Eötvös Loránd Tudományegyetem

Természettudományi Kar

Sulyán Tímea

Poisson eloszlás és alkalmazásai

BSc Elemző Matematikus Szakdolgozat

Témavezető:

Csiszár Villő

Valószínűségelméleti és Statisztika Tanszék

Budapest, 2016

Köszönetnyilvánítás

Ezúton szeretném megköszönni témavezetőmnek, Csiszár Villőnek a hasznos tanácsokat, észrevételeket, mellyel segítette a dolgozatom elkészítését. Hálás vagyok a konzultációkért valamint a türelméért és megértéséért. Köszönettel tartozom a családomnak és barátaimnak, hogy megteremtették a légkört az íráshoz valamint támogattak a szakdolgozat létrejöttében.

Tartalomjegyzék

Bevezetés

Siméon Denis Poisson ($1781-1840$) francia matematikus, fizikus, statisztikus. Élete során rengeteg ma is használatos eredményt ért el az alkalmazott matematika és a matematikai fizika területén is. Élete során körülbelül 300-400 matematikai értekezést publikált. Meglepő lehet ekkora mennyiség ismeretében, hogy egyszerre csak egy dologgal foglalkozott. A Poisson-eloszlás, 1837-ben tűnt fel először az írásaiban.

A valószínűségszámításban egyik leggyakrabban alkalmazott eloszlás a Poisson-eloszlás, melyet véletlenszerű időpontokban bekövetkező események számának meghatározására használunk egy adott időintervallumban. Ilyen esemény például az adatbázis szerverekhez beérkező lekérdezések, üzletbe érkező vevők, sajtóhibák mennyisége egy könyvben, forgalmas kereszteződésben történő balesetek, valamint a természeti katasztrófák száma is.

Szakdolgozatomban ennek az eloszlásnak a sokszínűségét vizsgálom. Az 1. fejezetben definiálom a Poisson eloszlás különböző változatait és a velük kapcsolatos mértékeket és példákat. A 2. fejezetben módszereket láthatunk arra, hogy hogyan használjuk más eloszlás közelítésére. A 3. fejezet a Poisson-eloszlás egy még érdekesebb felhasználását mutatja majd be a Poisson-folyamatokon keresztül, ahol az idő függvényében vizsgáljuk a modelljeinket.

1. fejezet

Poisson-eloszlás fajtái

1.1. Az egyszerű Poisson-eloszlás

Azt a különleges diszkrét eloszlást fogom bemutatni, melyben a változó megszámlálhatóan végtelen sok értéket vehet fel.

1.1. Definíció. (Poisson-eloszlás) *A ξ valószínűségi változót Poisson-eloszlásúnak nevezzük, ha lehetséges értékei: 0,1,... és*

$$p(k,\lambda) = P(\xi = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda}, \tag{1.1}$$

ahol $\lambda > 0$ rögzített és $k \in \mathbb{N}$.

Az 1.1-beli mennyiségeket k=0,1,... esetére összegezve és e^λ Taylor-sorát alkalmazva:

$$\sum_{k=0}^{\infty} p(k,\lambda) = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \cdot e^{-\lambda} = e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} \cdot e^{\lambda} = 1.$$

Tehát (1.1)-ben a pozitív számok eloszlást alkotnak, ezért elképzelhető egy olyan kísérlet, melynek során annak a valószínűsége, hogy pontosan k-szor következik be egy megfigyelt esemény, éppen $p(k,\lambda)$.

1.2. Tétel. *Ha a ξ valószínűségi változó Poisson-eloszlású, akkor várható értéke*

$$E(\xi) = \lambda. \tag{1.2}$$

Bizonyítás.

$$E(\xi) = \sum_{k=0}^{\infty} x_k \cdot p(k,\lambda) = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} =$$

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

Tehát valóban λ a várható értéke. □

1.3. Tétel. *Ha a ξ valószínűségi változó Poisson-eloszlású, akkor szórása*

$$D^2(\xi) = \lambda. \tag{1.3}$$

Bizonyítás.

$$E(\xi^2) = \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} e^{-\lambda} =$$

$$= \sum_{k=1}^{\infty} [(k-1)+1] \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \sum_{k=2}^{\infty} (k-1) \frac{\lambda^k}{(k-1)!} e^{-\lambda} + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} =$$

$$= \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda^2 + \lambda.$$

A szórásnégyzet:

$$D^2(\xi) = E(\xi^2) - E^2(\xi) = (\lambda^2 + \lambda) - \lambda^2 = \lambda,$$

amit kapni is akartunk eredményül. □

1.4. Definíció. (Generátorfüggvény) *Legyen $(a_0, a_1, \ldots)$ nemnegatív valós számok olyan sorozata, hogy $\sum_{k=0}^{\infty} a_k \leq 1$. Ekkor a sorozat generátorfüggvénye*

$$G(z) = \sum_{k=0}^{\infty} a_k z^k.$$

Néhány tulajdonsága:

1. $G'_\xi(1) = E(\xi)$

2. $G_\xi(z) = E(z^\xi)$

3. $G''_\xi(1) = E(\xi^2) - E(\xi)$

1.5. Definíció. (Valószínűségi változó generátorfüggvénye) *Ha p_k jelöli a $P(\xi = k)$ valószínűséget, akkor a ξ valószínűségi változóhoz tartozó generátorfüggvény*

$$G_\xi(z) = \sum_{k=0}^{\infty} p_k z^k.$$

1.6. Definíció. *A λ paraméterű Poisson-eloszlás generátorfüggvénye:*

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

A generátorfüggvénynek az 1. és 3. tulajdonságát felhasználva, a Poisson-eloszlás várható értéke és szórása könnyebben meghatározható:

$$G'_\xi(z) = (e^{-\lambda} e^{\lambda z})' = \lambda \cdot e^{-\lambda} e^{\lambda z}$$
$$E(\xi) = G'_\xi(1) = \lambda \cdot e^{-\lambda} e^{\lambda \cdot 1} = \lambda.$$

$$D^2(\xi) = G''_\xi(1) + G'_\xi(1) - \left[G'_\xi(1)\right]^2 = \lambda^2 e^{\lambda(1-1)} + \lambda e^{\lambda(1-1)} - \left(\lambda e^{\lambda(1-1)}\right)^2 =$$
$$= \lambda^2 + \lambda - \lambda^2 = \lambda.$$

A Poisson-eloszlás fontos tulajdonsága, hogy ha ξ_1 és ξ_2 független, Poisson-eloszlású valószínűségi változók λ_1 és λ_2 paraméterekkel, akkor a $\xi_1 + \xi_2$ összeg is Poisson-eloszlású $\lambda_1 + \lambda_2$ paraméterrel.

A következőkben megnézünk a valószínűségszámítással foglalkozó könyvek példáihoz hasonló feladatot.

1.7. Feladat. Egy 500 oldalas könyvben várhatóan 200 sajtóhiba található. Mekkora annak a valószínűsége, hogy 10 véletlenszerűen kiválasztott lapon nem lesz sajtóhiba, ha feltételezzük, hogy a sajtóhibák száma Poisson-eloszlású?

Megoldás. Az egy lapra jutó sajtóhibák száma $\frac{200}{500}$, tíz lapra $\frac{2}{5} \cdot 10 = 4 = \lambda$.

$$\mathrm{P}(10 \text{ lapon } 0 \text{ hiba}) = \frac{4^0}{0!} e^{-4} = 0,0183$$

A könyvből $1,83\%$ eséllyel tudunk kiválasztani véletlenszerűen 10 lapot, amely hibátlan.

1.8. Feladat. Az elmúlt évben Nógrád megyében a Központi Statisztikai Hivatal adatai szerint 649 baleset történt. Mennyi a valószínűsége, hogy két napon, hétfőn és pénteken is 3 baleset történik?

Megoldás. A balesetek átlagos száma egy napon $\lambda = \frac{649}{365} = 1,78$.

Független eseményekről van szó, mert a balesetek bekövetkezése nem függ egymástól így az sem, hogy melyik nap hány baleset történik. Éppen ezért

$$\mathrm{P}(\text{hétfőn is és pénteken is 3 baleset}) = \frac{1,78^3}{3!} e^{-1,78} \cdot \frac{1,78^3}{3!} e^{-1,78} = 0,0251$$

Tehát annak a valószínűsége, hogy a két legforgalmasabb nap mindegyikén 3 baleset történjen $2,5\%$.

1.2. Az összetett Poisson-eloszlás

A független valószínűségi változók összegének jelentős szerepe van. Ezen belül különösen annak, hogy az összeg tagjainak száma is valószínűségi változó.

1.9. Definíció. (Összetett Poisson-eloszlás) *Az $S_N = X_1 + \ldots + X_N$ véletlen tagszámú összegben, ha N Poisson-eloszlású, λ várható értékű, és X_j, $j = 1, 2, \ldots$ független, azonos eloszlású, nemnegatív egész értékű valószínűségi változókból áll, melyek függetlenek N-től is, akkor S_N eloszlása*

$$Q_S = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} Q^{*k}, \tag{1.4}$$

ahol $Q = Q_X$-t az összeg tagjaiban szereplő X_j eloszlások konvolúciójának nevezzük, melynek jele, Q^{*k}.

Az összetett Poisson-eloszlás várható értékét és szórásnégyzetét a következőképpen kapjuk. A teljes várható érték tételét felhasználva kapjuk a várható értéket

$$E(S_N) = E(N) \cdot E(X) = \lambda E(X)$$

A szórásnégyzet a teljes szórásnégyzet tételéből

$$D^2(S_N) = E(N)D^2(X) + D^2(N)E(X)^2 = E(N)E(X^2) = \lambda E(X^2).$$

1.10. Tétel. *Jelölje g_X az $X_1, X_2, \ldots$, G_S az S_N, G_N az N valószínűségi változók generátorfüggvényét. Ekkor*

$$G_S = G_N \circ G_X.$$

1.11. Definíció. *Egy $\boldsymbol{Q} = (q_0, q_1, \ldots)$ valószínűségeloszlás korlátlanul osztható, ha minden $n = 1, 2, \ldots$ esetén létezik olyan $\boldsymbol{P}_n = (p_{n,0}, p_{n,1}, \ldots)$ eloszlás, amelynek n-edik konvolúcióhatványa $\boldsymbol{Q}$-val egyenlő: $\boldsymbol{Q} = \boldsymbol{P}_n * \boldsymbol{P}_n * \cdots * \boldsymbol{P}_n$. Azaz, ha $\boldsymbol{Q}$ generátorfüggvényre teljesül, hogy minden n-re $\sqrt[n]{G(z)}$ is valószínűségi generátorfüggvény a $[0, 1]$ intervallumon.*

Ilyen például a Poisson-eloszlás. A λ paraméterű Poisson-eloszlás a λ/n paraméterű Poisson eloszlás n-edik konvolúcióhatványa.

Az 1.10 tétel szerint az összetett Poisson-eloszlás generátorfüggvénye

$$G_S(z) = e^{-\lambda+\lambda\cdot G_X(z)}.$$

Ebből látszik, hogy korlátlanul osztható:

$$\sqrt[n]{G_S(z)} = e^{\frac{\lambda}{n}(G_X(z)-1)}$$

szintén összetett Poisson-eloszlás generátorfüggvénye.

1.12. Tétel. (A generátorfüggvények folytonossági tétele) *Legyen $n = 1, 2, \ldots$ esetén az $(a_{n,0}, a_{n,1}, \ldots)$ számsorozat generátorfüggvénye $G_n(z)$.*

(a) Tegyük fel, hogy minden $k = 0, 1, \ldots$ esetén $\exists \lim_{n\to\infty} a_{n,k} =: a_k$. Ekkor $\lim_{n\to\infty} G_n(z) = G(z)$, $|z| < 1$, ahol $G(z)$ az $(a_0, a_1, \ldots)$ sorozat generátorfüggvénye. Sőt, a konvergencia egyenletes a $(-1, 1)$ intervallum minden zárt részintervallumán.

(b) Tegyük fel, hogy $G_n(z)$ konvergens minden $z \in (0, 1)$ pontban. Ekkor minden $k = 0, 1, \ldots$ esetén $\exists \lim_{n\to\infty} a_{n,k} =: a_k$, és $G_n(z)$ az $(a_0, a_1, \ldots)$ sorozat $G(z)$ generátorfüggvényéhez tart minden $z \in (-1, 1)$ pontban.

1.13. Tétel. *A diszkrét korlátlanul osztható eloszlások éppen az összetett Poisson-eloszlások.*

Bizonyítás. Legyen $H(z)$ egy korlátlanul osztható eloszlás generátorfüggvénye. Az azonosan 0 valószínűségi változó eloszlása korlátlanul osztható és összetett Poisson is, feltehetjük tehát, hogy $H(0) < 1$. Ha $H(0) = 0$ lenne, akkor $\sqrt[n]{H(0)} = 0$, és mivel $\sqrt[n]{H(z)}$ is generátorfüggvény, kiemelhető lenne belőle z, de akkor $H(z)$-ből már z^n, méghozzá bármilyen nagy n-re, ami nyilvánvalóan nem lehetséges. Tehát $0 < H(0) < 1$.

Tudjuk, hogy ha $G(z)$ generátorfüggvény, akkor $\widetilde{G}(z) = \frac{G(z)-G(0)}{1-G(0)}$ is az: azé az eloszlásé, amelyet úgy kapunk, hogy a G-hez tartozó eloszlás 0 indexű tagja helyére 0-t írunk, a többi tagot pedig úgy szorozzuk fel, hogy az összeg ismét 1 legyen. Más szóval, ha $G(z) = G_\xi(z) = E(z^\xi)$, akkor $\widetilde{G}(z) = E(z^\xi \mid \xi > 0)$, a ξ változó $\xi > 0$ feltétel melletti feltételes eloszlásának a generátorfüggvénye. Ezért minden n pozitív egészre

$$\frac{\sqrt[n]{H(z)} - \sqrt[n]{H(0)}}{1 - \sqrt[n]{H(0)}}$$

is generátorfüggvény. Ha $a > 0$, akkor

$$\sqrt[n]{a} = e^{\frac{1}{n}\log a} = 1 + \frac{1}{n}\left(\log a + o(1)\right),$$

amint $n \to \infty$, ezért

$$\frac{\sqrt[n]{H(z)} - \sqrt[n]{H(0)}}{1 - \sqrt[n]{H(0)}} \to \frac{\log H(z) - \log H(0)}{-\log H(0)}, \quad 0 < z < 1.$$

Az 1.12 tétel (b) része szerint a határérték is generátorfüggvény, de azt nem mondja ki, hogy valószínűségeloszlásé: tehát az együtthatók összege lehet, hogy kisebb, mint 1. A limeszfüggvény baloldali határértékét kiszámítva a $z = 1$ helyen ellenőrizhetjük ezt. Jelen esetben H folytonossága és $\log H(1) = 0$ miatt ez 1, vagyis

$$g(z) := \frac{\log H(z) - \log H(0)}{-\log H(0)}$$

is valószínűségeloszlás generátorfüggvénye.
Tehát a $H(z) = e^{\lambda(G(z)-1)}$, ahol $\lambda = -\log H(0) > 0$. $\square$

Az összetett Poisson-eloszlás alkalmazási területei:

A gyakorlatban azon belül az ökológiában széles körben alkalmazzák a következőképpen. Egy adott területen az állatcsaládok számát Poisson-eloszlású valószínűségi változónak tekintik. Az egyes almokban született állatok száma legyen egyforma F eloszlású, és az állatcsaládokat függetlennek feltételezik. Ekkor a területen született állatok száma összetett Poisson-eloszlású lesz.

A kockázati modellekben az egyes biztosítóintézetek működésénél három elemét különböztetjük meg a pénzforgalomnak

- egyes károk esetén a kifizetett összeg, más néven összkár,

- biztosítottak befizetése és a

- kezdeti tőke.

Az összetett kockázati modellek esetén a teljes veszteség mértékét szeretnénk meghatározni. Minden egyedhez (kötvény, biztosítás) több káresemény tartozhat. Az egyedekhez tartozó káresemények száma N, mely λ paraméterű Poisson-eloszlást követ. A teljes kárkifizetés nagyságának eloszlását azonosnak vesszük. Feltesszük, hogy a kárkifizetések nagyságát megadó $X_i, i = 1, 2, \ldots$ sorozat független az összeadandók számát megadó valószínűségi változótól, azaz N-től, valamint a sorozat tagjai egymástól is függetlenek. Ha az X_i változók azonos eloszlásúak, véges várható értékkel, N várható értékéről tudjuk, hogy véges ezért

$$S = \sum_{i=1}^{N} X_i.$$

A következő tétel biztosítja, hogyha a kárigényeket nagyság szerint rendezzük és az egyes csoportokon belül külön tekintve az összkár értékét, ismét összetett Poisson-eloszlásokat kapunk.

1.14. Tétel. (Összetett Poisson-eloszlás szétválasztása) *Tekintsünk olyan $Z_1, Z_2, \ldots$ valószínűségi változókat, melyek függetlenek, azonos Q eloszlásúak, továbbá függetlenek az N λ paraméterű Poisson-eloszlású változótól. Legyenek az $A_1, \ldots A_m \subset \mathbb{R}$ halmazok diszjunktak. Tegyük fel, hogy $Q(A_j) > 0$, $j = 1, \ldots, m$. Ekkor az*

$$N_k = \sum_{j=1}^{N} \chi_{\{Z_j \in A_k\}}, \quad k = 1, \ldots, m$$

valószínűségi változók független, $\lambda Q(A_k)$ paraméterű Poisson-eloszlású változók. Továbbá az

$$S_k = \sum_{j=1}^{N} Z_j \chi_{\{Z_j \in A_k\}}, \quad k = 1, \ldots, m$$

változók egymástól független, összetett Poisson-eloszlásúak.

Ezt még fel fogjuk használni a későbbiekben az összetett Poisson-folyamat során.

1.3. A keverék Poisson-eloszlás

A keverék eloszlásokra azért lehet szükségünk, hogy pontosabb modellt tudjunk készíteni. Ezt egy példán keresztül fogjuk megnézni.

1.15. Definíció. (Keverék eloszlás) *Jelölje* $\xi_1, \ldots, \xi_k$ *a k darab független diszkrét vagy folytonos valószínűségi változót. A k komponensű kevert eloszlás sűrűségfüggvénye*

$$f_\xi(x) = \sum_{j=1}^{k} \delta_j f_j(x).$$

A δ_j *jelöli, hogy mekkora súlyokkal vesszük be az egyes változókat az új eloszlásba és teljesül rájuk, hogy* $0 < \delta_j < 1$, *valamint* $\sum_{j=1}^{k} \delta_j = 1$. *Diszkrét eloszlások keverékéből diszkrét keverék eloszlást, folytonos eloszlások keverékéből pedig folytonos keverék eloszlást kapunk.*

Tekintsük az alábbi táblázatban szereplő adatokat, melyek $1900 - 2006$ között történt 7-es vagy annál nagyobb erősségű földrengések évenkénti számát adják meg.

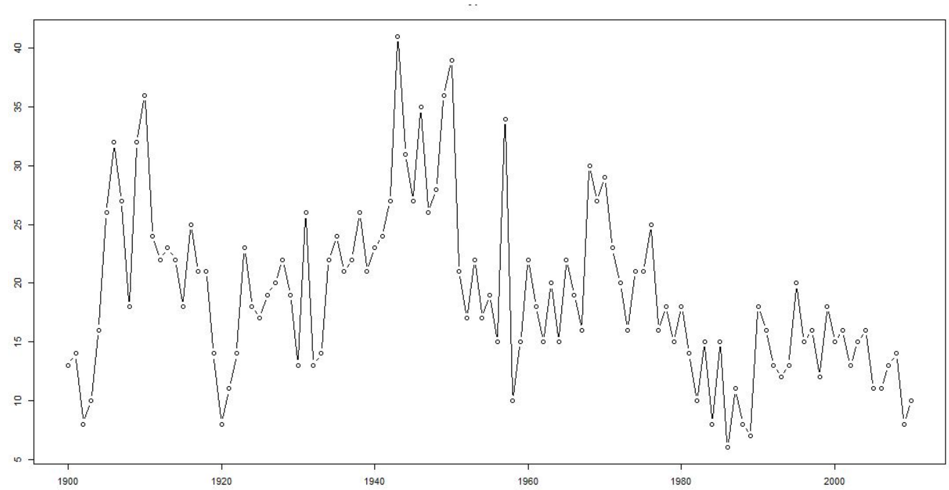
13	14	8	10	16	26	32	27	18	32	36	24	22	23	22	18	25	21
21	14	8	11	14	23	18	17	19	20	22	19	13	26	13	14	22	24
21	22	26	21	23	24	27	41	31	27	35	26	28	36	39	21	17	22
17	19	15	34	10	15	22	18	15	20	15	22	19	16	30	27	29	23
20	16	21	21	25	16	18	15	18	14	10	15	8	15	6	11	8	7
18	16	13	12	13	20	15	16	12	18	15	16	13	15	16	11	11	

Elsőként azt feltételeznénk, hogy az adatok Poisson-eloszlást követnek, melynek valószínűségfüggvénye

$$p(k, \lambda) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Ekkor a minta várható értékének és a szórásnégyzetének nagyjából meg kellene egyeznie, azonban a példában a minta szórásnégyzete $D^2 \approx 52$, a várható értéke pedig $E \approx 19$. Ez elég nagy eltérés, ezért alkalmazzuk a keverék eloszlást.

Az előző táblázatbeli adatokat ábrázolva is láthatjuk.



1.1. ábra. Földrengések száma (1900-2006)

1.16. Definíció. (Poisson-eloszlások keveréke) *Tekintsük a $\xi_1, \dots, \xi_k$ egymástól független, Poisson eloszlású valószínűségi változókat λ_i, $i = 1, \dots, k$ várható értékekkel. Az 1.15 definíció alapján jelölje $\delta_1, \dots, \delta_k$ a súlyokat és $p_1, \dots, p_k$ a valószínűségfüggvényüket. Legyen X eloszlása keverék eloszlás, mivel Poisson-eloszlások keverékéből áll össze, ezért X keverék Poisson-eloszlású lesz, melynek valószínűségfüggvénye*

$$P(X = x) = \sum_{i=1}^{k} P(X = x | C = i) P(C = i) = \sum_{i=1}^{k} \delta_i p_i(x).$$

Az X tehát k darab komponensből tevődik össze, ahol C értéke (1.5)-ben található.

komponens	1	2	...	k
valószínűségfüggvény	$p_1(x)$	$p_2(x)$	...	$p_k(x)$

$$C = \begin{cases} 1, & \delta_1 \text{ valószínűséggel} \\ 2, & \delta_2 \text{ valószínűséggel} \\ \vdots & \\ k, & \delta_k \text{ valószínűséggel.} \end{cases} \tag{1.5}$$

A keverék eloszlás várható értéke a komponens eloszlások várható értékének lineáris kombinációja:

$$E(X) = \sum_{i=1}^{k} P(C = i)E(X|C = i) = \sum_{i=1}^{k} \delta_i E(\xi_i).$$

1.17. Tétel. (Teljes szórásnégyzet tétel) *A tétel általánosan*

$$D^2(X) = \sum_{i=1}^{k} \delta_i D^2(\xi_i) + \sum_{i=1}^{k} \delta_i E(\xi_i)^2 - \left(\sum_{i=1}^{k} \delta_i E(\xi_i)\right)^2 =$$
$$= \sum_{i=1}^{k} \delta_i E(\xi_i^2) - \left(\sum_{i=1}^{k} \delta_i E(\xi_i)\right)^2.$$

A szórásnégyzet a két komponensű esetre még könnyen meghatározható analitikusan is az előző tétel alapján

$$D^2(X) = \delta_1 D^2(\xi_1) + \delta_2 D^2(\xi_2) + \delta_1\delta_2 \left(E(\xi_1) - E(\xi_2)\right)^2.$$

1.3.1. Paraméter becslés

Azokban az esetekben, amikor azt feltételezzük, hogy az adataink keverék Poisson-eloszlásból származnak, becsülnünk kell a λ_i és δ_i paramétereket. A paraméter becslést leggyakrabban maximum likelihood becsléssel végezzük el.

1.18. Definíció. (Maximum-likelihood módszer) *Tegyük fel, hogy* $x_1, x_2, \ldots, x_n$ *az adott minta, amelynek segítségével az ismeretlen a paramétert akarjuk becsülni. Ha az ismeretlen a paramétertől függő eloszlás mintaelemeinek közös sűrűségfüggvénye* $f(x, a)$*, akkor a független mintaelemek együttes sűrűségfüggvénye*

$$f(x_1, a) \cdot \ldots \cdot f(x_n, a) = \prod_{i=1}^{n} f(x_i, a),$$

ahol $x_1, \ldots, x_n$ *számok a mintaelemeknek a kísérlet során mért értékei. Ekkor a maximum likelihood-módszer szerint az a paraméter becslésének az* $x_1, \ldots, x_n$ *mintaelemeknek azt az* $\hat{a} = \hat{a}(x_1, x_2, \ldots, x_n)$ *függvényét nevezzük, melyre a* $\prod_{i=1}^{n} f(x_i, \hat{a})$ *szorzat a lehető legnagyobb értéket veszi fel, feltéve, hogy a maximum létezik és egyértelmű.*

Általában a k komponensű keverék eloszlás likelihood függvénye adott

$$L(\theta_1, \dots, \theta_k, \delta_1, \dots, \delta_k | x_1, \dots, x_n) = \prod_{j=1}^{n} \sum_{i=1}^{k} \delta_i p_i(x_j, \theta_i), \tag{1.6}$$

ahol $\theta_1, \dots, \theta_k$ a komponens eloszlások paraméterei, a $\delta_i, i = 1, \dots, k$ a súlyok, melyek összege 1. Az x_i-k pedig a megfigyelések. Ennek vesszük a logaritmusát, majd ezt követően a maximumát. A maximum likelihood becslést analitikusan végig számolni elég bonyolult, ezért numerikus módszerrel szokás elvégezni, amelyre található egy hasznos **R**-beli, *flexmix* csomag. A példánkra ezt alkalmazva, behelyettesítünk az 1.6 képletbe, majd maximalizáljuk a loglikelihood függvényt egy **R**-beli függvénnyel, az *nlm*-el.

A $\delta_i \geq 0$ és $\lambda_i \geq 0$, $(i = 1, \dots, k)$ paramétereket először át kell paraméterezni, hogy használni tudjuk az optimalizálás során. Az átparaméterezésre azért van szükségünk, mert az nlm függvény a valós számok teljes halmazán keresi a loglikelihood függvény maximum értékét, viszont a δ_i és λ_i értékekre van megszorítás. Legyenek

$$\eta_i = \log \lambda_i, \quad (i = 1, \dots, k)$$
$$\tau_i = \log \left(\frac{\delta_i}{1 - \sum_{j=2}^{k} \delta_j} \right), \quad (i = 2, \dots, k).$$

Így az $\eta_i, \tau_i \in \mathbb{R}$ tetszőlegesek.

Az eredeti paramétereket a következőképpen fejezhetjük ki:

$$\lambda_i = e^{\eta_i}, \quad (i = 1, \dots, k),$$
$$\delta_i = \frac{e^{\tau_i}}{1 + \sum_{j=2}^{k} e^{\tau_i}}, \quad (i = 2, \dots, k) \quad \text{és}$$
$$\delta_1 = 1 - \sum_{j=2}^{k} \delta_i.$$

Ezeket a változókat már tudjuk használni az *nlm* során.

modell	i	δ_i	λ_i	$-\log L$	várható érték	szórásnégyzet
$k=1$	1	1,000	19,364	391,9189	19,364	19,364
$k=2$	1	0,676	15,777	360,3690	19,364	46,182
	2	0,324	26,840			
$k=3$	1	0,278	12,736	356,8489	19,364	51,170
	2	0,593	19,785			
	3	0,130	31,629			
$k=4$	1	0,093	10,584	356,7337	19,364	51,638
	2	0,354	15,528			
	3	0,437	20,969			
	4	0,116	32,079			
megfigyelés					19,364	51,573

1.2. ábra. Poisson eloszlások keveréke a földrengés modellben.

Az 1.2 táblázatban találhatjuk a különböző esetekre, milyen értékeket kapunk. Látható, hogy $k = 3$ és $k = 4$ esetekben a keverék eloszlás jól közelíti a szórásnégyzetet. Azonban a $-\log L$ értéke $k = 4$ esetnél már nem változik jelentősen, valamint van egy változónk 0,09 súllyal, ami szintén nem annyira mérvadó. Így a négy változós eset már nem hoz annyival jobb eredményt, hogy megérje azt választanunk. Ez azt jelenti, hogy három egymástól független, 12,736; 19,785; 31,629 paraméterű Poisson-eloszlású valószínűségi változók keverékét kell használnunk 0,278; 0,593; 0,130 súlyokkal, hogy jól modellezhető legyen a probléma.

2. fejezet

Az eloszlás használata közelítésre

2.1. A binomiális eloszlás közelítése Poisson-eloszlással

A binomiális eloszlást szeretnénk úgy közelíteni, hogy az eltérés a lehető legkisebb legyen, azaz a binomiális és a másik eloszlás hányadosára 1 körüli értéket kapjunk. Erre a Poisson eloszlást szokás használni. Ez azért is célszerű mert a Poisson-eloszlásnak egyetlen paramétere van, míg a binomiális eloszlásnak az n és a p.

2.1. Definíció. (Binomiális eloszlás) *Legyen az n kísérletből álló Bernoulli-kísérletsorozatban p a jó és q a rossz eset valószínűsége, és $b(k;n,p)$ jelölje annak a valószínűségét, hogy a kísérletek során k jó és $n-k$ rossz eset következik be. Ekkor*

$$b(k;n,p) = P(\xi = k) = \binom{n}{k} p^k q^{n-k}, \tag{2.1}$$

$$\textit{ahol } 0 < p < 1, \quad q = 1-p \textit{ és } k = 0, \dots n.$$

Sokszor olyan Bernoulli-kísérletsorozattal foglalkozunk, ahol egymáshoz viszonyítva n nagy, p kicsi, és a

$$\lambda = np$$

szorzat nem túl nagy és nem is túl kicsi. Ezekben az esetekben $b(k;n,p)$ Poissontól származó becslését szokás használni.

2.2. Tétel. *Ha $n \to \infty$ esetén $p \to 0$ úgy, hogy közben $n \cdot p$ szorzat állandó marad, $np = \lambda > 0$, akkor $q = 1 - p$ jelöléssel*

$$\lim_{n\to\infty} \binom{n}{k} p^k q^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Bizonyítás. Ha $k = 0$, akkor

$$b(0; n, p) = (1-p)^n = \left(1 - \frac{\lambda}{n}\right)^n.$$

Mindkét oldal logaritmusát véve és Taylor-sorba fejtve,

$$\ln(b(0; n, p)) = n \cdot \ln\left(1 - \frac{\lambda}{n}\right) = -\lambda + \frac{\lambda^2}{2n} - \dots \quad .$$

Tehát, ha n elég nagy, akkor

$$b(0; n, p) \approx e^{-\lambda}.$$

Továbbá, hogy ha n elég nagy, akkor bármely rögzített k-ra

$$\frac{b(k; n, p)}{b(k-1; n, p)} = \frac{\lambda - (k-1)p}{kq} \approx \frac{\lambda}{k}.$$

Ezek alapján

$$b(1; n, p) \approx \lambda \cdot b(0; n, p) \approx \lambda e^{-\lambda},$$
$$b(2; n, p) \approx \frac{1}{2}\lambda \cdot b(1; n, p) \approx \frac{1}{2}\lambda^2 e^{-\lambda},$$

és általánosságban – amint teljes indukcióval belátható –

$$b(k; n, p) \approx \frac{\lambda^k}{k!} e^{-\lambda}. \tag{2.2}$$

Ez a binomiális eloszlás klasszikus közelítése a Poisson-eloszlással. □

2.3. Feladat. Tegyük fel, hogy az emberek egy százaléka balkezes. Becsüljük meg annak valószínűségét, hogy 200 emberből legalább 4 balkezes!

Megoldás. Tudjuk, hogy 1% balkezes, tehát 200 emberből várhatóan 2 balkezes $\Rightarrow \lambda = 2$.

P(legalább 4 ember balkezes)=1-P(legfeljebb 3 ember balkezes)=

=1-(P(0 balkezes)+P(1 balkezes)+P(2 balkezes)+P(3 balkezes))=

$= 1 - (\frac{2^0}{0!}e^{-2} + \frac{2^1}{1!}e^{-2} + \frac{2^2}{2!}e^{-2} + \frac{2^3}{3!}e^{-2}) = 0,143$

Annak a valószínűsége, hogy 200 emberből legalább 4 balkezes, körülbelül $14,3\%$.

A következőkben a binomiális eloszlás Poisson-féle közelítését használjuk, valamint együttes viselkedésüket vizsgáljuk, tehát $\lambda = np$.

2.4. Állítás. *Ha k monoton növekedve 0-tól végtelenhez tart, akkor az*

$$a_k = \frac{b(k;n,p)}{p(k;\lambda)} \tag{2.3}$$

hányados először nő, majd csökken, és maximumát a legnagyobb olyan k egész számra éri el, mely $(\lambda+1)$-nél nem nagyobb.

Bizonyítás. Tekintsük az $\frac{a_k}{a_{k-1}}$ hányadost:

$$\frac{b(k;n,p)}{p(k;\lambda)} \cdot \frac{p(k-1;\lambda)}{b(k-1;n,p)} = \frac{\binom{n}{k}p^k q^{n-k}}{e^{-\lambda}\frac{\lambda^k}{k!}} \cdot \frac{e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!}}{\binom{n}{k-1}p^{k-1}q^{n-k+1}} =$$

$$= \frac{\frac{n!}{k!\cdot(n-k)!}p^k q^{n-k}}{e^{-\lambda}\frac{\lambda^k}{k!}} \cdot \frac{e^{-\lambda}\frac{\lambda^{k-1}}{(k-1)!}}{\frac{n!}{(k-1)!(n-k+1)!}p^{k-1}q^{n-k+1}} = \frac{\frac{p}{k}\lambda^{k-1}}{\frac{\lambda^k\cdot q}{k\cdot(n-k+1)}} = \frac{p\lambda^{k-1}}{\frac{\lambda^k\cdot q}{n-k+1}} =$$

$$= p\cdot\lambda^{k-1}\cdot\frac{n-k+1}{q\cdot\lambda^k} = \frac{p(n-k+1)}{q}\cdot\lambda^{-1} = \frac{pn-pk+p}{q\lambda}$$

Ha ennek a törtnek a számlálója nagyobb a nevezőjénél,

$$\begin{aligned}
pn-pk+p &> \lambda q = \lambda(1-p)\\
\lambda-pk+p &> \lambda-\lambda p\\
-pk+p+\lambda p &> 0\\
p(1+\lambda) &> pk\\
1+\lambda &> k,
\end{aligned}$$

akkor (2.3) hányados addig nő, amíg $k < 1+\lambda$. $\square$

2.5. Állítás. *Ha k nő, akkor $b(k;n,p)$ először kisebb, majd nagyobb, majd ismét kisebb, mint $p(k;\lambda)$.*

Bizonyítás. Mindkét eloszlás tagjainak összege külön-külön 1, ezért nem lehetséges, hogy az egyik eloszlás mindig kisebb, mint a másik, mert akkor az összeg kisebb lenne 1-nél.

$$b(0;n,p) = \binom{n}{0} p^0 q^n = q^n = \left(1 - \frac{\lambda}{n}\right)^n$$
$$p(0,\lambda) = e^{-\lambda} \cdot \frac{\lambda^0}{0!} = e^{-\lambda} = e^{-np}$$

Felhasználva, hogy $1 + x < e^x$, ha $x \neq 0$,

$$1 - p < e^{-p},$$
$$(1-p)^n < e^{-np}$$

egyenlőtlenséget kapjuk. Azaz $b(0;n,p) < p(0,\lambda)$, ugyanakkor $b(k;n,p) = 0$, ha $k > n$, így elég nagy k-ra is $b(k;n,p) < p(k,\lambda)$. □

A binomiális-eloszlást szeretnénk úgy alulról és felülről becsülni, hogy az egyenlőtlenséget tovább alakítva majd felhasználva, a Poisson-eloszlás segítségével kapjunk egy újabb közelítést.

2.6. Állítás.

$$\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^{n-k} \geq b(k;n,p) \geq \frac{\lambda^k}{k!}\left(1 - \frac{k}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}. \tag{2.4}$$

Bizonyítás. Átírva az egyenlőtlenség közepét azt kapjuk, hogy

$$\frac{\lambda^k}{k!}\left(1 - \frac{\lambda}{n}\right)^{n-k} \geq \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \geq \frac{\lambda^k}{k!}\left(1 - \frac{k}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

Ezt leegyszerűsítve:

$$1 \geq \frac{n \cdot \ldots \cdot (n-k+1)}{n^k} \geq \left(1 - \frac{k}{n}\right)^k = \left(\frac{n-k}{n}\right)^k$$

A baloldali egyenlőtlenségben a tört számlálójában egy k tényezős szorzat áll. Ennek, ha minden tényezőjét n-nel felülről becsüljük, az egyenlőtlenség teljesül.

A jobb oldalon álló egyenlőtlenségben pedig a tört számlálójában minden tényezőt $(n-k)$-val alulról becsülve kapjuk a végeredményt. $\square$

A 2.3 feladatban becsültük a balkezesek számát, ennek a becslésnek a pontosságát láthatjuk a következő táblázatban a 2.6 egyenlőtlenség alapján.

k	$\frac{2^k}{k!}\left(1-\frac{2}{200}\right)^{200-k}$	$b(k;200;0,01)$	$\frac{2^k}{k!}\left(1-\frac{k}{200}\right)^k\left(1-\frac{2}{200}\right)^{200-k}$
0	0,134	0,1340	0,1340
1	0,2707	0,2707	0,2693
2	0,2734	0,2720	0,2680
3	0,1841	0,1814	0,1759
4	0,0930	0,0902	0,0858
5	0,0376	0,0357	0,0331
6	0,0126	0,0117	0,0105
7	0,0037	0,0033	0,0028
8	$9,2189\cdot 10^{-4}$	$8,0002\cdot 10^{-4}$	$6,6504\cdot 10^{-4}$
9	$2,0693\cdot 10^{-4}$	$1,7239\cdot 10^{-4}$	$1,3673\cdot 10^{-4}$
10	$4,1805\cdot 10^{-5}$	$3,3260\cdot 10^{-5}$	$2,5030\cdot 10^{-5}$

2.7. Állítás. *A 2.6 állítás felhasználásával következik, hogy*

$$p(k;\lambda)e^{k\cdot\frac{\lambda}{n}} \geq b(k;n,p) \geq p(k;\lambda)e^{\frac{-k^2}{n-k}-\frac{\lambda^2}{n-\lambda}}. \tag{2.5}$$

Bizonyítás. Elsőként tekintsük a bal oldali egyenlőtlenséget. Az előző feladatból tudjuk, hogy

$$\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k} \geq b(k;n,p). \tag{2.6}$$

Azt kell belátnunk, hogy

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

ekkor a 2.5 bal oldala biztosan teljesül. Átrendezve kell, hogy

$$e^{\frac{-n\lambda+k\lambda}{n}} \geq \left(1-\frac{\lambda}{n}\right)^{n-k}.$$

Felhasználva, hogy $1 - t \le e^{-t}$

$$1 - \frac{\lambda}{n} \le e^{-\frac{\lambda}{n}},$$
$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \le \left(e^{-\frac{\lambda}{n}}\right)^{n-k} = e^{\frac{-n\lambda + k\lambda}{n}}.$$

Ezzel a bal oldalt beláttuk.

A jobb oldalon ugyanezt a technikát alkalmazva akarjuk belátni, hogy

$$\frac{\lambda^k}{k!}\left(1 - \frac{k}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \ge \frac{\lambda^k}{k!} e^{-\lambda} e^{\frac{-k^2}{n-k} - \frac{\lambda^2}{n-\lambda}},$$

$$\left(1 - \frac{k}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \ge e^{-\lambda} e^{\frac{-k^2}{n-k} - \frac{\lambda^2}{n-\lambda}}. \tag{2.7}$$

Itt azt az egyenlőtlenséget használjuk a 2.7 egyenlőtlenség bal oldalán lévő tényezőkre külön–külön, hogy $1 - t \ge e^{\frac{-t}{1-t}}$

$$\left(1 - \frac{k}{n}\right)^k \ge \left(e^{\frac{-k/n}{1-(k/n)}}\right)^k = e^{\frac{-k^2/n}{1-(k/n)}} = e^{\frac{-k^2}{n-k}},$$

$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \ge \left(e^{\frac{-\lambda/n}{1-(\lambda/n)}}\right)^{n-k} = \left(e^{\frac{-\lambda}{n-\lambda}}\right)^{n-k} = e^{\frac{-\lambda(n-k)}{n-\lambda}}.$$

Azonban az

$$\left(1 - \frac{\lambda}{n}\right)^{n-k} \ge e^{-\lambda} e^{\frac{-\lambda^2}{n-\lambda}} = e^{\frac{-\lambda(n-\lambda)-\lambda^2}{n-\lambda}} = e^{\frac{-\lambda n}{n-\lambda}}$$

egyenlőtlenséget szeretnénk kapni. Tehát szeretnénk, ha

$$e^{\frac{-\lambda(n-k)}{n-\lambda}} \ge e^{\frac{-\lambda n}{n-\lambda}}.$$

egyenlőtlenség teljesülne. Az $e^{\frac{\lambda k}{n-\lambda}}$ hatványban a kitevő pozitív, tehát az egész hatvány egynél nagyobb. Ezt kihasználva:

$$e^{\frac{-n\lambda}{n-\lambda} + \frac{\lambda k}{n-\lambda}} = e^{\frac{-n\lambda}{n-\lambda}} \cdot e^{\frac{\lambda k}{n-\lambda}} \ge e^{\frac{-\lambda n}{n-\lambda}}.$$

Azaz a 2.5 egyenlőtlenség mindkét oldalát igazoltuk. □

Ebben az egyenlőtlenségben már megjelenik a Poisson-eloszlás, ahogy azt szerettük volna.

A 2.3 feladat adataira vizsgáljuk a becslés pontosságát, mégpedig úgy, hogy az állítás bal valamint jobb oldalán szereplő kifejezésekkel leosztjuk a binomiális eloszlást. Annál jobb a becslésünk minél közelebb van a hányados értéke 1-hez.

k	$\frac{b(k;200;0{,}01)}{p(k;2)e^{k\frac{2}{200}}}$	$\frac{b(k;200;0{,}01)}{p(k;2)e^{\frac{-k^2}{200-k}-\frac{2^2}{200-2}}}$
0	0,9900	1,0102
1	0,9900	1,0255
2	0,9851	1,0465
3	0,9753	1,0735
4	0,9607	1,1072
5	0,9416	1,1482
6	0,9181	1,1976
7	0,8906	1,2564
8	0,8595	1,3259
9	0,8251	1,4079
10	0,7880	1,5043

2.2. Poisson-közelítés párosítással

2.8. Definíció. *Legyenek* $\boldsymbol{P}_n = (p_{n,0}, p_{n,1}, \ldots)$, $n = 1, 2, \ldots$, *és* $\boldsymbol{Q} = (q_0, q_1, \ldots)$ *diszkrét valószínűségeloszlások. Azt mondjuk, hogy* $\boldsymbol{P}_n$ *gyengén konvergál* $\boldsymbol{Q}$*-hoz* ($\boldsymbol{P}_n \xrightarrow{w} \boldsymbol{Q}$)*, ha minden* $k = 0, 1, \ldots$ *esetén* $\lim_{n\to\infty} p_{n,k} = q_k$.

2.9. Definíció. *Legyen* $\boldsymbol{P} = (p_0, p_1, \ldots)$ *és* $\boldsymbol{Q} = (q_0, q_1, \ldots)$ *két diszkrét valószínűségeloszlás. Ekkor a variációs távolságuk*

$$||\boldsymbol{P} - \boldsymbol{Q}|| = \sum_{k=0}^{\infty} |p_k - q_k| \quad \textit{(ez a két számsorozat } l^1 \textit{ -távolsága).}$$

Azt mondjuk, hogy a $\boldsymbol{P}_n$ *diszkrét valószínűségeloszlások teljes variációban konvergálnak* $\boldsymbol{Q}$*-hoz, ha* $||\boldsymbol{P}_n - \boldsymbol{Q}|| \to 0$.

2.10. Lemma. $\boldsymbol{P}_n \xrightarrow{w} \boldsymbol{Q} \Leftrightarrow ||\boldsymbol{P}_n - \boldsymbol{Q}|| \to 0$*, azaz az előbbi két konvergenciafajta ekvivalens.*

2.11. Lemma. *Legyen* ξ *és* η *nemnegatív egész értékű. Ekkor az eloszlásaik variációs távolságára*

$$\sum_{k=0}^{\infty} |P(\xi = k) - P(\eta = k)| \leq 2 \cdot P(\xi \neq \eta).$$

Nézzük a bizonyítását, amit egy későbbi tétel során felhasználunk.

Bizonyítás. A szumma tagjaiban mindkét valószínűségből vonjuk ki $P(\xi = k, \eta = k)$-t.

$$\sum_{k=0}^{\infty} |P(\xi = k) - P(\eta = k)| = \sum_{k=0}^{\infty} |P(\xi = k, \eta \neq k) - P(\eta = k, \xi \neq k)| \leq$$

$$\leq \sum_{k=0}^{\infty} P(\xi = k, \eta \neq \xi) + \sum_{k=0}^{\infty} P(\eta = k, \xi \neq \eta) = 2P(\xi \neq \eta).$$

Amit eredményül akartunk kapni. □

Ezzel a lemmával már úgy adhatunk jó becslést két diszkrét eloszlás variációs távolságára, hogy keresünk hozzájuk olyan valószínűségi változókat, amelyek eloszlása éppen a két megadott, és amelyek minél nagyobb valószínűséggel megegyeznek egymással. Ezt az eljárást párosításnak nevezik.

Legyenek $A_1, A_2, \ldots, A_n$ független események. Legyen $P(A_i) = p_i$, és jelölje N azon események számát, amelyek teljesülnek: $N = I(A_1) + I(A_2) + \cdots + I(A_n)$. $E(N) = p_1 + \cdots + p_n$, legyen ez λ. Ekkor N eloszlása a következő mértékben közelíthető a λ paraméterű Poisson-eloszlással.

2.12. Tétel. (Le Cam)

$$\sum_{k=0}^{\infty} \left| P(N = k) - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq 2 \cdot \sum_{i=1}^{n} {p_i}^2.$$

Bizonyítás. A 2.10 lemmát alkalmazzuk. Egy-egy Poisson-eloszlású valószínűségi változót párosítunk külön-külön az $I(A_i)$ indikátorokhoz.

Legyenek $\xi_1, \xi_2, \ldots, \xi_n$ független valószínűségi változók, amelyek a $\{-1, 0, 1, \ldots\}$ értékeket vehetik fel. Legyen továbbá $\eta_i = I(\xi_i \geq 0)$ és $\zeta_i = \xi_i^+$. Ekkor $\eta_1, \eta_2, \ldots, \eta_n$ független indikátorok. A $\zeta_1, \zeta_2, \ldots, \zeta_n$ is függetlenek és ζ_i, p_i paraméterű Poisson-eloszlásúak. A $P(\xi_i = k)$, $P(\eta_i = k)$, $P(\zeta_i = k)$ valószínűségeket az alábbi táblázatban találjuk.

k	$\mathrm{P}(\xi_i = k)$	$\mathrm{P}(\eta_i = k)$	$\mathrm{P}(\zeta_i = k)$
-1	$1 - p_i$	0	0
0	$e^{-p_i} - 1 + p_i$	$1 - p_i$	e^{-p_i}
1	$\frac{{p_i}^1}{1!}e^{-p_i}$	p_i	$\frac{{p_i}^1}{1!}e^{-p_i}$
$k > 1$	$\frac{{p_i}^k}{k!}e^{-p_i}$	0	$\frac{{p_i}^k}{k!}e^{-p_i}$

Tudjuk, hogy $\eta_1 + \eta_2 + \cdots + \eta_n$ ugyanolyan eloszlású, mint N, $\zeta_1 + \zeta_2 + \cdots + \zeta_n$ pedig λ paraméterű Poisson-eloszlású, ezen felül

$$P(\eta_i \neq \zeta_i) = 1 - P(\eta_i = \zeta_i) = 1 - P(\eta_i = \zeta_i = 0) - P(\eta_i = \zeta_i = 1) =$$
$$1 - P(\xi_i = -1) - P(\xi_i = 1) = 1 - (1 - p_i) - e^{-p_i} \cdot p_i = p_i(1 - e^{-p_i}) \leq {p_i}^2$$

Itt megint azt az egyenlőtlenséget használtuk, hogy $e^x \geq 1 + x, x = -p_i$-re.

Tehát a

$$\sum_{k=0}^{\infty} \left| P(N = k) - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq 2P\left(\sum_{i=1}^{n} \eta_i \neq \sum_{i=1}^{n} \zeta_i \right) \leq 2 \sum_{i=1}^{n} P(\eta_i \neq \zeta_i) \leq 2 \sum_{i=1}^{n} {p_i}^2$$

közelítéssel meg tudtunk adni két olyan diszkrét eloszlást melyekre a feltételek teljesülnek és a becslés jól működik. $\square$

Ennél a közelítésnél, ha a mindegyik p_i-t p-nek választjuk, akkor a binomiális eloszlás és a Poisson-eloszlás variációs távolságát kapjuk. Azaz $p = \frac{\lambda}{n}$ esetén

$$\sum_{k=0}^{\infty} \left| \binom{n}{k} p^k q^{n-k} - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq 2 \cdot \sum_{1}^{n} p^2 = 2n \left(\frac{\lambda}{n} \right)^2 = 2 \cdot \frac{\lambda^2}{n}$$

3. fejezet

Folyamatok

Általában, ha folyamatokról beszélünk, az idővel valamilyen kapcsolatba hozzuk a történéseket. Mint például, mennyi ideig tartott egy vizsgált cselekvés? Valamennyi eltelt idő alatt hányszor következett be a megfigyelt esemény? Ebben a fejezetben ez utóbbi kérdésre választ adó modellekkel foglalkozunk.

3.1. A Poisson-folyamat

A folyamatot Siméon-Denis Poisson francia matematikusról nevezték el, amely egy véletlenszerű (sztochasztikus) folyamat és az időben véletlenszerűen bekövetkező eseményeket modellezi. Ilyen például az áruházba beérkező vevők száma, telefonközpontba beérkező hívások mennyisége, radioaktív bomlás stb. Ezt a folyamatot szokták pontfolyamatnak is nevezni, mert minden bekövetkezést az időtengely egy pontjának feleltetünk meg.
A Poisson-folyamaton belül megkülönböztetünk homogén és inhomogén változatot.

3.1.1. Homogén Poisson-folyamat

Azt vizsgáljuk, hogy tetszőleges t hosszúságú intervallumban mennyi az összes bekövetkezések $N(t)$ száma. Feltételezzük, hogy egy esemény bekövetkezésének valószínűsége bármely t hosszúságú intervallumban mindig ugyanannyi és független a múltbeli eseményektől.

Tegyük fel, hogy $N(t)$ monoton növő számláló folyamat, azaz egy természetes szám minden t-re. Továbbá az $N(t_i) - N(t_{i-1})$ valószínűségi változók függetlenek minden $t_0 < t_1 < \cdots < t_n$-re, tehát a folyamat független növekményű.

Vizsgáljuk meg a valószínűségét annak, hogy bizonyos időn belül 0; 1 vagy egynél több esemény következik be. Jelölje

$$P_n(t) = P(N(t) = n).$$

Azt mondjuk, hogy a folyamat az E_n állapotban van, ha 0 és t időpillanat között n darab esemény következett be.

Vegyük az egység hosszú időintervallum ekvidisztáns felosztását, ekkor $h = \frac{1}{n}$. Azon részintervallumok várható száma melyben van ugrás, azaz bekövetkezett esemény

$$n \cdot (1 - P_0(h)) = \frac{1}{h} \cdot (1 - P_0(h)).$$

A $h \to 0$ esetén az várható el, hogy ez az egységnyi hosszúságú intervallumon bekövetkező események várható értékéhez fog konvergálni, azaz feltehetjük, hogy

$$\frac{1}{h} \cdot (1 - P_0(h)) \to \lambda.$$

Tehát annak a valószínűsége, hogy egy részintervallumon nem következik be esemény,

$$P_0(h) = 1 - h\lambda + o(h), \tag{3.1}$$

ahol $o(h)$ egy h-nál kisebb mennyiséget jelent. A t-edik időpillanatig bekövetkezett események számát jelölje S_t. Innen csak a „szomszédos" S_{t+1} pontba tudunk ugrani a fizikai háttérből adódóan. Azaz egy részintervallumon egynél több ugrást tartalmazó h hosszúságú részintervallumok számának várható értéke 0-hoz kell, hogy tartson, tehát

$$n \cdot (1 - P_0(h) - P_1(h)) \to 0.$$

Ebből már megkapjuk azt is, hogy mennyi a valószínűsége, hogy egy esemény bekövetkezik h-n

$$\frac{1}{h}(1 - 1 - h\lambda + o(h) - P_1(h)) \to 0$$

$$P_1(h) = h\lambda + o(h) \tag{3.2}$$

Egynél több bekövetkezés valószínűsége pedig

$$o(h). \tag{3.3}$$

Ezek ismeretében már megmutatható, hogy

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}. \tag{3.4}$$

Tegyük fel, hogy $n \geq 1$ és a $t+h$ időpillanatban n esemény következett be, ennek a valószínűsége $P_n(t+h)$. Ez háromféleképpen következhet be:

a) A t-edik pillanatban már bekövetkezett n esemény, ekkor $[t, t+h]$-n már nem fog bekövetkezni egyetlen esemény sem. Ennek valószínűsége:

$$P_n(t)P_0(h) = P_n(t)(1 - \lambda h) + o(h)$$

b) A t-edik pillanatban $n-1$ darab esemény következett be eddig és $[t, t+h]$-n bekövetkezik még egy, aminek a valószínűsége

$$P_{n-1}(t)P_1(h) = P_{n-1}(t)\lambda h + O(h)$$

c) Ha ezeken kívül bármilyen más eset fordul elő, akkor $[t, t+h]$-n legalább két eseménynek kell teljesülnie, aminek a valószínűsége igen csak kicsi, az előbbiek alapján $o(h)$.

Így $P_n(t+h)$-t ennek a három független eseménynek az összegéből kapjuk

$$P_n(t+h) = P_n(t)(1-\lambda h) + P_{n-1}(t)\lambda h + o(h) = P_n(t) - P_n(t)\lambda h + P_{n-1}(t)\lambda h + o(h),$$

amelyet átalakítva

$$\frac{P_n(t+h) - P_n(t)}{h} = -P_n(t)\lambda + P_{n-1}(t)\lambda + \frac{o(h)}{h}.$$

Amennyiben h 0-hoz tart akkor az utolsó tag is 0-hoz tart, így

$$P_n'(t) = -P_n(t)\lambda + P_{n-1}(t)\lambda \qquad n \geq 1.$$

Az a) esetben lehetséges csak, hogy $n = 0$, ekkor

$$P_0(t+h) = P_0(t)(1 - \lambda h) + o(h) = P_0(t) - P_0(t)\lambda h + o(h).$$

$$P_0'(t) = -\lambda P_0(t).$$

Ennek és a $P_0(0) = 1$ segítségével láthatjuk, hogy $P_0(t) = e^{-\lambda t}$, valamint $P_1(0) = 0$ miatt $P_1(t) = \lambda t e^{-\lambda t}$.

Ezek ismeretében már a Possion-folyamat teljes definícióját is bevezethetjük.

3.1. Definíció. *Az $N(t)$ $(t \geq 0)$ Poisson-folyamat, ha*

- *$N(t)$ számláló folyamat, $N(t)$ természetes szám minden minden t-re és $N(t)$ monoton növő.*
- *Ha a $t_0 = 0 < t_1 < \cdots < t_n$ esetén az $N(t_i) - N(t_{i-1})$ valószínűségi változók függetlenek, tehát a folyamat független növekményű.*
- *Ha a $[t_{i-1}, t_i]$, $i = 0, \ldots n$ idő intervallumok azonos nagyságúak, akkor az $N(t_i) - N(t_{i-1})$ valószínűségi változók, Poisson-eloszlásúak λt $(t = t_i - t_{i-1})$ várható értékkel.*

Például, ha $t = 1.6; 2.7; 3.1; 3.7; 4.3; 5.2$ időpontokat, akkor az $N(5,2) - N(3,1)$ és az $N(2,7) - N(1,6)$ növekmények függetlenek, de az $N(5,2) - N(3,1)$ és $N(4,3) - N(3,7)$ növekmények már nem, mert az időintervallumok nem diszjunktak. Azonban, ha két intervallum végpontja megegyezik azokat függetlennek tekintjük.

A Poisson-folyamat tehát egy számláló folyamat, melyben két bekövetkezés időpontja között eltelt idő exponenciális eloszlású valószínűségi változó $\frac{1}{\lambda}$ várható értékkel. Ez látható abból is, amikor azt vizsgáljuk, hogy egy intervallumon 0 esemény következik be. A másik jellege a folyamatnak, hogy (3.1) – (3.3) valószínűségek nem függenek a korábbi, $N(t)$ állapottól.

3.1.2. Inhomogén Poisson-folyamat

Az előző részben (3.1) – (3.3) esetekben a λ egy konstans, t-től nem függő mennyiség volt. Azoknak az eseményeknek a modellezésekor, melyeknek az intenzitása változik időben, azoknál célszerű λ-t az időtől függővé tenni.

Tegyük fel továbbra is, hogy $N(t)$ számláló folyamat. Az inhomogén folyamatot hasonlóan definiáljuk mint a homogén esetben de nem ugyanazok a jellemzők rá. Az egymást nem átfedő intervallumok továbbra is függetlenek. A homogén esetben ugyanolyan nagyságú intervallumok ugyanolyan paraméterű eloszlást követtek, itt nem mondható el ugyanez. Az intenzitás időtől függ, ezért egy intervallumon bekövetkező események várható értékét az alábbiak szerint számoljuk

$$\Lambda(t_i; t_{i+1}) = \int_{t_i}^{t_{i+1}} \lambda(x)dx.$$

3.2. Példa. Tegyük fel, hogy N egy Poisson-folyamat egy adott $\lambda(t) = 2t$ intenzitásfüggvénnyel. Továbbá azt, hogy 100 esemény következett be $t = 1,2$-ig, azaz $N(1,2) = 100$. Az $M = N(1,2) = N(1,2) - N(0)$ egy Poisson-eloszlású valószínűségi változó, $\Lambda = \int_0^{1,2} 2zdz = 1,44$ várható értékkel. Ezt a megfigyelést használjuk föl ahhoz, hogy meghatározzuk az $N(2,6)$ eloszlását.
Azt már ismerjük, hogy a $[0; 1,2]$ intervallumon mi történt, így már csak az $[1,2; 2,6]$ intervallumot kell vizsgálnunk.

$$N(2,6) = [N(2,6) - N(1,2)] + [N(1,2) - N(0)].$$

Az első zárójelben lévő valószínűségi változó várható értékét a definíció alapján számolva kapjuk, hogy

$$\Lambda = \int_{1,2}^{2,6} 2zdz = 5,32.$$

Ezzel együtt már ismerünk mindent az $N(2,6)$ meghatározásához, mégpedig

$$[N(2,6)|N(1,2) = 100] \approx N + 100.$$

Ez megkönnyíti $N(2,6)$ várható értékének meghatározását

$$E[N(2,6)|N(1,2) = 100] = E[M + 100] = E[M] + 100 = 5,32 + 100 = 105,32$$

A szórásnégyzete pedig

$$D^2[N(2,6)|N(1,2) = 100] = D^2[M + 100] = D^2[M] = 5,32.$$

A valószínűséget ezek után már még könnyebb számolni. Számoljuk ki annak a valószínűségét, hogy $t = 2,6$ időpontig 115 esemény következik be.

$$P[N(2,6) = 115|N(1,2) = 100] =$$
$$= P[M + 100 = 115] = P[M = 15] = \frac{e^{-5,32} 5,32^{15}}{15!} = 2,8955 \cdot 10^{-4}.$$

3.3. Példa. Ahogy azt az előző példában is feltettük, legyen N egy Poisson-folyamat, $\lambda(t) = 2t$ intenzitásfüggvénnyel.
Feladatunk, hogy meghatározzuk $P[T_3 > 1,2]$-t, azaz mennyi a valószínűsége annak, hogy a harmadik esemény bekövetkezése $t = 1,2$ után történt.
A harmadik esemény $t = 1,2$ után történt ezért 1,2 időpontig legfeljebb 2 esemény

következett be, azaz $N(1,2) \leq 2$. Ez ekvivalens azzal, hogy $T_3 > 1,2$, ezért a valószínűségük is megegyezik. Az előző példa alapján $N(1,2)$ egy Poisson-eloszlású M valószínűségi változó 1,44 várható értékkel. Így

$$
\begin{aligned}
P[T_3 > 1,2] = P[N(1,2) \leq 2] = P[M \leq 2] = \\
= P[M = 0] + P[M = 1] + P[M = 2] = \\
= e^{-1,44} + e^{-1,44}\frac{1,44^1}{1!} + e^{-1,44}\frac{1,44^2}{2!} = 0,82375.
\end{aligned}
$$

Tehát 82%-os valószínűséggel az 1,2 időpont után fog bekövetkezni a harmadik esemény.

A homogén folyamatot sok esetben használjuk, azonban az inhomogén folyamat modellezésével pontosabb eredményt kapunk. A valóságban ilyen folyamattal modellezünk például:

- telefonközpontba beérkező hívásokat, ahol a hívások száma reggel 8 óra és délután 6 óra között gyakoribb, mint a nap többi időszakában,
- autópályáról érkező kocsik száma a fővárosba, reggel 8 óra körül megnő az intenzitás, aztán megint csökken,
- adatbázis rendszerben a tranzakciós folyamatokat,
- vendégek érkezése a gyorsétterembe különböző napszakokban sűrűsödik.

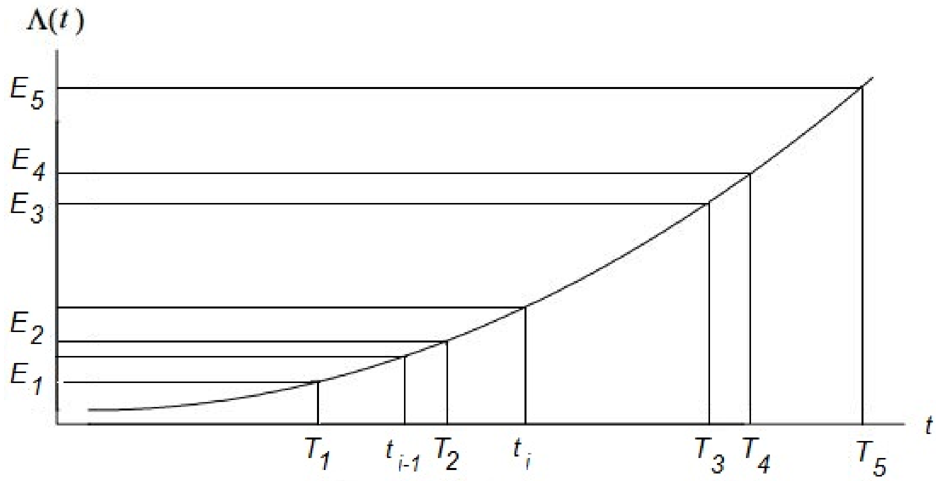
Sok módja van az inhomogén Poisson-folyamat generálásának, amiket röviden áttekintünk.

1. Az első, hogy a $\lambda = 1$ paraméterű homogén Poisson-folyamat időintervallumát átskálázzuk. Ha az $E_1, E_2, \ldots$ a bekövetkezett események időpontja a homogén folyamatban, akkor $\Lambda^{-1}(E_1), \Lambda^{-1}(E_2), \ldots$ pontok az inhomogén folyamat pontjait jelöli $\Lambda(t)$ intenzitásfüggvénnyel. Ezt láthatjuk a 3.1 ábrán. Tekintsünk egy időintervallumot, – legyen ez a $[t_{i-1}, t_i]$ – melyben található esemény, az intervallum átskálázását az ábráról leolvasva könnyű meghatározni.

A $t_{i-1}; t_i$ időpontokat a $\Lambda(t)$ tengelyre vetítve azok, azt a $\Lambda(t_{i-1})$ és $\Lambda(t_i)$ pontokban metszik. Innen

$$N(t_i) - N(t_{i-1}) = \tilde{N}(\Lambda(t_i)) - \tilde{N}(\Lambda(t_{i-1}))$$

valószínűségi változók Poisson-eloszlásúak $\Lambda(t_i) - \Lambda(t_{i-1}) = \int_{t_{i-1}}^{t_i} \lambda(x)dx$ paraméterrel.



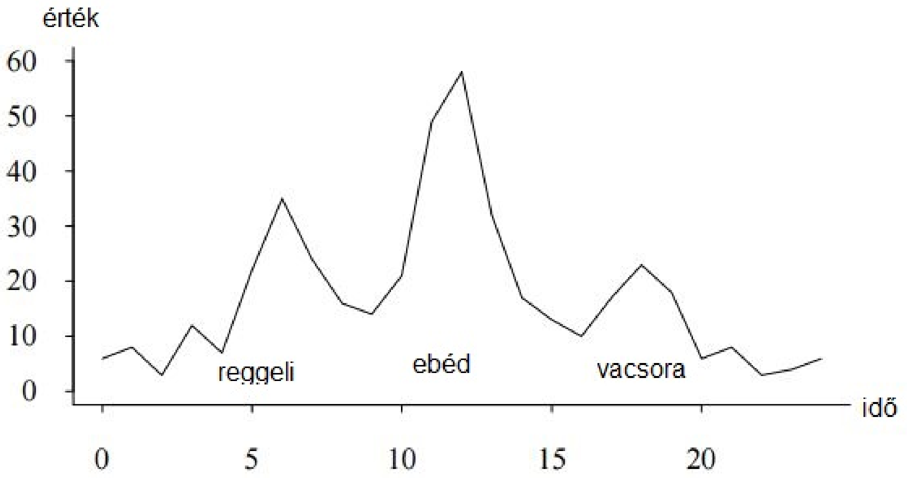
3.1. ábra. Homogén folyamat transzformációja

2. A második módszer csak említés szintjén: inhomogén folyamat előállítására a pontok közötti intervallumok egyenként való generálása. Adottak az $X_1 = x_1, X_2 = x_2, \ldots, X_i = x_i$ pontok és $x_1 < x_2 < \ldots < x_i$. Az $X_{i+1} - X_i$ intervallumok függetlenek minden i-re és az eloszlásfüggvényük

$$F(x) = 1 - e^{-[\Lambda(x_i+x) - \Lambda(x_i)]}.$$

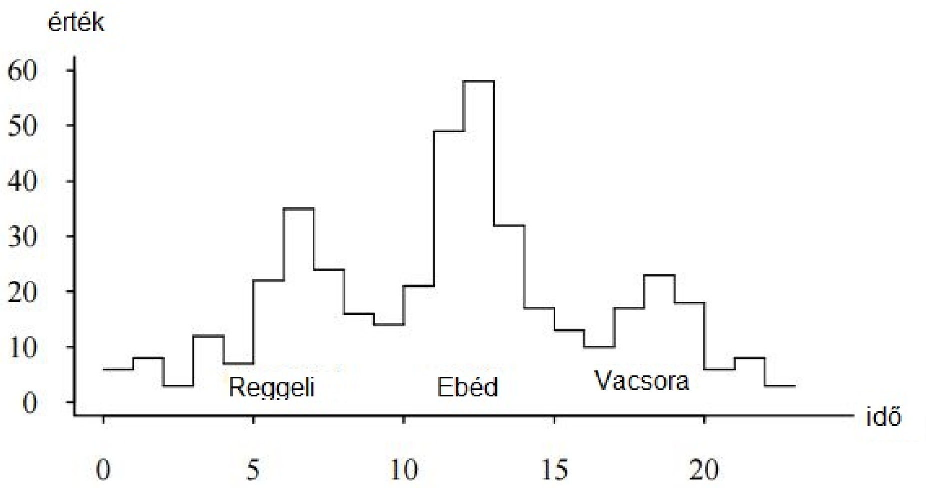
Intenzitás függvény becslése:

Tekintsük azt a példát, amikor a gyorsétteremben vásárlók érkezését figyeljük meg egy adott napon, ezt láthatjuk a 3.2 ábrán.



3.2. ábra. Fogyasztók érkezése egy nap

Vegyük egy vizsgált nap ekvidisztáns felosztását, ahol óránként megfigyeljük a folyamatunkat. A $\lambda(t)$ intenzitás függvényt úgy becsüljük, hogy szakaszonként konstans értéket vegyen föl. Ezzel a módszerrel a 3.3 ábrán látható becslését kapjuk a $\lambda(t)$-nek.

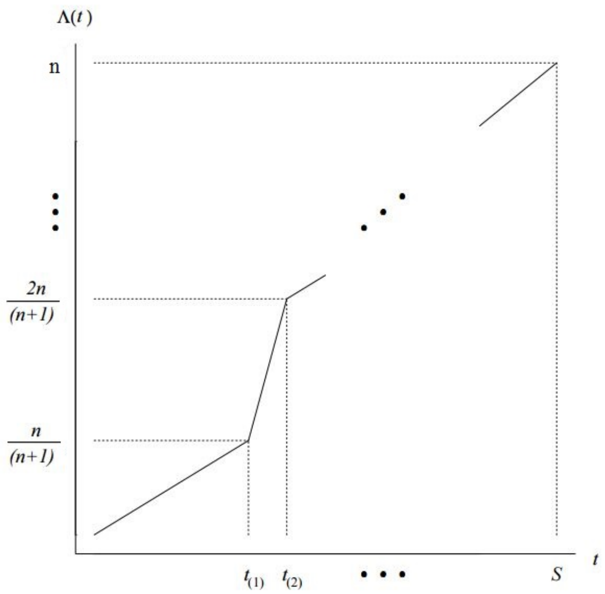


3.3. ábra.

A $\Lambda(t)$ intenzitás függvény becslésére azt alkalmazzuk, hogy nem a megfigyelések szerepelnek az időintervallumon, hanem maguk a bekövetkezések időpontjai. Két bekövetkezés időpontja között az intenzitás függvényt egy szakaszonként változó lineáris függvénnyel becsüljük.

$$\hat{\Lambda}(t) = \frac{i \cdot n}{(n+1)} + \left[\frac{n(t-t_i)}{(n+1)(t_{i+1}-t_i)}\right]$$
$$t_i < t \leq t_{i+1} \text{ és } i = 0, 1, \ldots, n.$$

Minél közelebb van egymáshoz két bekövetkezés annál meredekebb lesz azon az intervallumon a lineáris függvény.



3.4. ábra.

3.2. Az összetett Poisson-folyamat

Az összetett Poisson-folyamatot például a kockázati folyamatok modellezése során alkalmazzák, ezért most ezen keresztül vizsgáljuk. A már korábban tárgyalt összetett Poisson-eloszlás példái között szerepelt az összetett kockázati modell, ahol az S-el jelölt teljes kár eloszlása összetett Poisson-eloszlás volt. Az összkár értékét azonban gyakran az idő függvényében vizsgálják, ekkor a kárszámot megadó N változó az idő függvénye. Jelölje ezt $N(t)$, $t \geq 0$. Ezt hívják kárigény folyamatnak. Ekkor az összkár is az időtől függő

$$S(t) = \sum_{i=1}^{N(t)} Z_i.$$

A kár kifizetéséhez szükséges tartalék leírásának másik két eleme a díjbevételt megadó P_t folyamat a $[0, t]$ intervallumban, valamint a kezdeti tőke értéke u. Így kapjuk a rizikófolyamatot

$$U(t) = u + P(t) - S(t).$$

Klasszikus rizikófolyamatról beszélünk akkor, ha

$$\begin{aligned} P(t) = c \cdot t, &\qquad \text{ahol } c \text{ állandó} \\ N(t), &\qquad \lambda \text{ paraméterű Poisson-folyamat} \\ Z_i, i = 1, 2, \ldots &\text{ függetlenek és azonos eloszlásúak.} \end{aligned}$$

Ekkor az $S(t)$ folyamat, amely tehát a független, azonos eloszlású Z_j valószínűségi változók Poisson-tagszámú összege úgynevezett *összetett Poisson-folyamat.*

3.4. Definíció. *Egy folyamatot sztochasztikusan folytonosnak nevezünk, ha bármely* $\epsilon > 0$ *esetén* $P(|N(s) - N(t)| > \epsilon) \to 0$*, ha* $s \to t$*, minden rögzített* $t \geq 0$ *mellett.*

3.5. Tétel. *Ha az* $S(t)$*,* $t \geq 0$ *folyamat sztochasztikusan folytonos, független növekményű, az ugyanolyan növekményű időintervallumokhoz tartozó növekmények azonos eloszlásúak,* $S(0) = 0$*, akkor a folyamat összetett Posson-folyamat.*

A homogén Poisson-folyamattól abban tér el, hogy egyszerre több esemény is bekövetkezhet, azaz az ugrások száma nem feltétlenül 1. A növekmények eloszlása az időtartamtól függ.

Az 1.14 tételt általánosabban, Poisson-folyamatokra is alkalmazhatjuk. Tekintsük az $S(t) = \sum_{j=1}^{N(t)} Z_j$, $t \geq 0$összetett folyamatot. Tetszőleges $A \subset \mathbb{R}$ esetén nézhetjük, hogy melyek azok a $[0, t]$ időintervallumon bekövetkezett károsemények, melyekben a kár értéke az A halmazba esett. Azaz legyen

$$S(t)^A = \sum_{k=1}^{N(t)} Z_{k\chi\{Z_k \in A\}}.$$

Ez az $S(t)$ folyamat ritkítása, ami szintén összetett Poisson-folyamat. Diszjunkt A halmazok esetén a kapott $S_A(t)$ folyamatok egymástól függetlenek. Tehát, ha a kárigény nagysága szerint csoportosítjuk az összetett Poisson-folyamat ugrásait, egymástól független összetett Poisson-folyamatokat kapunk.

Összefoglalás

Láthattuk, hogy a Poisson-eloszlás széleskörűen alkalmazható a mindennapokban is és a hasznosságának csak egy kicsi hányadát olvashattuk. Az eloszlás bemutatott változatai különböző területen használatosak és adnak pontos értékeket.

A binomiális eloszlás közelítésére meglepően jó becsléseket kaptunk. Ezeket célszerű alkalmazni, hogy megkönnyítsük a számolásunkat, hiszen kevesebb paramétert kell felhasználni és az eltérés nem jelentős. A Poisson-eloszlást is szokták közelíteni, mégpedig a normális eloszlással.

Az utolsó fejezetben vált érdekesebbé a Poisson-eloszlás alkalmazása, amikor a folyamat során az időtől függővé tettük paraméterünket. Ez mondható a leghasznosabb alkalmazásának, ugyanis, ha valamilyen mindennapos véletlenszerű eseményeket szeretnénk megfigyelni, arra modellt építeni, az idő haladásával a bekövetkezések sűrűsége változik.

A szakdolgozatomban a balesetekről szóló adatok a www.ksh.hu, Siméon Denis Poisson életrajzi érdekességei a www.hu.wikipedia.org oldalról származnak.

Irodalomjegyzék

[1] Arató Miklós, Prokaj Vilmos, Zempléni András *Bevezetés a valószínűségszámításba és alkalmazásaiba: Példákkal, szimulációkkal*, 2013

[2] Csernyák László, *Valószínűségszámítás*, Nemzeti tankönyvkiadó, 2007

[3] D.J. Daley, D.Vere-Jones, *An Introduction to the Theory of Point Proesses: Volume I*, Springer, 2003

[4] D.R. Cox, Valerie Isham, *Point Processes*, Chapmann & Hall/CRC, 2000

[5] Denkinger Géza, *Valószínűségszámítási gyakorlatok*, Nemzeti tankönyvkiadó, 1999

[6] James W. Daniel, *Poisson Processes and mixture distibutions*, 2008

[7] Larry Leemis, *Estimating and Simulating Nonhomogeneous Poisson Processes*, 2003

[8] Michaletzky György, *Kockázati folyamatok jegyzet*, 2001

[9] Móri Tamás, *Generátorfüggvények jegyzet*, 2007

[10] Móri Tamás, *Poisson-approximáció párosítással jegyzet*, 2007

[11] Obádovics J. Gyula, *Valószínűségszámítás és matematikai statisztika*, Scolar Kiadó, 2001

[12] P.A.W. Lewis, G.S. Shedler, *Simulation of Nonhomogeneous Poisson Processes by thinning*, 1978

[13] Szűcs Gábor, *Kockázati folyamatok jegyzet*, 2015

[14] Walter Zucchini,Iain L. MacDonald, *Hidden Markov Models for Time Series: An Introduction Using R*, CRC Press, 2009

[15] William Feller, *Bevezetés a valószínűségszámításba és alkalmazásaiba*, Műszaki könyvkiadó, 1978