{1}=\frac{1}{5}$. $\operatorname{Var}\xi^4=E(\xi^4)^2-(E\xi^4)^2=E\xi^8-(E\xi^4)^2=\int_{-1}^{1}\frac{1}{2}x^8\,dx-\left(\int\frac{1}{2}x^4\,dx\right)^2=\frac{1}{9}-\left(\frac{1}{5}\right)^2=\frac{16}{225}$.

Második megoldás. Számoljuk ki ξ^4 $F(x)$ eloszlás és $f(x)$ sűrűségfüggvényét.

$$F(x)=P(\xi^4<x)=P(-x^{1/4}<\xi<x^{1/4})=x^{1/4},\qquad \text{ha } 0\le x\le 1,$$

$F(x)=0$, ha $x\le 0$ és $F(x)=1$, ha $x\ge 1$. Innen $f(x)=\frac{1}{4}x^{-3/4}$, ha $0\le x\le 1$, $f(x)=0$ egyébként. Ezért $E\xi=\int xf(x)\,dx=\int_0^1\frac{1}{4}x^{1/4}\,dx=\frac{1}{4}\cdot\frac{4}{5}=\frac{1}{5}$, $E\xi^2=\int x^2f(x)\,dx=\int_0^1\frac{1}{4}x^{5/4}\,dx=\frac{1}{4}\cdot\frac{4}{9}=\frac{1}{9}$, $\operatorname{Var}\xi=E\xi^2-(E\xi)^2=\frac{1}{9}-\left(\frac{1}{5}\right)^2=\frac{16}{225}$.