A Valószínűségszámítás I. előadássorozat hatodik témája.

Néhány fontos diszkrét eloszlás.

Ismertetem néhány fontos diszkrét eloszlás definicióját, és tárgyalom ezek legfontosabb tulajdonságait. Az eloszlások bevezetés előtt ismertetek néhány olyan példát, amelyekben azok megjelennek. Ezek közül néhánnyal már korábban is találkoztunk. Az első példa a következő.

a.) *A binomiális eloszlás.*

Tekintsünk először egy tipikus példát, amelyben a binomiális eloszlás megjelenik.

Dobjunk fel egy pénzdarabot n alkalommal egymástól függetlenül, és minden egyes dobásban legyen p, $0 \le p \le 1$, a fejdobás valószínűsége. Mi a valószínűsége annak, hogy pontosan k fejdobás következik be?

Ennek valószínűsége $\binom{n}{k}p^k(1-p)^{n-k}$. Valóban, $\binom{n}{k}$ olyan fej-írás sorozat létezik, amely k fej és $n-k$ írásjelet tartalmaz, és az adott feltételek mellett minden ilyen dobássorozat valószínűsége $p^k(1-p)^{n-k}$. Az ilyen jellegű feladatok leírására vezették be a binomiális eloszlás fogalmát.

A binomiális eloszlás definiciója. *Azt mondjuk, hogy a ξ valószínűségi változó binomiális eloszlású n és p paraméterrel, $n = 1, 2, \dots$, $0 \le p \le 1$, ha $P(\xi = k) = \binom{n}{k}p^k(1-p)^{n-k}$ minden $0 \le k \le n$ számra, és $P(\xi = k) = 0$ egyébként. Ezt az eloszlást szokás $B(n,p)$-vel jelölni.*

Egy $B(1,p)$ eloszlású ξ valószínűségi változót, azaz egy olyan ξ valószínűségi változót, amelyre $P(\xi = 0) = 1-p$, $P(\xi = 1) = p$ Bernoulli eloszlásúnak is szoktak nevezni.

Jegyezzük meg, hogy valóban eloszlást definiáltunk, azaz a nem-negatív valószínűségek összege egy, mert a binomiális tétel szerint

$$\sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} = (p + (1-p))^n = 1.$$

Jegyezzük meg, hogy a $B(n,p)$ és $B(m,p)$ binomiális eloszlások konvoluciója a $B(n+m,p)$ binomiális eloszlás.

Emlékeztetőül felidézem, hogy egy $\mathcal{P} = \{p_0, p_1, \dots\}$ és $\mathcal{Q} = \{q_0, q_1, \dots\}$ eloszlás konvoluciója az az $\mathcal{R} = \{r_1, r_2, \dots\}$ eloszlás amelynek elemeit az $r_n = \sum\limits_{k=0}^{n} p_k q_{n-k}$, $n = 1, 2, \dots$, képlet határozza meg. Ennek valószínűségi tartalma a következő: Ha ξ és η két független, nem negatív egész értékeket felvevő valószínűségi változó, amelyekre $P(\xi = n) = p_n$, $P(\eta = n) = q_n$, $n = 0, 1, 2, \dots$, akkor $\xi + \eta$ olyan nem negatív egész értékeket felvevő valószínűségi változó, amelyre $P(\xi + \eta = n) = r_n$, $n = 0, 1, 2, \dots$.

Valóban a fenti konvolucióról szóló állítás érvényes, mivel a tekintett két eloszlás konvoluciója egy olyan $p(k)$, $0 \le k \le n+m$ eloszlás, amelyre

$$
\begin{aligned}
p(k) &= \sum_{j=0}^{k} \binom{n}{j} p^j (1-p)^{n-j} \binom{m}{k-j} p^{k-j} (1-p)^{(m-(k-j))} \\
&= p^k (1-p)^{n-k} \sum_{j=0}^{k} \binom{n}{j} \binom{m}{k-j} = \binom{n+m}{k} p^k (1-p)^{n-k},
\end{aligned}
$$

mert

$$
\sum_{j=0}^{k} \binom{n}{j} \binom{m}{k-j} = \binom{n+m}{k}.
$$

Ez utóbbi állításnak kombinatorikus bizonyítását megadtam az első téma jegyzetében a 'Néhány a tárgyalt kombinatorikus kérdésekkel kapcsolatos eredmény' feladatsor 7. feladatának az ismertetésében. Sőt, e feladatsor 9. feladatában megadtam ezen azonosság egy általánosításának analitikus (az $(1+x)^\alpha$ függvény hatványsorának az alakján alapuló) bizonyítását is.

Két eloszlás konvolucióját kiszámolhatjuk az alább ismertetendő generátorfüggvény módszernek nevezett eljárás segítségével is. Ha adva van két $\mathcal{P} = \{p_0, p_1, \dots\}$ és $\mathcal{Q} = \{q_0, q_1, \dots\}$ eloszlás, akkor rendeljük hozzájuk az $F(x) = \sum\limits_{n=0}^{\infty} p_n x^n$ és $G(x) = \sum\limits_{n=0}^{\infty} q_n x^n$ hatványsorokat. Ezenkívül rendeljük a $\mathcal{P}$ és $\mathcal{Q}$ eloszlások $\mathcal{R} = \{r_0, r_1, \dots\}$ konvoluciójához a $H(x) = \sum\limits_{n=0}^{\infty} r_n x^n$ hatványsort. Ekkor teljesül az $F(x)G(x) = H(x)$ azonosság. Továbbá, ha ismerjük a $H(x)$ függvényt, akkor annak szukcesszív deriválásával majd a nulla érték behelyettesítésével kiszámolhatjuk a minket érdeklő r_n mennyiségeket. Innen az is következik, hogy ha $F(x)G(x) = H(x)$, akkor a $\mathcal{P}$ és $\mathcal{Q}$ eloszlások konvoluciója a $\mathcal{R}$ eloszlás.

Megmutatom azt is, hogyan lehet a binomiális eloszlások konvoluciójának alakjáról szóló állítást a generátorfüggvény módszer segítségével belátni. Számoljuk ki a $B(n,p)$ binomiális eloszlás $F_{n,p}(x)$ generátorfüggvényét.

$$
\begin{aligned}
F_{n,p}(x) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} x^k = (1-p)^n \sum_{k=0}^{n} \binom{n}{k} \left(\frac{px}{1-p}\right)^k \\
&= (1-p)^n \left[1 + \frac{px}{1-p}\right]^n = (1-p+px)^n.
\end{aligned}
$$

Innen

$$
F_{n,p}(x)F_{m,p}(x) = (1-p+px)^n (1-p+px)^m = (1-p+px)^{n+m} = F_{n+m,p}(x)
$$

minden $n \ge 1$, $m \ge 1$ egész és $0 \le p \le 1$ valós számra. Ez azt jelenti, hogy a $B(n,p)$ és $B(m,p)$ eloszlások konvoluciója a $B(n+m,p)$ eloszlás.

Feladat:

1.) Adjunk egyszerű, a binomiális eloszlás valószínűségi tartalmát kihasználó bizonyítást arra, hogy a $B(n,p)$ és $B(m,p)$ binomiális eloszlások konvoluciója a $B(n+m,p)$ binomiális eloszlás.

Megoldás: Elég megmutatni, hogy léteznek olyan független $B(n,p)$ paraméterű ξ és $B(m,p)$ paraméterű η binomiális eloszlású valószínűségi változók, amelyek $\xi+\eta$ összege $B(n+m,p)$ paraméterű binomiális eloszlású valószínűségi változó. Ennek érdekében tekintsük egy olyan pénzdarab $n+m$ darab egymás utáni független feldobását, amely p valószínűséggel esik a fej és $1-p$ valószínűséggel az írás oldalra. Legyen ξ az első n dobásban, η pedig az $n+1$-ik dobástól az $n+m$-ig dobásig tartó dobássorozatban megjelenő fej dobások száma. Ezek a ξ és η valószínűségi változók teljesítik a kívánt tulajdonságokat.

Számítsuk ki egy $B(n,p)$ eloszlású valószínűségi változó várható értékét és szórásnégyzetét. A legegyszerűbb módszer a következő: Legyenek $\xi_1,\dots,\xi_n$ független, $B(1,p)$ eloszlású valószínűségi változók. Ekkor $\xi=\sum\limits_{j=1}^{n}\xi_j$ $B(n,p)$ eloszlású valószínűségi változó, és $E\xi_j=1\cdot p+0\cdot(1-p)=p$, $E\xi_j^2=1\cdot p+0\cdot(1-p)=p$, $\operatorname{Var}\xi_j=E\xi_j^2-(E\xi_j)^2=p-p^2$ minden $j=1,\dots,n$ számra. Ezért a keresett várható érték és szórásnégyzet $E\xi=\sum\limits_{j=1}^{n}E\xi_j=np$ és $\operatorname{Var}\xi=\sum\limits_{j=1}^{n}\operatorname{Var}\xi_j=np(1-p)$.

Feladat:

2.) Számítsuk ki egy $B(n,p)$ eloszlású valószínűségi változó várható értékét és szórásnégyzetét az e fogalmakat definiáló összegek kiszámításának a segítségével.

Megoldás:

$$E\xi=\sum_{k=1}^{n}k\binom{n}{k}p^k(1-p)^{n-k},\qquad E\xi^2=\sum_{k=1}^{n}k^2\binom{n}{k}p^k(1-p)^{n-k},$$

és $\operatorname{Var}\xi=E\xi^2-(E\xi)^2$. Ezeket a kifejezéseket kell kiszámolnunk. Vegyük észre, hogy $k\binom{n}{k}=n\binom{n-1}{k-1}$, $k^2\binom{n}{k}=k\binom{n}{k}+k(k-1)\binom{n}{k}=n\binom{n-1}{k-1}+n(n-1)\binom{n-2}{k-2}$. Innen

$$\begin{aligned}E\xi&=n\sum_{k=0}^{n-1}\binom{n-1}{k}p^{k+1}(1-p)^{n-k-1}\\&=np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{(n-1)-k}=np(1+(1-p))^{n-1}=np,\end{aligned}$$

$$E\xi^2=n\sum_{k=0}^{n-1}\binom{n-1}{k}p^{k+1}(1-p)^{n-k-1}+n(n-1)\sum_{k=0}^{n-2}\binom{n-2}{k}p^{k+2}(1-p)^{n-k-2}$$

$$= np + n(n-1)p^2 \sum_{k=0}^{n-2} \binom{n-2}{k} p^k (1-p)^{(n-2)-k}$$
$$= np + n(n-1)p^2 (p + (1-p))^{n-1} = np + (n^2 - n)p^2,$$

és $\operatorname{Var}\xi = np + (n^2-n)p^2 - n^2p^2 = np(1-p)$.

A binomiális eloszlás természetes többdimenziós általánosítása a polinomiális eloszlás. Ennek megértéséhez tekintsük a következő feladatot. Adva van r urna. Ezekbe bedobunk összesen n golyót egymástól függetlenül. Az egyes golyók p_1 valószínűséggel esnek az első p_2 valószínűséggel a második, ... p_r valószínűséggel az r-ik urnába. Mi a valószínűsége annak, hogy a j-ik urnába k_j golyó esik, $1 \le j \le r$, $\sum\limits_{j=1}^{r} k_r = n$?

Ez a valószínűség $\frac{n!}{k_1!k_2!\cdots k_r!} p_1^{k_1} \cdots p_r^{k_r}$, mert $\frac{n!}{k_1!k_2!\cdots k_r!}$ ilyen dobássorozat van, és minden ilyen dobássorozat valószínűsége $p_1^{k_1} \cdots p_r^{k_r}$. (Itt a lehetséges dobássorozatok összeszámlálásánál megkülönböztettünk két olyan dobássorozatot, amelyekben ugyanannyi golyó esett az egyes urnákba, de más sorrendben.)

A polinomiális eloszlás definiciója. *A $(\xi_1, \dots, \xi_r)$ r változós véletlen vektor polinomiális eloszlású n, $n = 1, 2, \dots$, és p_j, $1 \le j \le r$, paraméterekkel, $\sum\limits_{j=1}^{r} p_j = 1$, ha*

$$P(\xi_1 = k_1, \xi_2 = k_2, \dots, \xi_r = k_r) = \frac{n!}{k_1!k_2!\cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$$

az olyan $0 \le k_j \le n$, $1 \le j \le r$ számokra, amelyekre $\sum\limits_{j=1}^{r} k_j = n$.

A polinomiális eloszlás valóban eloszlás, azaz $P(\xi_1 = k_1, \xi_2 = k_2, \dots, \xi_r = k_r) \ge 0$ minden lehetséges $k_1, \dots, k_r$ értékre, és $\sum P(\xi_1 = k_1, \xi_2 = k_2, \dots, \xi_r = k_r) = 1$. Ez utóbbi azonosságot vagy a polinomiális eloszlás valószínűségi tartalmának felhasználásával vagy a polinomiális tétel segítségével láthatjuk, mely szerint

$$\sum P(\xi_1 = k_1, \xi_2 = k_2, \dots, \xi_r = k_r) = \sum_{\substack{k_j \ge 0,\ j=1,\dots,r \\ k_1 + \cdots + k_r = n}} \frac{n!}{k_1!k_2!\cdots k_r!} p_1^{k_1} p_2^{k_2} \cdots p_r^{k_r}$$
$$= (p_1 + \cdots + p_r)^n = 1.$$

Feladat:

3.) Legyen $(\xi_1, \dots, \xi_r)$ polinomiális eloszlású véletlen vektor n és p_j, $1 \le j \le r$, $\sum\limits_{j=1}^{r} p_j = 1$ paraméterekkel. Ekkor $E\xi_j = np_j$, $\operatorname{Var}\xi_j = np_j(1-p_j)$, $1 \le j \le r$, és $\operatorname{Cov}(\xi_j, \xi_k) = -np_jp_k$, $1 \le j, k \le r$, $j \ne k$.

b.) *A negatív binomiális eloszlás és annak speciális esete, a geometriai eloszlás.*

Most is először egy olyan tipikus példát tekintek, amelyben ezek az eloszlások megjelennek.

Rögzítsünk egy r pozitív egész számot, és dobjunk fel egy pénzdarabot, amely p valószínűséggel esik a fej $1-p$ valószínűséggel az írás oldalra egymás után többször egymástól függetlenül, egészen addig amikor az r-ik fejdobás megjelenik. Mi a valószínűsége annak, hogy ez a $k+r$-ik dobás?

Ez a valószínűség $\binom{k+r-1}{r-1}(1-p)^k p^r$. Ez ugyanis azt jelenti, hogy az első $k+r-1$ dobásban pontosan $r-1$ fej és k írás dobás történt, aminek valószínűsége $\binom{k+r-1}{r-1}(1-p)^k p^{r-1}$, valamint a $k+r$-ik dobás fej, aminek a valószínűsége p, és független az előző eseménytől.

A negatív binomiális és geometriai eloszlás definiciója. *Azt mondjuk, hogy a ξ valószínűségi változó negatív binomiális eloszlású r, $r=1,2,\dots,$ és p, $0<p\le 1$, paraméterrel, ha ξ $r, r+1,\dots,$ értékeket vesz fel és*

$$P(\xi=k+r)=\binom{k+r-1}{r-1}(1-p)^k p^r, \quad k=0,1,2,\dots.$$

Az $r=1$ és p paraméterhez tartozó negatív binomiális eloszlást, azaz azt az eloszlást, amelyre $P(\xi=k+1)=p(1-p)^k$, $k=1,2,\dots,$ p paraméterű geometriai eloszlásnak is nevezik.

A második téma jegyzetében tárgyalt 3. és 5. feladat megoldási módszerének a segítségével megmutatható, hogy a negativ binomiális eloszlás valóban valószínűségeloszlás, azaz $\sum\limits_{k=0}^{\infty} P(\xi=k+r)=\sum\limits_{k=0}^{\infty}\binom{k+r-1}{r-1}(1-p)^k p^r=1$. E feladatokban ezt az állítást csak egy speciális esetben igazoltuk, (ha $k=3$, és $p=\frac16$), de az ott alkalmazott módszer alkalmazható az általános esetben is. Megmutatom, hogyan lehet az 5. feladat megoldásában használt módszert alkalmazni az általános esetben. A következő azonosságot fogjuk használni, amely a negatív binomiális eloszlás más tulajdonságainak a bizonyításában is hasznos.

$$\begin{aligned}\binom{k+r-1}{r-1}&=\binom{k+r-1}{k}=\frac{(k+r-1)(k+r-2)\cdots(r+1)r}{k!}\\&=\frac{r(r+1)\cdots(k+r-2)(k+r-1)}{k!}\\&=(-1)^k\frac{(-r)(-r-1)\cdots(-r-k+2)(-r-k+1)}{k!}=(-1)^k\binom{-r}{k}.\end{aligned} \tag{1}$$

Ezen azonosság alapján

$$P(\xi=k+r)=\binom{k+r-1}{r-1}(1-p)^k p^r=\binom{-r}{k}(-1)^k(1-p)^k p^r=\binom{-r}{k}(p-1)^k p^r,$$
$$k=0,1,2,\dots,$$

egy r és p paraméterű negatív binomiális eloszlásra. A negatív binomiális eloszlás fenti reprezentációja az oka ezen eloszlás elnevezésének. Ebből a képletből, illetve az $(1+x)^{-r}$ függvény $(1+x)^{-r}=\sum\limits_{k=0}^{\infty}\binom{-r}{k}x^k$, ha $|x|<1$ alakú Taylor sor előállításából az is következik, hogy

$$\sum_{k=0}^{\infty}\binom{k+r-1}{r-1}(1-p)^kp^r=p^r\sum_{k=0}^{\infty}\binom{-r}{k}(p-1)^k=p^r[(1+(p-1)]^{-r}=1,$$

amint állítottam.

Tárgyaljuk meg a negatív binomiális eloszlás néhány további fontos tulajdonságát.

Feladat:

4.) Bizonyítsuk be valószínűségi meggondolások segítségével, hogy egy r_1 és p paraméterű valamint egy r_2 és p paraméterű negatív binomiális eloszlás konvolúciója egy r_1+r_2 és p paraméterű negatív binomiális eloszlású valószínűségeloszlás. Következésképpen, egy r és p paraméterű negatív binomiális eloszlás előáll, mint r darab p paraméterű geometriai eloszlás konvoluciója.

Megoldás: Tekintsünk két egymásól független végtelen fej-írás dobássorozatot, amelyekben az egyes dobások p valószínűséggel esnek a fej és $1-p$ valószínűséggel az írás oldalra. Tekintsük azt a dobássorozatot, amely az első dobássorozat elemeiből áll, amíg az r_1-ik fejdobás megjelenik, majd a második dobássorozat elemeivel folytatódik egész addig, amíg abban az r_2-ik fejdobás megjelenik, majd a sorozat abbamarad. Jelölje ξ az így definiált sorozat hosszát, ξ_1 e sorozatnak az első végtelen dobássorozattal, ξ_2 pedig e sorozatnak a második dobássorozattal közös részének a hosszát. Ekkor ξ_1 és ξ_2 független valószínűségi változók, $\xi_1+\xi_2=\xi$, továbbá ξ_1, ξ_2 és ξ negatív binomiális eloszlású valószínűségi változók (r_1,p), (r_2,p), illetve (r_1+r_2,p) paraméterekkel. Innen következik a feladat állítása.

Lássuk be az előbbi feladat állítását közvetlen számolással is. Elég megmutatni azt, hogy egy r és p paraméterű negatív binomiális eloszlású és egy p paraméterű geometriai eloszlás konvoluciója egy $r+1$ és p paraméterű negatív binomiális eloszlás.

Ehhez azt kell belátni, hogy

$$\sum_{j=0}^{k}\binom{r+j-1}{r-1}(1-p)^jp^r(1-p)^{k-j}p=\binom{k+r}{r}(1-p)^kp^{r+1}.$$

Ezt az azonosságot könnyen látjuk, ha tudjuk, hogy érvényes a következő azonosság:

$$\sum_{j=0}^{k}\binom{j+r-1}{r-1}=\binom{k+r}{r}.$$

Az utóbbi azonosság viszont következik például a következő kombinatorikai érvelésből. A természetes számok halmazán $\binom{k+r}{r}$ módon jelölhetünk ki $r+1$ számot úgy, hogy a (nagyság szerint) $r+1$-ik szám a $k+r+1$ szám legyen, és az azonosság jobboldala ezzel egyenlő. Ezt viszont úgyis kiszámolhatjuk, hogy azt tekintjük, hány olyan elrendezés van, amelyben az r-ik kijelölt pont a $j+r$, az $r+1$-ik pedig a $k+r+1$ szám, majd összegezünk $0 \le j \le k$-ra. Mivel rögzített j-re az ilyen elrendezések száma $\binom{j+r-1}{r-1}$, innen következik a felírt azonosság.

Megmutatom, hogy hogyan lehet a 4. feladat állítását analitikus módon bebizonyítani. Tekintsük az r és p paraméterekhez tartozó $g_{r,p}(x)$ generátorfüggvényt, azaz a következő összeget:

$$g_{r,p}(x) = \sum_{k=0}^{\infty} \binom{k+r-1}{r-1} (1-p)^k p^r x^{k+r}.$$

Ha belátjuk, hogy ez az összeg konvergens például az $-1 < x < 1$ intervallumon, valamint $g_{r_1,p}(x) g_{r_2,p}(x) = g_{r_1+r_2,p}(x)$, akkor innen az ötödik téma ismertetésének végén tett észrevételekből következik az állítás. A $g_{r,p}(x)$ függvény viszont egyszerűen zárt alakra hozható az (1) azonosság segítségével. Valóban,

$$\begin{aligned} g_{r,p}(x) &= \sum_{k=0}^{\infty} (-1)^k \binom{-r}{k} (1-p)^k p^r x^{k+r} = (px)^r \sum_{k=0}^{\infty} \binom{-r}{k} (-(1-p)x)^k \\ &= (px)^r (1-(1-p)x)^{-r} = \left(\frac{px}{1-(1-p)x} \right)^r, \end{aligned}$$

ami az $(1+x)^{-r}$ függvény hatványsorának az alakjából látható. Innen viszont könnyen látható a generátorfüggvényekre felírt azonosság.

Számítsuk ki egy negatív binomiális eloszlású valószínűségi változó várható értékét és szórásnégyzetét. Ezt kiszámíthatjuk a generátorfüggvény deriválásának segítségével. Először egy más módszert választunk. Egy geometriai eloszlású valószínűségi változó várható értékét és szórásnégyzetét kiszámoljuk direkt módon, majd az általános esetet visszavezetjük erre.

Legyen ξ geometriai eloszlású valószínűségi változó p paraméterrel, azaz legyen $P(\xi = k+1) = p^k(1-p)$, $k = 0, 1, \dots$. Ekkor

$$E\xi = \sum_{k=0}^{\infty} (k+1)(1-p)^k p = p \sum_{k=0}^{\infty} k(1-p)^{k-1},$$

$E\xi^2 = p \sum\limits_{k=0}^{\infty} k^2 (1-p)^{k-1}$, és $\operatorname{Var} \xi = E\xi^2 - (E\xi)^2$. Deriváljuk kétszer a $\sum\limits_{k=0}^{\infty} x^k = \frac{1}{1-x}$, $|x| < 1$ azonosságot. Azt kapjuk, hogy

$$\frac{1}{(1-x)^2} = \sum_{k=1}^{\infty} k x^{k-1},$$

$$\frac{2}{(1-x)^3} = \sum_{k=2}^{\infty} k(k-1) x^{k-2}.$$

Innen $x = 1-p$ helyettesítéssel $\frac{1}{p^2} = \sum\limits_{k=1}^{\infty} k(1-p)^{k-1}$, $E\xi = p\sum\limits_{k=1}^{\infty} k(1-p)^{k-1} = \frac{p}{p^2} = \frac{1}{p}$, $\sum\limits_{k=0}^{\infty} k^2(1-p)^{k-1} = (1-p)\sum\limits_{k=2}^{\infty} k(k-1)(1-p)^{k-2} + \sum\limits_{k=1}^{\infty} k(1-p)^{k-1} = \frac{2(1-p)}{p^3} + \frac{1}{p^2}$, $E\xi^2 = \frac{2(1-p)}{p^2} + \frac{1}{p}$, $\operatorname{Var}\xi = \frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}$.

Feladat:

5.) Számítsuk ki egy n és p paraméterű negatív binomiális eloszású valószínűségi változó várható értékét és szórásnégyzetét annak az ismeretnek a segítségével, hogy egy p paraméterű geometriai eloszlású valószínűségi változó várható értéke $\frac{1}{p}$ és szórásnégyzete $\frac{1-p}{p^2}$.

Megoldás: Tekintsünk n független geometriai eloszlású $\xi_1, \dots, \xi_r$ valószínűségi változót, és legyen $S = \sum\limits_{j=1}^{n} \xi_j$. Ekkor S negatív binomiális eloszlású valószínűségi változó n és p paraméterekkel. Továbbá, $ES = nE\xi_1 = \frac{n}{p}$, $\operatorname{Var} S = \frac{n(1-p)}{p^2}$.

Tárgyaljuk meg, hogyan lehet egy negatív binomiális eloszlású valószínűségi változó várható értékét és szórásnégyzetét kiszámolni a generátorfüggvénye ismeretében. Sőt, lássunk be egy olyan formulát, amely általánosabb esetben is alkalmazható.

Legyen $\mathcal{P} = \{p_k\colon\ k = 1, 2, \dots\}$, $p_k \ge 0$, $k = 0, 1, 2, \dots$, $\sum\limits_{k=0}^{\infty} p_k = 1$, valószínűség eloszlás a nem negatív egész számokon. Vezessük be az ezen eloszláshoz tartozó

$$g(x) = g_{\mathcal{P}}(x) = \sum_{k=0}^{\infty} p_k x^k.$$

generátorfüggvényt. Vegyük észre, hogy a $g(x)$ függvényt definiáló hatványsor konvergens a $-1 < x < 1$ intervallumban. Továbbá kétszeri deriválás és az $x = 1$ (formális) helyettesítés azt sugallja, hogy

$$g'(x) = \sum_{k=1}^{\infty} k p_k x^{k-1}, \quad g''(x) = \sum_{k=2}^{\infty} k(k-1) p_k x^{k-2},$$

valamint

$$g'(1) = \sum_{k=1}^{\infty} k p_k, \quad g''(1) = \sum_{k=2}^{\infty} k(k-1) p_k,$$

és ezen azonosságok segítségével kiszámolhatjuk a keresett várható értéket és szórásnégyzetét. Ugyanis $E\xi = \sum\limits_{k=0}^{\infty} k p_k = g'(1)$, $E\xi^2 = \sum\limits_{k=0}^{\infty} k^2 p_k = \sum\limits_{k=2}^{\infty} k(k-1) p_k + \sum\limits_{k=1}^{\infty} k p_k = g''(1) + g'(1)$, ahonnan $\operatorname{Var}\xi = E\xi^2 - (E\xi)^2 = g''(1) + g'(1) - g'(1)^2$.

Felmerül a kérdés, szabad-e az alábbi számolásokat végrehajtani. A problémát az okozza, hogy hatványsorokat szabad tagonként deriválni a konvergenciatartomány belsejében, de mi az $x = 1$ helyettesítést hajtottuk végre, és lehet, hogy a $g(x)$ függvény hatványsora nem terjeszthető ki egy a $(-1, 1)$ intervallumnál nagyobb intervallumra. A negatív binomiális eloszlás generátorfüggvénye esetén egy ilyen kiterjesztés lehetséges, de érdemes az általános esetet is tekinteni, amikor nem lehet ezt a lehetőséget kizárni.

Az analízis bizonyos eredményei biztosítják a fenti számolás jogosságát. Egyrészt igaz az, hogy ha egy $h(x) = \sum\limits_{k=0}^{\infty} a_k x^k$ hatványsor konvergens egy $(-A, A)$ intervallumban, és a $h(A) = \sum\limits_{k=0}^{\infty} a_k A^k$ összeg konvergens, akkor $\lim\limits_{x \to A} h(x) = h(A)$. Másrészt, ha a $h(x)$ Taylor-sor a_k együtthatói nem negatívak, akkor $\lim\limits_{x \to A} h(x) = h(A)$. Speciálisan, ha ebben az esetben $\lim\limits_{x \to A} h(x) < \infty$ akkor $h(A) < \infty$. Ezeknek az eredményeknek az alkalmazásával $A = 1$ és $h(x) = g'(x)$ illetve $h(x) = g''(x)$ választással meg lehet mutatni a fenti számolások jogosságát. Érdemes megjegyezni, hogy az általános esetben a $g'(1) = \lim\limits_{x \to 1} g'(x)$, $g''(1) = \lim\limits_{x \to 1} g''(x)$ képletet kell alkalmaznunk annak érdekében, hogy a $g'(1)$ és $g''(1)$ számokat értelmezni tudjuk.

Megjegyzés: Az, hogy egy hatványsor hogyan viselkedik a kovergenciakörének szélén, a komplex függvénytan egyik nehéz és fontos kérdése. Annak érdekében, hogy lássunk egy egyszerű példát, amely rávilágíthat arra, hogy vigyázni kell a formális számolások során, tekintsük az $h(x) = \frac{1}{1+x} = \sum\limits_{k=0}^{\infty} (-1)^k x^k$ hatványsort. Ekkor $h(x)$ konvergens a $-1 < x < 1$ intervallumban, $h(1) = \frac{1}{2}$, és $h(x)$ hatványsora az $x = 1$ pontban, a $\sum\limits_{k=0}^{\infty} (-1)^k$ sor, divergens. Ez a példa azért nem mond ellen a fent elmondottaknak, mert ebben a hatványsorban vannak negatív együtthatók is.

Feladatok:

6.) Lássuk be, hogy amennyiben egy ξ valószínűségi változó generátorfüggvénye valamely $g(x)$ függvény, akkor $E\xi = g'(1)$, $\operatorname{Var}\xi = g''(1) + g'(1) - g'(1)^2$, ahol az általános esetben a $g'(1) = \lim\limits_{x \to 1} g'(x)$ és $g''(1) = \lim\limits_{x \to 1} g''(x)$ képletek definiálják ezeket a mennyiségeket. Számoljuk ki egy r és p paraméterű negatív binomiális eloszlású valószínűségi változó várható értékét és szórásnégyzetét, felhasználva, hogy e valószínűségi változó generátorfüggvénye a $\left(\frac{px}{1-(1-p)x}\right)^r$ függvény.

Megoldás: A $g(x) = \sum\limits_{k=0}^{\infty} p_k x^k$ függvény két egymásutáni deriválása és az $x = 1$ helyettesítés adja, hogy $g'(1) = \sum\limits_{k=1} k p_k = E\xi$, $g''(1) = \sum\limits_{k=1} k(k-1) p_k = E\xi(\xi - 1)$. Mint azt az előadáson megtárgyaltuk az analízis bizonyos eredményei lehetővé teszik a fent alkalmazott tagonkénti deriválást. Ezért $E\xi = g'(1)$, $E\xi^2 = g''(1) + g'(1)$, és $\operatorname{Var}\xi = g''(1) + g'(1) - g'(1)^2$.

Az r és p paraméterű negatív binomiális eloszlás generátorfüggvénye

$$g(x) = \left(\frac{px}{1-(1-p)x}\right)^r.$$

Innen

$$g'(x) = rp\left(\frac{px}{1-(1-p)x}\right)^{r-1} \frac{1}{(1-(1-p)x)^2},$$

$$\begin{aligned} g''(x) =& r(r-1)p^2\left(\frac{px}{1-(1-p)x}\right)^{r-2} \frac{1}{(1-(1-p)x)^4} \\ &+ rp\left(\frac{px}{1-(1-p)x}\right)^{r-1} \frac{2(1-p)}{(1-(1-p)x)^3}. \end{aligned}$$

Behelyettesítéssel $E\xi = \frac{r}{p}$, $g''(1) = \frac{r^2+r}{p^2} - \frac{2r}{p}$, $\operatorname{Var}\xi = g''(1) + g'(1) - g'(1)^2 = \frac{r(1-p)}{p^2}$.

7.) Mutassunk példát olyan $\mathcal{P} = \{p_k\colon\ k = 1,2,\dots\}$, $p_k \ge 0$, $k = 0,1,2,\dots$, $\sum\limits_{k=0}^{\infty} p_k = 1$, valószínűség eloszlásra a nem negatív egész számokon, amelynek $g(x) = \sum\limits_{k=0}^{\infty} p_k x^k$ generátorfüggvénye semmilyen $\varepsilon > 0$ szám esetén nem terjeszthető ki a $-1-\varepsilon < x < 1+\varepsilon$ intervallumra.

Megoldás: Legyen $\alpha > 1$ tetszőleges szám. Ekkor $C(\alpha) = \sum\limits_{k=1}^{\infty} \frac{1}{k^\alpha} < \infty$, és $p_k = p_k(\alpha) = \frac{1}{C(\alpha)k^\alpha}$, $k = 1,2,\dots$, valószínűségeloszlás. Ennek generátor függvénye a $\sum\limits_{k=1}^{\infty} p_k x^k = \frac{1}{C(\alpha)} \sum\limits_{k=1}^{\infty} \frac{x^k}{k^\alpha}$ semmilyen $\varepsilon > 0$ számra nem konvergál a $-1-\varepsilon < x < 1+\varepsilon$ intervallumon, mert minden $x > 1$ számra $\lim\limits_{k\to\infty} \frac{x^k}{k^\alpha} = \infty$.

8.) Legyen ξ geometriai eloszlású valószínűségi változó, azaz legyen $P(\xi = k) = p(1-p)^{k-1}$, $k = 1,2,\dots$. (Itt a geometriai eloszlásnak az eredeti definiciótól formálisan eltérő, de azzal ekvivalens jellemzését adtuk meg. A $P(\xi = k)$ valószínűségeket adtuk meg $k = 1,2,\dots$ értékekre a $P(\xi = k+1)$, $k = 0,1,2,\dots$ valószínűségek helyett.) Lássuk be, hogy ξ teljesíti a következő (diszkrét) örökifjú tulajdonságot.

$$P(\xi = k+l | \xi > l) = p(1-p)^{k-1} = P(\xi = k), \quad l = 0,1,2, \ \ k = 1,2,\dots. \tag{2}$$

Adjuk meg ennek az azonosságnak a valószínűségszámítási magyarázatát is.

Megoldás:

$$\begin{aligned} P(\xi = k+l|\xi > l) &= \frac{P(\xi = k+l)}{\sum\limits_{j=1}^{\infty} P(\xi = l+j)} = \frac{p(1-p)^{k+l-1}}{\sum\limits_{j=1}^{\infty} p(1-p)^{l+j-1}} \\ &= \frac{(1-p)^{k-1}}{\sum\limits_{j=0}^{\infty} (1-p)^j} = p(1-p)^{k-1} = P(\xi = k). \end{aligned}$$

Tekintsük egy a fej-oldalára p valószínűséggel eső pénzérme végtelen egymás utáni feldobását. A most bebizonyított azonosság baloldalán annak a feltételes valószínűsége áll, hogy az l-ik dobás után még k dobást kell tenni ahhoz, hogy az első fejdobás megjelenjen, feltéve, hogy az első l dobásban, nem volt fejdobás. Mivel az első l dobás és az utána következő dobások eredményei egymástól függetlenek, és az egyes dobások fej-oldalra való esésének ugyanannyi a valószínűsége, ezért ez a feltételes valószínűség megegyezik annak a valószínűségével, hogy a k-ik dobás az első fejdobás. A (2) azonosság ezt a tényt fejezi ki.

Megjegyzés. Összegezve a (2) formulát rögzített l-re és minden $k > m$-re valamilyen rögzített m nem negatív egész számra azt kapjuk, hogy minden $l > 0$ és $m > 0$ egész számra $P(\xi > l + m|\xi > l) = P(\xi > m)$. Ezt az azonosságot lehet úgy interpretálni, hogy ha valakinek az élettartama geometriai eloszlású, azaz annak a valószínűsége, hogy a k időpontban fog meghalni $P(\xi = k) = p(1-p)^k$, $k = 1, 2, \dots$, akkor annak a valószínűsége, hogy az illető az l időpont után még legalább m ideig fog élni feltéve, hogy az l életkort megélte ugyanannyi, mint annak a valószínűsége, hogy egy hasonló eloszlású újszülött legalább m ideig fog élni. Ezért ezt az azonosságot szokás (diszkrét) örökifjú tulajdonságnak nevezni. A diszkrét jelző itt arra utal, hogy a geometriai eloszlás csak (nem negatív) egész számokra teljesíti ezt az azonosságot. Később látni fogjuk, hogy van egy olyan eloszlás, az exponenciális eloszlás, amely teljesíti ennek az azonosságnak egy természetes általánosítását minden (pozitív) valós számra. Ezt hívják örökifjú tulajdonságnak, ami az exponenciális eloszlásnak egy nagyon fontos tulajdonsága.

c.) *A hipergeometrikus eloszlás.*

Tekintsünk először egy tipikus példát, amelyben ilyen eloszlások megjelennek.

Adva van egy urnában M piros és $N - M$ fehér golyó. Visszatevés nélkül kihúzunk n, $n \le N$, golyót. Mi annak a valószínűsége, hogy k piros golyót húztunk ki?

Ez a valószínűség $\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$. Valóban, ha megkülönböztetjük az egyes golyókat, akkor $\binom{N}{n}$ különböző húzáseredmény alakulhat ki. (Nem különböztetünk meg két húzáseredményt, ha ugyanazokat a golyókat húztuk ki csak más sorrendben.) Kissé részletesebben kifejtve: Az urnában levő golyókat számozzuk meg 1-től N-ig úgy, hogy a piros golyók kapják az $1, \dots, M$ és a fehér golyók az $M+1, \dots, N$ számokat. Egy n hosszúságú visszatevés nélküli húzássorozat eredménye egy az $1, \dots, N$ számokból álló n hosszú számsorozat, amelynek minden tagja különböző. A különböző húzássorozatok számát számoljuk ki, ha azonosítunk két olyan sorozatot, amelyben ugyanazok a számok szerepelnek csak más sorrendben. Ezután azon n hosszúságú húzássorozatok számát számoljuk össze, amelyekben k piros, azaz az $1, \dots, M$ számok valamelyikével indexezett golyó van.) Olyan húzássorozat, amelyben az M piros golyóból k-t, az $N - M$ fehér golyóból pedig $n - k$-t húzunk $\binom{M}{k}\binom{N-M}{n-k}$ van. Mivel ezek azok az n hosszú húzássorozatok, amelyekben pontosan k piros golyót húzunk, és az egyes húzássorozatok valószínűsége megegyezik, innen következik az állítás.

A hipergeometrikus eloszlás definiciója. *Azt mondjuk, hogy a ξ valószínűségi változó hipergeometrikus eloszlású N, M és n paraméterekkel, ahol N, M és n nem* *negatív egész számok, $M < N$, $n < N$, ha ξ valamely egész k értéket vesz fel, amelyre $0 \le k \le n$, és*

$$P(\xi = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad 0 \le k \le n.$$

Ilyen módon valóban eloszlást definiáltunk. A $\sum\limits_{k=0}^{n} P(\xi = k) = 1$ azonossággal ekvivalens $\sum\limits_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k} = \binom{N}{n}$ azonosságot (más jelöléssel) beláttuk a binomiális eloszlás vizsgálatában.

A következő feladat egy tipikus probléma, ahol a hipergeometrikus eloszlás megjelenik.

Számoljuk ki annak valószínűségét, hogy a lottóban pontosan három találatunk lesz.

E feladat megoldását tartalmazza a 2. téma ismertetésének 6. feladata. Ezért ezt nem kell újra tárgyalnom. Oldjuk meg a következő feladatokat.

Feladatok:

9.) Számoljuk ki egy N, M és n paraméterű hipergeometrikus eloszlású ξ valószínűségi változó várható értékét és szórásnégyzetét, azaz egy olyan valószínűségi változó várható értékét és szórásnégyzetét, amelyre

$$P(\xi = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad 0 \le k \le n.$$

Mutassuk meg, hogy $E\xi = n\frac{M}{N}$, $\operatorname{Var}\xi = \frac{nM}{N}\left(1 - \frac{M}{N}\right)\frac{n-n}{N-1}$.

Megoldás: Tekintsünk egy urnát, amely M piros és $N - M$ fehér golyót tartalmaz, és húzzunk ki n golyót visszatevés nélkül. Definiáljuk a következő ξ_j, $1 \le l \le n$, valószínűségi változókat. Legyen $\xi_j = 1$, ha a j-ik húzás piros, $\xi_j = 0$, ha a j-ik húzás fehér. Ekkor a megoldandó feladat ekvivalens az $S = \sum\limits_{j=1}^{n} \xi_j$ összeg várható érékének és szórásnégyzetének a kiszámolásával. Viszont már korábbi eredményekből következik, hogy $E\xi_j = \frac{M}{N}$, $\operatorname{Var}\xi_j = \frac{M}{N} - \left(\frac{M}{N}\right)^2$ minden $1 \le j \le N$ indexre, továbbá, $\operatorname{Cov}(\xi_j, \xi_k) = \frac{M}{N}\frac{M-1}{N-1} - \left(\frac{M}{N}\right)^2$ minden $1 \le j, k \le n$, $j \ne k$ számpárra. A várható érték additivitásából következik, hogy $ES = \frac{nM}{N}$. (Itt nincs szükség az összeadandók függetlenségének a feltételezésére.) Az összeg szórásnégyzetének kiszámítására tanult általános képletből, (amikor az összeadandók nem feltétlenül függetlenek) és a fenti formulákból következik, hogy

$$\begin{aligned}\operatorname{Var}S &= n\operatorname{Var}\xi_1 + n(n-1)\operatorname{Cov}(\xi_1, \xi_2)\\ &= n\left(\frac{M}{N} - \left(\frac{M}{N}\right)^2\right) + n(n-1)\left(\frac{M}{N}\frac{M-1}{N-1} - \left(\frac{M}{N}\right)^2\right)\end{aligned}$$

$$= n\frac{M}{N}\left(1-\frac{M}{N}-(n-1)\frac{N-M}{(N-1)N}\right)$$
$$= n\frac{M}{N}\left(1-\frac{M}{N}\right)\left(1-\frac{n-1}{N-1}\right) = n\frac{M}{N}\left(1-\frac{M}{N}\right)\frac{N-n}{N-1}.$$

Ha egy urnahúzásban N és $N-M$ nagy, azaz sok fehér és piros golyó van az urnában, és fix számú golyót kihúzunk, akkor a húzáseredmény szempontjából alig van jelentősége annak, hogy a kihúzott golyókat visszadobjuk-e vagy sem. Ilyen jellegű állítást fogalmaz meg a következő (egyszerű) feladat.

10.) Ha $N\to\infty$, $\lim\limits_{N\to\infty}\frac{M}{N}=p$, $0<p<1$, n rögzített egész szám, akkor

$$\lim_{N\to\infty}\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k}p^k(1-p)^{n-k} \quad \text{minden } 0\le k\le n \text{ számra.}$$

Megoldás:

$$\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$
$$= \frac{n!}{k!(n-k)!}\frac{M(M-1)\cdots(M-k+1)}{N(N-1)\cdots(N-k+1)}\frac{(N-M)\cdots(N-M-n+k+1)}{(N-k)\cdots(N-n+1)}$$
$$\to \binom{n}{k}p^k(1-p)^{n-k} \text{ ha } N\to\infty,$$

mert $\frac{M(M-1)\cdots(M-k+1)}{N(N-1)\cdots(N-k+1)}\to p^k$, és $\frac{(N-M)\cdots(N-M-n+k+1)}{(N-k)\cdots(N-n+1)}\to(1-p)^{n-k}$, ha $N\to\infty$.

A hipergeometrikus eloszlás természetes többváltozós általánosítása a polihipergeometrikus eloszlás. Ennek megértése érdekében tekintsük a következő feladatot: Egy urnában r különböző színű golyó van, N_1 1-es, N_2 2-es, … N_r r-es színű golyó. Legyen $N=\sum\limits_{j=1}^{r}N_j$. Ezekből a golyókból kihúzunk n-et visszatevés nélkül. Mi annak a valószínűsége, hogy k_1 1-es, k_2 2-es, … k_r r-es színű golyót húzunk ki?

A válasz:

$$\frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N}{n}},$$

ha $n=\sum\limits_{j=1}^{r}k_r$, egyébként pedig nulla. Ennek az állításnak a bizonyítása hasonlóan történhet, mint a hipergeometrikus eloszlás bevezetése előtt tekintett feladaté. Ennek a feladatnak az alapján vezették be a következő fogalmat.

A polihipergeometrikus eloszlás definiciója. *Azt mondjuk, hogy a* $(\xi_1,\dots,\xi_r)$ *véletlen vektor polihipergeometrikus eloszlású* $N_1,\dots N_r$ *és* n *paraméterekkel, ahol* N_j,

$1 \le j \le r$, és n nem negatív egész számok, és $N = \sum\limits_{j=1}^{r} N_j$ jelöléssel $n < N$, ha olyan $(k_1, \dots, k_r)$ egész számokból álló vektorok az értékei, amelyekre $0 \le k_j \le N_j$ minden $1 \le j \le r$ indexre, és

$$P(\xi_1 = k_1, \dots, \xi_r = k_r) = \frac{\binom{N_1}{k_1}\binom{N_2}{k_2}\cdots\binom{N_r}{k_r}}{\binom{N}{n}}, \quad \textit{ha } n = \sum_{j=1}^{r} k_j,$$

és $P(\xi_1 = k_1, \dots, \xi_r = k_r) = 0$, ha $n \ne \sum\limits_{j=1}^{r} k_j$.

Mivel a polihipergeometrikus eloszlás számunkra kevésbé fontos, ezért egy ilyen eloszlású véletlen vektor várható értékének és kovarianciáinak kiszámolását csak a (nem kötelező tananyagot tartalmazó) 2. kiegészítésben tárgyalom.

d.) *A Poisson eloszlás.*

A Poisson eloszlást közvetlenül fogom definiálni. Azok a tulajdonságai, amelyek miatt fontos szerepet játszik a valószínűségszámításban ezen eloszlás tárgyalása során fognak kiderülni.

A Poisson eloszlás definiciója. *Egy ξ valószínűségi változó λ paraméterű Poisson eloszlású, ha ξ nem negatív egész értékeket vesz fel, és*

$$P(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \dots$$

Ez valóban valószínűség eloszlás, mert

$$\sum_{k=0}^{\infty} P(\xi = k) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Számítsuk ki egy λ paraméterű Poisson eloszlású valószínűségi változó generátorfüggvényét. Ez

$$\sum_{k=0}^{\infty} P(\xi = k) x^k = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} x^k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda x)^k}{k!} = e^{-\lambda} e^{\lambda x} = e^{\lambda(x-1)}.$$

A következő állítást fontossága miatt fogalmazzuk meg Tétel formájában.

Tétel független Poisson eloszlású valószínűségi változók összegének eloszlásáról. *Ha ξ és η két független, λ illetve μ paraméterű Poisson eloszlású valószínűségi változó, akkor $\xi + \eta$ $\lambda + \mu$ paraméterű Poisson eloszlású valószínűségi változó.*

Bizonyítás:

$$\begin{aligned}P(\xi+\eta=k)&=\sum_{j=0}^{k}P(\xi=j,\eta=k-j)=\sum_{j=0}^{k}P(\xi=j)P(\eta=k-j)\\&=\sum_{j=0}^{k}\frac{\lambda^j}{j!}e^{-\lambda}\frac{\mu^{k-j}}{(k-j)!}e^{-\mu}=\frac{e^{-(\lambda+\mu)}}{k!}\sum_{j=0}^{k}\frac{k!}{j!(k-j)!}\lambda^j\mu^{(k-j)}\\&=\frac{e^{-(\lambda+\mu)}}{k!}(\lambda+\mu)^k.\end{aligned}$$

Feladat:

11.) Mutassuk meg, hogy a Poisson eloszlás generátorfüggvényének alakjából látható, hogy ha ξ és η két független, λ illetve μ paraméterű Poisson eloszlású valószínűségi változó, akkor $\xi+\eta$ $\lambda+\mu$ paraméterű Poisson eloszlású valószínűségi változó.

Megoldás: Mivel egy λ paraméterű Poisson eloszlású valószínűségi változó generátorfüggvénye $e^{\lambda(x-1)}$, egy μ paraméterű Poisson eloszlású valószínűségi változó generátorfüggvénye $e^{\mu(x-1)}$, és egy $\lambda+\mu$ paraméterű Poisson eloszlású valószínűségi változó generátorfüggvénye $e^{(\lambda+\mu)(x-1)}$ a feladat állítása következik az

$$e^{\lambda(x-1)}e^{\mu(x-1)}=e^{(\lambda+\mu)(x-1)}$$

azonosságból.

A fenti eredmény következménye, hogy amennyiben véges sok független, Poisson eloszlású valószínűségi valószínűségi változónak vesszük az összegét, az ismét Poisson eloszlású lesz, amelynek paramétere az egyes valószínűségi változók paramétereinek az összege. A következő feladat állítása, amelynek érdekes következményei vannak, tekinthető úgy, mint ennek az állításnak a megfordítása. Abban ugyanis egy alkalmas konstrukció segítségével egy Poisson eloszlású valószínűségi változót bontunk fel független, kisebb paraméterű Poisson eloszlású valószínűségi változók összegére.

Feladat:

12.) Legyen adva k darab urna, és ezekbe dobjunk be véletlen ξ számú golyót, ahol ξ Poisson eloszlású valószínűségi változó $\lambda>0$ paraméterrel. Legyenek az egyes dobások eredményei egymástól és a ξ valószínűségi változótól függetlenek. Tegyük fel továbbá, hogy minden egyes dobásnál a golyó $p_j\ge 0$ valószínűséggel esik a j-ik urnába, $j=1,\dots,k$, $\sum\limits_{j=1}^{k}p_j=1$. Jelölje η_j a j-ik urnába eső golyók számát. Ekkor az η_j, $j=1,\dots,k$, valószínűségi változók függetlenek, és η_j Poisson eloszlású λp_j paraméterrel, $j=1,\dots,k$.

Megoldás:

$$\begin{aligned}P(\eta_1 = l_1, \dots, \eta_k = l_k) &= P(\xi = l_1 + \cdots + l_k)\frac{(l_1 + \cdots + l_k)!}{l_1! \cdots l_k!} p_1^{l_1} \cdots p_k^{l_k} \\ &= \frac{\lambda^{(l_1+\cdots+l_k)}}{l_1! \cdots l_k!} p_1^{l_1} \cdots p_k^{l_k} e^{-\lambda} \\ &= \frac{\lambda^{l_1} \cdots \lambda^{l_k}}{l_1! \cdots l_k!} p_1^{l_1} \cdots p_k^{l_k} e^{-\lambda(p_1+\cdots+p_k)} = \prod_{j=1}^{k} \frac{(\lambda p_j)^{l_j}}{l_j!} e^{-\lambda p_j}\end{aligned}$$

tetszőleges $l_1 \ge 0, \dots, l_k \ge 0$ egész számokra. Innen adódik az állítás.

Lássuk be az előző feladat segítségével a következő állítást:

13.) Legyen adva egy ξ Poisson eloszlású valószínűségi változó λ paraméterrel. Dobjunk le egymástól és a ξ valószínűségi változótól függetlenül véletlenül egyenletesen ξ darab pontot az egységintervallumra, azaz tegyük fel, hogy minden pont $b - a$ valószínűséggel esik valamely $[a, b] \subset [0, 1]$ intervallumba. Ekkor a $[0, 1]$ intervallum tetszőleges felbontására $[s_0, s_1]$, $[s_1, s_2]$, ... , $[s_{k-1}, s_k]$ diszjunkt intervallumokra, $0 = s_0 < s_1 < \cdots < s_k = 1$ igaz az, hogy az egyes intervallumokba eső pontok száma egymástól független Poisson eloszlású valószínűségi változó $s_j - s_{j-1}$, $1 \le j \le k$, paraméterrel.

Megoldás: Tekintsük a következő urnamodellt. Veszünk ξ számú golyót, tehát annyit, ahány ledobott pontot vettünk az előző feladatban. Tegyük az l-ik golyót a j-ik urnába, ha a l-ik pont az $[s_{j-1}, s_j]$ intervallumba esett, $1 \le j \le k$. Akkor az előző feladat eredménye alapján az egyes urnákba eső golyók száma egymástól függetetlen Poisson eloszlású valószínűségi változó $s_j - s_{j-1}$, $1 \le j \le k$, paraméterrel. Innen következik a feladat állítása.

Megjegyzés. Az előző feladat eredménye egyszerű módszert ad úgynevezett Poisson folyamatok konstruálására. Ez azonban nem témája ennek az előadássorozatnak.

Számítsuk ki egy ξ λ paraméterű Poisson eloszlású valószínűségi változó várható értékét és szórásnégyzetét.

$$E\xi = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda+\lambda} = \lambda,$$

$$\begin{aligned}E\xi^2 = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} &= \sum_{k=2}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} + \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= e^{-\lambda}\lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda = \lambda^2 + \lambda,\end{aligned}$$

és $\operatorname{Var} \xi = E\xi^2 - (E\xi)^2 = (\lambda + \lambda^2) - \lambda^2 = \lambda$.

Megjegyzem, hogy a Poisson eloszlás várható értékére és szórásnégyzetére kapott eredmények összhangban vannak azzal a ténnyel, hogy független Poisson eloszlású valószínűségi változók összege olyan Poisson eloszlású valószínűségi változó, amelynek paramétere az összeadandók paramétereinek az összege. Ugyanis mind a várható érték mind a szórásnégyzet additív független valószínűségi változók összegzése esetén.

Feladat:

14.) Legyenek ξ_j, $j = 1, \dots, r$, független Poisson eloszlású valószínűségi változók λ_j, $1 \le j \le r$, paraméterekkel. Mutassuk meg, hogy

$$P\left(\xi_1 = k_1, \dots, \xi_r = k_r \,\middle|\, \sum_{j=1}^{r} \xi_j = n\right) = \frac{n!}{k_1! \cdots k_r!} \prod_{j=1}^{r} \frac{\lambda_j^{k_j}}{\left(\sum\limits_{s=1}^{r} \lambda_s\right)^{k_j}},$$

ha $\sum\limits_{j=1}^{r} k_j = n$. Azaz a $\xi_1, \dots, \xi_r$ vektor feltételes eloszlása feltéve, hogy $\sum\limits_{j=1}^{r} k_j = n$ a polinomiális eloszlás n és $p_j = \frac{\lambda_j}{\sum\limits_{s=1}^{r} \lambda_s}$, $1 \le j \le r$, paraméterekkel.

Megoldás:

$$\begin{aligned} P\left(\xi_1 = k_1, \dots, \xi_r = k_r \,\middle|\, \sum_{j=1}^{r} \xi_j = n\right) &= \frac{P(\xi_1 = k_1, \dots, \xi_r = k_r)}{P\left(\sum_{j=1}^{r} \xi_j = n\right)} \\ &= \frac{P(\xi_1 = k_1) \cdots P(\xi_r = k_r)}{P\left(\sum_{j=1}^{r} \xi_j = n\right)}, \end{aligned}$$

mert $\{\omega\colon\ \xi_1(\omega) = k_1, \dots, \xi_r(\omega) = k_r)\} \subset \{\omega\colon\ \sum_{j=1}^{r} \xi_j(\omega) = n\}$, és a $\xi_1, \dots, \xi_r$ valószínűségi változók függetlenek. Innen

$$\begin{aligned} P\left(\xi_1 = k_1, \dots, \xi_r = k_r \,\middle|\, \sum_{j=1}^{r} \xi_j = n\right) &= \frac{\prod\limits_{j=1}^{r} \frac{\lambda_j^{k_j}}{k_j!} e^{-\lambda_j}}{\frac{\left(\sum\limits_{s=1}^{r} \lambda_s\right)^{k_1 + \cdots + k_r} e^{-(\lambda_1 + \cdots + \lambda_r}}{n!}} \\ &= \frac{n!}{k_1! \cdots k_r!} \prod_{j=1}^{r} \frac{\lambda_j^{k_j}}{\left(\sum\limits_{s=1}^{r} \lambda_s\right)^{k_j}}. \end{aligned}$$

A következő eredmény a binomiális eloszlás Poisson közelítéséről szól.

Lemma a binomiális eloszlás Poisson közelítéséről. *Minden $n = 1, 2, \dots$ számra tekintsünk egy S_n binomiális eloszlású valószínűségi változót $B(n, p_n)$ eloszlással, azaz*

n és p_n paraméterekkel. Tegyük fel továbbá, hogy a $\lambda_n = np_n$, $n = 1, 2, \dots$, számok teljesítik a $\lim\limits_{n\to\infty} \lambda_n = \lambda$ feltételt valamilyen $\lambda > 0$ számmal. Ekkor

$$\lim_{n\to\infty} P(S_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

minden $k = 0, 1, 2, \dots$ számra.

Bizonyítás:

$$P(S_n = k) = \binom{n}{k} \left(1 - \frac{\lambda_n}{n}\right)^{n-k} \left(\frac{\lambda_n}{n}\right)^k =$$
$$\frac{\lambda_n^k}{k!} \left(1 - \frac{\lambda_n}{n}\right)^n \left(1 - \frac{\lambda_n}{n}\right)^{-k} \frac{n(n-1)\cdots(n-k+1)}{n^k} \to \frac{\lambda^k}{k!} e^{-\lambda},$$

ha $n \to \infty$ minden $k = 0, 1, 2, \dots$ számra, mert

$$\frac{\lambda_n^k}{k!} \to \frac{\lambda^k}{k!}, \quad \left(1 - \frac{\lambda_n}{n}\right)^n \to e^{-\lambda}, \quad \left(1 - \frac{\lambda_n}{n}\right)^{-k} \to 1,$$

és $\frac{n(n-1)\cdots(n-k+1)}{n^k} \to 1$, ha $n \to \infty$.

Az előző lemma eredménye azt mondja, hogy ha tekintünk n függeten, egyforma eloszlású $\xi_j^{(n)}$, $1 \le j \le n$, valószínűségi változót, amelyekre $P(\xi_j^{(n)} = 0) = 1 - P(\xi_j^{(n)} = 1) = \lambda_n$, majd vesszük ezek $S_n = \sum\limits_{j=1}^n \xi_j^{(n)}$ összegét, akkor $n \to \infty$ és $n\lambda_n \to \lambda$ esetén az S_n összegek eloszlása tart a λ paraméterű Poisson eloszláshoz.

Korábbi eredményeinkből az is következik, hogy olyan $T_n = \sum\limits_{j=1}^n \eta_j^{(n)}$ alakú $n = 1, 2, \dots$, véletlen összegekre, amelyekben rögzített n számra az $\eta_j^{(n)}$ valószínűségi változók függetlenek és Poisson eloszlásúak $\frac{\lambda_n}{n}$ paraméterrel, és $\lim\limits_{n\to\infty} \lambda_n = \lambda > 0$ szintén teljesül a $\lim\limits_{n\to\infty} P(T_n = k) = \frac{\lambda^k}{k!} e^{-\lambda}$ reláció minden k egész számra. Ekkor ugyanis T_n Poisson eloszlású λ_n paraméterrel.

Az előbb tekintett két példa speciális esete egy általánosabb határeloszlástételnek Poisson határeloszlással. Ismertetem ezt az eredményt, majd megtárgyalom annak következményeit. Megmutatom, hogy ezen eredmény alapján miért természetes feltenni, hogy bizonyos jelenségekben megjelenő véletlen mennyiségek (pl. csillaghulláskor a lehulott csillagok száma vagy rádióaktív bomlásban a széthasadt uránatomok száma) Poisson eloszlású.

Határeloszlástétel Poisson határeloszlással. *Legyen adva minden rögzített $n = 1, 2, \dots$ számra $\xi_1^{(n)}, \dots, \xi_n^{(n)}$ független egyforma eloszlású, nem negatív egész értékeket* *felvevő valószínűségi változók olyan sorozata, mely sorozatok teljesítik a következő feltételeket:*

a) $\lim\limits_{n\to\infty} nP\left(\xi_1^{(n)}=1\right)=\lambda>0.$

b) $\lim\limits_{n\to\infty} nP\left(\xi_1^{(n)}\ge 2\right)=0.$

Ekkor az $S_n=\sum\limits_{j=1}^{n}\xi_j^{(n)}$, $n=1,2,\dots$, *véletlen összegekre teljesül a*

$$\lim_{n\to\infty} P(S_n=k)=\frac{\lambda^k}{k!}e^{-\lambda}$$

reláció minden k nem negatív egész számra.

Megjegyzés: Érvényes e tétel állításának megfelelő általánosítása. Alkalmas feltételek mellett független nem negatív egész értékű, de nem feltétlenül azonos eloszlású valószínűségi változók összegének az eloszlása konvergál egy Poisson eloszláshoz. Egy ilyen a fenti tételhez hasonlóan bizonyítható állítást megfogalmazok nem kötelező házi feladat formájában.

Ennek az eredménynek több különböző bizonyítása ismeretes. Az egyik tanulságos bizonyítás a generátorfüggvények módszerén alapul. Itt ezt a bizonyítást fogom ismertetni.

A tétel bizonyításának gondolata: Tekintsük az S_n valószínűségi változók

$$G_n(x)=\sum_{j=0}^{\infty}P(S_n=j)x^j,\quad n=1,2,\dots,$$

generátorfüggvényeit. A generátorfüggvény módszer azt sugallja, hogy lássuk be a $\lim\limits_{n\to\infty} G_n(x)=G(x)$ relációt valamilyen $-A<x<A$, $A>0$, intervallumban, ahol $G(x)=e^{\lambda(x-1)}$, a λ paraméterű Poisson eloszlás generátorfüggvénye. Ugyanis, mivel hatványsorok konvergenciájából következik a hatványsorok deriváltjainak a konvergenciája is, és hatványsorokat szabad tagonként deriválni, e relációból következik a $\lim\limits_{n\to\infty}\left.\frac{d^kG_n(x)}{dx^k}\right|_{x=0}=\left.\frac{dG^k(x)}{dx^k}\right|_{x=0}=\lambda^k e^{-\lambda}$ összefüggés minden $k=0,1,2,\dots$ számra. (Ennek az érvnek a jogosságát kissé részletesebben tárgyaltam az 5. téma ismertetésében. Az indoklás felhasználja az analízis néhány alapvető, de nem triviális eredményét.) Azaz ebből a relációból következik, hogy $k!\lim\limits_{n\to\infty}P(S_n=k)=k!\frac{\lambda^k}{k!}e^{-\lambda}$ minden $k=0,1,\dots$ számra.

A bizonyítandó reláció igazolásának érdekében vegyük észre, hogy a $G_n(x)$ függvényt felírhatjuk (a generátorfüggvények tulajdonságai miatt) a következő alakban: $G_n(x)=g_n^n(x)$ minden $-1<x<1$ számra, ahol $g_n(x)=\sum\limits_{j=0}^{\infty}P(\xi_1^{(n)}=j)x^j$, a $\xi_1^{(n)}$ valószínűségi változó generátorfüggvénye.

A fenti azonosságot felhasználva, majd a bizonyítandó relációban logaritmust véve elég megmutatni azt, hogy $\lim\limits_{n\to\infty} n\log g_n(x)=\lambda(x-1)$. Továbbá, mivel $P\left(\xi_1^{(n)}=0\right)=1-P\left(\xi_1^{(n)}=1\right)-\sum\limits_{j=2}^{\infty}P\left(\xi_1^{(n)}=j\right)$, ezért

$$g_n(x)=1+P\left(\xi_1^{(n)}=1\right)(x-1)+\sum_{j=2}^{\infty}P\left(\xi_1^{(n)}=j\right)\left(x^j-1\right).$$

Ezért az a) és b) feltételek miatt azt várjuk, hogy a $g_n(x)\sim 1+\frac{\lambda}{n}(x-1)$ formula jó közelítés. Mivel $\log(1+u)\sim u$ kis u számokra, ezért természetes azt várni, hogy $n\log g_n(x)\sim\frac{\lambda}{n}(x-1)$ jó közelítés, és $\lim\limits_{n\to\infty}\log g_n(x)=\lambda(x-1)$, ahonnan következik a Tétel állítása. A Tétel bizonyítását befejezzük, ha igazoljuk a fenti közelítések jogosságát.

A Tétel bizonyításának befejezése. Vegyük észre, hogy minden $\varepsilon>0$ számra létezik olyan $n_0=n_0(\varepsilon)$ küszöbindex, amelyre $\left|g_n(x)-\left(1+\frac{\lambda}{n}(x-1)\right)\right|<\frac{\varepsilon}{n}$, ha $n\ge n_0$ és $|x|<1$. Valóban a $g_n(x)=1+P\left(\xi_1^{(n)}=1\right)(x-1)+\sum\limits_{j=2}^{\infty}P\left(\xi_1^{(n)}=j\right)\left(x^j-1\right)$ azonosság teljesül. Ezenkívül a b) relációból következik, hogy

$$\left|\sum_{j=2}^{\infty}P\left(\xi_1^{(n)}=j\right)\left(x^j-1\right)\right|\le 2\sum_{j=2}^{\infty}P\left(\xi_1^{(n)}=j\right)=2P\left(\xi_1^{(n)}\ge 2\right)\le\frac{\varepsilon}{2n},$$

és az a) relációból pedig az, hogy $\left|P\left(\xi_1^{(n)}=1\right)(x-1)-\frac{\lambda}{n}(x-1)\right|<\frac{\varepsilon}{2n}$, ha $n\ge n_0$ és $|x|<1$. Ezért igaz a fenti azonosság.

Továbbá, mivel mint azt például a $\log(1+x)$ függvény Taylor sorfejtéséből lehet látni, $|\log(1+u)-u|<u^2$, ha $|u|<\frac{1}{2}$, ezért a fenti egyenlőtlenségből $u=g_n(x)-1$ választással kapjuk, hogy $|\log g_n(x)-(g_n(x)-1)|<\frac{2\varepsilon}{n}$, és $\left|\log g_n(x)-\frac{\lambda(x-1)}{n}\right|<\frac{3\varepsilon}{n}$, ha $n\ge n_1$, és $|x|\le 1$ alkalmas $n_1=n_1(\varepsilon)$ küszöbindexre. Mivel ez az állítás igaz minden $\varepsilon>0$ számra, ezért $\lim\limits_{n\to\infty} n\log g_n(x)=\lambda(x-1)$, $\lim\limits_{n\to\infty} g_n(x)^n=e^{\lambda(x-1)}$, ha $x<1$. Viszont láttuk, hogy innen következik a Tétel állítása.

Nem kötelező házi feladat.

Legyen

$$\begin{matrix}\xi_{1,1}\dots,\xi_{1,n_1}\\ \vdots\qquad\vdots\\ \xi_{k,1}\dots,\xi_{k,n_k}\\ \vdots\qquad\vdots\end{matrix}$$

szériasorozat, azaz tegyük fel, hogy az egy sorban álló valószínűségi változók függetlenek. Tegyük fel továbbá, hogy e valószínűségi változók teljesítik a következő feltételeket:

1.) A $\xi_{k,j}$ valószínűségi változók nem negatív egész értékeket vesznek fel.

2.) $P(\xi_{k,j} = 1) = \lambda_{k,j}$, $\lim\limits_{k\to\infty} \sum\limits_{j=1}^{n_k} \lambda_{k,j} = \lambda > 0$.

3.) $\sup\limits_{1\le j\le n_k} \lambda_{k,j} \to 0$, ha $k \to \infty$, és $\sum\limits_{j=1}^{n_k} P(\xi_{k,j} \ge 2) \to 0$, ha $k \to \infty$.

Ekkor az $S_k = \sum\limits_{j=1}^{n_k} \xi_{k,j}$ valószínűségi változók eloszlásban konvergálnak a λ paraméterű Poisson eloszláshoz, ha $k \to \infty$, azaz $\lim\limits_{k\to\infty} P(S_k = l) = \frac{\lambda^l}{l!}e^{-\lambda}$ minden $l = 0, 1, 2, \dots$ számra.

A most bizonyított határeloszlástétel szemléletes tartalma: Tekintsük például a csillaghullást. Hány hullócsillagot látunk egy adott időintervallumban, (mondjuk egy óra alatt) egy nyár éjszakai megfigyelésen? Szeretnénk tudni, hogy a lehullott csillagok (véletlen száma) milyen valószínűségi törvényeknek tesz eleget.

Ezt megértendő, osszuk fel az egy óra időintervallumot rövid ΔT hosszúságú időintervallumokra. A lehullott csillagok száma e rövid ΔT időintervallumokban lehullott csillagok számának az összege. Feltehetjük, hogy diszjunkt időintervallumokban lehullott csillagok száma egymástól független, és a különböző rövid intervallumokban lehulló csillagok száma hasonló valószínűségi törvényeket teljesít. Annak a valószínűsége, hogy egy rövid időintervallumban lehull egy csillag nagyon kicsi, és arányos az időintervallum hosszával. Annak a valószínűsége, hogy egy kis időintervallumban kettő vagy még több csillag is lehull, még ehhez képest is elhanyagolhatóan kicsi. Ez azt jelenti, hogy természetes feltenni, hogy teljesülnek az előbb megfogalmazott tétel feltételei. Ezért az alkalmazható, és innen következik, hogy egy adott időintervallumban lehullott csillagok száma Poisson eloszlású. Hasonló érvelés alkalmazható sok más hasonló esetben. Ez magyarázza meg, hogy miért különösen fontos a Poisson eloszlás.

A diszkrét eloszlású valószínűségi változókról szóló ismertetést egy a hipergeometrikus eloszlással kapcsolatos statisztikai problémával zárom az 1. kiegészítésben. Ez a példa azért is érdekes lehet, mert természetes módon megismertet minket a matematikai statisztika néhány olyan fontos fogalmával, mint a becslés vagy a konfidenciaintervallum.

1. kiegészítés. *Egy a hipergeometrikus eloszlással kapcsolatos statisztikai problémáról.*

Tekintsük először a következő problémát:

Feladat:

Egy tóban 3000 hal van. Véletlenül kihalásznak belőle 1000 darabot, és ezekre piros pöttyöt festenek és visszaengedik őket. Ezután ismét kifognak véletlenül 1000 halat. Mi annak a valószínűsége, hogy a kifogott halak között 100 megfestett van?

Megoldás: $\frac{\binom{1000}{100}\binom{2000}{900}}{\binom{3000}{1000}}$, mert ennyi annak a valószínűsége, hogy 1000 hal kiválasztása esetén az 1000 megfestett halból 100-at a 2000 meg nem festett halból pedig 900 halat választunk.

Megjegyzés: A gyakorlatban előforduló kérdés ennek a fordítottja. Elvégezzük a fenti kisérletet, megszámoljuk a második fogásban kifogott megfestett halak számát, és ebből próbálunk a tóban levő halak ismeretlen számára következtetni. Milyen eljárás segítségével tudjuk ezt jól megtenni?

Érdemes ezt a problémát részletesebben megtárgyalni. Nem tudjuk, hogy hány hal van a tóban. De ennek megbecslése érdekében a következő eljárást alkalmazhatjuk:

Végezzünk két fogást, az első fogásban kifogott halakat jelöljuk meg, és számoljuk meg, hogy a második fogásban hány megjelölt és megjelöletlen halat fogtunk. Ennek alapján meg akarjuk állapitani, hogy hány hal lehet összesen a tóban. Ezt természetesen csak bizonyos (véletlentől függő) pontossággal tudjuk meghatározni. Az ilyen tipusú feladatok tipikusak a matematikai statisztikában, az ilyen problémák vizsgálatát nevezik becsléselméletnek. Világos, hogy a feladatban szereplő adatok esetén nem valószínű, hogy 1000-nél alig több hal van a tóban, mert akkor sokkal több megjelölt hal lenne a második fogásban. Az hogy rengeteg, mondjuk 1 000 000 hal lenne a tóban szintén nem túl valószínű, mert akkor sokkal kevesebb megjelölt hal lenne a második fogásban. A matematikai statisztikában kidolgoztak egy általános elvet, egy maximum likelihood módszernek nevezett eljárást, amely nagyon általános feltételek mellett jó módszert ad a minket érdeklő mennyiség becslésére. Megtárgyaljuk, hogy milyen eredményt ad ez a jelen esetben is alkalmazható módszer. Tekintsünk egy kissé általánosabb problémát. Vezessük be a következő jelöléseket:

x jelöli a tóban lévő halak (ismeretlen) számát.

n jelöli az első fogásban kifogott (és megjelölt) halak számát.

r jelöli a második fogásban kifogott halak számát.

k jelöli a második fogásban kifogott, előzöleg megjelölt halak számát.

Annak valószínűsége, hogy adott (ismeretlen) x és n, r számok esetén pontosan k megjelölt halat fogunk ki

$$q_k(x,n,r) = \frac{\binom{n}{k}\binom{x-n}{r-k}}{\binom{x}{r}}.$$

Tekintsük az ismeretlen x szám (maximum likelihood) becslésének azt az x számot, amelyre a $q_k(x,n,r)$ mennyiség (rögzített n, k és r számok mellett) maximális.

Határozzuk meg a fenti feladatban a maximum likelihood becslést.

Némi számolás mutatja, hogy

$$\frac{q_k(x,n,r)}{q_k(x-1,n,r)} = \frac{x-n}{x-n-r+k} \cdot \frac{x-r}{x} = \frac{x^2-rx-nx+rn}{x^2-rx-nx+kx}.$$

Ez a tört kisebb mint egy, ha $rn < kx$, nagyobb mint egy, ha $rn > kx$. Ezért a becslés $rn = kx$, azaz $x = \frac{rn}{k}$, pontosabban az e számot közrefogó egész számok valamelyike. Valóban $x < \frac{rn}{k}$ esetében a $q_k(x,n,r)$ függvény (mint az x változó függvénye rögzített k, n és r paraméterekkel) monoton nő, $x > \frac{rn}{k}$ esetében pedig a $q_k(x,n,r)$ függvény monoton csökken.

Természetes kérdés az, hogy az így kapott becslés valóban jó-e. Azt nem várhatjuk, hogy az adott becslés teljesen pontos. A természetes elvárás az, hogy meg tudjuk adni az x pontnak viszonylag egy kis környezetét, egy olyan $[x-a, x+a]$ intervallumot, amelyre igaz, hogy annak valószínűsége, hogy a halak valódi száma ebbe az intervallumba esik nagyobb, mint egy előírt egyhez közeli szám. A matematikai statisztikában az adott tulajdonsággal rendelkező (véletlen) intervallumot konfidencia (megbízhatósági) intervallumnak hívják.

Ahhoz, hogy ilyen konfidenciaintervallumot tudjunk szerkeszteni szükség van bizonyos valószínűségi változók eloszlásának jobb ismeretére, és ez a valószínűségszámítás egyik alapvető feladata. Jegyezzük meg, hogy a most vizsgált feladatban, ha a tóban x számú hal van, és az első fogásban n, a második fogásban pedig r halat fogunk ki, akkor a második fogásban kifogott véletlen k számú megjelölt hal számának a várható értéke $Ek = \frac{rn}{x}$. Ahhoz, hogy vizsgálni tudjuk mennyire jó a becslés, hogyan lehet jó konfidenciaintervallumot konstruálni, azt kell megértenünk, hogy mekkora az ingadozása a k valószínűségi változónak a várható értéke körül. Ezért érdemes egy hipergeometrikus eloszlás szórásnégyzetét kiszámolni. További értékes információkat nyerhetünk, ha tételeket bizonyítunk hipergeometrikus eloszlások aszimptotikus eloszlására akkor, amikor a benne szereplő paraméterek nagyok. Bár ezzel a kérdéssel nem fogunk foglalkozni, hasonló problémákat fogunk tárgyalni, amelyek vizsgálatában ilyen jellegű kérdések megoldása hasznos.

2. kiegészítés. *A polihipergeometrikus eloszlás várható értéke és szórásnégyzete.*

Először a következő lemmát látom be.

Lemma 1. *Legyenek* $\xi_1, \dots, \xi_L$ *és* $\eta_1, \dots, \eta_M$ *valószínűségi változók ugyanazon a valószínűségi mezőn. Ekkor*

$$\mathrm{Cov}\left(\sum_{j=1}^{L}\xi_j, \sum_{k=1}^{M}\eta_k\right) = \sum_{j=1}^{L}\sum_{k=1}^{M}\mathrm{Cov}\,(\xi_j, \eta_k).$$

Bizonyítás.

$$\begin{aligned}\mathrm{Cov}\left(\sum_{j=1}^{L}\xi_j, \sum_{k=1}^{M}\eta_k\right) &= E\left(\sum_{j=1}^{L}\xi_j \cdot \sum_{k=1}^{M}\eta_k\right) - \left(E\sum_{j=1}^{L}\xi_j\right)\left(E\sum_{k=1}^{M}\eta_k\right)\\ &= \sum_{j=1}^{L}\sum_{k=1}^{M}E\xi_j\eta_k - \sum_{j=1}^{L}\sum_{k=1}^{M}E\xi_jE\eta_k = \sum_{j=1}^{L}\sum_{k=1}^{M}\mathrm{Cov}\,(\xi_j, \eta_k).\end{aligned}$$

Ennek alapján

Lemma 2. *Legyen* $(\xi_1, \dots, \xi_r)$ *polihipergeometrikus eloszlású véletlen vektor* $N_1, \dots N_r$ *és* n *paraméterekkel. Ekkor* $E\xi_j = n\frac{N_j}{N}$, $\mathrm{Var}\,\xi_j = n\frac{N_j}{N}\left(1 - \frac{N_j}{N}\right)\frac{N-n}{N-1}$ *minden* $1 \le j \le r$ *indexre, és* $\mathrm{Cov}\,(\xi_j, \xi_k) = n\frac{(n-N)N_jN_k}{(N-1)N^2}$ *minden* $1 \le j, k \le r$, $j \ne k$ *indexpárra.*

Bizonyítás. Számítsuk ki először a $\mathrm{Cov}\,(\xi_j, \xi_k)$ kovarianciafüggvényt. Ennek érdekében tekintsünk egy urnát, benne N_1 1-es, N_2 2-es, ... , N_r r-es színű golyót, és húzzunk ki belőlük n golyót visszatevés nélkül. Legyen $N = \sum\limits_{k=1}^{r} N_r$, és vezessük be a következő $\eta_{l,s}$, $1 \le l \le n$, $1 \le s \le r$, valószínűségi változókat: $\eta_{l,s} = 1$, ha az l-ik húzásban s-es színű golyót húzunk, és $\eta_{l,s} = 0$, ha az l-ik húzás eredménye más. Definiáljuk a $\xi_s = \sum\limits_{l=1}^{n}\eta_{l,s}$ véletlen összegeket, $1 \le s \le r$. Ekkor a $\mathrm{Cov}\,(\xi_j, \xi_k)$ kovarianciafüggvényt kell kiszámolnunk az előbb definiált valószínűségi változókkal.

Számoljuk ki először a $\mathrm{Cov}\,(\eta_{l,j}, \eta_{l',k})$ kovarianciafüggvényeket. Külön kell választani az $l = l'$ és $l \ne l'$ eseteket. Mind a két esetben használhatjuk azt a tényt, hogy hasonlóan a két színű golyót tartalmazó urnamodellhez annak, hogy milyen valószínűséggel húzok bizonyos előírt színű golyókat adott húzásban nem függ attól, hogy hanyadik húzást tekintettük.

Az $l = l'$ esetben

$$\begin{aligned}\mathrm{Cov}\,(\eta_{l,j}, \eta_{l,k}) &= P\,(\eta_{l,j} = 1, \eta_{l,k} = 1) - P\,(\eta_{l,j} = 1)\,P\,(\eta_{l,k} = 1)\\ &= -P\,(\eta_{1,j} = 1)\,P\,(\eta_{1,k} = 1) = -\frac{N_jN_k}{N^2}.\end{aligned}$$

Ha $l \neq l'$, akkor

$$\begin{aligned}
\operatorname{Cov}\left(\eta_{l,j}, \eta_{l',k}\right) &= P\left(\eta_{l,j}=1, \eta_{l',k}=1\right) - P\left(\eta_{l,j}=1\right) P\left(\eta_{l',k}=1\right) \\
&= P\left(\eta_{1,j}=1, \eta_{2,k}=1\right) - P\left(\eta_{1,j}=1\right) P\left(\eta_{1,k}=1\right) \\
&= \frac{N_j N_k}{N(N-1)} - \frac{N_j N_k}{N^2}.
\end{aligned}$$

Innen az 1. lemma eredménye alapján

$$\operatorname{Cov}\left(\xi_j, \xi_k\right) = -n\frac{N_j N_k}{N^2} + n(n-1)\left(\frac{N_j N_k}{N(N-1)} - \frac{N_j N_k}{N^2}\right) = n\frac{(n-N)N_j N_k}{(N-1)N^2}.$$

Az $E\xi_j$ és $\operatorname{Var}\xi_j$ mennyiségeket is ki lehet számítani hasonlóan, de erre nincs szükség. Ha csak azt vesszük figyelembe, hogy az l-ik húzásban j-es vagy más színű golyót húztunk-e, azaz csak az $\eta_{j,l}$ valószínűségi változókkal dolgozunk, akkor rövid meggondolás után láthatjuk, hogy ugyanazt az eredményt kapjuk, mint a (két színnel rendelkező) hipergeometrikus eloszlást modellező urnamodell esetén $M = N_j$ és $N - M = N - N_j$ paraméterekkel. Ezért $E\xi_j = n\frac{N_j}{N}$, $\operatorname{Var}\xi_j = n\frac{N_j}{N}\left(1 - \frac{N_j}{N}\right)\frac{N-n}{N-1}$.