Valószínűségszámítás 2.

Csiszár Villő

Tartalomjegyzék

1. Matematikai statisztika

A statisztika egyik ága a leíró statisztika. Ekkor a megfigyelt adatokat áttekinthető formában ábrázoljuk, pl. hisztogrammal (oszlopdiagrammal), kördiagrammal, egyéb grafikonokkal. Másrészt az adatokból kiszámítunk néhány fontos, jellemző értéket, pl. az átlagot (mintaközepet), tapasztalati szórást, szélsőértékeket.

A *matematikai statisztika* alapfeladata: egy véletlen jelenség mechanizmusát (pl. az őt leíró valószínűségi változó eloszlását) nem ismerjük, de megfigyeléseket végezve, a megfigyelésekből szeretnénk rá következtetni. A következő témakörökkel fogunk foglalkozni.

Becsléselmélet: A valószínűségi változó valamilyen jellemzőjét szeretnénk a mintából megbecsülni, illetve a becslés hibáját meghatározni. Minél pontosabb, megbízhatóbb becslést keresünk. Példák:

- Egy munkáltatót egy titkárnő által gépelt szövegekben előforduló hibák száma érdekli. Pl. a hibák átlagos száma és a hibaszám szórása. A munkáltató 30 darab, közel azonos hosszúságú, a titkárnő által legépelt szövegben megszámolja a hibákat. Ésszerű feltenni, hogy a hibák száma Poisson eloszlású, de az eloszlás paramétere (λ) ismeretlen. A megfigyelések alapján szeretne következtetni λ-ra, ebből a várható érték és a szórás már kiszámolható. Másrészt, a várható értéket és a szórást becsülheti a Poisson feltételezés nélkül is.

- Egy fonalgyárban a fonalszakadásokat vizsgálják. Annak a valószínűségét szeretnék megbecsülni, hogy a fonal egy 8 órás műszak alatt egyszer sem szakad el. Ennek érdekében 20 fonalszál mindegyikéről feljegyzik, hogy mennyi idő múlva szakad el. Ésszerű feltenni, hogy a fonalak élettartama exponenciális eloszlású (örökifjú tulajdonságú), de λ ismeretlen.

- Egy kosarazó 10-szer kosárra dob. Betalál $\to 1$ pont, nem talál be $\to 0$ pont. Kapott pontszám egy dobásból: $\mathrm{Ind}(p)$, ahol a találat valószínűsége p ismeretlen, ezt szeretnénk megbecsülni.

- Hétfőtől péntekig naponta megnézzük egy város áramfogyasztását. Ez feltehetőleg normális eloszlású, de m, σ ismeretlen.

- Hétfőtől péntekig megmérjük, hogy mennyit kell várni a buszra. Feltehető, hogy ez egyenletes eloszlású $[0, b]$ intervallumon, ahol b ismeretlen.

Hipotézisvizsgálat: A jelenséggel kapcsolatban van egy előzetes feltételezésünk, amelyet tesztelni szeretnénk. Ha a megfigyeléseink összeegyeztethetők a feltevéssel, elfogadjuk azt, ha viszont ellentmondanak neki, akkor elutasítjuk a feltevést. Jó döntési eljárást keresünk. Példák:

- egészségügyben: gyógyszerek, kezelések hatásosságának bizonyítása;

- ipar: selejtarány ellenőrzése: le kell-e a gépsort cserélni?

- irodalomtudomány: 2 szövegről el kell dönteni, hogy ugyanaz írta-e őket;

- szociológia: pártpreferencia és iskolázottság között van-e összefüggés?

Először definiáljuk pontosan a statisztikai mezőt, melyben dolgozni fogunk.

1.1. Definíció. Az $(\Omega, \mathcal{A}, \mathcal{P})$ hármast *statisztikai mezőnek* hívjuk, ahol Ω nemüres halmaz (eseménytér), $\mathcal{A}$ σ-algebra (események családja), $\mathcal{P}$ pedig a szóbajöhető valószínűségi mértékek családja. Azaz

$$\mathcal{P} = \{P_\vartheta | \vartheta \in \Theta\},$$

ahol P_ϑ egy lehetséges valószínűségi mérték. A Θ halmazt *paramétertérnek* nevezzük, ennek valamelyik eleme a jelenséget leíró valódi paraméter, de nem tudjuk, hogy melyik.

Legtöbbször Θ véges dimenziós euklideszi tér részhalmaza, ekkor azt mondjuk, hogy paraméteres a feladat. Θ lehet ennél jóval „nagyobb", pl.: ha $\mathcal{P}$ az összes lehetséges valószínűségi mérték, ekkor nemparaméteres a feladat. A ϑ paraméter valódi értékét tehát nem ismerjük, csak azt tudjuk, hogy a Θ halmaz valamelyik eleme. Célunk, hogy a megfigyeléseink alapján megbecsüljük ϑ értékét.

1.2. Definíció. Az $\mathbf{X} = (X_1, \ldots, X_n) : \Omega \to \mathcal{X} \subseteq \mathbb{R}^n$ valószínűségi változót n elemű mintának nevezzük. Itt $\mathcal{X}$ a mintatér (a minta lehetséges értékeinek halmaza), n pedig a minta nagysága vagy elemszáma. Az X_i koordináták a minta elemei.

A minta az, amit megfigyelünk, és amely egyáltalán információt ad nekünk arról, hogy mi is lehet az ismeretlen paraméter értéke. Ha másképp nem mondjuk, akkor fel fogjuk tenni, hogy ugyanazt a véletlen jelenséget figyeljük meg n-szer, egymástól függetlenül.

1.3. Definíció. $\mathbf{X}$ *független elemű* minta, ha X_i-k az összes P_ϑ szerint függetlenek. $\mathbf{X}$ *azonos eloszlású* minta, ha X_i-k az összes P_ϑ szerint azonos eloszlásúak.

A minta eloszlásfüggvényeinek családja $\{F_{n;\vartheta} |\ \vartheta \in \Theta\}$, ahol

$$F_{n;\vartheta}(x_1, \ldots, x_n) = P_\vartheta(X_1 < x_1, \ldots, X_n < x_n).$$

Ha $\mathbf{X}$ *n elemű, független, azonos eloszlású minta*, akkor az eloszlásfüggvény szorzattá bomlik:

$$F_{n;\vartheta}(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{1;\vartheta}(x_i),$$

ahol $F_{1;\vartheta}$ az X_i koordináták közös eloszlásfüggvénye (egy elemű minta esetén az 1 indexet nem mindig írjuk ki, tehát $F_{1;\vartheta} = F_\vartheta$). Emlékezzünk rá, hogy az eloszlásfüggvény helyett diszkrét esetben a

$$p_{n;\vartheta}(x_1, \ldots, x_n) = P_\vartheta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

valószínűségeket, abszolút folytonos esetben pedig az

$$f_{n;\vartheta}(x_1, \ldots, x_n)$$

sűrűségfüggvényt is használhatjuk. Független, azonos eloszlású minta esetén ezek is szorzatra bomlanak.

1.1. Példa. (titkárnő) A minta: $\mathbf{X} = (X_1, \ldots, X_{30}) : \Omega \to \mathbb{N}_0^{30}$, ahol X_i az i. szövegben talált hibák száma. $\mathbf{X}$ független, azonos eloszlású minta, a mintaelemek szóbajöhető eloszlásai: $X_i \sim \text{Poisson}(\vartheta)$, a paramétertér $\Theta = (0, \infty) \subset \mathbb{R}$, azaz egyparaméteres feladatról van szó. A valószínűségeket részletesen kiírva

$$p_{30;\vartheta}(x_1, x_2, \ldots, x_{30}) = \prod_{i=1}^{30} p_{1;\vartheta}(x_i) = \prod_{i=1}^{30} e^{-\vartheta} \frac{\vartheta^{x_i}}{x_i!} = e^{-30\vartheta} \frac{\vartheta^{\sum x_i}}{\prod x_i!}. \quad \blacksquare$$

1.4. Definíció. Az mintatéren megadott $T : \mathcal{X} \to \mathbb{R}^k$ függvényt, illetve magát a $T = T(\mathbf{X})$ valószínűségi változót *(k-dimenziós) statisztikának* nevezzük.

Gyakran használt statisztikák a mintaátlag, a tapasztalati medián, a legkisebb mintaelem, a tapasztalati szórás. Finom megkülönböztetés, hogy ha a mintát valószínűségi változónak tekintjük, akkor a $T(\mathbf{X})$ statisztika is valószínűségi változó, ha viszont a minta egy konkrét $\mathbf{x}$ realizációjával dolgozunk, akkor $T(\mathbf{x})$ egy k-dimenziós vektor.

1.2. Példa. Néhány gyakran használt statisztika ($\mathbf{X}$ mindenhol n elemű minta):

1. $T(\mathbf{X}) = (X_1^{(n)}, X_2^{(n)}, \ldots, X_n^{(n)})$ a *rendezett minta*, ahol $X_1^{(n)} \le X_2^{(n)} \le \ldots \le X_n^{(n)}$. Például az $\mathbf{x} = (2, 4, 1, 3)$ minta realizációra $T(\mathbf{x}) = (1, 2, 3, 4)$.

2. $T(\mathbf{X}) = X_n^{(n)} - X_1^{(n)}$ a *mintaterjedelem.*

3. $T(\mathbf{X}) = \overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ a *mintaátlag.*

4. $T(\mathbf{X}) = \begin{cases} X_{\frac{n+1}{2}}^{(n)} & \text{ha } n \text{ páratlan} \\ \dfrac{X_{\frac{n}{2}}^{(n)} + X_{\frac{n}{2}+1}^{(n)}}{2} & \text{ha } n \text{ páros} \end{cases}$ a *tapasztalati medián.*

5. $T(\mathbf{X}) = S_X^2 = \dfrac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ a *tapasztalati szórásnégyzet.*

6. $T(\mathbf{X}) = \dfrac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|$ az *átlagos abszolút eltérés.* ■

Korábban definiáltuk a tapasztalati eloszlást, és a hozzá tartozó tapasztalati eloszlásfüggvényt:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(X_i < x), \quad x \in \mathbb{R}.$$

1.1. Tétel. ***(Glivenko, A statisztika alaptétele.)*** *Legyenek $X_1, \ldots, X_n, \ldots$ független, azonos F eloszlásfüggvényű valószínűségi változók. Ekkor az $\hat{F}_n$ tapasztalati eloszlásfüggvény 1 valószínűséggel egyenletesen tart F-hez, azaz*

$$P(\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right| = 0) = 1.$$

A tétel jelentése az, hogy ha elég sok megfigyelést végzünk, akkor tetszőleges pontossággal visszakapjuk a valódi eloszlást. Azt könnyű belátni, hogy minden rögzített $x \in \mathbb{R}$-re

$$P(\lim_{n\to\infty} \left|\hat{F}_n(x) - F(x)\right| = 0) = 1,$$

hiszen ez éppen a nagy számok erős törvénye az $Y_i = \mathrm{I}(X_i < x)$ független, azonos eloszlású valószínűségi változókra. Megjegyezzük még, hogy a $\sup_{x\in\mathbb{R}} \left|\hat{F}_n(x) - F(x)\right|$ maximális eltérés nagyságrendje $1/\sqrt{n}$.

1.1. Feladat. Legyen $\mathbf{X}$ független, azonos eloszlású minta, a koordináták közös eloszlásfüggvénye F. Számoljuk ki $\hat{F}_n(x)$ várható értékét és szórásnégyzetét!
Megoldás: $n\hat{F}_n(x) \sim \mathrm{Bin}(n, F(x))$. Így $E(\hat{F}_n(x)) = F(x)$ és $D(\hat{F}_n(x)) = \sqrt{F(x)(1-F(x))/n}$. ■

1.1. Elégséges statisztika

Az $\mathbf{X}$ minta információt tartalmaz arról, hogy melyik $\vartheta \in \Theta$ az igazi paraméter, hiszen a $P_\vartheta(\mathbf{X} = \mathbf{x})$ valószínűség függ ϑ-tól (bizonyos ϑ-kra nagy a valószínűsége, hogy ezt a mintát kapjuk, másokra kisebb). A $T(\mathbf{X})$ statisztika is hordoz információt, hiszen a $P_\vartheta(T(\mathbf{X}) = t)$ valószínűség is függ ϑ-tól. Az eredeti minta általában több információt tartalmaz a paraméterről, mint a belőle kiszámolt statisztika.

Nézzük a diszkrét minta esetét! A

$$P_\vartheta(\mathbf{X} = \mathbf{x}) = P_\vartheta(\mathbf{X} = \mathbf{x}, T(\mathbf{X}) = T(\mathbf{x})) = P_\vartheta(T(\mathbf{X}) = T(\mathbf{x})) \cdot P_\vartheta(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = T(\mathbf{x}))$$

felírásból látszik, hogy ha a $P_\vartheta(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))$ feltételes valószínűség valójában már nem függ az ismeretlen ϑ paramétertől, akkor a minta nem tartalmaz több információt, mint a statisztika. Vagyis a $T(\mathbf{X})$ statisztika pont annyi információt hordoz az ismeretlen paraméterre nézve, mint az eredeti $\mathbf{X}$ minta.

Ezt úgy is elképzelhetjük, hogy gyűjtöttünk egy nagy $\mathbf{X}$ mintát, de nem volt kapacitásunk tárolni, ezért kiszámoltuk belőle az egyszerűbb $T(\mathbf{X})=t$ statisztikát, és csak ezt tartottuk meg, az eredeti mintát eldobtuk. Ha most valaki mégis számon kérné rajtunk az eredeti mintát, akkor a mintának a statisztikára vett feltételes eloszlásból tudunk egy $\mathbf{Y}$ mintát generálni (mivel ez a feltételes eloszlás ismert, nem függ ϑ-tól). Az így kapott $\mathbf{Y}$ minta pont olyan eloszlású lesz, mint az eredeti $\mathbf{X}$, tehát nem fogunk „lebukni", hogy csaltunk.

1.5. Definíció. Legyen $\mathbf{X}=(X_1,\dots,X_n)$ diszkrét minta az $(\Omega,\mathcal{A},\mathcal{P})$ statisztikai mezőn. Azt mondjuk, hogy a $T(\mathbf{X})$ statisztika *elégséges* a ϑ paraméterre, ha minden $\mathbf{x},t$ párra a $P_\vartheta(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t)$ feltételes valószínűség nem függ ϑ-tól.

1.3. Példa. Legyen $X_i\sim\mathrm{Ind}(p)$, ahol $0\le p\le 1$ ismeretlen paraméter. Belátjuk, hogy $\sum_i X_i$ elégséges statisztika p-re, azaz elég feljegyezni, hogy összesen hány 1-es van a mintában, ezzel nem veszítünk információt p-ről. A definíció alapján számolunk:

$$P_p(\mathbf{X}=\mathbf{x}|\sum_{i=1}^n X_i=t)=\begin{cases}0 & \text{ha } \sum_{i=1}^n x_i\neq t,\\ \dfrac{\prod_{i=1}^n p^{x_i}(1-p)^{1-x_i}}{\binom{n}{t}p^t(1-p)^{n-t}}=\dfrac{p^{\overbrace{\sum x_i}^{=t}}(1-p)^{\overbrace{n-\sum x_i}^{=n-t}}}{\underbrace{\binom{n}{t}p^t(1-p)^{n-t}}_{\mathrm{binom};\sum \mathrm{Ind}}}=\dfrac{1}{\binom{n}{t}} & \text{ha } \sum_{i=1}^n x_i=t.\end{cases}$$

Azaz a kapott feltételes valószínűség tényleg nem függ p-től. Ha például egy cinkelt érmén a fejdobás ismeretlen valószínűsége p, és erre szeretnénk következtetni, akkor n érmedobásból csak az a „lényeg", hogy hány fej volt, az nem számít, hogy a fejek és az írások milyen sorrendben jöttek ki. A fenti számolásból látszik, hogy bármi is a p paraméter értéke, adott t számú fej mellett mindegyik $\binom{n}{t}$ sorrend egyformán valószínű. ■

1.2. Tétel. ***(Neyman faktorizációs tétele.)*** *Legyen $\mathbf{X}$ diszkrét eloszlású minta. A $T(\mathbf{X})$ statisztika akkor és csak akkor elégséges, ha találhatók olyan h és g_ϑ függvények, melyekre $P_\vartheta(\mathbf{X}=\mathbf{x})=$ $=h(\mathbf{x})\cdot g_\vartheta(T(\mathbf{x}))$.*

Bizonyítás.
$\Rightarrow$: Tegyük fel, hogy $T(\mathbf{X})$ elégséges statisztika. Ekkor

$$P_\vartheta(\mathbf{X}=\mathbf{x})=P(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))\cdot P_\vartheta(T(\mathbf{X})=T(\mathbf{x}))=h(\mathbf{x})\cdot g_\vartheta(T(\mathbf{x})),$$

felhasználva, hogy az első tényező nem függ ϑ-tól.
$\Leftarrow$: Most tudjuk, hogy $P_\vartheta(\mathbf{X}=\mathbf{x})=h(\mathbf{x})\cdot g_\vartheta(T(\mathbf{x}))$, meg kell mutatni, hogy $T(\mathbf{X})$ elégséges statisztika.

$$P_\vartheta(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=t)=\frac{P_\vartheta(\mathbf{X}=\mathbf{x},T(\mathbf{X})=t)}{P_\vartheta(T(\mathbf{X})=t)}=\frac{P_\vartheta(\mathbf{X}=\mathbf{x})}{\sum\limits_{\mathbf{y}:T(\mathbf{y})=t}P_\vartheta(\mathbf{X}=\mathbf{y})}=$$
$$\frac{h(\mathbf{x})\cdot g_\vartheta(t)}{\sum\limits_{\mathbf{y}:T(\mathbf{y})=t}h(\mathbf{y})\cdot g_\vartheta(T(\mathbf{y}))}=\frac{h(\mathbf{x})}{\sum\limits_{\mathbf{y}:T(\mathbf{y})=t}h(\mathbf{y})},$$

ha $T(\mathbf{x})=t$, egyébként pedig a feltételes valószínűség nulla. ■

A tételnek az a jelentősége, hogy módszert ad arra, hogyan lehet elégséges statisztikát találni.

1.4. Példa. Legyen $X_1, \dots, X_n \sim \text{Poisson}(\lambda)$, keressünk elégséges statisztikát λ-ra!

$$p_{n;\lambda}(\mathbf{x}) = P_\lambda(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda} \cdot \frac{\lambda^{\sum x_i}}{\prod_{i=1}^n x_i!} = \underbrace{\frac{1}{\prod_{i=1}^n x_i!}}_{h(x)} \cdot \underbrace{e^{-n\lambda} \cdot \lambda^{\sum x_i}}_{g_\lambda(\underbrace{\sum x_i}_{:=T(\mathbf{x})})},$$

azaz a mintaelemek összege elégséges statisztika. ■

Abszolút folytonos mintára az előző definíció nem működik, mivel sok T statisztikára a $\{T(\mathbf{X}) = t\}$ esemény minden t-re 0 valószínűségű, így a feltételes valószínűség nem értelmes. Neyman faktorizációs tétele viszont egy olyan állítást fogalmaz meg, ami abszolút folytonos esetben is értelmes, ha a valószínűség helyett sűrűségfüggvényt írunk.

1.6. Definíció. Legyen $\mathbf{X}$ abszolút folytonos minta, sűrűségfüggvényeinek családja legyen $f_{n;\vartheta}(\mathbf{x})$. A $T(\mathbf{X})$ statisztika *elégséges* a ϑ paraméterre, ha létezik a sűrűségfüggvénynek $f_{n;\vartheta}(\mathbf{x}) = h(\mathbf{x}) \cdot g_\vartheta(T(\mathbf{x}))$ alakú faktorizációja.

1.5. Példa. Legyen $X_i \sim \text{E}(0, b)$, ahol az intervallum jobboldali b végpontja ismeretlen paraméter. Próbáljunk faktorizálni!

$$f_{n;b}(\mathbf{x}) = \prod_{i=1}^n f_{1;b}(x_i) = \prod_{i=1}^n \frac{1}{b} \text{I}(0 \le x_i \le b) = \underbrace{\text{I}(x_1^{(n)} \ge 0)}_{h(\mathbf{x})} \cdot \underbrace{\frac{1}{b^n} \cdot \text{I}(x_n^{(n)} \le b)}_{g_b(x_n^{(n)})},$$

tehát $T(\mathbf{X}) = X_n^{(n)}$ elégséges statisztika. ■

Nyilván, ha T elégséges, akkor annak egy kölcsönösen egyértelmű S függvénye is az, sőt minden olyan S statisztika elégséges, amelyből T kiszámolható. A gyakorlatban minél egyszerűbb, úgynevezett *minimális elégséges* statisztikát keresünk. Ennek a fogalomnak adható precíz matematikai definíciója, de ezzel most nem foglalkozunk.

1.2. Feladat. Keressünk elégséges statisztikát a normális eloszlás paramétereire! Tehát a mintaelemek eloszlása $X_i \sim N(m, \sigma^2)$, és (m, σ^2) ismeretlen kétdimenziós paramétervektor.

$$f_{n;m,\sigma^2}(\mathbf{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum (x_i-m)^2} =$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \cdot e^{-\frac{1}{2\sigma^2}(\sum x_i^2 - 2m\sum x_i + nm^2)} = 1 \cdot g_{m,\sigma^2}(\sum x_i, \sum x_i^2),$$

tehát a $h(\mathbf{x}) = 1$ választással látszik, hogy $T(\mathbf{X}) = (\sum X_i, \sum X_i^2)$ kétdimenziós elégséges statisztika. Ugyanígy elégséges lenne $S(\mathbf{X}) = (\overline{X}, S_X^2)$, hiszen T és S kölcsönösen egyértelmű függvényei egymásnak. ■

1.2. Becslések és jóságuk

Vizsgáljuk meg, hogy egy megfigyelt mintából hogyan lehet a paraméter $\psi(\vartheta)$ függvényét becsülni (gyakran magát a paramétert kell becsülni, de nem mindig)! Nem remélhetjük, hogy a pontos értéket eltaláljuk, de azt igen, hogy jól meg tudjuk közelíteni. A $\psi(\vartheta)$ mennyiség *becslése* alatt valamely $T(\mathbf{X})$ statisztikát értünk. Azért vezetünk be egy új elnevezést a $T(\mathbf{X})$ statisztikára, mert most úgy gondolunk rá, mint a $\psi(\vartheta)$ mennyiséget jól közelítő becslésre. Mit várunk el a $T(\mathbf{X})$ becsléstől?

1. A $T(\mathbf{X})$ becslés nagyjából $\psi(\vartheta)$ körül ingadozzék.
2. $T(\mathbf{X})$ minél kevésbé ingadozzék $\psi(\vartheta)$ körül, azaz a becslés legyen minél pontosabb.
3. Tegyük fel, hogy minden n mintaelemszámra van egy $T_n(X_1,\dots,X_n)$ becslésünk. Szeretnénk, ha ez egyre jobban megközelítené a valódi $\psi(\vartheta)$ értéket, ha a minta elemszáma végtelenhez tart.

Formalizáljuk az első elvárást!

1.7. Definíció. A $T(\mathbf{X})$ becslés *torzítatlan* $\psi(\vartheta)$-ra, ha

$$E_\vartheta(T(\mathbf{X})) = \psi(\vartheta) \quad \forall \vartheta \in \Theta.$$

Általában a $T(\mathbf{X})$ becslés *torzítása* a $b_T(\vartheta) = E_\vartheta(T(\mathbf{X})) - \psi(\vartheta)$ függvény.

Korábban láttuk, hogy a mintaátlag torzítatlan becslés a várható értékre, a korrigált tapasztalati szórásnégyzet pedig a szórásnégyzetre.

1.6. Példa. Indikátor eloszlású mintánál ($p \in (0,1)$) keressünk torzítatlan becslést $\psi(p) = \frac{1}{p}$-re! Belátható, hogy $\frac{1}{\overline{X}}$ *nem* torzítatlan, sőt, $1/p$-t nem lehet torzítatlanul becsülni. Belátjuk ugyanis, hogy $\psi(p)$-t akkor és csak akkor lehet n elemű indikátor-mintából torzítatlanul becsülni, ha $\psi(p)$ p-nek legfeljebb n-edfokú polinomja. Legyen ugyanis T tetszőleges becslés.

$$E_p(T(X_1,\dots,X_n)) = \sum_{\mathbf{x}\in\{0,1\}^n} T(x_1,\dots,x_n)\cdot p^{\sum x_i}\cdot(1-p)^{n-\sum x_i},$$

ez pedig legfeljebb n-edfokú polinomja p-nek. Másrészt p^k egy torzítatlan becslése: $\mathrm{I}(X_1=1,\dots X_k==1)$, $(k\le n)$. ■

1.8. Definíció. $T_n(X_1,\dots,X_n)$ aszimptotikusan torzítatlan becsléssorozat $\psi(\vartheta)$-ra, ha $\forall\vartheta\in\Theta$-ra

$$E_\vartheta(T_n(X_1,\dots,X_n)) \to \psi(\vartheta) \quad (n\to\infty).$$

A gyakorlatban, ha elég nagy a minta, akkor általában egy aszimptotikusan torzítatlan becslés is megfelelő. Például a tapasztalati szórásnégyzet aszimptotikusan torzítatlan a szórásnégyzetre.

Formalizáljuk most a 2. elvárást!

1.9. Definíció. Legyenek T_1, T_2 torzítatlanok $\psi(\vartheta)$-ra. Ekkor azt mondjuk, hogy T_1 *hatásosabb* T_2-nél, ha $D^2_\vartheta(T_1) \le D^2_\vartheta(T_2)$ minden $\vartheta\in\Theta$-ra. A T (torzítatlan) becslés *hatásos*, ha minden torzítatlan becslésnél hatásosabb.

Fontos, hogy két becslés nem biztos, hogy összehasonlítható hatásosság szempontjából, hiszen lehet, hogy bizonyos ϑ-kra $D^2_\vartheta(T_1) < D^2_\vartheta(T_2)$, másokra viszont $D^2_\vartheta(T_1) > D^2_\vartheta(T_2)$.
Nem torzítatlan becslések esetén az átlagos négyzetes veszteséget, azaz az $E_\vartheta[(T-\psi(\vartheta))^2]$ mennyiséget akarhatjuk minimalizálni.

1.3. Tétel. *Ha T_1 és T_2 is hatásos, akkor 1 valószínűséggel megegyeznek, azaz $P_\vartheta(T_1 = T_2) = 1$ minden $\vartheta\in\Theta$ esetén.*

Bizonyítás. A torzítatlanság miatt $E_\vartheta(T_1) = E_\vartheta(T_2) = \psi(\vartheta)$, és mivel mindkét becslés hatásos, $D^2_\vartheta(T_1) = D^2_\vartheta(T_2)$ minden ϑ-ra. Legyen most

$$T = \frac{T_1+T_2}{2}.$$

Egyrészt T is torzítatlan, hiszen $E_\vartheta(T) = \psi(\vartheta)$, másrészt T_1 hatásossága miatt

$$D^2_\vartheta(T_1) \le D^2_\vartheta(T) = \frac{1}{4}\left(D^2_\vartheta(T_1) + D^2_\vartheta(T_2) + 2\mathrm{cov}_\vartheta(T_1,T_2)\right) = \frac{1}{2}D^2_\vartheta(T_1) + \frac{1}{2}\mathrm{cov}_\vartheta(T_1,T_2),$$

azaz $D^2_\vartheta(T_1) \le \mathrm{cov}_\vartheta(T_1,T_2)$. Átosztva kapjuk, hogy

$$1 \le \frac{\mathrm{cov}_\vartheta(T_1,T_2)}{D_\vartheta(T_1)D_\vartheta(T_2)} = R_\vartheta(T_1,T_2),$$

azaz az ismert tétel szerint $T_1 = aT_2 + b$ teljesül 1 valószínűséggel. A várható értékek és szórások egyezése miatt azonban $a = 1$ és $b = 0$ lehet csak. ■

1.4. Tétel. *Legyen $\mathbf{X} = (X_1,\dots,X_n)$ független, azonos eloszlású minta. Legyen $\psi(\vartheta) = E_\vartheta(X_i)$, továbbá tegyük fel, hogy $D^2_\vartheta(X_i) < \infty$ minden ϑ-ra. Ekkor $\overline{X}$ hatásosabb becslése $\psi(\vartheta)$-nak minden $\sum_{i=1}^n c_iX_i$ alakú torzítatlan becslésnél.*

Bizonyítás. Vegyük először észre, hogy $\sum_{i=1}^n c_iX_i$ akkor és csak akkor torzítatlan, ha $\sum_{i=1}^n c_i = 1$. Számítsuk ki a szórásnégyzeteket!

$$D^2_\vartheta(\overline{X}) = \frac{D^2_\vartheta(X_i)}{n}, \quad D^2_\vartheta(\sum c_iX_i) = \sum D^2_\vartheta(c_iX_i) = (\sum c_i^2)D^2_\vartheta(X_i).$$

Azt kell tehát belátni, hogy

$$\frac{1}{n} \le \sum_{i=1}^n c_i^2.$$

A számtani és négyzetes közép közötti egyenlőtlenségből

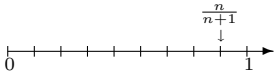
$$\sqrt{\frac{\sum c_i^2}{n}} \ge \frac{\sum c_i}{n} = \frac{1}{n}, \text{ azaz } \sum c_i^2 \ge \frac{n}{n^2} = \frac{1}{n}. \quad ■$$

1.7. Példa. Legyen $X_i \sim E(0,b)$, és vizsgáljuk a következő két becslést b-re:

$$T_1 = \frac{n+1}{n}X^{(n)}_n, \quad T_2 = 2\cdot\overline{X}.$$

T_2 nyilván torzítatlan, és T_1 is az: legyen ugyanis $X_i = b\cdot Y_i$, ahol $Y_i \sim E(0,1)$, ekkor $X^{(n)}_n = b\cdot Y^{(n)}_n$, tehát elég az $E(0,1)$ eloszlással foglalkozni. $Y^{(n)}_n$ eloszlásfüggvénye: $P(Y^{(n)}_n < t) = t^n$, $Y^{(n)}_n$ sűrűségfüggvénye: $n\cdot t^{n-1}$, tehát

$$E(Y^{(n)}_n) = \int_0^1 t\cdot nt^{n-1}\,\mathrm{d}t = n\cdot\frac{1}{n+1} = \frac{n}{n+1}.$$



Ebből $E_b(T_1) = b$. Melyik becslés a hatásosabb?

$$D^2_b(T_2) = 4\cdot\frac{D^2_b(X_i)}{n} = \frac{4}{n}\cdot\frac{b^2}{12} = \frac{b^2}{3n}.$$

Másrészt

$$D^2(Y^{(n)}_n) = \int_0^1 t^2\cdot n\cdot t^{n-1}\,\mathrm{d}t - \left(\frac{n}{n+1}\right)^2 = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2 =$$
$$= \frac{n(n+1)^2 - n^2(n+2)}{(n+1)^2(n+2)} = \frac{n}{(n+1)^2(n+2)}.$$

Továbbá $D_b^2(X_n^{(n)}) = b^2 D_b^2(Y_n^{(n)})$, azaz

$$D_b^2(T_1) = \left(\frac{n+1}{n}\right)^2 \cdot D_b^2(X_n^{(n)}) = \frac{b^2}{n(n+2)}.$$

Kaptuk tehát, hogy T_1 hatásosabb T_2-nél (minden n-re). ■

Formalizáljuk végül a 3. elvárást!

1.10. Definíció. A $T_n(X_1, \ldots, X_n)$ becléssorozat *konzisztens* $\psi(\vartheta)$-ra, ha $T_n \to \psi(\vartheta)$ sztochasztikusan ($n \longrightarrow \infty$), azaz

$$P_\vartheta(|T_n - \psi(\vartheta)| > \varepsilon) \overset{n\to\infty}{\longrightarrow} 0 \quad \forall \vartheta \in \Theta.$$

1.8. Példa. a) A mintaátlag konzisztens becslés a várható értékre: $\overline{X} \to E_\vartheta(X_i)$ sztochasztikusan (ez a nagy számok gyenge törvénye).
b) Ha $E_\vartheta(T_n) = \psi(\vartheta)$ (azaz T_n torzítatlan becslés, de az is elég lenne, hogy aszimptotikusan torzítatlan) és $D_\vartheta^2(T_n) \to 0$ ($n \to \infty$), akkor T_n konzisztens, mivel

$$P_\vartheta(|T_n - \psi(\vartheta)| > \varepsilon) \overset{\text{Cseb}}{\leq} \frac{D_\vartheta^2(T_n)}{\varepsilon^2} \longrightarrow 0.$$

c) Legyen $X_i \sim E(0, \vartheta)$, és $T_n = \frac{n+1}{n} X_n^{(n)}$. Ez a fentiek szerint konzisztens becléssorozat, hiszen

$$D_\vartheta^2\left(\frac{n+1}{n} \cdot X_n^{(n)}\right) = \frac{\vartheta^2}{n(n+2)} \longrightarrow 0. \quad ■$$

1.3. Maximum likelihood becslés

Korábban megnéztük, hogy milyen általános becslési módszerek léteznek. Megismerkedtünk a tapasztalati becslésekkel, a momentum módszerrel, és a maximum likelihood módszerrel. Ez utóbbit, fontossága miatt, ismétlésként tekintsük újra át, illetve ismerkedjünk meg vele közelebbről!

1.11. Definíció. Legyen az $\mathbf{X} = (X_1, \ldots, X_n)$ minta eloszlásfüggvényeinek családja $F_{n;\vartheta}$. Ekkor az $L_n(\vartheta) = L_n(\vartheta, \mathbf{x})$ *likelihood függvényt* a következőképpen definiáljuk: abszolút folytonos minta esetén $L_n(\vartheta; \mathbf{x}) = f_{n;\vartheta}(\mathbf{x})$, diszkrét minta esetén $L_n(\vartheta; \mathbf{x}) = p_{n;\vartheta}(\mathbf{x})$.

1.12. Definíció. Legyen $\mathbf{x} = (x_1, ..., x_n)$ egy független, azonos eloszlású minta realizációja, Θ a paramétertér. Ekkor a ϑ paraméter *maximum likelihood (ML) becslése* $\hat{\vartheta}$, ha

$$L_n(\hat{\vartheta}) = \max\{L_n(\vartheta) : \vartheta \in \Theta\},$$

azaz a likelihood függvény maximumhelye. Tehát az a paraméter lesz a becslésünk, amely mellett a mintánk valószínűsége (illetve sűrűsége) a lehető legnagyobb. A paraméter egy $\psi(\vartheta)$ függvényének ML becslése $\psi(\hat{\vartheta})$.

Elképzelhető, hogy a keresett maximumhely nem létezik, vagy ha létezik, nem egyértelmű. Ha $L_n(\vartheta)$ differenciálható, akkor a maximumhelyet a $(\ln L_n)'(\vartheta) = 0$ *likelihood egyenlet* megoldásaként szokás keresni.

1.5. Tétel. *Ha létezik ML becslés, akkor az megadható a T elégséges statisztika függvényeként.*

Bizonyítás. A faktorizációs tétel alapján $L_n(\vartheta) = h(\mathbf{x}) \cdot g_\vartheta(T(\mathbf{x}))$, ahol a $h(\mathbf{x})$ tényező nem játszik szerepet a ϑ szerinti maximumhely keresésében. Azaz a maximumhelyek halmaza csak $T(\mathbf{x})$-től függ. ■

A ML becslés általában nem torzítatlan. Azonban a következő tétel szerint bizonyos erős feltételek mellett a ML becslésnek „jó” aszimptotikus tulajdonságai vannak. Ez az egyik oka annak, hogy ez a módszer a legelterjedtebb a gyakorlatban.

1.6. Tétel. *Bizonyos feltételek mellett elég nagy n-re a $\hat{\vartheta}_n$ ML becslés létezik, konzisztens, aszimptotikusan torzítatlan, továbbá aszimptotikusan normális eloszlású:*

$$\sqrt{n} \cdot (\hat{\vartheta}_n - \vartheta) \to N(0, \sigma^2(\vartheta))$$

eloszlásban ($n \to \infty$), ahol a $\sigma^2(\vartheta)$ aszimptotikus szórásnégyzet a „lehető legkisebb".

1.9. Példa. $X_i \sim E(a, b)$, adjuk meg a paraméterek ML becslését!

$$L_n(a, b; \mathbf{x}) = \left(\frac{1}{b-a}\right)^n \cdot \mathrm{I}(a \le x_i \le b \quad \forall i),$$

azaz a legkisebb olyan $[a, b]$ intervallumot keressük, amely mindegyik megfigyelést tartalmazza. Ennek megoldása $\hat{a} = x_1^{(n)}$, $\hat{b} = x_n^{(n)}$. ■

1.10. Példa. $X_i \sim \mathrm{N}(m, \sigma^2)$, adjuk meg a paraméterek ML becslését!

$$L_n(m, \sigma; \mathbf{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}}\right)^n \cdot \frac{1}{\sigma^n} \cdot e^{-\frac{\sum(x_i-m)^2}{2\sigma^2}},$$

ez a paraméterek differenciálható függvénye. Vegyünk logaritmust:

$$\ln L_n(m, \sigma; \mathbf{x}) = \ln\left(\frac{1}{\sqrt{2\pi}}\right)^n - n \ln \sigma - \frac{\sum(x_i - m)^2}{2\sigma^2}.$$

A likelihood egyenlet most két egyenletből áll, ugyanis mindkét parciális deriváltnak nullának kell lenni:

$$\frac{\partial}{\partial m} \ln L_n(m, \sigma; \mathbf{x}) = \frac{\sum(x_i - m)^2}{\sigma^2} = 0$$

és

$$\frac{\partial}{\partial \sigma} \ln L_n(m, \sigma; \mathbf{x}) = -\frac{n}{\sigma} - \frac{2\sum(x_i - m)^2}{2\sigma^3} = 0.$$

A két egyenlet megoldása $\hat{m} = \overline{x}$ és $\hat{\sigma}^2 = S_x^2$, és ezek valóban maximumhelyet határoznak meg. ■

1.4. Bayes becslések

Az eddigiekben nem vettük figyelembe a becslés előtt esetlegesen már meglévő információkat a paraméter hozzávetőleges értékéről (ezt *a priori* információnak nevezzük). Ha például egy cinkelt érmével 10 dobásból csupa fej jött ki, akkor a fejdobás ismeretlen p valószínűségének ML becslése $\hat{p} = 1$, ami nem tűnik reálisnak. Hogyan tudnánk a becslésbe belekalkulálni azt az előzetes feltevést, hogy a fejdobás valószínűsége azért $1/2$ körül van nagy eséllyel?

Erre ad módszert a Bayes-i hozzáállás. Feltesszük, hogy a paraméter maga is egy valószínűségi változó, azaz van egy *a priori eloszlása*. A paraméter adott értéke mellett tudjuk a minta eloszlását. Ez a kettő már meghatározza a paraméter és a minta együttes eloszlását. Az együttes eloszlásból pedig megkaphatjuk a paraméter *a posteriori eloszlását*, vagyis az a priori eloszlásnak a megfigyelt minta által módosított változatát.

1.11. Példa. Valaki feldobott néhány szabályos érmét, de nem tudjuk, hányat. Azt tudjuk, hogy 3 érmén lett fej az eredmény. Tegyük fel, hogy az érmék ismertetlen ϑ számának a priori eloszlása egyenletes az $\{1, 2, 3, 4, 5, 6\}$ halmazon. Mi lesz az a posteriori eloszlás?
Jelölje a ϑ paraméter egy lehetséges értékét t. Az a priori valószínűségek:

$$q(t) = P(\vartheta = t) = \frac{1}{6}, \quad t = 1, 2, 3, 4, 5, 6.$$

A megfigyelésünk $x = 3$, ennek valószínűsége a t paraméterérték mellett:

$$f_t(x) = f_t(3) = P_t(X = 3) = P(X = 3|\vartheta = t) = \binom{t}{3}\left(\frac{1}{2}\right)^t.$$

A Bayes tétel segítségével kapjuk meg az a posteriori eloszlást:

$$q^*(t|x) = q^*(t|3) = P(\vartheta = t|X = 3) = \frac{P(X = 3|\vartheta = t)P(\vartheta = t)}{\sum_s P(X = 3|\vartheta = s)P(\vartheta = s)}.$$

Ezeket az értékeket kiszámolva, a következő a posteriori valószínűségeket kapjuk:

t	1	2	3	4	5	6
$q^*(t\|3)$	0	0	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{5}{16}$	$\frac{5}{16}$

A következő kérdés, hogy mi legyen a ϑ paraméter becslése? Az tűnik jó választásnak, ha az a posteriori eloszlás valamilyen középértékét vesszük, pl. a várható értékét vagy a mediánját. A leggyakrabban a várható értéket szokás venni, vagyis a

$$\hat{\vartheta} = T(\mathbf{x}) = E(\vartheta|\mathbf{X} = \mathbf{x})$$

érték lesz a ϑ paraméter Bayes becslése.

Legyen egy T becslés *a priori rizikója* (négyzetes veszteségfüggvény mellett)

$$E\left[(T(\mathbf{X}) - \vartheta)^2\right],$$

ez azt méri, hogy várhatóan mennyi a paraméter és annak becslése közötti négyzetes eltérés. Megmutatható, hogy ezt a rizikót éppen a Bayes becslés minimalizálja. A módszer abban az esetben is működik, ha a paraméter egy $\psi(\vartheta)$ függvényét szeretnénk becsülni, ennek Bayes becslése

$$\hat{\psi}(\vartheta) = T(\mathbf{x}) = E(\psi(\vartheta)|\mathbf{X} = \mathbf{x}).$$

Visszatérve a 1.11. Példára, a feldobott érmék számának Bayes becslése

$$\hat{\vartheta} = T(3) = E(\vartheta|\mathbf{X} = 3) = \sum_t t \cdot q^*(t|3) = \frac{77}{16} = 4{,}81.$$

A Bayes becslés egyik hátrányára is rávilágít a példa: nyilván lehetetlen, hogy 4,81 érmét dobtak fel, tehát olyan becslést kaptunk, ami nincs is benne a paramétertérben. Ez pl. a maximum likelihood becslésnél nem fordulhat elő (a ML becslés ebben a példában nem egyértelmű, $\hat{\vartheta} = 5$ és $\hat{\vartheta} = 6$ egyaránt megfelel). Ha az a posteriori eloszlás mediánját választjuk becslésnek, akkor a $\hat{\vartheta} = 5$ becslést kapjuk.

1.12. Példa. Egy doboz desszertben 18, külsőre egyforma édesség van, de kétféle töltéssel: van köztük nugátos és van marcipános. Nem tudjuk, hány marcipános van a dobozban. 8 darabot megettünk, ebből 5 volt marcipános. Adjunk Bayes-becslést a dobozban eredetileg levő marcipános desszertek számára, ha az a priori eloszlás 18 rendű és 1/3 paraméterű binomiális!

Jelölje most is az ismeretlen paramétert (hány marcipános volt eredetileg) ϑ, annak egy lehetséges értékét pedig t. Az a priori eloszlás szerint $\vartheta \sim \mathrm{Bin}(18, 1/3)$, azaz a valószínűségek:

$$q(t) = P(\vartheta = t) = \binom{18}{t}(1/3)^t(2/3)^{18-t}.$$

Most a megfigyelésünk $x = 5$, ennek valószínűsége a t paraméterérték mellett:

$$f_t(x) = f_t(5) = P_t(X = 5) = P(X = 5|\vartheta = t) = \frac{\binom{t}{5}\binom{18-t}{3}}{\binom{18}{8}}.$$

A Bayes tétel segítségével kapjuk meg az a posteriori eloszlást:

$$q^*(t|x) = q^*(t|5) = P(\vartheta = t|X = 5) = \frac{P(X = 5|\vartheta = t)P(\vartheta = t)}{\sum_s P(X = 5|\vartheta = s)P(\vartheta = s)} =$$

$$K(x)\binom{10}{t-5}(1/3)^{t-5}(2/3)^{10-(t-5)},$$

ahol $K(x)$ egy t-t nem tartalmazó konstans. Felismerhetjük, hogy a mintára feltételesen a ϑ paraméter eloszlása $5 +$ $+ \mathrm{Bin}(10, 1/3)$, ez tehát az a posteriori eloszlás. A dobozban eredetileg lévő marcipános desszertek számának Bayes becslése

$$\hat{\vartheta} = T(5) = E(\vartheta|\mathbf{X} = 5) = 5 + 10 \cdot \frac{1}{3} = \frac{25}{3} = 8{,}33.$$

Az eredmény nem meglepő: előzetes feltételezésünk (az a priori eloszlás) éppen azt fejezi ki, hogy a 18 „kosárkát" a dobozban úgy töltötték meg, hogy mindegyikbe, egymástól függetlenül, 1/3 valószínűséggel tettek marcipánost, és 2/3 valószínűséggel nugátost. Megettünk 8-at, ebből 5 volt marcipános, tehát ennyi már biztosan volt a dobozban. A maradék 10 édességről viszont nem tudtunk meg semmit, hiszen azok függetlenek a megevett 8-tól. Így azokról továbbra is csak azt tudjuk, hogy mindegyik 1/3 valószínűséggel marcipános. ■

A fenti példában a paraméter és a minta is diszkrét értékű. A gyakorlatban előforduló példákban természetesen lehetséges, hogy a minta és/vagy a paraméter folytonos skálán mozog. Mi a feltételes valószínűségeket csak pozitív valószínűségű feltétel esetén definiáltuk, de ez a definíció kiterjeszthető az általános esetre is.

Ha például a paraméter folytonos, a minta diszkrét, akkor legyen $q(t)$ a paraméter a priori sűrűségfüggvénye, $f_t(\mathbf{x})$ a minta eloszlása a t paraméterérték mellett, ekkor az a posteriori sűrűségfüggvény:

$$q^*(t|\mathbf{x}) = \frac{f_t(\mathbf{x})q(t)}{\int f_s(\mathbf{x})q(s)\mathrm{d}s} = K(\mathbf{x})f_t(\mathbf{x})q(t),$$

és a Bayes becslés ennek várható értéke:

$$\hat{\vartheta} = \int t q^*(t|\mathbf{x})\,\mathrm{d}t = K(\mathbf{x}) \int t f_t(\mathbf{x})q(t)\,\mathrm{d}t\,.$$

A nevezetes diszkrét eloszlások esetén az ismeretlen paraméter legtöbbször vagy $[0, 1]$-beli (a binomiális, geometriai, negatív binomiális eloszlás p varamétere) vagy pozitív (a Poisson eloszlás λ paramétere). Milyen abszolút folytonos eloszlásokat ismerünk ezeken a tartományokon?

A [0,1] intervallumon az egyenletes eloszlást ismerjük. Ennek általánosítása a béta eloszlások családja. Az Béta(a, b) eloszlás sűrűségfüggvénye

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1,$$

egyébként pedig 0. Itt a és b pozitív paraméterek, $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x$ pedig normáló konstans. Az $a = b = 1$ választással visszakapjuk az egyenletes eloszlást.

Hogy néznek ezek ki? (Ábra)

Az egyenletes rendezett minta k-adik tagja Béta$(k, n+1-k)$ eloszlású. Ennek vázlatos levezetése a következő: jelölje a k-adik rendezett mintaelem eloszlásfüggvényét $F^{(k)}(x)$. Ekkor pici h-ra

$$F^{(k)}(x+h) - F^{(k)}(x) = P(x < X_k^{(n)} < x + h) \approx nh \cdot \binom{n-1}{k-1} x^{k-1} \cdot (1-x-h)^{n-k},$$

ahol elhanyagoltuk annak a valószínűségét, hogy az $(x, x+h)$ intervallumba legalább két mintaelem esik (ennek a valószínűségnek a nagyságrendje h^2). A sűrűségfüggvény az eloszlásfüggvény deriváltja, tehát

$$f^{(k)}(x) = \lim_{h\to 0} \frac{F^{(k)}(x+h) - F^{(k)}(x)}{h} = n\binom{n-1}{k-1} x^{k-1} \cdot (1-x)^{n-k}.$$

Ebből megkaptuk, hogy pozitív egész a, b esetén a béta eloszlás normáló konstansa

$$B(a,b) = \frac{(a-1)!\,(b-1)!}{(a+b-1)!}.$$

Számítsuk ki a béta eloszlás várható értékét (pozitív egész paraméterek esetén)! Legyen $X \sim$ $\sim$ Béta(a, b).

$$E(X) = \int_0^1 x \cdot \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = \frac{B(a+1,b)}{B(a,b)} = \frac{a}{a+b}.$$

Hasonlóan számítható ki, hogy szórásnégyzete

$$D^2(X) = \frac{ab}{(a+b)^2(a+b+1)}.$$

1.13. Példa. Térjünk vissza a szakasz elején tárgyalt példára. Egy érmével tízszer dobunk, és szeretnénk az eredmény alapján megbecsülni a fejdobás ismeretlen ϑ valószínűségét. Azt kaptuk, hogy a fejdobások száma $x = 10$, vagyis csupa fejet dobtunk. Legyen az a priori eloszlás Béta(2,2), ennek várható értéke 1/2, szórása 0,22. Ekkor az a posteriori sűrűségfüggvény

$$q^*(t|10) = K(10)f_t(10)q(t) = K(10) \cdot t^{10} \cdot \frac{1}{6}t(1-t) = C \cdot t^{11}(1-t)^1,$$

ahol C konstans. Felismerjük, hogy ez a Béta$(12, 2)$ eloszlás sűrűségfüggvénye, azaz a Bayes becslés

$$\hat{\vartheta} = \frac{12}{12+2} = 0{,}86. \quad \blacksquare$$

Nézzük most meg, milyen eloszlást ismerünk a pozitív félegyenesen? Eddig az exponenciális eloszlást tanultuk, ebből konvolúcióval lehet új eloszlásokat csinálni. Először számítsuk ki két független, azonos paraméterű exponenciális eloszlású valószínűségi változó összegének eloszlását:

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x = \int_0^z \lambda e^{-\lambda x} \cdot \lambda e^{-\lambda(z-x)}\,\mathrm{d}x = \lambda^2 z e^{-\lambda z}.$$

Ezt iterálva megkapjuk, hogy n darab független, λ paraméterű exponenciális eloszlású valószínűségi változó összegének sűrűségfüggvénye:

$$f(x) = \frac{1}{(n-1)!}\lambda^n x^{n-1} e^{-\lambda x} \quad x > 0.$$

Ezt nevezzük n rendű, λ paraméterű gamma eloszlásnak, jelölésben $X \sim \text{Gamma}(n, \lambda)$. nyilván $E(X) = \frac{n}{\lambda}$ és $D^2(X) = \frac{n}{\lambda^2}$. Általánosabban is definiálható a gamma eloszlás, mégpedig úgy, hogy a rendje tetszőleges pozitív szám. A Gamma(a, λ) eloszlás sűrűségfüggvénye:

$$f(x) = \frac{1}{\Gamma(a)}\lambda^a x^{a-1} e^{-\lambda x} \quad x > 0,$$

ahol $\Gamma(a)$ olyan normáló konstans, hogy az $f(x)$ függvény integrálja 1 legyen. A várható érték és szórásnégyzet képlete érvényben marad, csak n helyébe a-t kell írni.

A definícióból nzilvánvaló, hogy egész n, m esetén a Gamma(n, λ) és a Gamma(m, λ) eloszlás konvolúciója a Gamma$(n + m, \lambda)$ eloszlás. Általában is érvényes, hogy azonos paraméterű gamma eloszlások konvolúciója (ugyanolyan paraméterű) gamma, a rendek összeadódnak.

1.14. Példa. Legyen az $\mathbf{X}$ minta Poisson eloszlású, ahol a λ paraméter a priori eloszlása Gamma(a, μ). Ekkor az a posteriori eloszlás is Gamma lesz, mégpedig $a+\sum_i x_i$ renddel és $\mu+n$ paraméterrel, vagyis a λ Bayes becslése

$$\hat{\lambda} = \frac{a + \sum_i x_i}{\mu + n}.$$

Látható, hogy a Bayes becslés az a priori információ és a minta kombinációja: az a priori várható érték a/μ, a minta tapasztalati várható értéke $\sum_i x_i/n$, a Bayes becslés pedig valahol a kettő között van. $\blacksquare$

1.5. Intervallum becslések

Eddig úgynevezett *pontbecslésekkel* foglalkoztunk, azaz a paramétert (vagy annak függvényét) egyetlen értékkel becsültük meg. A pontbecslés bizonytalanságát a becslés szórásával fejezhetjük ki. Ha például a becslésről belátható, hogy aszimptotikusan normális eloszlású, akkor az igazi paraméter kb. 95% valószínűséggel a pontbecslés körüli 2·szórás sugarú intervallumban van. A becslésben rejlő bizonytalanságot kifejezhetjük úgy is, hogy a paramétert nem egy értékkel, hanem egy intervallummal becsüljük.

1.13. Definíció. Legyen $\mathbf{X} = (X_1, \ldots, X_n)$ minta F_ϑ eloszlásból, ahol ϑ valós paraméter. Azt mondjuk, hogy a $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ intervallum *legalább* $1-\alpha$ *megbízhatósági szintű konfidenciaintervallum* ϑ*-ra* (röviden KI$(1-\alpha)$, ha

$$P_\vartheta\left(T_1(\mathbf{X}) < \vartheta < T_2(\mathbf{X})\right) \geq 1-\alpha \quad \forall \vartheta.$$

Megjegyzés: A KI pontos megbízhatósági szintje

$$\inf_{\vartheta\in\Theta} \left\{ P_\vartheta(\vartheta \in (T_1, T_2)) \right\}.$$

1.7. Tétel. *Ha* (T_1, T_2) *KI*$(1-\alpha)$ ϑ*-ra, akkor* (S_1, S_2) *KI*$(1-\alpha)$ $\psi(\vartheta)$*-ra, ahol*

$$S_1 = \inf\left\{ \psi(\vartheta) \,|\, \vartheta \in (T_1, T_2) \right\} \quad S_2 = \sup\left\{ \psi(\vartheta) \,|\, \vartheta \in (T_1, T_2) \right\}. \quad \blacksquare$$

1.15. Példa. Legyen $\mathbf{X}$ az $\mathrm{E}(0, \vartheta)$ eloszlásból származó n elemű minta. Adjunk KI-t ϑ-ra! Láttuk, hogy a legnagyobb mintaelem elégséges statisztika, tehát ésszerűnek tűnik ennek függvényében keresni a KI-t. Mivel $X_n^{(n)}$ biztosan kisebb ϑ-nál, ezt választhatjuk a KI bal végpontjának, a jobb végpontot pedig keressük $c_n X_n^{(n)}$ alakban!

$$1-\alpha = P_\vartheta(X_n^{(n)} < \vartheta < c_n X_n^{(n)}) = P_\vartheta(X_n^{(n)} > \frac{\vartheta}{c_n}) = 1 - P_\vartheta(X_n^{(n)} < \frac{\vartheta}{c_n}) = 1 - \left(\frac{\vartheta/c_n}{\vartheta}\right)^n = 1 - \frac{1}{c_n^n}.$$

Ennek megoldása $c_n = 1/\alpha^{1/n}$. $\blacksquare$

Az egyik legfontosabb (és legszebb) eset a normális eloszlás várható értékére KI konstruálása, ismert vagy ismeretlen szórás mellett. Nézzük meg ezeket!

1.16. Példa. Legyen $X_i \sim N(m, \sigma^2)$, ahol σ ismert, m ismeretlen. Adjunk m-re $KI(1-\alpha)$-t! Kiindulásként vegyük észre, hogy $\overline{X} \sim N(m, \frac{\sigma^2}{n})$, azaz $\sqrt{n} \cdot \dfrac{\overline{X} - m}{\sigma} \sim N(0{,}1)$. Ezért

$$P\left(\left|\sqrt{n} \cdot \frac{\overline{X} - m}{\sigma}\right| < u_{\frac{\alpha}{2}}\right) = 1-\alpha,$$

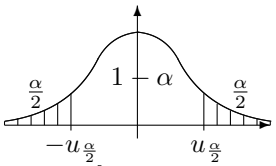
ahol u_α az az érték, melyre $\Phi(u_\alpha) = 1-\alpha$, ezt táblázatból nézhetjük ki. Érdemes bevezetnünk erre egy elnevezést.

1.14. Definíció. Legyen F egy tetszőleges eloszlásfüggvény, és legyen $0 < \gamma < 1$. Az eloszlás γ-kvantilise x_γ, ha

$$F(x_\gamma) \leq \gamma \text{ és } F(x_\gamma + 0) \geq \gamma.$$

(Emlékeztető: $F(x+0)$ a függvény jobboldali határértéke az x pontban.) Belátható, hogy a γ-kvantilis mindig létezik, de nem feltétlenül egyértelmű.

Ezzel a definícióval u_α a standard normális eloszlás $(1-\alpha)$-kvantilise.



Ebből megkaphatjuk a KI alsó és felső határát:

$$\frac{\sqrt{n}}{\sigma}\cdot\left|\overline{X}-m\right|<u_{\frac{\alpha}{2}} \iff \left|\overline{X}-m\right|<\frac{\sigma\cdot u_{\frac{\alpha}{2}}}{\sqrt{n}} \iff \overline{X}-\frac{\sigma\cdot u_{\frac{\alpha}{2}}}{\sqrt{n}}<m<\overline{X}+\frac{\sigma\cdot u_{\frac{\alpha}{2}}}{\sqrt{n}}.$$

Azaz kaptuk, hogy

$$T_1=\overline{X}-\frac{\sigma\cdot u_{\frac{\alpha}{2}}}{\sqrt{n}},\quad T_2=\overline{X}+\frac{\sigma\cdot u_{\frac{\alpha}{2}}}{\sqrt{n}}.$$

A KI hossza három dologtól függ: 1) a kívánt megbízhatóság (nagyobb megbízhatósághoz hosszabb KI kell), 2) az ismert szórás (ha a mintaelemek szórása nagyobb, a KI is hosszabb lesz), 3) a minta elemszáma (nagyobb minta esetén rövidebb a KI). ■

Ha a σ szórás nem ismert, akkor nehezebb dolgunk van. A megoldáshoz meg kell ismernünk két új eloszlást, a χ^2-eloszlást és a t-eloszlást.

1.15. Definíció. Legyenek $X_i\sim N(0,1)$ függetlenek, és $Y=\sum_{i=1}^{n}X_i^2$. Az Y valószínűségi változó eloszlását *n szabadságfokú khi-négyzet eloszlásnak* nevezzük, jelölés: $Y\sim\chi_n^2$. Továbbá $\sqrt{Y}$ eloszlását *n szabadságfokú khi eloszlásnak* nevezzük, jelölés: $\sqrt{Y}\sim\chi_n$.

Valójában a khi-négyzet eloszlás nem új. Számítsuk ki az 1 szabadsági fokú khi-négyzet eloszlás sűrűségfüggvényét! Ehhez először az eloszlásfüggvényt számoljuk ki, legyen tehát $Y=X^2$, ahol X standard normális. Legyen $x>0$.

$$F(x)=P(Y<x)=P(|X|<\sqrt{x})=P(-\sqrt{x}<X<\sqrt{x})=\Phi(\sqrt{x})-\Phi(-\sqrt{x})=2\Phi(\sqrt{x})-1.$$

Deriválással kapjuk a sűrűségfüggvényt:

$$f(x)=F'(x)=2\varphi(\sqrt{x})\frac{1}{2\sqrt{x}}=\frac{1}{\sqrt{2\pi x}}e^{-x/2}=\frac{1}{\sqrt{2\pi}}x^{-\frac{1}{2}}e^{-\frac{1}{2}x},$$

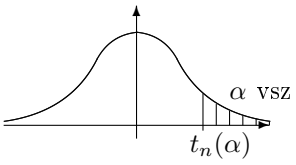
ami éppen a Gamma$(\frac{1}{2},\frac{1}{2})$ eloszlás sűrűségfüggvénye. Ebből konvolúcióval következik, hogy a χ_n^2 eloszlás megegyezik a Gamma$(\frac{n}{2},\frac{1}{2})$ eloszlással.

1.16. Definíció. Legyen $X\sim N(0,1)$, $Y\sim\chi_n$ függetlenek. Legyen $Z=\sqrt{n}\cdot\dfrac{X}{Y}$. Ekkor a Z valószínűségi változó eloszlását *n szabadságfokú t eloszlásnak, vagy n szabadságfokú Student eloszlásnak* nevezzük, jelölés: $Z\sim t_n$.

A t_n eloszlás sűrűségfüggvénye is kiszámítható, de pontos alakjára nem lesz szükségünk. Könnyen látszik, hogy a sűrűségfüggvény szimmetrikus, azaz $E(Z)=0$ $(n>1)$. Megmutatható, hogy $D^2(Z)=\frac{n}{n-2}$ $(n>2)$. A t_n eloszlás $n\to\infty$ esetén a standard normálishoz tart, de vastagabb a farka (sűrűségfüggvénye nagy x-re kb. $c_nx^{-(n+1)}$).

A következő fontos és érdekes tételt nem bizonyítjuk.

1.8. Tétel. (Fisher-Bartlett) *Legyen $X_i\sim N(m,\sigma^2)$ független minta $(i=1,\dots,n)$. Ekkor $\overline{X}$ és S_n^* függetlenek, és $\dfrac{(n-1)S_n^{*2}}{\sigma^2}=\dfrac{nS_n^2}{\sigma^2}\sim\chi_{n-1}^2$.*



1. ábra. A t_n eloszlás sűrűségfüggvénye, és a $t_n(\alpha)$ kritikus érték.

Megjegyzés. Legyen $X_i \sim N(m, \sigma^2)$, és $Y_i = (X_i - m)/\sigma$. Ekkor $\dfrac{(n-1)S_n^{*2}}{\sigma^2} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$. Mivel az $(Y_i - \overline{Y})$ valószínűségi változók nem függetlenek (összegük nulla), eggyel csökken a szabadsági fok. Érdekesség, hogy a mintaátlag és a tapasztalati szórásnégyzet függetlensége karakterizálja a normális eloszlást.

1.17. Példa. Legyen $X_i \sim N(m, \sigma^2)$, és most m, σ ismeretlenek. Adjunk m-re $KI(1-\alpha)$-t!
A Fisher-Bartlett tétel szerint

$$\sqrt{n} \cdot \frac{\overline{X} - m}{S_n^*} = \sqrt{n-1} \cdot \frac{\sqrt{n} \cdot \dfrac{\overline{X} - m}{\sigma} \sim \mathrm{N}(0,1)}{\dfrac{\sqrt{n-1} \cdot S_n^*}{\sigma} \sim \chi_{n-1}} \sim t_{n-1}.$$

Azaz

$$P\left(\left| \sqrt{n} \cdot \frac{\overline{X} - m}{S_n^*} \right| < t_{n-1}(\alpha/2) \right) = 1 - \alpha,$$

ahol a $t_{n-1}(\alpha/2)$ kritikus érték a t_{n-1} eloszlás $(1-\alpha/2)$-kvantilise. Ebből a KI

$$T_1 = \overline{X} - \frac{S_n^* \cdot t_{n-1}(\frac{\alpha}{2})}{\sqrt{n}} \quad T_2 = \overline{X} + \frac{S_n^* \cdot t_{n-1}(\frac{\alpha}{2})}{\sqrt{n}}. \quad \blacksquare$$

1.18. Példa. Legyen $X_i \sim \mathrm{Ind}(p)$. Adjunk p-re hozzávetőleges (aszimptotikus) $KI(1-\alpha)$-t!
Kiindulásként vegyük észre, hogy $\overline{X}$ jó becslés p-re, és $\overline{X} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$, ha n elég nagy.

$$\sqrt{n} \cdot \frac{\overline{X} - p}{\sqrt{p(1-p)}} \approx \mathrm{N}(0,1) \Rightarrow P\left(\frac{\sqrt{n}}{\sqrt{p(1-p)}} \cdot \left|\overline{X} - p\right| < u_{\frac{\alpha}{2}} \right) \approx 1 - \alpha.$$

A zárójelben álló egyenlőtlenséget kell most p-re átrendezni. Más módszer: a nevezőben lévő p helyébe $\overline{X}$-et írunk, és a

$$\frac{\sqrt{n}}{\sqrt{\overline{X}(1-\overline{X})}} \cdot \left|\overline{X} - p\right| < u_{\frac{\alpha}{2}}$$

egyenlőtlenséget rendezzük át. $\blacksquare$

1.6. Statisztikai próbák és jóságuk

A statisztikai próba olyan eljárás, mellyel eldöntjük, hogy a megfigyeléseink alapján egy előzetes feltételezésünk (hipotézisünk) tartható-e, vagy a megfigyelések ellentmondanak a feltételezésnek. Nézzünk egy példát!

Tegyük fel, hogy egy gyárban minőségellenőrzést végzünk, azaz megvizsgáljuk, hogy a gyártott termékek minősége megfelelő-e. Előzetes feltételezésünk szerint a gyártási folyamat rendben van, ez

számunkra mondjuk azt jelenti, hogy a termékek legfeljebb 5%-a selejtes. A feltételezés ellenőrzéséhez 25 véletlenszerűen választott terméket megvizsgálunk, és ha legfeljebb 3 selejtes van köztük, akkor a feltételezést elfogadjuk. Ellenkező esetben a feltételezést elvetjük. Kérdés, hogy jó-e ez az eljárás?

Mivel döntésünk egy véletlenszerűen választott mintára épül, teljes bizonyossággal nem tudjuk eldönteni, hogy a feltételezésünk helyes-e. Kétféle hibát követhetünk el: ha igaz a feltételezés, mégis elutasítjuk, akkor *elsőfajú hibát vétünk*, ha nem igaz a feltételezés, mégis elfogadjuk, akkor *másodfajú hibát vétünk*. Mekkora ezen hibák valószínűsége? Kiszámításához az egyszerűség kedvéért tegyük fel, hogy a mintát visszatevéssel vesszük. Továbbá jelölje p a valódi (ismeretlen) selejtarányt.

Először tegyük fel, hogy a feltételezés igaz, azaz $p \leq 0{,}05$.

$$P_p(\text{elsőfajú hiba}) = P_p(\geq 4 \text{ selejt}) = \sum_{k=4}^{25} \binom{25}{k} p^k \cdot (1-p)^{25-k}.$$

Ez a valószínűség akkor a legnagyobb, ha $p = 0{,}05$, így az elsőfajú hiba valószínűsége legfeljebb

$$\alpha = \sup_{p \leq 0{,}05} P_p(\geq 4 \text{ selejt}) = P_{0{,}05}(\geq 4 \text{ selejt}) = 0{,}034.$$

Most tegyük fel, hogy a feltételezés hamis, azaz $p > 0{,}05$. A másodfajú hiba valószínűsége

$$P_p(\leq 3 \text{ selejt}) = \sum_{k=0}^{3} \binom{25}{k} p^k \cdot (1-p)^{25-k}.$$

Ha például a selejtarány $p = 0{,}1$, akkor 0,763 valószínűséggel fogjuk a feltételezést tévesen elfogadni.

Definiáljuk most formálisan az alapfogalmakat!

1.17. Definíció. Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mező, tehát $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, ahol Θ a paramétertér. Alapfeltevésünk az, hogy a valódi paraméter a teljes Θ paramétertér egy Θ_0 részhalmazába esik. Vagyis tekintsük a paramétertér $\Theta = \Theta_0 \cup \Theta_1$ diszjunkt felbontását, ezzel a hipotézisek a következő alakba írhatók:

nullhipotézis: $H_0 : \vartheta \in \Theta_0$

ellenhipotézis: $H_1 : \vartheta \in \Theta_1$.

A nullhipotézis ellenőrzésének érdekében gyűjtünk egy $\mathbf{X} = (X_1, ..., X_n)$ mintát (ez legtöbbször független, azonos eloszlású), a minta lehetséges értékeinek halmaza az $\mathcal{X}$ *mintatér*.

Statisztikai próba alatt egy döntési eljárást értünk, azaz a gyűjtött minta alapján döntést kell hoznunk, hogy elfogadjuk-e a nullhipotézist, vagy elutasítjuk. Ezt úgy formalizálhatjuk, hogy a mintatérnek vesszük egy $\mathcal{X} = \mathcal{X}_e \cup \mathcal{X}_k$ diszjunkt felbontását. A két részhalmaz elnevezése:

$\mathcal{X}_e$: *elfogadási tartomány*

$\mathcal{X}_k$: *kritikus (elutasítási) tartomány.*

Ha a minta az elfogadási tartományba esik, azaz $\mathbf{X} \in \mathcal{X}_e$, akkor H_0-t elfogadjuk, ellenkező esetben, azaz ha $\mathbf{X} \in \mathcal{X}_k$, akkor H_0-t elutasítjuk.

Az bevezető példában $\Theta = [0{,}1]$, $\Theta_0 = [0{,}0.05]$, $\Theta_1 = (0.05{,}1]$. A minta: $\mathbf{X} = (X_1, \ldots, X_{25})$, ahol $X_i = 1$, ha az i-edik kiválasztott termék selejtes, $X_i = 0$, ha az i-edik kiválasztott termék hibátlan. Így $\mathcal{X} = \{0{,}1\}^{25}$. A próbát meghatározó tartományok:

$$\mathcal{X}_e = \{\mathbf{x} \in \mathcal{X} : \sum_{i=1}^{25} x_i \leq 3\}, \quad \mathcal{X}_k = \{\mathbf{x} \in \mathcal{X} : \sum_{i=1}^{25} x_i \geq 4\}.$$

Nagyon fontos, hogy a két hipotézis szerepe nem egyenrangú. Az alapfeltevést csak nagyon indokolt esetben szeretnénk elutasítani, ezért az elsőfajú hiba súlyosabbnak számít, mint a másodfajú. Az elsőfajú hiba maximális valószínűségét szokás megadni, emellett természetesen a másodfajú hiba esélyének minimalizására törekszünk. Ebből kifolyólag a döntések értelmezése is különböző:

– H_0-t elfogadjuk: nem jelenti azt, hogy igaz, csak azt, hogy nincs okunk elutasítani.

– H_0-t elutasítjuk: komoly bizonyítékot találtunk arra, hogy H_0 nem igaz.

Például egy új gyógyszer vizsgálatánál a gyógyszer hatásosságára keresünk bizonyítékot, ezért a hipotézisek: H_0: a gyógyszer nem hatásos, H_1: a gyógyszer hatásos.

Folytassuk a definíciókat!

1.18. Definíció. Elsőfajú hiba: H_0 igaz, mégis elutasítjuk. Ennek valószínűsége: $P_\vartheta(\mathbf{X} \in \mathcal{X}_k)$, ahol $\vartheta \in \Theta_0$. A próba terjedelme:

$$\alpha = \sup_{\vartheta \in \Theta_0} P_\vartheta(\mathbf{X} \in \mathcal{X}_k).$$

Másodfajú hiba: H_0 hamis, mégis elfogadjuk. Ennek valószínűsége: $P_\vartheta(\mathbf{X} \in \mathcal{X}_e)$, ahol $\vartheta \in \Theta_1$. A másodfajú hiba valószínűsége helyett gyakran inkább az erőt szokás felírni, mely a helyes döntés valószínűsége, ha H_0 nem igaz. Képlettel, a próba erőfüggvénye:

$$\beta(\vartheta) = 1 - P_\vartheta(\mathbf{X} \in \mathcal{X}_e) = P_\vartheta(\mathbf{X} \in \mathcal{X}_k), \quad \vartheta \in \Theta_1.$$

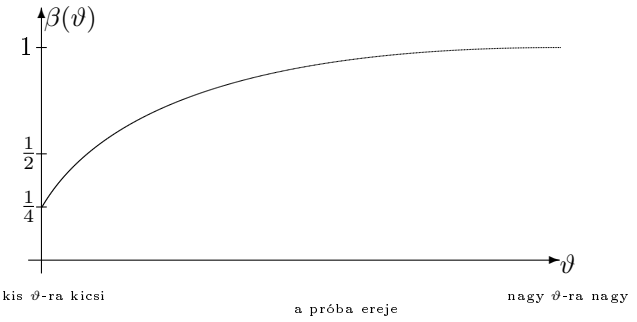
Másképp $\beta(\vartheta_1)$ a próba ereje a $H_1 : \vartheta = \vartheta_1$ ellenhipotézissel szemben.

Ha egy próbasorozatot vizsgálunk, azaz minden n mintaelemszámra van egy $(\mathcal{X}_e^n, \mathcal{X}_k^n)$ tartományokkal definiált próbánk, akkor ezt jelölhetjük a terjedelemben és az erőfüggvényben is: α helyett α_n-t, β helyett β_n-t írhatunk.

1.19. Példa. Legyen $X \sim E(-\vartheta, 1 + 2\vartheta)$ egyetlen megfigyelés, ahol $\vartheta \geq 0$ ismeretlen.
$H_0 : \vartheta = 0$ (egyszerű hipotézis – Θ_0 egyelemű halmaz – nem kell sup a terjedelemben)
$H_1 : \vartheta > 0$ (összetett hipotézis – Θ_1 többelemű halmaz).
A próba: $\mathcal{X}_e = (0{,}1, 0{,}85)$, $(\mathcal{X} = \mathbb{R})$.

$$\alpha = P_0(X \in \mathcal{X}_k) = 1 - P_0(X \in \mathcal{X}_e) = 1 - P_0(0{,}1 < X < 0{,}85) = 1 - 0{,}75 = \underset{\sim\sim}{0{,}25}.$$

erőfüggvény: $\beta(\vartheta) = P_\vartheta(X \in \mathcal{X}_k) = 1 - P_\vartheta(0{,}1 < X < 0{,}85) = 1 - \frac{0{,}75}{1+3\vartheta}$.



■

Mikor jó a próba?

1) Torzítatlan: a próba ereje legalább akkora, mint a terjedelme:
$\beta(\vartheta) \geq \alpha \quad \forall \vartheta \in \Theta_1$.

2) Erős: Az $(\mathcal{X}_e, \mathcal{X}_k)$ próba egyenletesen erősebb, mint a $(\mathcal{X}'_e, \mathcal{X}'_k)$ próba, ha
$\beta(\vartheta) = P_\vartheta(X \in \mathcal{X}_k) \geq \beta'(\vartheta) = P_\vartheta(X \in \mathcal{X}'_k) \quad \forall \vartheta \in \Theta_1$.

3) Konzisztens: Az $(\mathcal{X}_e^n, \mathcal{X}_k^n)$ legfeljebb α terjedelmű konzisztens próbasorozat, ha
(terjedelem) $\alpha_n \leq \alpha \quad \forall n$ és
$\beta_n(\vartheta) \overset{n\to\infty}{\longrightarrow} 1 \quad \forall \vartheta \in \Theta_1$.

1.19. Definíció. A $(\mathcal{X}_e, \mathcal{X}_k)$ próba egyenletesen legerősebb, ha minden más, legfeljebb ekkora terjedelmű próbánál egyenletesen erősebb.

A statisztikusok számos, a gyakorlatban lépten-nyomon előforduló hipotézisvizsgálati feladatra kidolgoztak „jó" próbákat. Ezek a klasszikus próbák.

1.7. A normális eloszlás paramétereire vonatkozó próbák

Az egyik leggyakrabban előforduló eloszlás a normális eloszlás, melyet a várható értéke és a szórása jellemez. Ezekre a paraméterekre három típusú próbát tanulunk. Ezek a típusok:

1. A várható értékre vonatkozó próba, ha a szórás ismert $\rightarrow$ u-próba.
2. A várható értékre vonatkozó próba, ha a szórás ismeretlen $\rightarrow$ t-próba.
3. A szórásra vonatkozó próba, ha a várható érték ismeretlen (vagy akár ismert) $\rightarrow$ F-próba.

A próbák ezen belül még különbözhetnek aszerint, hogy egymintásak vagy kétmintásak, illetve az ellenhipotézis jellege szerint (egyoldali vagy kétoldali ellenhipotézis). Ezek a próbák egyenletesen legerősebbek a legfeljebb ekkora terjedelmű torzítatlan próbák között.

Egymintás u-próba

Legyen $X_1, \ldots, X_n \sim \mathrm{N}(m, \sigma^2)$ ahol σ ismert, m ismeretlen.
A hipotézisek:

$$a)\ \begin{matrix} H_0: & m = m_0 \\ H_1: & m \neq m_0 \end{matrix} \qquad b)\ \begin{matrix} H_0: & m \leq m_0 \\ H_1: & m > m_0 \end{matrix} \qquad c)\ \begin{matrix} H_0: & m \geq m_0 \\ H_1: & m < m_0 \end{matrix} \tag{1}$$

A próbastatisztika:

$$u = \frac{\overline{X} - m_0}{\sigma} \cdot \sqrt{n} \overset{H_0}{\sim} N(0,1).$$

Ezért ha a kívánt terjedelem α, akkor a kritikus tartomány:

$$a)\, \mathcal{X}_k = \{|u| > u_{\frac{\alpha}{2}}\} \quad b)\, \mathcal{X}_k = \{u > u_\alpha\} \quad c)\, \mathcal{X}_k = \{u < -u_\alpha\} \tag{2}$$

ahol u_δ a standard normális eloszlás $(1-\delta)$-kvantilise, ezt a $\Phi(x)$ függvény táblázatából keressük ki.

Mj.: (1)-ben az a) esetben kétoldali, a b) és c) esetekben egyoldali ellenhipotézisről beszélünk.
Hogy néz ki a próba erőfüggvénye (kétoldali ellenhipotézisre)? Vezessük be a $\Delta_n(m) := \frac{m-m_0}{\sigma}\sqrt{n}$ jelölést.

$$\beta_n(m) = P_m(|u| > u_{\frac{\alpha}{2}}) = P_m\left(\left|\frac{\overline{X} - m_0}{\sigma}\sqrt{n}\right| > u_{\frac{\alpha}{2}}\right) = 1 - P_m\left(-u_{\frac{\alpha}{2}} < \frac{\overline{X} - m_0}{\sigma}\sqrt{n} < u_{\frac{\alpha}{2}}\right) =$$

$$1 - P_m\left(-u_{\frac{\alpha}{2}} < \frac{\overline{X} - m}{\sigma}\sqrt{n} + \frac{m - m_0}{\sigma}\sqrt{n} < u_{\frac{\alpha}{2}}\right) = 1 - P_m\left(-u_{\frac{\alpha}{2}} - \Delta_n(m) < \frac{\overline{X} - m}{\sigma}\sqrt{n} < u_{\frac{\alpha}{2}} - \Delta_n(m)\right) =$$

$$1 - \Phi(u_{\frac{\alpha}{2}} - \Delta_n(m)) + \Phi(-u_{\frac{\alpha}{2}} - \Delta_n(m)) = \Phi(-u_{\frac{\alpha}{2}} + \Delta_n(m)) + \Phi(-u_{\frac{\alpha}{2}} - \Delta_n(m)),$$

ahol felhasználtuk, hogy $\frac{\overline{X}-m}{\sigma}\sqrt{n} \sim N(0,1)$. Az erőfüggvény kapott képletéből leolvasható, hogy
1) $\beta_n(m)$ folytonos
2) $\beta_{n-1}(m) < \beta_n(m) \quad (m \neq m_0)$ és $\beta_n(m) \overset{n\to\infty}{\longrightarrow} 1$ (a próba konzisztens)
3) $\beta_n(m) > \alpha \quad (m \neq m_0)$ (erő nagyobb mint a terjedelem - a próba torzítatlan)
4) $\lim\limits_{m\to\pm\infty} \beta_n(m) = 1$

1.20. Példa. Legyen $x_1, \ldots, x_{16}$ 16 elemű minta ismeretlen m várható értékű, $\sigma = 1$ szórású normális eloszlásból. Hipotéziseink: $H_0 : m = 0$, $H_1 : m \neq 0$, és $\alpha = 0{,}1$ terjedelem mellett szeretnénk dönteni. Tegyük fel, hogy a megfigyelt mintára $\overline{x} = 0{,}1$. Hogyan döntünk?
Egymintás u-próbát kell végezni. A próbastatisztika: $u = \frac{0{,}1-0}{1}\sqrt{16} = 0{,}4$. A kritikus értéket meghatározó egyenlet:

$$\Phi(u_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2} = 0{,}95 \Rightarrow u_{\frac{\alpha}{2}} = 1{,}65.$$

Mivel $|0{,}4| \leq 1{,}65$, elfogadjuk H_0-t, azaz feltehető, hogy a minta az N(0,1) eloszlásból származik.

Keressük most meg a legnagyobb terjedelmet, ami mellett még elfogadjuk H_0-t, azaz azt az α értéket, amikor az u próbastatisztika értéke éppen az elfogadási és a kritikus tartomány határán van. Az ezt meghatározó egyenlet:

$$u_{\frac{\alpha}{2}} = |0{,}4| \;\Rightarrow\; \Phi(u_{\frac{\alpha}{2}}) = \Phi(0{,}4) = 0{,}66 = 1 - \frac{\alpha}{2} \;\Rightarrow\; \frac{\alpha}{2} = 0{,}34 \;\Rightarrow\; \alpha = 0{,}68.$$

Ezt szokás p-értéknek hívni, és ha számítógéppel végzünk próbákat, akkor mindig ezt a p-értéket kapjuk meg. Értelmezése tehát: a nullhipotézist a p-értéknél kisebb terjedelem mellett elfogadjuk, nagyobb terjedelem mellett viszont elutasítjuk. Azaz minél kisebb a p-érték, annál *szignifikánsabb* az eredmény, tehát annál erősebb bizonyítékot szolgáltatnak az adatok arra, hogy a nullhipotézis nem igaz. ■

A kétmintás u-próbára akkor van szükség, ha két független mintasorozatunk van ismert szórású normális eloszlásokból, és a várható értéküket szeretnénk összehasonlítani.

Kétmintás u-próba

Legyenek $X_1, \ldots, X_{n_1} \sim N(m_1, \sigma_1^2)$ és $Y_1, \ldots, Y_{n_2} \sim N(m_2, \sigma_2^2)$ független minták, ahol σ_1, σ_2 ismert, m_1, m_2 ismeretlenek.
A hipotézisek:

$$a)\ \begin{matrix} H_0 : \ m_1 = m_2 \\ H_1 : \ m_1 \neq m_2 \end{matrix} \qquad b)\ \begin{matrix} H_0 : \ m_1 \leq m_2 \\ H_1 : \ m_1 > m_2 \end{matrix} \qquad c)\ \begin{matrix} H_0 : \ m_1 \geq m_2 \\ H_1 : \ m_1 < m_2 \end{matrix} \tag{3}$$

A próbastatisztika:

$$u = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \overset{H_0}{\sim} N(0{,}1).$$

Tehát a kritikus tartományok ugyanazok, mint egymintás esetben, azaz (2) adja meg őket.

Ha nem ismerjük a szórást, akkor t-próbára van szükségünk, a próbastatisztikát és annak eloszlását ugyanúgy kapjuk meg, mint a konfidencia intervallum konstruálásánál.

Egymintás t-próba

Legyen $X_1, \dots, X_n \sim N(m, \sigma^2)$ ahol m, σ ismeretlen.
A hipotézisek: ugyanazok, mint az egymintás u-próbánál (1).
A próbastatisztika:

$$t = \frac{\overline{X} - m_0}{S_n^*} \cdot \sqrt{n} \overset{H_0}{\sim} t_{n-1},$$

ahol $S_n^* = \sqrt{\frac{1}{n-1}\sum (X_i - \overline{X})^2}$.
Jelölje a szbadsági fokot f, tehát $f = n - 1$.
A kritikus tartomány:

$$a)\, \mathcal{X}_k = \{|t| > t_f(\tfrac{\alpha}{2})\} \quad b)\, \mathcal{X}_k = \{t > t_f(\alpha)\} \quad c)\, \mathcal{X}_k = \{t < -t_f(\alpha)\} \tag{4}$$

ahol a $t_f(\delta)$ kritikus érték a t_f eloszlás $(1-\delta)$-kvantilise.
A $t_f(\frac{\alpha}{2})$-t és a $t_f(\alpha)$-t a „t-próba kritikus értékei" című táblázatból keressük ki, az oszlopok fölött kell figyelni arra, hogy egyoldali vagy kétoldali próbánk van.

A kétmintás t-próbára akkor van szükség, ha két független mintasorozatunk van ismeretlen szórású normális eloszlásokból, és a várható értéküket szeretnénk összehasonlítani. Bár a szórásokat nem ismerjük, a próba akkor működik, ha feltehető, hogy a két minta szórása megegyezik.

Kétmintás t-próba

Legyenek $X_1, \dots, X_{n_1} \sim N(m_1, \sigma^2)$ és $Y_1, \dots, Y_{n_2} \sim N(m_2, \sigma^2)$ független minták, ahol m_1, m_2 és σ ismeretlenek.
A hipotézisek: ugyanazok, mint a kétmintás u-próbánál (3).
A próbastatisztika:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n_1-1){S_{n_1}^*}^2 + (n_2-1){S_{n_2}^*}^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot (n_1+n_2-2)}{n_1+n_2}} \overset{H_0}{\sim} t_{n_1+n_2-2}.$$

Jelölje a szbadsági fokot f, tehát $f = n_1 + n_2 - 2$.
A kritikus tartomány ugyanaz, mint az egymintás esetben (4).

Vezessük le, hogyan jön ki a kétmintás t-próbánál a próbastatisztika. Egyrészt $m_1 = m_2$ esetén

$$\overline{X} - \overline{Y} \sim N(0, \sigma^2/n_1 + \sigma^2/n_2), \text{ azaz } \frac{1}{\sigma}(\overline{X} - \overline{Y})\sqrt{\frac{n_1 \cdot n_2}{n_1+n_2}} \sim N(0,1).$$

Másrészt teljesül, hogy

$$\frac{1}{\sigma^2}\left[(n_1-1){S_{n_1}^*}^2 + (n_2-1){S_{n_2}^*}^2\right] \sim \chi^2_{n_1-1+n_2-1},$$

mivel a tagok külön-külön $\chi^2_{n_1-1}$ illetve $\chi^2_{n_2-1}$ eloszlásúak, és függetlenek. Mivel pedig az előző két képletben felírt valószínűségi változók függetlenek is, hányadosuk a szabadsági fok gyökével beszorozva valóban t eloszlású lesz.

Ha a két minta szórása szignifikánsan különbözik, akkor a fenti próbát kissé módosítani kell, ezt vagy szintén t-próbának, vagy Welch-próbának hívják. A módosítás abból áll, hogy most a próbastatisztika

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{{S^*_{n_1}}^2}{n_1} + \frac{{S^*_{n_2}}^2}{n_2}}} \overset{H_0}{\approx} t_f,$$

ahol az f szabadsági fok

$$f = \frac{(g_1 + g_2)^2}{\frac{g_1^2}{n_1 - 1} + \frac{g_2^2}{n_2 - 1}},$$

és $g_i = {S^*_{n_i}}^2 / n_i$. Ha f nem egész, akkor kerekítjük.

A szórásra vonatkozó próbához szükségünk lesz az F eloszlásra.

1.20. Definíció. Ha X f_1 szabadsági fokú, Y pedig f_2 szabadsági fokú, egymástól független khinégyzet eloszlású valószínűségi változók, akkor a $Z = \frac{X/f_1}{Y/f_2}$ valószínűségi változó (f_1, f_2) szabadsági fokú F-eloszlású, jelölésben $Z \sim F_{f_1,f_2}$. Itt f_1 a számláló szabadsági foka, f_2 a nevező szabadsági foka. $(E(Z) = \frac{f_2}{f_2 - 2}.)$

A kétmintás F-próbára akkor van szükség, ha két független mintasorozatunk van normális eloszlásokból, és a szórásukat szeretnénk összehasonlítani (például azért, mert utána kétmintás t-próbát szeretnénk végezni a várható értékükre).

Kétmintás F-próba

Legyenek $X_1, \ldots, X_{n_1} \sim N(m_1, \sigma_1^2)$ és $Y_1, \ldots, Y_{n_2} \sim N(m_2, \sigma_2^2)$ független minták, ahol m_1, m_2 és σ_1, σ_2 ismeretlenek.
A hipotézisek:

$$a)\ \begin{matrix} H_0: \ \sigma_1 = \sigma_2 \\ H_1: \ \sigma_1 \neq \sigma_2 \end{matrix} \qquad b)\ \begin{matrix} H_0: \ \sigma_1 \leq \sigma_2 \\ H_1: \ \sigma_1 > \sigma_2 \end{matrix} \qquad c)\ \begin{matrix} H_0: \ \sigma_1 \geq \sigma_2 \\ H_1: \ \sigma_1 < \sigma_2 \end{matrix}$$

A próbastatisztika:

$$F = \frac{{S^*_{n_1}}^2}{{S^*_{n_2}}^2} \overset{H_0}{\sim} F_{n_1-1, n_2-1}.$$

Jelölje a szabadsági fokokat $f_1 = n_1 - 1$ és $f_2 = n_2 - 1$.
A kritikus tartomány:

$$a)\mathcal{X}_k = \{F < F_{f_1,f_2}(1 - \frac{\alpha}{2}) \text{ vagy } F > F_{f_1,f_2}(\frac{\alpha}{2})\}$$
$$b)\mathcal{X}_k = \{F > F_{f_1,f_2}(\alpha)\} \quad c)\mathcal{X}_k = \{F < F_{f_1,f_2}(1 - \alpha)\},$$

ahol az $F_{f_1,f_2}(\delta)$ kritikus érték az F_{f_1,f_2} eloszlás $(1 - \delta)$-kvantilise.
A kritikus értékeket az „F-próba kritikus értékei” című táblázatból keressük ki.

A próbastatisztika H_0 melletti eloszlása:

$$F = \frac{{S^*_{n_1}}^2}{{S^*_{n_2}}^2} = \frac{\frac{1}{n_1 - 1} \cdot \left(\frac{(n_1 - 1){S^*_{n_1}}^2}{\sigma_1^2} \right)}{\frac{1}{n_2 - 1} \cdot \left(\frac{(n_2 - 1){S^*_{n_1}}^2}{\sigma_2^2} \right)} \sim F_{n_1-1, n_2-1}.$$

A próba praktikusabb formája kétoldali ellenhipotézisre:

$$F < F_{f_1,f_2}(1 - \alpha/2) \Leftrightarrow \frac{1}{F_{f_1,f_2}(1 - \alpha/2)} < \frac{1}{F} \sim F_{f_2,f_1}, \text{ ezért } F_{f_2,f_1}(\alpha/2) = \frac{1}{F_{f_1,f_2}(1 - \alpha/2)}.$$

Így a kritikus tartomány ekvivalens alakja:

$$\mathcal{X}_k = \left\{ \frac{1}{F} > F_{f_2,f_1}(\alpha/2) \text{ vagy } F > F_{f_1,f_2}(\alpha/2) \right\}.$$

A gyakorlatban használt α-kra a kritikus érték 1-nél nagyobb, ezért elég F és $1/F$ közül a nagyobbikat összehasonlítani a megfelelő kritikus értékkel.

Egymintás F-próba (vagy χ^2-próba)

Legyen $X_1, \ldots, X_n \sim N(m, \sigma^2)$ ahol m, σ ismeretlen.
A hipotézisek:

$$a) \begin{array}{l} H_0: \ \sigma = \sigma_0 \\ H_1: \ \sigma \neq \sigma_0 \end{array} \qquad b) \begin{array}{l} H_0: \ \sigma \leq \sigma_0 \\ H_1: \ \sigma > \sigma_0 \end{array} \qquad c) \begin{array}{l} H_0: \ \sigma \geq \sigma_0 \\ H_1: \ \sigma < \sigma_0 \end{array}$$

A próbastatisztika:

$$\chi^2 = (n-1)\frac{S_n^{*2}}{\sigma_0^2} \overset{H_0}{\sim} \chi^2_{n-1}.$$

Jelölje a szabadsági fokot $f = n - 1$.
A kritikus tartomány:

$$a)\mathcal{X}_k = \{\chi^2 < \chi^2_f(1-\tfrac{\alpha}{2}) \text{ vagy } \chi^2 > \chi^2_f(\tfrac{\alpha}{2})\} \quad b)\mathcal{X}_k = \{\chi^2 > \chi^2_f(\alpha)\} \quad c)\mathcal{X}_k = \{\chi^2 < \chi^2_f(1-\alpha)\},$$

ahol a $\chi^2_f(\delta)$ kritikus érték χ^2_f eloszlás $(1-\delta)$-kvantilise.
A kritikus értékeket a „χ^2-próba kritikus értékei" című táblázatból keressük ki.
Ehelyett végezhetünk az

$$F = \frac{S_n^{*2}}{\sigma_0^2} \overset{H_0}{\sim} F_{n-1,\infty}$$

statisztikára F-próbát (az $F_{n-1,\infty}$ eloszlás a χ^2_{n-1} eloszlás átskálázott változata).

1.21. Példa. Kétféle altató (A és B) hatásosságát tesztelték 10 betegen. Az alábbi táblázat azt mutatja, hogy az altató mennyivel növelte meg a betegek éjszakai alvásidejét (órában mérve).

Beteg sorszáma	A altató	B altató	különbség
1	1.9	0.7	1.2
2	0.8	−1.6	2.4
3	1.1	−0.2	1.3
4	0.1	−1.2	1.3
5	−0.1	−0.1	0
6	4.4	3.4	1
7	5.5	3.7	1.8
8	1.6	0.8	0.8
9	4.6	0	4.6
10	3.4	2	1.2

Vajon van-e szignifikáns különbség a két gyógyszer hatásossága között ($\alpha = 0.01$ terjedelem mellett)?
Hipotézisek: $H_0: m_1 = m_2$, $H_1: m_1 \neq m_2$. A két minta azonban nem független, mert ugyanazokon a betegeken próbálták ki mind a két gyógyszert. Vegyük ezért az $A - B$ különbséget, és teszteljük egymintás t-próbával a $H_0: m = 0$, $H_1: m \neq 0$ hipotéziseket!
$S_n^* = 1.23$, $\overline{X} = 1.58$, $t = \frac{\overline{X} - m_0}{S_n^*}\sqrt{n} = 4.06$. A kritikus érték: $t_9(0.005) = 3.35$, így, mivel $|4.06| > 3.35$,

a nullhipotézist elvetjük, azaz a két gyógyszer hatásossága között szignifikáns különbség van.
Tegyük most fel, hogy a két gyógyszert más-más 10 betegen tesztelték (de továbbra is a fenti táblázat adatait használjuk). Ekkor a két minta független, kétmintás t-próba végezhető. $\overline{X} = 2.33$, $\overline{Y} = 0.75$, ${S^*}_A{}^2 = 4$, ${S^*}_B{}^2 = 3.8$. A kétféle gyógyszer hatásának szórásai feltételezhetően egyenlőek, de ellenőrizhetjük is F-próbával: $F' = 1.1$, míg a kritikus érték $F_{9,9}(0.025) = 4.03$. A t-próba próbastatisztikája:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(10-1){S^*}_A{}^2 + (10-1){S^*}_B{}^2}} \cdot \sqrt{\frac{10 \cdot 10 \cdot (10+10-2)}{10+10}} = 1.78.$$

A kritikus érték $t_{18}(0.01/2) = 2.89$. Mivel $|1.78| < 2.89$ elfogadjuk H_0-t, azaz nincs rá bizonyíték, hogy az egyik gyógyszer hatásosabb a másiknál. ■

1.8. Khi-négyzet próbák

1.9. Tétel. *(biz. nélkül) Legyen $A_1, A_2, \ldots, A_r$ teljes eseményrendszer, jel. $P(A_i) = p_i$. n darab független megfigyelésből jelölje ν_i az A_i esemény gyakoriságát. Ekkor a*

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i}$$

valószínűségi változó $n \longrightarrow \infty$ esetén az $r-1$ szabadsági fokú χ^2 eloszláshoz tart (eloszlásban). Általánosabban, ha a p_i valószínűségek $s(< r-1)$ db ismeretlen paramétertől függenek, akkor jelölje $\hat{p}_i$ a paraméterek ML-becslését. Ekkor a

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - n \cdot \hat{p}_i)^2}{n \cdot \hat{p}_i}$$

valószínűségi változó $n \longrightarrow \infty$ esetén az $r-s-1$ szabadsági fokú χ^2 eloszláshoz tart (eloszlásban).

A tétel alapján aszimptotikus próba végezhető, ami azt jelenti, hogy a próba terjedelme közelítőleg α lesz, ha n elég nagy.

χ^2-próba (tiszta eset)

Legyen $A_1, A_2, \ldots, A_r$ teljes eseményrendszer.
A hipotézisek:

$$H_0 : P(A_i) = p_i \quad i = 1, \ldots, r, \quad H_1 : \exists i : \ P(A_i) \neq p_i.$$

n darab független megfigyelésből jelölje ν_i az A_i esemény gyakoriságát.
A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - n \cdot p_i)^2}{n \cdot p_i} \overset{H_0}{\approx} \chi^2_{r-1}.$$

Jelölje a szabadsági fokot $f = r - 1$.
A kritikus tartomány:

$$\mathcal{X}_k = \{\chi^2 > \chi^2_f(\alpha)\}, \tag{5}$$

ahol a $\chi^2_f(\alpha)$ kritikus érték a χ^2_f eloszlás $(1-\alpha)$-kvantilise.
A kritikus értéket a „χ^2-próba kritikus értékei” című táblázatból keressük ki.

Vegyük észre, hogy a kritikus tartományba a próbastatisztika azon értékeit tettük, melyek az ellenhipotézis esetén fordulnak inkább elő, azaz a nagy értékeket.

χ^2-próba (becsléses eset)

Legyen $A_1, A_2, \ldots, A_r$ teljes eseményrendszer.
A hipotézisek:

$$H_0 : \exists \vartheta : \ P(A_i) = p_i(\vartheta) \forall i, \quad H_1 : \nexists \vartheta : \ P(A_i) = p_i(\vartheta) \forall i,$$

ahol ϑ egy s dimenziós paramétervektor. n darab független megfigyelésből jelölje ν_i az A_i esemény gyakoriságát, valamint legyen $\hat{\vartheta}$ a ϑ paramétervektor ML becslése, és $\hat{p}_i = p_i(\hat{\vartheta})$.
A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - n \cdot \hat{p}_i)^2}{n \cdot \hat{p}_i} \overset{H_0}{\approx} \chi^2_{r-s-1}.$$

Jelölje a szabadsági fokot $f = r - s - 1$.
A kritikus tartomány: (5), azaz ugyanaz, mint az előbb.

Mj.: Mivel a próba aszimptotikus, vigyáznunk kell arra, hogy a minta elemszáma elég nagy legyen. Pl. megkövetelhetjük, hogy az összes várt érték (np_i, ill. $n\hat{p}_i$) legalább 5 legyen. Ha ez nem teljesül, akkor a kis várt gyakoriságokkal rendelkező eseményeket összevonjuk.

1.8.1. Illeszkedésvizsgálat

Khi-négyzet próbával ellenőrizhetjük, hogy egy minta egy adott eloszlásból származhat-e. Mivel a khi-négyzet próbában egy véges teljes eseményrendszer szerepel, a próbát inkább diszkrét eloszlások illeszkedésének vizsgálatához szokás használni. Mindenesetre a feltételezett eloszlás értékkészletét véges sok csoportra kell osztanunk, hogy a próbát elvégezhessük.

1.22. Példa. Kockával dobunk. Nullhipotézisünk: H_0 : a kocka szabályos, azaz $P(A_i) = \frac{1}{6}; \quad i = 1, \ldots, 6$.
A megfigyelt gyakoriságok táblázata ($n = 60$):

értékek	1	2	3	4	5	6
ν_i	8	7	14	12	10	9
$n \cdot p_i$	10	10	10	10	10	10

A próbastatisztika:

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(14-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(9-10)^2}{10} = 3.4.$$

A kritikus érték ($f = r - 1 = 5$): $\chi^2_5(0.1) = 9.24$. Mivel $3.4 < 9.24$, H_0-t elfogadjuk, azaz a kocka szabályosnak tekinthető. ■

1.23. Példa. 400-szor feljegyeztük, hogy egy hibás kapcsoló hányadik próbálkozásra gyújtotta fel a villanyt. H_0 : $X_i \sim \text{Geo}(p)$ valamilyen p-re. Az adatok:

	1	2	3	4
ν_i	324	57	14	5

A minta átlaga: $\overline{X} = \dfrac{1 \cdot 324 + 2 \cdot 57 + 3 \cdot 14 + 4 \cdot 5}{400} = \dfrac{500}{400} = \dfrac{5}{4}$, ebből a paraméter ML becslése $\hat{p} = \frac{1}{\overline{X}} = 0.8$. A geometriai eloszlás pozitív egész értékeket vehet fel, tehát a lehetséges értékek egy

csoportosítása: $A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, A_4 = \{4, \ldots\}$. A négy csoport becsült valószínűsége:

$$\hat{p}_1 = 0.8,\ \hat{p}_2 = 0.2 \cdot 0.8 = 0.16,\ \hat{p}_3 = 0.2^2 \cdot 0.8 = 0.032,\ \hat{p}_4 = 1 - \hat{p}_1 - \hat{p}_2 - \hat{p}_3 = 0.008.$$

A várt értékek táblázata tehát:

	1	2	3	≥ 4
$n \cdot \hat{p}_i$	320	64	12.8	3.2

A próbastatisztika:

$$\chi^2 = \frac{(324-320)^2}{320} + \frac{(57-64)^2}{64} + \frac{(14-12.8)^2}{12.8} + \frac{(5-3.2)^2}{3.2} = 1.95.$$

A szabadsági fok: $f = r - s - 1 = 4 - 1 - 1 = 2$. Ha a terjedelem $\alpha = 0.05$, akkor a kritikus érték $\chi^2_2(0.05) = 5.99$, és mivel $1.95 < 5.99$, így H_0-t elfogadjuk. (Tulajdonképpen még a harmadik és negyedik csoportot is össze lehetne vonni.) ■

Ha a feltételezett eloszlás folytonos, akkor a teljes számegyenest kell intervallumokra, illetve félegyenesekre bontani. Az intervallumok megválasztásához néhány jó tanács:
1)Az intervallumok száma ne legyen se túl kevés (nem elég erős a próba, a mintában lévő információ nagy része elveszik), se túl sok (a χ^2 közelítés sérül).
2) Az osztópontokat úgy válasszuk, hogy az intervallumok p_i valószínűségei közel egyformák legyenek.
Összefoglalva, az illeszkedésvizsgálat lépései:
1) Ha becsléses esettel van dolgunk, akkor a paramétereket megbecsüljük a mintából ML módszerrel.
2) A lehetséges értékkészletet véges sok csoportba osztjuk (TER-t hozunk létre).
3) Kiszámoljuk a csoportok várt gyakoriságát.
4) Ha vannak kicsi (pl. 5-nél kisebb) várt gyakoriságok, akkor összevonunk csoportokat.
5) Kiszámoljuk a próbastatisztikát, kikeressük a szabadsági foknak és a terjedelemnek megfelelő kritikus értéket.
6) Összehasonlítva a kettőt, levonjuk a következtetést.

1.8.2. Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket:
az 1. szempont szerint r osztály van: $A_1, \ldots, A_r$,
a 2. szempont szerint s osztály van: $B_1, \ldots, B_s$.
Nullhipotézisünk: H_0 : a két szempont független egymástól, azaz $P(A_i \cap B_j) = P(A_i) \cdot P(B_j) = p_i q_j$ minden i, j-re. Az ellenhipotézis pedig az, hogy a két szempont összefügg.
n darab független megfigyelésből jelölje ν_{ij} az $A_i \cap B_j$ esemény gyakoriságát, valamint legyen $\nu_{i\bullet} = \sum_{j=1}^{s} \nu_{ij}$ az A_i gyakorisága és $\nu_{\bullet j} = \sum_{i=1}^{r} \nu_{ij}$ a B_j gyakorisága. Általában p_i, q_j nem ismertek. Ekkor először megbecsüljük őket a mintából ML-módszerrel:

$$\hat{p}_i = \frac{\nu_{i\bullet}}{n}, \quad \hat{q}_j = \frac{\nu_{\bullet j}}{n}.$$

Összesen $r-1$ darab p_i paramétert becsültünk (mivel $p_r = 1 - \sum_{i=1}^{r-1} p_i$ már adódik), és $s-1$ darab q_j paramétert. Így

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\nu_{ij} - n \cdot \hat{p}_i \hat{q}_j)^2}{n \cdot \hat{p}_i \hat{q}_j} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(\nu_{ij} - \frac{\nu_{i\bullet}\nu_{\bullet j}}{n})^2}{\frac{\nu_{i\bullet}\nu_{\bullet j}}{n}} \overset{H_0}{\approx} \chi^2_f,$$

ahol a szabadsági fok: $f = r \cdot s - (r - 1 + s - 1) - 1 = (r-1)(s-1)$.

Mj.: Legyen $r = s = 2$. Ekkor a próbastatisztika egyszerűbb alakra hozható:

$$\chi^2 = \frac{n \cdot (\nu_{11}\nu_{22} - \nu_{12}\nu_{21})^2}{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}}.$$

1.24. Példa. 200 emberről feljegyezték, hogy szőke-e, és hogy kékszemű-e.

haj/szem	kék	más	
szőke	30	20	50
más	70	80	150
	100	100	200

Függetlennek tekinthető-e a hajszín és a szemszín ($\alpha = 0.05$)?
A várt értékek táblázata:

haj/szem	kék	más
szőke	25	25
más	75	75

A próbastatisztika: $\chi^2 = \frac{200 \cdot (30 \cdot 80 - 70 \cdot 20)^2}{100 \cdot 100 \cdot 50 \cdot 150} = 2.67$.
Szabadsági fok: $(2-1) \cdot (2-1) = 1$, kritikus érték: $\chi^2_1(0.05) = 3.84$.
Tehát a szemszín és a hajszín függetlennek tekinthető. ■

1.8.3. Homogenitásvizsgálat

Legyen X és Y két valószínűségi változó, és közös értékkészletüket bontsuk fel az $A_1, \ldots, A_r$ osztályokra.
$H_0 : X$ és Y eloszlása megegyezik, azaz $P(X \in A_i) = P(Y \in A_i)$ minden i-re.
Az X-et n-szer megfigyelve, legyen ν_i az A_i osztály gyakorisága. Hasonlóan, Y-t m-szer megfigyelve, legyen μ_i az A_i osztály gyakorisága.
Az i. osztály valószínűségének ML becslése: $\hat{p}_i = \frac{\nu_i + \mu_i}{n+m}$.
A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{r} \frac{(\nu_i - n\hat{p}_i)^2}{n\hat{p}_i} + \sum_{i=1}^{r} \frac{(\mu_i - m\hat{p}_i)^2}{m\hat{p}_i} \overset{H_0}{\approx} \chi^2_f,$$

ahol a szabadsági fok: $f = (r-1) + (r-1) - (r-1) = r - 1$, hiszen mindkét összeg szabadsági foka (ha nem becsült paraméterek lennének) $r-1$, viszont $r-1$ paramétert becsültünk, amit le kell vonni.
A fenti statisztika átalakítása után kapjuk, hogy

$$\chi^2 = \sum_{i=1}^{r} \frac{\left(\frac{\nu_i}{n} - \frac{\mu_i}{m}\right)^2}{\nu_i + \mu_i} \cdot n \cdot m \overset{H_0}{\approx} \chi^2_{r-1}.$$

1.25. Példa. Két kockával dobunk. Tekinthető-e a két kocka egyformának ($\alpha = 0{,}05$)?
Az adatok:

érték	1	2	3	4	5	6	
ν_i (1. kocka)	7	11	8	10	8	6	$r = 6$
μ_i (2. kocka)	16	11	20	19	18	16	$n = 50$
							$m = 100$

A próbastatisztika:

$$\chi^2 = \sum_{i=1}^{6} \frac{\left(\frac{\nu_i}{50} - \frac{\mu_i}{100}\right)^2}{\nu_i + \mu_i} \cdot 50 \cdot 100 = 3.58.$$

A kritikus érték: $\chi^2_{6-1}(0.05) = 11.1$. Mivel $3.58 < 11.1$, H_0-t elfogadjuk, a két kocka egyformának tekinthető. ■

1.9. Folytonos eloszlású mintára a χ^2-próba helyett alkalmazható próbák

Láttuk, hogy a χ^2-próba háromféle feladat tesztelésére alkalmas. Mindhárom típus alkalmazható folytonos eloszlású mintákra, ha azokat diszkretizáljuk. Azonban a χ^2-próba alapvetően diszkrét jellegű, így felmerül a kérdés, hogy folytonos eloszlású mintákra nem lehet-e jobb próbákat konstruálni. Az alábbiakban bemutatunk néhányat ezek közül.

1.9.1. Illeszkedésvizsgálat

Legyen $X_1, \dots, X_n$ folytonos F eloszlásból vett minta. A hipotézisek: $H_0 : F = F_0$, $H_1 : F \neq F_0$.

Kolmogorov-Szmirnov próba: Készítsük el a minta $\hat{F}_n$ tapasztalati eloszlásfüggvényét, és tekintsük a következő próbastatisztikát:

$$D_n = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_0(x) \right|.$$

Ennek kiszámítása a következő: mivel $X_i^{(n)}$ és $X_{i+1}^{(n)}$ között a tapasztalati eloszlásfüggvény konstans $\frac{i}{n}$, az F_0 eloszlásfüggvény pedig monoton nő, kapjuk, hogy

$$D_n = \max_{0 \le i \le n} \left[\max \left(|F_0(X_i^{(n)}) - i/n|, |F_0(X_{i+1}^{(n)}) - i/n| \right) \right].$$

Kijött az is, hogy ennek H_0 melletti eloszlása nem függ F_0-tól, hiszen ha X_i eloszlásfüggvénye F_0, akkor $F_0(X_i) \sim E(0,1)$. Továbbá $\sqrt{n} D_n$ aszimptotikus eloszlása meghatározható:

$$P(\sqrt{n} D_n < y) \overset{n \to \infty}{\longrightarrow} K(y) = \sum_{i=-\infty}^{+\infty} (-1)^i \cdot e^{-2i^2 y^2} \quad (y > 0).$$

A fenti K eloszlásfüggvényhez tartozó eloszlás az ún. *Kolmogorov eloszlás.* Azaz ha n elég nagy, akkor a Kolmogorov eloszlásból számolhatunk kritikus értéket (pl. $\alpha = 0.05$ terjedelem mellett akkor utasítjuk el H_0-t, ha $\sqrt{n} D_n > 1.36$), ha pedig n kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

Mj.: Vizsgálhatjuk a $H_1 : F(x) > F_0(x)\ \forall x$ vagy a $H_1 : F(x) < F_0(x)\ \forall x$ egyoldali ellenhipotéziseket is, ekkor a próbastatisztika

$$D_n^+ = \sup_{x \in \mathbb{R}} \left(\hat{F}_n(x) - F_0(x) \right), \text{ illetve } D_n^- = \sup_{x \in \mathbb{R}} \left(F_0(x) - \hat{F}_n(x) \right).$$

Ezekre is igaz, hogy H_0 melletti eloszlásuk nem függ F_0-tól. Továbbá $\sqrt{n} D_n^{\pm}$ aszimptotikus eloszlása meghatározható:

$$P(\sqrt{n} D_n^{\pm} < y) \overset{n \to \infty}{\longrightarrow} K_1(y) = 1 - e^{-2y^2} \quad (y > 0).$$

A fenti K_1 eloszlásfüggvényhez tartozó eloszlást nevezhetjük *Szmirnov eloszlásnak.* Azaz ha n elég nagy, akkor a Szmirnov eloszlásból számolhatunk kritikus értéket (pl. $\alpha = 0.05$ terjedelem mellett akkor utasítjuk el H_0-t, ha $\sqrt{n} D_n^{\pm} > 1.22$), ha pedig n kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

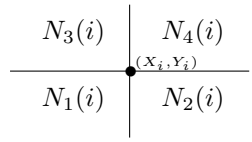
1.9.2. Függetlenségvizsgálat

Legyen (X_i, Y_i) egy folytonos $H(x,y)$ eloszlásfüggvényű eloszlásból származó n elemű minta. Jelölje a H-hoz tartozó marginális eloszlásokat $F(x) = \lim_{w \to \infty} H(x,w)$ és $G(y) = \lim_{z \to \infty} H(z,y)$. A nullhipotézis: H_0 : a két koordináta független, azaz $H(x,y) = F(x)G(y)\ \forall x, y$.

Blum-Kiefer-Rosenblatt próba:

Legyen minden i-re

$$\begin{array}{rcl} N_1(i) & = & |\{\, j : X_j < X_i \text{ és } Y_j < Y_i \,\}| \\ N_2(i) & = & |\{\, j : X_j \ge X_i \text{ és } Y_j < Y_i \,\}| \\ N_3(i) & = & |\{\, j : X_j < X_i \text{ és } Y_j \ge Y_i \,\}| \\ N_4(i) & = & |\{\, j : X_j \ge X_i \text{ és } Y_j \ge Y_i \,\}| \end{array}$$



Azaz $N_\ell(i)$ darab pont esik az (X_i, Y_i) pont által meghatározott ℓ.-ik síknegyedbe.

Próbastatisztika:

$$B_n = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{N_1(i)}{n} \cdot \frac{N_4(i)}{n} - \frac{N_2(i)}{n} \cdot \frac{N_3(i)}{n} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} (\hat{H}_n(X_i, Y_i) - \hat{F}_n(X_i) \hat{G}_n(Y_i))^2,$$

ahol $\hat{F}_n, \hat{G}_n, \hat{H}_n$ a tapasztalati eloszlásfüggvények:

$$\hat{H}_n(X_i, Y_i) = N_1(i)/n, \quad \hat{F}_n(X_i) = (N_1(i) + N_3(i))/n, \quad \hat{G}_n(Y_i) = (N_1(i) + N_2(i))/n.$$

Belátható, hogy B_n H_0 melletti eloszlása nem függ az F, G marginális eloszlásoktól, továbbá nB_n aszimptotikus eloszlása meghatározható. Azaz ha n elég nagy, akkor az aszimptotikus eloszlásból számolhatunk kritikus értéket (pl. $\alpha = 0.05$ terjedelem mellett akkor utasítjuk el H_0-t, ha $nB_n > > 0.058$), ha pedig n kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

Kendall próba:
Tekintsük a következő próbastatisztikát:

$$K_n = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[I((X_i - X_j)(Y_i - Y_j) \geq 0) - I((X_i - X_j)(Y_i - Y_j) < 0) \right].$$

Szavakkal elmondva, K_n a rendezett ponptpárok száma, mínusz a nem rendezett pontpárok száma, azaz kétszer a rendezett pontpárok száma, mínusz az összes pontpár száma. Belátható, hogy K_n H_0 melletti eloszlása nem függ az F, G marginális eloszlásoktól. Ha ugyanis az (X_i, Y_i) minta helyett az $(F(X_i), G(Y_i))$ mintából számolnánk ki K_n-et, ugyanazt az értéket kapnánk, hiszen F, G monoton növő függvények. Viszont $F(X_i)$ és $G(Y_i)$ már független $E(0,1)$ eloszlásúak.

Megmutatható az is, hogy K_n aszimptotikusan normális eloszlású, azaz ha n elég nagy, akkor u-próba végezhető, ha pedig n kicsi, akkor külön táblázat tartalmazza a kritikus értékeket. Az u-próbához standardizálni kell K_n-et:
a) K_n várható értéke H_0 mellett:

$$E_0(I((X_i - X_j)(Y_i - Y_j) \geq 0)) = P_0((X_i - X_j)(Y_i - Y_j) \geq 0) = P_0(X_i - X_j \geq 0 \text{ és } Y_i - Y_j \geq 0) + P_0(X_i - X_j \leq 0 \text{ és } Y_i - Y_j \leq 0) = 0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5.$$

Hasonlóan, $E_0(I((X_i - X_j)(Y_i - Y_j) < 0)) = 0.5$, azaz $E_0(K_n) = 0$.
b) K_n szórásnégyzete H_0 mellett:

$$D_0^2(K_n) = \frac{n(n-1)(2n+5)}{18}.$$

Megjegyezzük, hogy a próba csak a $\tau \neq 0$ típusú ellenhipotézisekre konzisztens, ahol

$$\tau = 2P((X_1 - X_2)(Y_1 - Y_2) > 0) - 1$$

a Kendall-féle függőségi együttható ($|\tau| \leq 1$, ha X és Y függetlenek, akkor $\tau = 0$, de fordítva nem igaz).

1.9.3. Homogenitásvizsgálat

$X_1, \ldots, X_n$ és $Y_1, \ldots, Y_m$ független minták valamilyen folytonos eloszlásokból (az eloszlásfüggvények F, illetve G).
$H_0 : F = G$, $H_1 : F \neq G$.

Kolmogorov-Szmirnov próba:
Készítsük el a mintákból az $\hat{F}_n, \hat{G}_m$ tapasztalati eloszlásfüggvényeket, és tekintsük a következő próbastatisztikát:

$$D_{n,m} = \sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - \hat{G}_m(x) \right|.$$

Megmutatható, hogy $D_{n,m}$ H_0 melletti eloszlása nem függ a két minta közös eloszlásától, és $\sqrt{\frac{m \cdot n}{m+n}} \cdot D_{n,m}$ aszimptotikusan Kolmogorov eloszlású. Azaz ha n, m elég nagyok, akkor a Kolmogorov

eloszlásból számolhatunk kritikus értéket , ha pedig n, m kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

Mann-Whitney-Wilcoxon próba:
Tekintsük a következő próbastatisztikát:

$$W_{n,m} = \sum_{i=1}^{n} \sum_{j=1}^{m} I(X_i \geq Y_j).$$

1.10. Tétel. *Legyen Z az egyesített minta, ennek elemszáma $N := n + m$. Vegyük a $Z_1^{(N)} < \cdots < Z_N^{(N)}$ rendezett mintát, és jelölje r_i, hogy $X_i^{(n)}$ hanyadik legkisebb elem ebben a rendezett mintában. Ekkor $W_{n,m} = r_1 + \cdots + r_n - \frac{n(n+1)}{2}$.*

Bizonyítás.
$X_1^{(n)}$ $r_1 - 1$ db Y_j-nél nagyobb, $X_2^{(n)}$ $r_2 - 2$ db Y_j-nél nagyobb, és általában, $X_i^{(n)}$ $r_i - i$ db Y_j-nél nagyobb. Ezeket összeadva kapjuk az állítást. ■

Megmutatható, hogy $W_{n,m}$ H_0 melletti eloszlása nem függ a két minta közös eloszlásától, és aszimptotikusan normális eloszlású. Az első állítás azért igaz, mert H_0 mellett az X-ek és Y-ok minden sorrendje egyformán valószínű, azaz $1/\binom{n+m}{n}$ a valószínűsége annak, hogy az X-ek adott rangokat foglalnak el, $W_{n,m}$ pedig ezeknek a rangoknak a függvénye. A második állítás pedig azért igaz, mert az $I(X_i \geq Y_j)$ változók, bár nem függetlenek, viszonylag gyenge összefüggést mutatnak.

Ez alapján, ha n, m elég nagyok, akkor u-próba végezhető, ha pedig n, m kicsi, akkor külön táblázat tartalmazza a kritikus értékeket. Az u-próbához standardizálni kell $W_{n,m}$-et:
a) $W_{m,n}$ várható értéke H_0 mellett:
$E_0(I(X_i \geq Y_j)) = P_0(X_i \geq Y_j) = \frac{1}{2} \Rightarrow E_0(W_{n,m}) = \frac{n \cdot m}{2}$
b) $W_{n,m}$ szórásnégyzete H_0 mellett:

$$D_0^2(W_{n,m}) = \frac{n \cdot m}{4} \cdot \frac{n + m + 1}{3}.$$

Megjegyezzük, hogy a próba csak a $P(X > Y) \neq 1/2$ típusú ellenhipotézis esetén konzisztens.

1.10. Lineáris modell, szórásanalízis

A lineáris modellben azt feltételezzük, hogy a megfigyelésünk (Y) valamilyen általunk ismert, az egyedre jellemző értékek ($x_j,\ j = 1, \ldots, p$) lineáris függvénye, valamilyen (mérési) hibával terhelten. Azaz

$$Y = a_1 x_1 + a_2 x_2 + \ldots + a_p x_p + \epsilon = x^T a + \epsilon,$$

ahol feltesszük, hogy az ϵ valószínűségi változó várható értéke nulla. Az $a = (a_1, \ldots, a_p)^T$ oszlopvektor tartalmazza az ismeretlen paramétereket. A konstans tagot úgy építhetjük be a modellbe, hogy feltesszük például, hogy $x_1 = 1$, ekkor a_1 a konstans tag. Az $x = (x_1, \ldots, x_p)^T$ oszlopvektor tartalmazza az egyedre jellemző értékeket (például kor, magasság, jövedelem, de lehet kategórikus változó is, pl. nem (férfi = 0, nő = 1), családi állapot (házas = 0, nem házas = 1), stb.) Feltesszük, hogy van egy n elemű mintánk, azaz megfigyeltük az

$$Y_i = a_1 x_{i1} + a_2 x_{i2} + \ldots + a_p x_{ip} + \epsilon_i,\ i = 1, \ldots, n$$

értékeket, ahol most azt is feltesszük, hogy az ϵ_i változók függetlenek, és mindegyiknek a szórásnégyzete ugyanannyi (jelölje ezt σ^2). A modellt a következő tömör mátrixos alakba írhatjuk:

$$Y = Xa + \epsilon,$$

ahol Y $n \times 1$-es oszlopvektor, X $n \times p$ méretű mátrix, a $p \times 1$-es oszlopvektor, ϵ pedig $n \times 1$-es oszlopvektor. Az a paramétervektor legkisebb négyzetes becslése az az $\hat{a}$ vektor, melyre $||Y - Xa||^2$

minimális. Világos, hogy ekkor $X\hat{a}$ az Y vektornak a $V = \{Xb : b \in \mathbb{R}^p\}$ lineáris altérre vett merőleges vetülete, azaz $Y - X\hat{a}$ merőleges az összes Xb alakú vektorra:

$$0 = (Xb)^T(Y - X\hat{a}) = b^T X^T(Y - X\hat{a}) \quad \forall b \in \mathbb{R}^p,$$

ami csak úgy lehet, ha $X^TY = X^TX\hat{a}$. Ha X^TX invertálható (ezt fel szokás tenni), akkor a paramétervektor legkisebb négyzetes becslése

$$\hat{a} = (X^TX)^{-1}X^TY.$$

Mivel $E(Y) = Xa$, ezért $E(\hat{a}) = a$, azaz ez torzítatlan becslés, és Y-nak lineáris függvénye. Továbbá az is igaz, hogy a hiba szórásnégyzetének torzítatlan becslése

$$\hat{\sigma}^2 = \frac{||Y - X\hat{a}||^2}{n - p}.$$

Ha azt is feltesszük, hogy a hibák normális eloszlásúak, azaz $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, akkor a fenti $\hat{a}$ egyben az a paraméter maximum likelihood becslése is. Ugyanebben a modellben σ^2 maximum likelihood becslése

$$\hat{\sigma}^2 = \frac{||Y - X\hat{a}||^2}{n},$$

ez viszont nem torzítatlan.

A szórásanalízis a lineáris modell egy speciális esete. A legegyszerűbb esetével, az egyszempontos szórásanalízissel foglalkozunk. Ez az eljárás arra használható, hogy több független mintát hasonlítsunk össze egymással.

Feltesszük, hogy k csoportunk van (például k különböző táppal etetett csirkék súlyát mérjük meg), az i-edik csoportban n_i megfigyelésünk van, és $n = \sum_{i=1}^k n_i$. Mindegyik megfigyelés normális eloszlású és függetlenek, az i-edik csoport j-edik megfigyelése

$$Y_{ij} = a_i + \epsilon_{ij} \sim \mathrm{N}(a_i, \sigma^2).$$

Látszik, hogy ez valóban a lineáris modell speciális esete: az Y vektort úgy kapjuk, hogy az Y_{ij} értékeket egymás alá írjuk:

$$Y = (Y_{11}, Y_{12}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}, \ldots, Y_{k1}, \ldots, Y_{kn_k})^T.$$

Az X modellmátrix pedig $n \times k$ méretű lesz, méghozzá az X_{ij}-nek megfelelő sorban az i-edik elem 1-es, a többi pedig 0. Ezzel valóban $Y = Xa + \epsilon$, ahol ϵ egy n-dimenziós oszlopvektor, melynek koordinátái függetlenek és $\mathrm{N}(0, \sigma^2)$ eloszlásúak. Így az a paramétervektor maximum likelihood becslését az előzőek szerint megkaphatjuk. Arra vagyunk kíváncsiak, hogy az összes csoportnak ugyanannyi-e a várható értéke, azaz

$$H_0 : a_1 = a_2 = \cdots = a_k.$$

A $k = 2$ esetben ezt a hipotézist a kétmintás t-próbával tesztelhetjük. $k > 2$ esetben a hipotézist akkor fogjuk elutasítani, ha a csoportok közötti szóródás lényegesen nagyobb, mint a csoportokon belüli. Erre használják a következő szórásfelbontó táblázatot:

szóródás oka	négyzetösszeg	szabadsági fok	tapasztalati szórásnégyzet
csoportok közötti	Q_a	$k-1$	$s_a^2 = \frac{Q_a}{k-1}$
csoportokon belüli	Q_e	$n-k$	$s_e^2 = \frac{Q_e}{n-k}$
teljes	Q	$n-1$	–

Jelölje $\overline{Y}_{i\bullet}$ az i-edik csoport megfigyeléseinek átlagát, $\overline{Y}_{\bullet\bullet}$ pedig mind az n megfigyelés átlagát. Ekkor

$$Q_a = \sum_{i=1}^k n_i(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2, \quad Q_e = \sum_{i=1}^k \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{i\bullet})^2, \quad Q = \sum_{i=1}^k \sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2 = Q_a + Q_e.$$

Megmutatható, hog a fenti nullhipotézis esetén az $F = \frac{s_a^2}{s_e^2}$ statisztika $F_{k-1,n-k}$ eloszlású, azaz akkor vetjük el a nullhipotézist, ha F értéke nagyobb, mint az $F_{k-1,n-k}$ eloszlás $(1-\alpha)$-kvantilise.

2. Sztochasztikus folyamatok

Ebben a szakaszban időben lejátszódó véletlen folyamatokkal foglalkozunk. Az időparaméter lehet diszkrét vagy folytonos is. A vizsgált folyamatok a következők lesznek: diszkrét illetve folytonos idejű Markov láncok, szimmetrikus bolyongás és Wiener folyamat, felújítási folyamatok, elágazó folyamatok.

2.1. Diszkrét idejű Markov láncok

Tekintsünk egy megszámlálható sok csúcspontú, irányított gráfot úgy, hogy minden élre egy nemnegatív szám van írva, és minden csúcs kimenő éleire írt számok összege 1 (hurokél is megengedett). Ezen a gráfon bolyongunk az élekre írt valószínűségek szerint. Ha egységnyi időközönként lépünk akkor diszkrét paraméterű Markov láncot kapunk. Ennek meghatározó tulajdonsága, hogy a lánc jövőbeli fejlődése csak a pillanatnyi állapottól függ, a múlttól nem.

A folyamattal kapcsolatban a következő kérdések merülhetnek fel: Honnan hová lehet eljutni? Mekkora eséllyel érünk vissza a kiindulási helyünkre? Mekkora az esélye, hogy végtelen sokszor visszatérünk? Átlagosan mennyi idő alatt érünk vissza a kiindulási helyre, vagy általánosabban egy másik csúcsba? Vannak-e elnyelő csúcsok? Ha igen, mekkora eséllyel nyelődünk el bennük? Van-e stacionárius kezdeti eloszlás? Hány? Tart-e a bolyongás egy stacionárius eloszláshoz? Milyen gyorsan? Érvényes-e valamilyen NSzT? Érvényes-e valamilyen CHT?

Definiáljuk először pontosan a Markov láncot!

2.1. Definíció. *Legyen adott az $\{X_n\}_{n\in\mathbb{N}}$ folyamat, ahol $X_n : \Omega \to I$ valószínűségi változók az $(\Omega, \mathcal{A}, P)$ valószínűségi mezőn, I pedig megszámlálható halmaz. A folyamatot diszkrét idejű (homogén) Markov láncnak nevezzük, ha*

- *A folyamat Markov-tulajdonságú, azaz minden $n_1 < \cdots < n_k < m$ és minden $i_1, \ldots, i_k, j \in I$ esetén*
$$P(X_m = j \mid X_{n_1} = i_1, \ldots, X_{n_k} = i_k) = P(X_m = j \mid X_{n_k} = i_k).$$
- *A Markov folyamat stacionárius (vagy homogén) átmenetvalószínűségű, azaz minden $i, j \in I$ esetén*
$$P(X_{n+1} = j \mid X_n = i) = p_{ij},$$
n-től függetlenül.

A $P = (p_{ij})$ mátrixot átmenetmátrixnak nevezzük (nem keverendő össze a valószínűséggel), elemei az átmenetvalószínűségek. Az X_0 eloszlását kezdeti eloszlásnak hívjuk, és $p = (p_i)$-vel jelöljük. Egy adott állapotsorozat valószínűsége a Markov tulajdonság szerint:

$$\begin{aligned} &P(X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = \\ &= \quad P(X_0 = i_0)P(X_1 = i_1|X_0 = i_0)\cdots P(X_n = i_n|X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) = \\ &= \quad p_{i_0}p_{i_0 i_1}\cdots p_{i_{n-1}i_n}. \end{aligned}$$

A P átmenetmátrix úgynevezett sztochasztikus mátrix, azaz $p_{ij} \geq 0$ és minden sor összege 1.

2.2. Állítás. *Legyenek $p_{ij}^{(n)} = P(X_{n+m} = j|X_m = i)$ az n-edrendű átmenetvalószínűségek ezekre teljesül a Chapman-Kolmogorov egyenlőség:*

$$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(n)} p_{kj}^{(m)}.$$

Ez azt jelenti, hogy a a $p_{ij}^{(n)}$ mennyiségek éppen a P^n mátrix megfelelő elemei.

Bizonyítás.

$$P(X_{n+m}=j \mid X_0=i)=\sum_k P(X_{n+m}=j, X_n=k \mid X_0=i)=$$
$$\sum_k P(X_{n+m}=j \mid X_n=k, X_0=i)P(X_n=k \mid X_0=i)=\sum_k p_{kj}^{(m)}p_{ik}^{(n)}.$$

■

2.3. Definíció. *1. Azt mondjuk, hogy az i állapotból elérhető j, $(i \to j)$, ha van olyan $n \geq 0$, hogy $p_{ij}^{(n)} > 0$. Ez reflexív ($p_{ii}^{(0)} = 1$) és tranzitív (Chapman-Kolmogorov) reláció.*

2. Azt mondjuk, hogy i és j közlekednek, ha $i \to j$ és $j \to i$. Ez ekvivalenciareláció, tehát osztályokra bontja az állapotteret (csak az átmenetmátrixtól függ). A Markov lánc irreducibilis, ha egyetlen osztályból áll.

3. Az i állapot lényeges, ha $i \to j$ esetén $j \to i$ is teljesül.

Az állapotokra értelmezett valamely tulajdonság osztálytulajdonság, ha egy osztálynak vagy minden eleme ilyen tulajdonságú, vagy egy sem. Triviálisan látszik, hogy a lényegesség osztálytulajdonság, azaz beszélhetünk lényeges és lényegtelen osztályokról. (Biz.: Tegyük fel, hogy i lényeges, j lényegtelen, és $i \to j$. Létezik k, hogy $j \to k$, de $k \not\to j$. Másrészt $i \to j \to k$, tehát i lényegessége miatt $k \to i \to j$, ami ellentmondás.) A lényeges osztályokból nem lehet kijutni (mert akkor vissza is tudnánk jönni, azaz osztályon belül maradnánk), a lényegtelenekből viszont igen: ha elhagytuk őket, akkor többé már nem térhetünk vissza. A lényegtelen osztályok között parciális rendezés van: $C >> D$, ha $i \in C, j \in D$ esetén $i \to j$ (tranzitív, reflexív, antiszimmetrikus).

2.4. Definíció. *Az $\{n > 0 : p_{ii}^{(n)} > 0\}$ halmaz legnagyobb közös osztója az i periódusa, jelölése $d(i)$. Ha a halmaz üres, akkor a periódust nem értelmezzük. Ha $d(i) = 1$, akkor az állapot aperiodikus.*

2.5. Állítás. *Egy osztály minden állapotának ugyanannyi a periódusa.*

Bizonyítás. Legyen $i, j \in C$ azonos osztálybeliek. Ekkor létezik n, m, hogy $p_{ij}^{(n)} > 0$, $p_{ji}^{(m)} > 0$. Ha valamely s-re $p_{jj}^{(s)} > 0$, akkor $p_{ii}^{(n+s+m)} > 0$, $p_{ii}^{(n+2s+m)} > 0$. Emiatt $d(i)|n+s+m$ és $d(i)|n+2s+m$, amiből $d(i)|s$ következik. $d(i)$ tehát közös osztója az ilyen s számoknak, azaz $d(i)|d(j)$. Mivel i és j szerepe felcserélhető, az állítást beláttuk. ■

Azt is meg lehet mutatni, hogy ha C egy $d > 1$ periódusú osztály, akkor C felbomlik d darab $C_0, C_1, \ldots, C_{d-1}$ részosztályra úgy, hogy a lánc a C_i osztályból a C_{i+1} osztályba léphet csak át (a részosztályok számozását modulo d értjük).

Vezessük be a következő jelöléseket:

$$f_{ij}^{(0)} = 0, \quad f_{ij}^{(n)} = P(X_n = j,\, X_k \neq j :\ k = 1,2,\ldots n-1 \mid X_0 = i) \quad n \geq 1,$$

Az $f_{ij}^{(n)}$ mennyiség tehát annak valószínűsége, hogy az i állapotból indulva a lánc először az n. lépésben ér el a j állapotba. Legyen még

$$f_{ij}^{*} = \sum_{n=1}^{\infty} f_{ij}^{(n)}, \ \text{ és } \ g_{ij} = P(X_n = j \text{ végtelen sok } n\text{-re } \mid X_0 = i).$$

2.6. Definíció. *Az i állapot visszatérő vagy rekurrens, ha $f_{ii}^{*} = 1$, egyébként pedig átmeneti vagy tranziens. Ha i visszatérő, akkor az átlagos visszatérési idő $m_i = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$. Az i állapot pozitív rekurrens, ha $m_i < \infty$, ha pedig $m_i = \infty$, akkor nulla rekurrens.*

Könnyen látszik, hogy a nem lényeges állapotok tranziensek, hiszen az i nem lányeges állapotból indulva, a lánc pozitív valószínűséggel előbb hagyja el az osztályt, mint hogy visszatérne i-be. Ha viszont a lánc már kilépett az i osztályából, akkor soha nem fog visszatérni i-be. Fordítva általában nem igaz, azaz a tranziens állapotok lehetnek lényegesek. Az is látszik, hogy ha i rekurrens állapot, akkor $g_{ii} = 1$, ha pedig i tranziens, akkor $g_{ii} = 0$. Ugyanis annak valószínűsége, hogy a lánc legalább n-szer visszatér i-be, éppen $(f_{ii}^*)^n$, g_{ii} pedig ennek a határértéke.

Igaz még a következő két összefüggés:

$$1.\ g_{ij} = f_{ij}^* g_{jj}, \quad 2.\ \text{ha } g_{ii} = 1 \text{ és } f_{ij}^* > 0, \text{ akkor } g_{ij} = 1.$$

Ebből következik, hogy a visszatérőség osztálytulajdonság.

2.7. Tétel. *Az i állapot akkor és csak akkor visszatérő, ha $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$.*

A bizonyítás előtt egy nagyon hasznos kis lemmát látunk be, mely a Toeplitz szummációs tétel speciális esete.

2.8. Lemma. *(Nörlund) Legyenek $a_n \geq 0$ és b_n valós sorozatok, és $\lim_{n\to\infty} b_n = b$. Ha $\lim_{n\to\infty} a_n / \sum_{k=0}^{n} a_k = 0$, akkor*

$$\lim_{n\to\infty} \frac{\sum_{k=0}^{n} a_k b_{n-k}}{\sum_{k=0}^{n} a_k} = b.$$

Bizonyítás. A bizonyítást arra az esetre végezzük el, amikor b véges, a végtelen eset is hasonlóan intézhető el. Legyen B olyan, hogy $|b_n - b| < B$ minden n-re, és adott ϵ-hoz N olyan küszöbindex, hogy $n \geq N$ esetén $|b_n - b| < \epsilon$.

$$|\sum_{k=0}^{n} a_k(b_{n-k} - b)| \leq (\sum_{k=0}^{n-N} a_k)\epsilon + (\sum_{k=n-N+1}^{n} a_k)B.$$

Ebből

$$\limsup_{n\to\infty} \left| \frac{\sum_{k=0}^{n} a_k b_{n-k}}{\sum_{k=0}^{n} a_k} - b \right| \leq \epsilon + B \lim_{n\to\infty} \frac{\sum_{k=n-N+1}^{n} a_k}{\sum_{k=0}^{n} a_k} = \epsilon.$$

■

Megjegyzés: Ha az a_n sorozat korlátos, akkor biztosan teljesíti a lemma feltételét.

Bizonyítás. (2.7 Tételé.)

$$\sum_{r=1}^{n} p_{ij}^{(r)} = \sum_{r=1}^{n} \sum_{k=0}^{r-1} f_{ij}^{(r-k)} p_{jj}^{(k)} = \sum_{k=0}^{n-1} p_{jj}^{(k)} \sum_{s=1}^{n-k} f_{ij}^{(s)} = \sum_{k=0}^{n} a_k b_{n-k},$$

ahol

$$a_k = p_{jj}^{(k)}, \quad b_k = \sum_{s=1}^{k} f_{ij}^{(s)}, \quad b_0 = 0.$$

Ekkor az előző lemmát alkalmazva, $b = f_{ij}^*$, és

$$\lim_{n\to\infty} \frac{\sum_{m=1}^{n} p_{ij}^{(m)}}{\sum_{m=0}^{n} p_{jj}^{(m)}} = f_{ij}^*. \tag{6}$$

Ha ezt az $i = j$ esetre alkalmazzuk, akkor megkapjuk, hogy i akkor és csak akkor visszatérő, ha $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$. ■

Mindezek segítségével kiszámolható például az, hogy az egydimenziós bolyongás akkor és csak akkor visszatérő, ha szimmetrikus.

2.9. Példa. Bolyongás a számegyenesen. Legyen a jobbra lépés valószínűsége p, a balra lépésé $q = 1-p$ $(p,q>0)$. A lánc irreducibilis, periódusa 2. Vizsgáljuk a visszatérőséget! Elég a 0 állapottal foglalkozni.

$$p_{00}^{(2n)} = \binom{2n}{n} p^n (1-p)^n.$$

Felírható, hogy

$$\sum_{n=0}^{\infty} \binom{2n}{n} x^n = (1-4x)^{-1/2}, \text{ ha } 0 < 4x < 1.$$

Ha tehát $p \neq 1/2$, akkor az átmenetvalószínűségek sorösszege $|1-2p|^{-1} < \infty$, azaz a lánc tranziens. Szimmetrikus esetben viszont a lánc rekurrens (A Stirling formula szerint $n! \sim \sqrt{2\pi n}(\frac{n}{e})^n$, és ebből $p_{00}^{(2n)} \sim \frac{1}{\sqrt{\pi n}}$). Tranziens esetben érdekesek az f_{ij}^* valószínűségek. Tegyük fel először, hogy $i > j$. Nyilvánvaló, hogy $f_{ij}^* = (f_{10}^*)^{i-j}$. (6) szerint

$$f_{10}^* = \frac{\sum_{n=1}^{\infty} p_{10}^{(n)}}{\sum_{n=0}^{\infty} p_{00}^{(n)}}.$$

Mármost

$$p_{10}^{(2n+1)} = \binom{2n+1}{n} p^n (1-p)^{n+1} = \frac{1}{2p} \binom{2n+2}{n+1} p^{n+1} (1-p)^{n+1},$$

ezért

$$f_{10}^* = \frac{\frac{1}{2p}\left(\frac{1}{|1-2p|} - 1\right)}{\frac{1}{|1-2p|}} = \frac{1}{2p}(1-|1-2p|) = \begin{cases} 1, & \text{ha} \quad p < 1/2 \\ (1-p)/p, & \text{ha} \quad p > 1/2 \end{cases}.$$

Hasonlóan járhatunk el, ha $i < j$, ehhez csak az

$$f_{01}^* = \begin{cases} p/(1-p), & \text{ha} \quad p < 1/2 \\ 1, & \text{ha} \quad p > 1/2 \end{cases}$$

mennyiség kell. Végül pedig $f_{ii}^* = f_{00}^*$, és

$$f_{00}^* = \frac{\sum_{n=1}^{\infty} p_{00}^{(2n)}}{\sum_{n=0}^{\infty} p_{00}^{(2n)}} = 1 - |1-2p| = \begin{cases} 2p, & \text{ha} \quad p < 1/2 \\ 2(1-p), & \text{ha} \quad p > 1/2 \end{cases}.$$

Tegyük fel, hogy j átmeneti állapot, i pedig tetszőleges. A (6) képlet szerint ekkor $p_{ij}^{(n)} \to 0$, hiszen a tört nevezője véges összeghez tart, határértéke pedig egy nulla és egy közötti szám, tehát a számlálója is véges összeghez kell tartson. Ha tehát a Markov lánc már jó ideje tart, akkor elhanyagolható valószínűséggel tartózkodik a tranziens j állapotban, függetlenül attól, hogy melyik i állapotból indult. Vajon rekurrens állapot esetén mit mondhatunk erről a határértékről?

2.10. Tétel. *Ha i rekurrens állapot d periódussal, akkor* $\lim_{n\to\infty} p_{ii}^{(nd)} = \frac{d}{m_i}$.

Bizonyítás. Vegyük először észre, hogy elég a $d = 1$ esetet igazolni. Ha ugyanis a periódus $d > 1$, akkor az $\{X_0, X_d, X_{2d}, \ldots\}$ láncra áttérve, ott már i aperiodikus állapot lesz, és az átlagos visszatérési idő nyilván az eredetinek d-edrésze.

Az egyszerűség kedvéért vezessük be a következő jelöléseket: $p_n = p_{ii}^{(n)}$, $f_n = f_{ii}^{(n)}$. Legyen még $r_n = \sum_{k=n+1}^{\infty} f_k$ annak a valószínűsége, hogy az első n lépés alatt nem térünk vissza i-be. Igazak a következő összefüggések:

$$\sum_{k=0}^{\infty} r_k = \sum_{k=1}^{\infty} k f_k = m_i, \quad \sum_{k=0}^{n} r_k p_{n-k} = 1,$$

a második összefüggés úgy adódik, hogy a lánc lehetséges meneteit az n. lépésig felbontjuk aszerint, hogy hányadik lépésben járt utoljára az i állapotban (éppen az $n-k$. lépésben). Ha tudnánk, hogy p_n konvergens, akkor a Nörlund lemma alapján a határértékére azonnal adódna $1/m_i$.

Annak bizonyítását, hogy a sorozat konvergens, nem részletezzük, mivel meghaladja ennek a jegyzetnek a kereteit. ■

Azt kaptuk tehát, hogy az i állapot akkor és csak akkor pozitív rekurrens, ha a $\lim_{n\to\infty} p_{ii}^{(nd)}$ határérték pozitív. Az is könnyen látszik, hogy a pozitivitás osztálytulajdonság: ha i és j ugyanabban a rekurrens osztályban vannak, akkor van olyan n és m, hogy $p_{ij}^{(n)} > 0$ és $p_{ji}^{(m)} > 0$. A $p_{ii}^{(m+kd+n)} \geq$ $\geq p_{ij}^{(n)} p_{jj}^{(kd)} p_{ji}^{(m)}$ kifejezésben k-val végtelenhez tartva azt kapjuk, hogy ha j pozitív állapot, akkor szükségképpen i is az.

Most már könnyen bebizonyítható a következő általános tétel.

2.11. Tétel. *Legyen i, j két tetszőleges állapot, és jelölje j periódusát d. Ekkor $r = 1, 2, \ldots, d$ esetén*

$$\lim_{n\to\infty} p_{ij}^{(nd+r)} = f_{ij}^*(r)\frac{d}{m_j},$$

ahol $f_{ij}^*(r) \geq 0$ és $\sum_{r=1}^{d} f_{ij}^*(r) = f_{ij}^*$. (Tranziens állapotra legyen $m_j = \infty$.)

Bizonyítás. Legyen $f_{ij}^*(r) = \sum_{n=0}^{\infty} f_{ij}^{(nd+r)}$. Ekkor

$$p_{ij}^{(nd+r)} = \sum_{k=0}^{n} f_{ij}^{(kd+r)} p_{jj}^{(nd-kd)}.$$

A Nörlund lemmából azonnal következik az állítás, $a_k = f_{ij}^{(kd+r)}$, $b_k = p_{jj}^{(kd)}$ szereposztással. ■

Nézzük meg az általános tételünk néhány speciális esetét!

- Ha j tranziens vagy nulla rekurrens, akkor $p_{ij}^{(n)} \to 0$.
- Ha i és j ugyanabban az aperiodikus, pozitív rekurrens osztályban vannak, akkor $p_{ij}^{(n)} \to \frac{1}{m_j}$.
- Ha i és j ugyanabban a d periódusú, pozitív rekurrens osztályban vannak, méghozzá $j \in C_r(i)$, akkor $p_{ij}^{(nd+r)} \to \frac{d}{m_j}$.

2.12. Példa. Az egydimenziós szimmetrikus bolyongás nulla rekurrens, mivel $p_{00}^{(2n)} \to 0$.

2.13. Példa. Tegyük fel, hogy egy társasjátékban minden körben annyi mezőt ugrik előre a bábunk, ahányat egy dobókockával dobtunk. Jelölje p_n annak esélyét, hogy rálépünk az n-edik mezőre. Mutassuk meg, hogy $p_n \to 2/7$.

A következőkben megvizsgáljuk, hogy mik a Markov lánc egyensúlyi állapotai, ezeket stacionárius eloszlásoknak fogjuk hívni.

2.14. Definíció. *Legyen P egy átmenetmátrix. A $\pi = (\pi_i)_{i\in I}$ eloszlás stacionárius, ha $\pi_i = \sum_{k\in I} \pi_k p_{ki}$ minden $i \in I$-re, azaz a π kezdeti eloszlású, P átmenetvalószínűségű X_n Markov láncra $P(X_n = i) = \pi_i$ minden $i \in I, n \geq 0$-ra. Ez utóbbi esetben azt mondjuk, hogy a Markov lánc stacionárius (ez megfelel a szokásos definíciónak, azaz hogy a véges dimenziós eloszlások eltolás-invariánsak).*

Először vizsgáljuk meg, hogy egy osztályon belül mik a stacionárius eloszlást meghatározó egyenletrendszer megoldásai, azaz tegyük fel, hogy a Markov lánc irreducibilis.

2.15. Tétel. *Irreducibilis Markov láncban a stacionárius eloszlás egyenletrendszerének abszolút konvergens megoldásai $u_i = c/m_i$ (c tetszőleges konstans).*

Bizonyítás. A technikai nehézségek elkerülése érdekében a tételt csak véges állapotterű, aperiodikus láncra bizonyítjuk. Először belátjuk, hogy $u_i = c/m_i$ megoldás. Ha a lánc átmeneti vagy nulla rekurrens, akkor $u_i = 0$, mely nyilván megoldás. Legyen most a lánc pozitív rekurrens. Felírható, hogy

$$p_{ii}^{(n)} = \sum_{k \in I} p_{ik}^{(n-1)} p_{ki}.$$

Tartson n végtelenhez:

$$\frac{1}{m_i} = \lim p_{ii}^{(n)} = \sum_{k \in I} \lim p_{ik}^{(n-1)} p_{ki} = \sum_{k \in I} \frac{1}{m_k} p_{ki},$$

azaz $u_i = c/m_i$ valóban megoldás.

Ezután meg kell mutatni, hogy nincs más megoldás. Legyen u_i egy tetszőleges megoldás. Ekkor

$$u_i = \sum_{k \in I} u_k p_{ki} = \sum_{k \in I} \sum_{j \in I} u_j p_{jk} p_{ki} = \sum_{j \in I} u_j \sum_{k \in I} p_{jk} p_{ki} = \sum_{j \in I} u_j p_{ji}^{(2)},$$

ezt iterálva kapjuk, hogy minden n-re

$$u_i = \sum_{k \in I} u_k p_{ki}^{(n)}.$$

Ha most az $n \to \infty$ határértéket vesszük, akkor

$$u_i = \sum_{k \in I} u_k \lim p_{ki}^{(n)} = \left(\sum_{k \in I} u_k \right) \frac{1}{m_i} = \frac{c}{m_i}.$$

■

Látható tehát, hogy ha az irreducibilis lánc átmeneti vagy nulla rekurrens, akkor nincs stacionárius eloszlása, pozitív rekurrens esetben azonban van. Keressük meg, hogy milyen c konstans esetén kapunk eloszlást, azaz milyen c-re lesz az $u_i = c/m_i$ számok összege 1? Legyen $\pi_i = 1/m_i$, az előző bizonyítás utolsó képletébe visszahelyettesítve ezt a megoldást:

$$\pi_i = \left(\sum_{k \in I} \pi_k \right) \pi_i,$$

vagyis $\sum_{k \in I} \pi_k = 1$, így π a Markov lánc egyértelmű stacionárius eloszlása. Ha a Markov lánc aperiodikus is, akkor minden i, j párra $p_{ij}^{(n)} \to \pi_j$, azaz tetszőleges állapotból indulva, X_n eloszlása tart a stacionárius eloszláshoz.

Ha most olyan Markov láncokat nézünk, melyeknek több osztálya van, akkor a stacionárius eloszlás létezésének szükséges és elégséges feltétele, hogy legyen legalább egy pozitív rekurrens osztály. Ha több ilyen osztály van, akkor a stacionárius eloszlás már nem egyértelmű. Jelölje a pozitív rekurrens osztályokat $D_\alpha : \ \alpha \in A$, és legyen $D = \cup_\alpha D_\alpha$ a pozitív állapotok halmaza. Ekkor a π eloszlás akkor és csak akkor stacionárius, ha

$$\pi_i = \begin{cases} 0 & \text{ha} \quad i \notin D, \\ \lambda_\alpha / m_i & \text{ha} \quad i \in D_\alpha, \end{cases}$$

ahol $\lambda_\alpha \geq 0$, $\sum_\alpha \lambda_\alpha = 1$, vagyis a Markov lánc stacionárius eloszlásait a pozitív osztályokon egyértelműen létező stacionárius eloszlások keverékeként kapjuk meg.

Legyen most a Markov láncunk megint irreducibilis, pozitív rekurrens, és legyen stacionárius, azaz $P(X_n = i) = \pi_i$ minden n, i esetén. Korábban láttuk, hogy a Markov tulajdonság szimmetrikus

a múltat és a jövőt tekintve, tehát ha visszafelé nézzük a láncot, akkor is Markov láncot kapunk, azaz minden N-re az $Y_n = X_{N-n}$ $(n = 0, \dots, N)$ sorozat (véges) stacionárius Markov láncot alkot. Számítsuk ki ennek az új Markov láncnak az egylépéses q_{ij} átmenetvalószínűségeit!

$$q_{ij} = P(Y_1 = j | Y_0 = i) = P(X_{N-1} = j | X_N = i) = \frac{P(X_{N-1} = j, X_N = i)}{P(X_N = i)} = \frac{\pi_j p_{ji}}{\pi_i}.$$

Azt mondjuk, hogy az $\{X_n\}_{n \geq 0}$ Markov lánc megfordítható, ha megfordítása önmaga. Ez tehát azzal ekvivalens, hogy stacionáius eloszlása kielégíti a

$$\pi_i p_{ij} = \pi_j p_{ji} \quad \forall i, j$$

egyenleteket.

2.16. Példa. Vegyünk egy bolyongást a nemnegatív számokon, ahol a nullában egy visszaverő fal van ($p_{01} = 1$). Legyen a jobbra lépés esélye $p < 1/2$, a balra lépésé $q = 1 - p$. A lánc irreducibilis, periódusa $d = 2$. Korábbi számolásunk alapján a lánc visszatérő. Belátjuk, hogy pozitív rekurrens. Próbáljuk ugyanis megoldani a fenti – megfordítható Markov lánc stacionárius eloszlására vonatkozó – egyenleteket!

$$\pi_0 \cdot 1 = \pi_1 q, \quad \pi_{i-1} p = \pi_i q \quad i > 1.$$

Az a megoldás, melynek összege 1:

$$\pi_0 = \frac{1 - 2p}{2 - 2p}, \quad \pi_i = \pi_0 \frac{p^{i-1}}{q^i}, \; i \geq 1.$$

2.17. Példa. (Ehrenfest diffúziós modell) Képzeljünk el egy tartályt, benne N darab molekulát. A tartályt képzeletben két egyenlő részre osztjuk, és azt vizsgáljuk, hogy hány molekula van a két félben. A diffúziót úgy modellezzük, hogy minden lépésben egy véletlenszerűen választott molekula átmegy a másik felébe a tartálynak. Jelölje X_n, hogy n lépés után hány molekula van a tartály első felében. Ez nyilván Markov lánc a $\{0, 1, \dots, N\}$ állapottéren, és

$$p_{i,i-1} = \frac{i}{N}, \quad p_{i,i+1} = \frac{N - i}{N}.$$

Próbáljuk ismét megoldani a megfordítható Markov lánc stacionárius eloszlására vonatkozó egyenleteket!

$$\pi_{i-1} \frac{N - i + 1}{N} = \pi_i \frac{i}{N},$$

ami a $\pi_i = \frac{N-i+1}{i} \pi_{i-1}$ rekurziót adja. Ennek megoldása éppen az N rendű, $1/2$ paraméterű binomiális eloszlás.

A stacionárius eloszlást különösen könnyű megadni abban az esetben, ha az átmenetmátrix minden oszlopának összege is 1 (vagyis duplán sztochasztikus). Ekkor ugyanis az egyenletes eloszlás jó lesz.

Ha a Markov lánc véges állapotterű, akkor (i) minden rekurrens osztály pozitív, és (ii) minden tranziens osztály lényegtelen, és a lánc 1 valószínűséggel előbb-utóbb elhagyja. Ugyanis (i)-hez legyen C rekurrens osztály. Ekkor minden $i \in C$-re és $n \geq 1$-re

$$1 = \sum_{j \in C} p_{ij}^{(n)}.$$

Ha viszont az osztály nulla lenne, akkor $\lim_{n \to \infty} p_{ij}^{(n)} = 0$ lenne minden $j \in C$-re, ami C végessége miatt ellentmond a fentieknek. (ii)-hez legyen C tranziens osztály. Minden $i \in C$-re és $n \geq 1$-re

$$1 = \sum_{j \in C} p_{ij}^{(n)} + \sum_{j \notin C} p_{ij}^{(n)},$$

amiből

$$1 = \sum_{j\in C} \lim_{n\to\infty} p_{ij}^{(n)} + \lim_{n\to\infty} \sum_{j\notin C} p_{ij}^{(n)} = \lim_{n\to\infty} P(X_n \notin C | X_0 = i) = P(\exists n : X_n \notin C | X_0 = i).$$

A fentiekből adódik, hogy véges állapottér esetén az irreducibilis Markov lánc pozitív rekurrens, valamint, hogy minden Markov láncnak van legalább egy pozitív osztálya.

2.18. Példa. Kártyakeverés: hányszor kell megkeverni egy pakli kártyát, hogy a lapok sorrendje teljesen véletlenszerű legyen? Ha a pakli egy keverése azt jelenti, hogy a lapok sorrendjét valamilyen permutáció szerint átrendezzük, akkor a Markov lánc átmenetmátrixa könnyen láthatóan duplán sztochasztikus lesz. Ebben az esetben tehát valóban az egyenletes eloszlás lesz a stacionárius eloszlás, vagyis elég sok keverés után a pakli összes lehetséges sorrendje (nagyjából) egyformán valószínű lesz. Hogy ténylegesen hány keverésre van szükség, az függ a pakli nagyságától és a keverés módszerétől. Ha például egy 52 lapos paklit kettéválasztunk (hogy hol, azt Bin$(52, 1/2)$ eloszlás szerint választjuk), majd a két részt „egymásba pörgetjük" (az összes lehetséges egymásba pörgetést azonos valószínűséggel választva), akkor nagyjából 7 keverés után tekinthető a pakli jól megkevertnek.

2.2. Szimmetrikus bolyongás

A szimmetrikus bolyongásról a Valószínűségszámítás 1. tárgyban már tanultunk. Emlékeztetőül, legyen

$$S_0 = 0, \quad S_n = S_{n-1} + X_n \ (n \geq 1),$$

ahol az X_i változók függetlenek, és $P(X_i = 1) = P(X_i = -1) = 1/2$. A szimmetrikus bolyongás diszkrét idejű Markov lánc az $I = \mathbb{Z}$ állapottéren. Beláttuk, hogy visszatérő, de a visszatérés várható ideje végtelen, azaz nulla visszatérő.

Most további két érdekes kérdésre keressük a választ: ha egy adott időpontig vizsgáljuk a bolyongást (mondjuk az első $2n$ lépést tekintjük), akkor a teljes időtartam hányad részénél van a nullába való utolsó visszatérés, illetve a teljes idő hányad részét tölti a bolyongás a pozitív oldalon? Mindkét kérdésre meglepő lesz a válasz.

Emlékeztetőül, a következő jelöléseket használjuk:

$$u_{2n} = P(S_{2n} = 0) = \binom{2n}{n} \frac{1}{2^{2n}}, \quad f_{2n} = P(S_1 \neq 0, S_2 \neq 0, \ldots, S_{2n-1} \neq 0, S_{2n} = 0).$$

Az első visszatérés valószínűségét a tükrözési elv segítségével meghatároztuk:

$$f_{2n} = \frac{1}{2n} u_{2n-2}.$$

2.1. Lemma. Igazak a következő összefüggések:

1. $P(S_1 \neq 0, S_2 \neq 0, \ldots, S_{2n} \neq 0) = u_{2n}$.
2. $P(S_1 > 0, S_2 > 0, \ldots, S_{2n} > 0) = u_{2n}/2$.
3. $P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n} \geq 0) = u_{2n}$.

Bizonyítás. 1. Felhasználva u_{2n} képletét, egyszerű átalakításokkal adódik, hogy

$$f_{2n} = \frac{1}{2n} u_{2n-2} = \frac{1}{2n-1} u_{2n} = u_{2n-2} - u_{2n}.$$

Ezért

$$P(S_1 \neq 0, S_2 \neq 0, \ldots, S_{2n} \neq 0) = \sum_{k=n+1}^{\infty} f_{2k} = \sum_{k=n+1}^{\infty} (u_{2k-2} - u_{2k}) = u_{2n}.$$

2. Mivel azoknak az utaknak, melyek a $2n$-edik lépésig nem térnek vissza a nullába, fele végig pozitív, fele pedig végig negatív, ezért az előző állítás alapján készen vagyunk.

3. Nézzük annak valószínűségét, hogy egy $2n$ hosszú út végig pozitív. Az első lépés „felfelé" lépés kell legyen, ennek valószínűsége 1/2. Ha most az (1,1) pontot választjuk új origónak, akkor az a feltételünk, hogy a hátralevő $2n-1$ hosszú út ebben az új koordinátarendszerben végig nemnegatív legyen. Ezért

$$P(S_1 > 0, S_2 > 0, \ldots, S_{2n} > 0) = \frac{1}{2} P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n-1} \geq 0).$$

Ha viszont $S_{2n-1} \geq 0$, akkor $S_{2n} \geq 0$ is teljesül, ezért

$$P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n-1} \geq 0) = P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n-1} \geq 0, S_{2n} \geq 0).$$

Ezért az előző állítás alapján készen vagyunk.

■

2.19. Tétel. *Annak a valószínűsége, hogy egy $2n$ hosszú bolyongás során a nullába való utolsó visszatérés a $2k$ időpontban következett be:*

$$\alpha_{2k,2n} = P(S_{2k} = 0, S_{2k+1} \neq 0, S_{2k+2} \neq 0, \ldots, S_{2n} \neq 0) = u_{2k} u_{2n-2k}, \quad k = 0, \ldots, n.$$

Bizonyítás. Az összes út száma 2^{2n}. A kedvező utak száma: olyan $2k$ hosszú útból, melyre $S_{2k} = 0$, éppen $2^{2k} u_{2k}$ darab van, ezt kell folytatni olyan $2n-2k$ hosszú szakasszal, mely nem tér vissza nullába. Ez utóbbiból az előző lemma első állítása szerint $2^{2n-2k} u_{2n-2k}$ darab van. Ezeket egymással tetszőlegesen kombinálva, a $2n$ hosszú kedvező utak száma

$$2^{2k} u_{2k} 2^{2n-2k} u_{2n-2k} = 2^{2n} u_{2k} u_{2n-2k}.$$

■

A tétel következménye az a nem nyilvánvaló tény, hogy az utolsó visszatérés eloszlása szimmetrikus: $\alpha_{2k,2n} = \alpha_{2n-2k,2n}$. Rögzítsük most n-et, és jelölje U_n a nullába való utolsó visszatérés időpontját. Ennek a valószínűségi változónak az eloszlása tehát

$$P(U_n = 2k) = \alpha_{2k,2n}, \quad k = 0, 1, \ldots, n.$$

Az eloszlás szimmetriája miatt $E(U_n) = n$. A meglepő viszont az, hogy ez az eloszlás nem a várható értéke körül koncentrálódik, épp ellenkezőleg, a várható érték körüli értékek a legkevésbé valószínűek! Például $n = 10$ esetén az $\alpha_{2k,20}$ valószínűségek $k = 0, 1, 2, 3, 4, 5$-re:

$$0{,}1762 \quad 0{,}0927 \quad 0{,}0736 \quad 0{,}0655 \quad 0{,}0617 \quad 0{,}0606.$$

A Stirling formula alapján felírható u_{2n} aszimpotikája: $u_{2n} \sim \frac{1}{\sqrt{\pi n}}$. Ezért, ha n és k is elég nagy, akkor

$$\alpha_{2k,2n} \sim \frac{1}{\pi\sqrt{k(n-k)}} = \frac{1}{n} \frac{1}{\pi\sqrt{\frac{k}{n}(1-\frac{k}{n})}} = \frac{1}{n} f\left(\frac{k}{n}\right),$$

ahol $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ a (0,1) intervallumon értelmezett függvény. Ebből közelítést kaphatunk a nullába való utolsó visszatérés eloszlásfüggvényére:

$$P(U_n < x \cdot 2n) = \sum_{k:\frac{k}{n}<x} \alpha_{2k,2n} \approx \frac{1}{n} \sum_{k:\frac{k}{n}<x} f\left(\frac{k}{n}\right) \approx \int_0^x f(y)\mathrm{d}y.$$

Szerencsére az f függvény primitív függvénye explicit megadható: $\frac{2}{\pi}\arcsin\sqrt{y}$, így kapjuk, hogy

$$P(U_n < x \cdot 2n) \approx \frac{2}{\pi}\arcsin\sqrt{x}.$$

Ezért U_n eloszlását n-edrendű diszkrét arkusz-szinusz eloszlásnak nevezik, a 2.19. Tételt pedig arkusz-szinusz törvénynek is hívhatjuk. Nagy n-re azt kapjuk, hogy a nullába való utolsó visszatérés kb. 10% eséllyel az út első 3%-ába esik, kb. 20% eséllyel az út első 10%-ába esik, illetve kb. 30% eséllyel az út első 20%-ába esik. Ezek az eredmények elsőre meglepőek, ellentmondanak az intuíciónknak, hogy a szimmetrikus bolyongás „gyakran" visszatér az x-tengelyre.

Vizsgáljunk meg most egy másik kérdést, mégpedig azt, hogy az útnak hányad része fut a pozitív, és hányad része a negatív oldalon.

2.20. Tétel. *Annak a valószínűsége, hogy egy $2n$ hosszú bolyongásnak $2k$ lépése van a pozitív oldalon:*

$$b_{2k,2n} = u_{2k}u_{2n-2k}, \quad k = 0, \ldots, n.$$

Vagyis a pozitív oldalon megtett lépések számának ugyanaz az eloszlása, mint a nullába való utolsó visszatérésnek. (Ebben az esetben az utak tükrözésével nyilvánvaló, hogy a pozitív oldalon tett lépések számának eloszlása szimmetrikus kell legyen.)

Bizonyítás. n szerinti indukcióval bizonyítunk, az $n = 1$ eset triviális. Tegyük fel, hogy $2n$-nél rövidebb utakra már beláttuk az állítást.

A $2n$ hosszú utakra rátérve, először is $b_{2n,2n} = P(S_1 \geq 0, S_2 \geq 0, \ldots, S_{2n} \geq 0) = u_{2n}$, ahogy korábban láttuk, és szimmetria miatt $b_{0,2n} = u_{2n}$ is teljesül. Legyen most $1 \leq k \leq n-1$, tehát olyan utakat nézünk, melyek a pozitív és negatív oldalon is járnak. Legyen a nullába való első visszatérés időpontja $2r$, ekkor $1 \leq r \leq n-1$. Két eset lehet:
1. eset: az első $2r$ lépés a pozitív oldalon van, ekkor nyilván $r \leq k$, és a megfelelő valószínűség

$$\frac{1}{2}f_{2r}b_{2k-2r,2n-2r}.$$

(f_{2r} az esélye, hogy az első visszatérés időpontja $2r$, $\frac{1}{2}$ az esélye, hogy ez a $2r$ hosszú kezdőszakasz pozitív, és $b_{2k-2r,2n-2r}$ az esélye, hogy a hátralévő szakasznak még $2k-2r$ lépése van a pozitív oldalon.)
2. eset: az első $2r$ lépés a negatív oldalon van, ekkor nyilván $r \leq n-k$, és a megfelelő valószínűség

$$\frac{1}{2}f_{2r}b_{2k,2n-2r}.$$

Ezért

$$b_{2k,2n} = \sum_{r=1}^{k}\frac{1}{2}f_{2r}b_{2k-2r,2n-2r} + \sum_{r=1}^{n-k}\frac{1}{2}f_{2r}b_{2k,2n-2r}.$$

A jobboldalon szereplő b mennyiségek már mind $2n$-nél rövidebb utakra vonatkoznak, azaz használhatjuk az indukciós feltevést:

$$b_{2k,2n} = \frac{1}{2}\sum_{r=1}^{k}f_{2r}u_{2k-2r}u_{2n-2k} + \frac{1}{2}\sum_{r=1}^{n-k}f_{2r}u_{2k}u_{2n-2k-2r}.$$

Felhasználva a Valószínűségszámítás 1. tárgyban már látott

$$u_{2k} = \sum_{r=1}^{k}f_{2r}u_{2k-2r}$$

összefüggést (a $2k$ hosszú, nullába visszatérő utakat aszerint csoportosítjuk, hogy mikor volt a nullába való első visszatérés), kapjuk, hogy

$$\sum_{r=1}^{k} f_{2r} u_{2k-2r} u_{2n-2k} = u_{2n-2k} \sum_{r=1}^{k} f_{2r} u_{2k-2r} = u_{2n-2k} u_{2k},$$

és hasonlóan

$$\sum_{r=1}^{n-k} f_{2r} u_{2k} u_{2n-2k-2r} = u_{2k} \sum_{r=1}^{n-k} f_{2r} u_{2n-2k-2r} = u_{2k} u_{2n-2k}.$$

A kettőt összerakva épp a bizonyítandó képletet kapjuk. ■

Összevetésként nézzük most meg azt az esetet, amikor rögzített végpontú bolyongást vizsgálunk. Tekintsük tehát a (0,0) pontból a $(2n,0)$ pontba vezető utakat. Ilyenből $\binom{2n}{n}$ darab van, és ezek mind egyforma valószínűségűek.

2.21. Tétel. *Annak a valószínűsége, hogy egy $2n$ hosszú, nullába érkező bolyongás $2k$ lépést tesz meg a pozitív oldalon:*

$$\beta_{2k,2n} = \frac{1}{n+1}, \quad k = 0, \ldots, n.$$

Tehát a pozitív oldalon megtett lépések számának minden lehetséges értéke egyformán valószínű! A tétel bizonyításától eltekintünk, az előzőhöz hasonlóan, indukcióval lehet belátni.

Történeti érdekességként nézzük meg, hogy 1876-ban Galton hogyan elemzett adatokat. Egy növényfaj magasságát szerették volna megnövelni valamilyen kezeléssel. Tegyük fel, hogy n növényt kezeltek, másik n-et viszont nem. A kezelt növények magassága $a_1 > a_2 > \ldots > a_n$, míg a kezeletleneké $b_1 > b_2 > \ldots > b_n$ lett. Azt kell eldönteni, hogy hatásos-e a kezelés. Galton megszámolta, hogy hány esetben teljesült az $a_i > b_i$ esemény, azaz kiszámolta a következő statisztikát:

$$G_n = |\{i : \ a_i > b_i\}|.$$

Ha G_n értéke nagy, az a kezelés hatásosságára utal. Galton adatainál $n = 15$ volt, és $G_n = 13$ adódott. Galton úgy találta, hogy ez elegendő bizonyíték a kezelés hatásosságára.

Tegyük fel, hogy a kezelés teljesen hatástalan, azaz a kezelt és a kezeletlen növények magassága ugyanolyan eloszlású. Vegyük az összeöntött $2n$ elemű mintát, és rendezzük nagyság szerint csökkenő sorrendbe. Majd feleltessünk meg ennek egy $2n$ hosszú utat: ha az összeöntött minta i-edik eleme egy kezelt növény adata, akkor az i-edik lépésben felfelé lépjünk egyet, ellenkező esetben lefelé. Így egy $2n$ hosszú, nullába érkező utat kapunk, és a homogenitás nullhipotézise mellett minden ilyen út egyformán valószínű. Azt kell észrevenni, hogy $a_i > b_i$ pontosan akkor következik be, ha a most definiált út $(2i-1)$-edik és $2i$-edik lépése a pozitív oldalon van. Vagyis $G_n = k$ azzal ekvivalens, hogy az útnak pontosan $2k$ lépése van a pozitív oldalon, azaz a nullhipotézis mellett

$$P(G_n = k) = \frac{1}{n+1}, \quad k = 0, \ldots, n.$$

Vagyis azt kaptuk, hogy teljesen hatástalan kezelés esetén is több, mint 6% az esélye, hogy $G_{15} = 15$, vagyis a kezelt csoport i-edik legmagasabb növénye magasabb, mint a kezeletlen csoport megfelelő példánya, minden i-re. Tehát a $G_{15} = 13$ érték egyáltalán nem mondható meggyőző bizonyítéknak a kezelés hatásosságára.

Galton eljárása nem elég erős, viszont az alapötlet nem rossz: vegyük az ömlesztett mintát, és nézzük meg, hogy ezen belül hányadik pozíciókban helyezkednek el a kezelt csoport adatai. Ez az alapja a korábban tanult Wilcoxon próbának is.

2.3. Folytonos idejű Markov láncok

Továbbra is olyan folyamatokat vizsgálunk, melyek jövőbeli fejlődése csak a pillanatnyi állapottól függ, a múltbeliektől nem. Azonban most az idő folytonosan telik, azaz minden $t \geq 0$ esetén adott az $X_t : \Omega \to I$ valószínűségi változó, mely a folyamat állapotát adja meg a t időpillanatban. Legyen

$$p_{ij}(t) = P(X_{s+t} = j | X_s = i), \quad t \geq 0,$$

ahol az átmenetvalószínűség megint nem függ s-től. A $p_{ij}(t)$ értékekből alkotott mátrixot jelölje $P(t)$ ($P(0)$ az egységmátrix). Ekkor $P(t)$ sztochasztikus mátrix, és a Markov tulajdonság miatt teljesül a Chapman-Kolmogorov összefüggés, azaz $P(s+t) = P(s)P(t)$. Ha adottak a $P(t)$ mátrixok valamilyen $[0, \epsilon)$ intervallumon, akkor a Chapman-Kolmogorov egyenletek miatt $P(t)$ tetszőleges $t > 0$ értékre kiszámítható. Ha ugyanis n elég nagy, akkor $t/n < \epsilon$, és

$$P(t) = P(t/n)^n.$$

Mivel most nincs „legrövidebb lépésköz," ezért bonyolultabb a helyzet, mint a diszkrét paraméter esetén, ahol elegendő volt az egylépéses átmenetmátrixot megadni. Azonban ha a $P(t)$ mátrixcsaládot ismerjük a nulla környezetében, az már meghatározza a családot minden t-re.

Megmutatható, hogy bizonyos feltételek mellett (melyek az alkalmazásokban mindig teljesülnek, és most nem részletezzük őket) a $p_{ij}(t)$ függvények szépek: folytonosak, sőt deriválhatók a $(0, \infty)$ félegyenesen, sőt a 0-ban is létezik a jobboldali deriváltjuk. Legyen ez a jobboldali derivált $p'_{ij}(0) = = q_{ij}$, azaz

$$\begin{aligned} \lim_{t \to 0} \frac{p_{ii}(t) - 1}{t} &= q_{ii} (\leq 0), \\ \lim_{t \to 0} \frac{p_{ij}(t) - 0}{t} &= q_{ij} (\geq 0), \ i \neq j. \end{aligned}$$

2.1. Definíció. A folytonos idejű $\{X_t\}_{t \geq 0}$ Markov lánc (infinitezimális) generátora a $Q = P'(0)$ mátrix. Ennek főátlójában a $q_{ii} = -q_i$ nempozitív értékek állnak, a főátlón kívül pedig a q_{ij} nemnegatív értékek ($i \neq j$). Mivel a $P(t)$ mátrix minden sorösszege 1, deriválva azt kapjuk, hogy a Q generátormátrix minden sorösszege nulla:

$$\forall i : \quad \sum_j q_{ij} = 0, \text{ más alakban } \sum_{j \neq i} q_{ij} = q_i.$$

A nullában vett derivált azt fejezi ki, hogy kis idő alatt mekkora valószínűséggel jut a lánc az i állapotból a j-be:

$$\begin{aligned} p_{ij}(h) = P(X_{t+h} = j | X_t = i) &= q_{ij} h + o(h), \ j \neq i \\ p_{ii}(h) = P(X_{t+h} = i | X_t = i) &= 1 - q_i h + o(h), \end{aligned}$$

ahol $o(h)$ olyan mennyiséget jelöl, amely h-val osztva nullához tart, ha $h \to 0$.

Jelölje most T_i, hogy a lánc mennyi ideig tartózkodik az i állapotban. A Markov tulajdonság szerint, ha a lánc az i állapotban van, akkor az, hogy még mennyi ideig fog itt tartózkodni, nem függ a múlttól, azaz attól sem, hogy mióta van már az i állapotban. Ez pont azt jelenti, hogy T_i eloszlása örökifjú kell legyen! Ezt adja a következő pongyola levezetés is: tegyük fel, hogy $X_0 = i$. Ekkor

$$P(T_i > x) = P(X_t = i, 0 \leq t \leq x) = \lim_{n \to \infty} p_{ii}(x/n)^n = \lim_{n \to \infty} \left(1 - q_i \frac{x}{n} + o\left(\frac{x}{n}\right)\right)^n = e^{-q_i x}.$$

Vagyis T_i exponenciális eloszlású q_i paraméterrel. Adott Q generátor esetén a következőképpen képzelhetjük el a Markov láncot: a kiinduló állapotot (jelölje ezt i) a kezdeti eloszlás szerint választjuk, majd ott tartózkodunk $\mathrm{Exp}(q_i)$ eloszlású ideig. Ekkor átugrunk valamelyik másik állapotba, méghozzá a j állapotot q_{ij}/q_i valószínűséggel választjuk. Ezt folytatjuk tovább.

Ugyanezt a konstrukciót elmondhatjuk úgy is, hogy ha egy ugrás az i állapotba vitt, akkor az összes többi állapotban azonnal ketyegni kezd egy-egy óra, a j állapot órája $\mathrm{Exp}(q_{ij})$ eloszlású idő után csörög. Amikor az első óra megszólal, átugrunk a hozzá tartozó állapotba. Ezt bizonyítja a következő lemma.

2.2. Lemma. Ha $X_i \sim \mathrm{Exp}(\lambda_i)$ függetlenek, akkor $Y = \min X_i$ is exponenciális eloszlású, $\lambda = \sum \lambda_i$ paraméterrel. Továbbá

$$P(Y = X_i) = \frac{\lambda_i}{\lambda}.$$

Bizonyítás. Az első állítás bizonyítása: minden $x > 0$ esetén

$$P(Y > x) = P(X_i > x \ \forall i) = \prod_i P(X_i > x) = \prod_i e^{-\lambda_i x} = e^{-\sum_i \lambda_i x} = e^{-\lambda x},$$

vagyis Y valóban a kívánt exponenciális eloszlású. A második állításra rátérve, akkor lesz a minimum értéke X_i, ha X_i kisebb az összes többi változó minimumánál. Legyen tehát $Z = \min_{j:j\neq i} X_j$. Ezzel a jelöléssel az $\{Y = X_i\}$ esemény azzal ekvivalens, hogy $X_i < Z$. Az első állítás szerint Z is exponenciális eloszlású, $\mu = \sum_{j:j\neq i} \lambda_j$ paraméterrel, és független X_i-től. Kiintegrálva,

$$P(X_i < Z) = \int_0^\infty \int_x^\infty \lambda_i e^{-\lambda_i x} \mu e^{-\mu y} \mathrm{d}y \, \mathrm{d}x = \int_0^\infty \lambda_i e^{-\lambda_i x} e^{-\mu x} \, \mathrm{d}x = \frac{\lambda_i}{\lambda_i + \mu} = \frac{\lambda_i}{\lambda}.$$

■

Határozzuk most meg a $p_{ij}(t)$ függvények deriváltját! Rögtön mátrixos alakba írva:

$$P'(t) = \lim_{h \searrow 0} \frac{P(t+h) - P(t)}{h} = P(t) \lim_{h \searrow 0} \frac{P(h) - P(0)}{h} = P(t)P'(0) = P(t)Q,$$

ahol felhasználtuk a $P(t+h) = P(t)P(h)$ Chapman-Kolmogorov összefüggést. Ha most fordítva, a $P(t+h) = P(h)P(t)$ egyenlőséget használjuk, akkor hasonlóan kapjuk, hogy

$$P'(t) = \lim_{h \searrow 0} \frac{P(t+h) - P(t)}{h} = \left(\lim_{h \searrow 0} \frac{P(h) - P(0)}{h} \right) P(t) = P'(0)P(t) = QP(t).$$

Ezt a két differenciálegyenlet-rendszert nevezik Kolmogorov-féle differenciálegyenleteknek, mégpedig

$$\begin{aligned} P'(t) = P(t)Q \quad &: \quad \text{előre (forward) egyenlet} \\ P'(t) = QP(t) \quad &: \quad \text{hátra (backward) egyenlet} \end{aligned}$$

Az elnevezést az magyarázza, hogy pl. az előre egyenletnél, amikor azt vizsgáltuk, hogy mekkora valószínűséggel jut a lánc $t+h$ idő alatt i-ből j-be, akkor azt aszerint bontottuk fel, hogy i-ből hová jut a lánc t idő alatt, majd előrefelé menve az időben, hozzátettünk még egy kis h hosszú intervallumot, így kapva meg a $(t+h)$-t. A hátrafelé egyenletnél viszont a t hosszú intervallum elé tettük a kis h hosszú szakaszt, azaz hátrafelé mentünk az időben.

Megjegyezzük, hogy a fenti levezetésben szerepel egy összegzés és egy határérték felcserélése, amely nem mindig tehető meg. Belátható, hogy a hátrafelé egyenletek mindig teljesülnek, az előrefelé egyenletek teljesüléséhez azonban további technikai feltételekre van szükség.

A fő feladatunk ezek után az, hogy adott Q generátormátrix mellett megoldjuk a Kolmogorov-féle differenciálegyenleteket, ezzel megkapva a $P(t)$ $(t \geq 0)$ átmenetmátrix-családot. Szép esetekben a differenciálegyenleteknek egyértelműen létezik a megoldásuk, és az így kapott $P(t)$ család sztochasztikus mátrixokból áll, melyek eleget tesznek a Chapman-Kolmogorov egyenleteknek is, azaz valóban egy Markov láncot határoznak meg.

Példaként nézzük meg az úgynevezett születési folyamatokat. Ezek olyan Markov láncok, melyek állapottere a természetes számok halmaza, és a Q infinitezimális generátorban csak a főátló és a fölötte lévő elemek nem nullák, azaz

$$q_i = q_{i,i+1} = \lambda_i, \quad i = 0,1,\dots.$$

λ_i az i állapothoz tartozó születési intenzitás, ezekről feltesszük, hogy mind pozitívak. Ez tehát egy olyan folyamatot fog meghatározni, mely egy adott i állapotban $\mathrm{Exp}(\lambda_i)$ ideig tartózkodik, majd átugrik az $(i+1)$ állapotba, és így tovább. Tegyük fel, hogy $X_0 = 0$, és legyen $r_n(t) = P(X_t = n) = p_{0n}(t)$, valamint $r(t) = (r_0(t), r_1(t), \dots)$. Ismerjük $r(0)$-t, kérdés, hogy meg tudjuk-e határozni $r(t)$-t? Írjuk fel az előre egyenleteket! Az első egyenlet $r_0'(t) = -\lambda_0 r_0(t)$, melynek az $r_0(0) = 1$ kezdeti feltételt kielégítő egyértelmű megoldása $r_0(t) = e^{-\lambda_0 t}$. Legyen most $n \geq 1$:

$$r_n'(t) = \lambda_{n-1} r_{n-1}(t) - \lambda_n r_n(t),$$

és $r_n(0) = 0$ a kezdeti feltétel. Próbáljunk rekurzívan haladni! Legyen $v_n(t) = e^{\lambda_n t} r_n(t)$, erre $v_n'(t) = e^{\lambda_n t} \lambda_{n-1} r_{n-1}(t)$. Ennek megoldása

$$v_n(t) = e^{\lambda_n t} r_n(t) = \lambda_{n-1} \int_0^t e^{\lambda_n x} r_{n-1}(x) dx,$$

azaz

$$r_n(t) = \lambda_{n-1} e^{-\lambda_n t} \int_0^t e^{\lambda_n x} r_{n-1}(x) dx. \tag{7}$$

Hasonlóan kaphatjuk meg a $p_{ij}(t)$ átmenetvalószínűségeket is, hiszen az i állapotból indulva is egy születési folyamatunk van, csak át kell indexelni az állapotokat. Tehát megkaptuk, hogy az előrefelé egyenleteknek egyértelműen létezik a megoldásuk. Kérdés, hogy vajon sztochasztikus mátrixot kapunk-e a megoldásból? Nem feltétlenül! Belátható, hogy a kapott $P(t)$ mátrixok akkor és csak akkor lesznek sztochasztikusak, ha $\sum_{n=0}^{\infty} 1/\lambda_n = \infty$. Ha a $P(t)$ mátrix nem sztochasztikus, az szemléletesen azt jelenti, hogy a lánc véges idő alatt kimegy a végtelenbe, azaz felrobban.

Fontos speciális eset a Poisson folyamat, amikor mindegyik születési intenzitás megegyezik, $\lambda_i = \lambda$. Ekkor tehát mindegyik állapotban $\mathrm{Exp}(\lambda)$ ideig tartózkodik a lánc, mielőtt tovább ugrana az $(i+1)$ állapotba. A fenti rekurzív megoldást elkezdve kiírogatni, hamar felismerhető, hogy ebben az esetben

$$p_{ij}(t) = e^{-\lambda t} \frac{(\lambda t)^{j-i}}{(j-i)!}, \text{ ha } j \geq i,$$

azaz t idő alatt a folyamat ugrásainak száma $\mathrm{Poisson}(\lambda t)$ eloszlású.

Másik speciális eset a Yule folyamat, ennek állapottere a pozitív egész számok halmaza, és $\lambda_i = i\lambda$, $i = 1, 2, \dots$. Ez a folyamat egy olyan populáció növekedését írja le, amikor minden egyed, egymástól függetlenül, λ intenzitással osztódik ketté. Ha ismét elkezdjük a fenti rekurziót kiszámolni, láthatóvá válik, hogy a megoldás

$$p_{1j}(t) = e^{-\lambda t}(1 - e^{-\lambda t})^{j-1}.$$

Azt kaptuk tehát, hogy egyetlen egyedből indulva, t idő elteltével a populáció mérete $\mathrm{Geo}(e^{-\lambda t})$ eloszlású. Ha k egyedből indulunk, akkor ők egymástól függetlenül szaporodnak, vagyis a $P(t)$ átmenetmátrix k-adik sorában a $\mathrm{Negbin}(k, e^{-\lambda t})$ eloszlás jelenik meg.

Vizsgáljuk most meg az állapotok osztályozását! Folytonos idejű Markov láncoknál is mondhatjuk, hogy a j állapot elérhető i-ből, mégpedig akkor, ha van $t \geq 0$, melyre $p_{ij}(t) > 0$. Két állapot akkor érintkezik, ha kölcsönösen elérhetők egymásból. Ez nyilván ekvivalenciareláció, mely osztályokra bontja az állapotteret. A lényegesség ugyanúgy definiálható, mint diszkrét időben, viszont a periódusnak folytonos időben nincs értelme.

Gyakran érdemes vizsgálni a Markov lánc diszkretizáltját: ez azt jelenti, hogy választunk egy $h > 0$ időegységet, és csak időegységenként nézünk rá a láncra. Formálisan tehát definiálhatjuk az

$$\tilde{X}_n = X_{nh}, \quad n = 0, 1, \dots$$

folyamatot, mely egy diszkrét idejű Markov lánc az I állapottéren, átmenetmátrixa $\tilde{P} = P(h)$. Ha i-ből elérhető j a diszkretizált láncban, akkor a folytonosban is, és a következő lemma szerint ez visszafelé is igaz.

2.22. Lemma. *$p_{ii}(t) > 0$ minden t-re, és $p_{ij}(t_0) > 0$ esetén $p_{ij}(t) > 0$ minden $t \geq t_0$.*

Bizonyítás. Az első állítás azért igaz, mert $p_{ii}(t) \geq p_{ii}(t/n)^n$ minden n-re, és mivel $p_{ii}(t)$ folytonos és határértéke a nullában 1, ezért elég nagy n-re $p_{ii}(t/n) > 0$. A második állítás pedig triviális a

$$p_{ij}(t) \geq p_{ij}(t_0)p_{jj}(t - t_0) > 0$$

egyenlőtlenség miatt. ■

Megjegyezzük, hogy ennél több is igaz: ha $p_{ij}(t_0) > 0$ valamely t_0-ra, akkor $p_{ij}(t) > 0$ minden pozitív t-re. Tehát az állapotok osztályozása és lényegessége megegyezik az összes diszkretizáltban és a folytonos idejű láncban. Továbbá a diszkretizáltakban minden állapot aperiodikus ($p_{ii}(h) > 0$).

Térjünk rá a visszatérőség vizsgálatára! Emlékezzünk rá, hogy diszkrét időben az i állapot akkor és csak akkor volt visszatérő, ha $\sum_{n=0}^{\infty} p_{ii}^{(n)} = \infty$. Vagyis a folytonos idejű Markov lánc h-diszkretizáltjában i akkor és csak akkor visszatérő, ha $\sum_{n=0}^{\infty} p_{ii}(nh) = \infty$. Megmutatható, hogy a $p_{ii}(t)$ függvény integrálja alulról és felülről közelíthető ilyen összegekkel, azaz minden h-ra van olyan $\delta(h) > 0$ szám, hogy minden N-re

$$\delta(h)h \sum_{n=0}^{N-1} p_{ii}(nh) \leq \int_0^{Nh} p_{ii}(t)dt \leq \frac{1}{\delta(h)} h \sum_{n=1}^{N} p_{ii}(nh).$$

A folytonos idejű Markov láncban mondjuk azt, hogy az i állapot visszatérő, ha $\int_0^{\infty} p_{ii}(t)dt = \infty$. Ezzel a definícióval az állapotok visszatérősége nem függ attól, hogy a folytonos idejű láncban, vagy valamelyik diszkretizáltban tekintjük.

Egy szemléletesebb interpretációja a visszatérőségnek a következő. Tekintsünk egy i állapotot, és tegyük fel, hogy $X_0 = i$. Legyen $S_i = \{t : X_t = i\}$ azoknak az időpontoknak a halmaza, amikor a lánc az i állapotban van (ez intervallumok uniója), és jelölje a számegyenesen a hosszúságot ℓ. Megmutatható, hogy ha i visszatérő, akkor $P(\ell(S_i) = \infty) = 1$, ha viszont i átmeneti, akkor $P(\ell(S_i) = = \infty) = 0$.

Diszkrét esetben a visszatérő állapotokat tovább osztottuk pozitív visszatérő és nulla visszatérő állapotokra. Legyen most a folytonos idejű láncunk irreducibilis és rekurrens. A h-diszkretizált lánc akkor és csak akkor lesz pozitív rekurrens, ha a $\lim_{n\to\infty} p_{ij}(nh)$ határérték pozitív. Megmutatható, hogy a folytonos láncban is létezik a $\lim_{t\to\infty} p_{ij}(t)$ határérték (és ennek persze meg kell egyeznie a $p_{ij}(nh)$ sorozatok határértékével, bármely h-ra). Vagyis a láncunk nulla rekurrens, ha $\lim_{t\to\infty} p_{ij}(t) = = 0$ minden i, j párra, és pozitív rekurrens, ha $\lim_{t\to\infty} p_{ij}(t) = \pi_j > 0$ minden i, j párra, ahol π az egyértelműen létező stacionárius eloszlás.

Hogyan tudjuk vajon a stacionárius eloszlást a generátor mátrixból kiszámítani? A stacionárius eloszlást definiáló $\pi^T P(t) = \pi^T$ (minden $t \geq 0$) egyenletet a nullában deriválva:

$$0 = (\pi^T)' = (\pi^T P(t))' = \pi^T P'(0) = \pi^T Q,$$

azaz a $\pi^T Q = 0$ egyenletet kaptuk.

Csakúgy, mint a diszkrét esetben, most is ellenőrizhető, hogy ha valamely π-re $\pi_i q_{ij} = \pi_j q_{ji}$ minden i, j párra teljesül, akkor $\pi^T Q = 0$.

Fontos folyamatok az úgynevezett születési-halálozási folyamatok. Ezek állapottere a természetes számok halmaza, és Q olyan mátrix, melyben $q_{i,i+1} = \lambda_i > 0$ a születési intenzitások, és $q_{i,i-1} = \mu_i > > 0$ a halálozási intenzitások, $q_{ii} = -(\lambda_i + \mu_i)$, és minden más elem nulla. Az így definiált Markov lánc nyilván irreducibilis.

Legyen $\rho_0 = 1$, $\rho_n = \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_i}$. A $\pi_i q_{ij} = \pi_j q_{ji}$ egyenleteket a születési-halálozási folyamatokra felírva, $\pi_i \lambda_i = \pi_{i+1}\mu_{i+1}$ adódik, melynek megoldása $\pi_i = \rho_i \pi_0$.

Legyen most még $\tilde{\rho}_0 = 1$, $\tilde{\rho}_n = \prod_{i=1}^{n} \frac{\mu_i}{\lambda_i}$.

2.23. Tétel. (Karlin-McGregor tétel) *Tekintsünk egy születési-halálozási folyamatot a természetes számokon, és legyen $\lambda_i, \mu_i > 0$ minden szóbajövő i-re.*
(i) Akkor és csak akkor létezik a Kolmogorov-egyenleteknek egyértelmű (szubsztochasztikus) megoldása, ha

$$S = \sum_{i=0}^{\infty} \rho_i = \infty \text{ vagy } \tilde{S} = \sum_{i=0}^{\infty} \tilde{\rho}_i = \infty.$$

Ebben az esetben a lánc
(ii) tranziens, ha $S = \infty$ és $\tilde{S} < \infty$,
(iii) nulla rekurrens, ha $S = \infty$ és $\tilde{S} = \infty$,
(iv) pozitív rekurrens, ha $S < \infty$ és $\tilde{S} = \infty$.

2.24. Példa. (M/M/1 sor) Egy rendszerbe λ-Poisson folyamat szerint érkeznek igények, melyeket egy kiszolgáló egység szolgál ki (érkezési sorrendben). A kiszolgálási idő eloszlása $\text{Exp}(\mu)$. Ha X_t jelöli azt, hogy a t időpontban hány igény tartózkodik a rendszerben, születési-halálozási folyamatot kapunk, melyre $\lambda_i = \lambda$, $\mu_i = \mu$. Az eddigiek alapján $\lambda > \mu$ esetén a lánc tranziens, $\lambda = \mu$ esetben nulla rekurrens, $\lambda < \mu$ esetén pozitív rekurrens. Utóbbi esetben a stacionárius eloszlás

$$\pi_i = \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right), \quad i = 0,1,\ldots.$$

Stacionaritás esetén a rendszerben tartózkodó igények várható száma $\frac{\lambda}{\mu - \lambda}$.

2.25. Példa. (M/M/∞ sor) Az előző példán annyit módosítsunk, hogy végtelen sok kiszolgáló egység van, tehát minden beérkező igényt azonnal elkezdhetünk kiszolgálni. Most a $\lambda_i = \lambda$, $\mu_i = i\mu$ paraméterek szerepelnek a generátor mátrixban. Ez a lánc pozitív rekurrens lesz, stacionárius eloszlása $\text{Poisson}(\lambda/\mu)$.

2.4. Felújítási folyamatok

Ebben a szakaszban olyan folytonos idejű folyamatokkal foglalkozunk, melyek bizonyos események, „felújítások" számát vizsgálják. Tegyük fel, hogy a folyamatot egy felújítás időpontjában kezdjük el nézni, azaz a $t = 0$ időpontban történt egy felújítás (ez a nulladik felújítás). A felújítások között eltelt X_i időtartamokról feltesszük, hogy függetlenek, azonos eloszlásúak, és $P(X_i > 0) = 1$. Még pontosabban, legyen X_i az $(i-1)$-edik és az i-edik felújítás között eltelt idő, ekkor az n-edik felújítás időpontja:

$$S_n = \sum_{i=1}^{n} X_i, \text{ ha } n \geq 1, \quad S_0 = 0.$$

Jelölje $N(t)$ a $(0,t]$ intervallumban a felújítások számát, ekkor $N(0) = 0$ és

$$N(t) = \max\{n : \ S_n \leq t\} \quad t > 0.$$

Az $\{N(t)\}_{t \geq 0}$ folyamatot nevezzük felújítási folyamatnak. Legyen még $M(t) = E(N(t))$ a felújítások várható száma $(0,t]$-ben, ezt nevezzük felújítási függvénynek.

Felújítási folyamatokat lehet alkalmazni például készletezési feladatokban, sorbanállási problémáknál, részecskeszámláló berendezések vizsgálatánál. Egy diszkrét idejű, pozitív rekurrens Markov láncban például egy adott állapotba való visszatérések időpontjai felújítási folyamatot alkotnak.

2.5. A Poisson folyamat

A legegyszerűbb felújítási folyamat a Poisson folyamat, mellyel már a Markov láncoknál is találkoztunk. Ekkor $X_i \sim \text{Exp}(\lambda)$. Az exponenciális eloszlás örökifjú tulajdonságából kaptuk, hogy ilyenkor a

diszjunkt intervallumokban történt felújítások száma független egymástól (azaz a folyamat független növekményű), és egy s hosszúságú intervallumban a felújítások számának eloszlása Poisson($s\lambda$):

$$N(t+s) - N(t) \sim \text{Poisson}(s\lambda) \quad \forall t \geq 0, s > 0,$$

valamint minden n-re

$$N(t_n) - N(t_{n-1}), N(t_{n-1}) - N(t_{n-2}), \ldots, N(t_1) - N(t_0) \text{ függetlenek, ha } 0 = t_0 < t_1 < \ldots < t_n.$$

Érdekes észrevétel, hogy ha egy Poisson folyamatot kiritkítunk (azaz minden felújítási pontot, egymástól és a folyamattól függetlenül p valószínűséggel tartunk meg, $1-p$ valószínűséggel pedig kitöröljük), akkor szintén Poisson folyamatot kapunk. Az is igaz, hogy ha két független, λ illetve μ paraméterű (intenzitású) Poisson folyamatot egyesítünk, akkor $\lambda + \mu$ paraméterű Poisson folyamatot kapunk. Továbbá a Poisson folyamat minden pillanatban újra kezdődik, azaz ha a t_0 időpontban kezdjük el nézni a folyamatot, akkor $\{N(t_0+t) - N(t_0)\}_{t \geq 0}$ is Poisson folyamat.

Bizonyítás nélkül jegyezzük meg, hogy Poisson folyamat esetén az $N(t) = n$ feltétel mellett ez az n felújítás úgy helyezkedik el a $(0,t]$ intervallumban, mintha n pontot egyenletes eloszlás szerint választottunk volna. Ezt támasztja alá például az alábbi számolás $s < t$-re és $k \leq n$-re:

$$P(N(s) = k | N(t) = n) = \frac{P(N(s) = k, N(t) - N(s) = n - k)}{P(N(t) = k)} = \\ \frac{e^{-\lambda s} \frac{(\lambda s)^k}{k!} e^{-\lambda(t-s)} \frac{(\lambda(t-s))^{n-k}}{(n-k)!}}{e^{-\lambda t} \frac{(\lambda t)^n}{n!}} = \binom{n}{k} \left(\frac{s}{t}\right)^k \left(1 - \frac{s}{t}\right)^{n-k},$$

azaz az $\{N(t) = n\}$ feltétel mellett $N(s) \sim \text{Bin}(n, s/t)$.

A Poisson folyamat esetén a felújítási függvény nyilván $M(t) = \lambda t$.

2.6. Általános eredmények

A felújítási folyamattal kapcsolatban az egyik lényeges észrevétel, hogy

$$\{N(t) \geq k\} \iff \{S_k \leq t\}.$$

Mivel S_k független, azonos eloszlásúak összege, így S_k eloszlását jól ismerjük. Például könnyen lehet nagy számok törvényét bizonyítani a felújítási folyamatra. Vegyük ugyanis észre, hogy

$$S_{N(t)} \leq t < S_{N(t)+1},$$

ahol $S_{N(t)}$ jelöli a $(0,t]$ intervallumban az utolsó felújítás időpontját, $S_{N(t)+1}$ pedig a t utáni első felújítás időpontját. Tegyük fel a továbbiakban, hogy $\mu = E(X_i)$ véges. Ekkor felhasználva, hogy $\lim_{t\to\infty} N(t) = \infty$, a nagy számok erős törvénye alapján

$$\mu = \lim_{t\to\infty} \frac{S_{N(t)}}{N(t)} \leq \lim_{t\to\infty} \frac{t}{N(t)} \leq \lim_{t\to\infty} \frac{S_{N(t)+1}}{N(t)} = \mu.$$

Megkaptuk tehát, hogy

$$\lim_{t\to\infty} \frac{N(t)}{t} = \frac{1}{\mu} \quad 1 \text{ valószínűséggel.}$$

Bizonyítás nélkül megjegyezzük, hogy ha $D^2(X_i) = \sigma^2 < \infty$, akkor centrális határeloszlás-tétel is érvényes:

$$\frac{N(t) - t/\mu}{\sqrt{t}\frac{\sigma}{\mu^{3/2}}} \to \text{N(0,1) eloszlásban.}$$

2.26. Tétel. (Wald azonosság)

$$E(S_{N(t)+1}) = \mu(M(t)+1).$$

Bizonyítás. $S_{N(t)+1} = \sum_{i=1}^{N(t)+1} X_i$, ez tehát egy véletlen tagszámú összeg. Ismert, hogy ha a véletlen tagszámú összeg tagszáma független az összeadandóktól, akkor az összeg várható értéke megegyezik egy tag várható értékének és a tagok számának várható értékének szorzatával. Esetünkben a tagok száma nem független a tagoktól, az eredmény mégis igaz:

$$S_{N(t)+1} = \sum_{i=1}^{\infty} X_i I(S_{i-1} \le t),$$

ahol az X_i és az $I(S_{i-1} \le t)$ valószínűségi változók függetlenek. Ezért

$$E(S_{N(t)+1}) = \sum_{i=1}^{\infty} E(X_i)E(I(S_{i-1} \le t)) = \mu \sum_{i=1}^{\infty} P(S_{i-1} \le t) =$$
$$\mu \sum_{i=1}^{\infty} P(N(t) \ge i-1) = \mu(1 + \sum_{i=0}^{\infty} P(N(t) > i)) = \mu(1+M(t)).$$

A bizonyításban felhasználtuk, hogy ha Z nemnegatív egész értékű valószínűségi változó, akkor $E(Z) = \sum_{i=0}^{\infty} P(Z > i)$. ■

Például a Poisson folyamat esetén az örökifjú tulajdonság miatt $E(S_{N(t)+1}) = t + \mu$, és $M(t) = t/\mu$, tehát valóban teljesül az egyenlőség. Ugyanakkor például általában $E(S_{N(t)}) \neq \mu M(t)$, a Poisson folyamaton ellenőrizve:

$$t = \mu M(t) > E(S_{N(t)}),$$

hiszen $S_{N(t)} < t$ 1 valószínűséggel.

2.27. Tétel. (Elemi felújítási tétel) *Ha $E(X_i) = \mu < \infty$, akkor*

$$\lim_{t\to\infty} \frac{M(t)}{t} = \frac{1}{\mu}.$$

Bizonyítás. ■

A felújítási elmélet központi eredménye az úgynevezett felújítási tétel. Ez egy mély, nehéz tétel, így nem fogjuk bizonyítani. Kimondásához is szükségünk van némi előkészületre.

...

2.28. Tétel. *Tegyük fel, hogy X_i eloszlása nem rácsos. Ekkor minden $h > 0$ esetén*

$$\lim_{t\to\infty} (M(t) - M(t-h)) = \frac{h}{\mu},$$

azaz ha a folyamat már régóta folyik, akkor egy h hosszú intervallumban a felújítások várható száma körülbelül h/μ.

Most három érdekes valószínűségi változó határeloszlását fogjuk vizsgálni. A felújítások nyelvén fogalmazva, ha a t időpillanatban nézünk rá a rendszerre, akkor az éppen működő alkatrészhez a következő három élettartamot rendelhetjük hozzá:

γ_t	=	$S_{N(t)+1} - t$	hátralévő élettartam
δ_t	=	$t - S_{N(t)}$	eddigi/pillanatnyi élettartam
β_t	=	$S_{N(t)+1} - S_{N(t)}$	teljes élettartam

2.29. Tétel. *Tegyük fel, hogy X_i eloszlása nem rácsos, jelölje eloszlásfüggvényét $F(x)$. Ekkor a következő eloszlásbeli konvergenciák teljesülnek:*

$$\lim_{t\to\infty} P(\gamma_t < z) = \lim_{t\to\infty} P(\delta_t < z) = \frac{1}{\mu}\int_0^z (1-F(x))dx = H(z), \quad z \geq 0.$$

A jobboldalon álló $H(z)$ eloszlásfüggvény abszolút folytonos, a hozzá tartozó sűrűségfüggvény

$$h(z) = H'(z) = \frac{1-F(z)}{\mu}.$$

A hátralévő, illetve a pillanatnyi élettartam aszimptotikus várható értéke (véges szórás esetén)

$$\int_0^\infty (1-H(z))dz = \frac{\mu^2+\sigma^2}{2\mu}.$$

Továbbá

$$\lim_{t\to\infty} P(\beta_t < z) = \frac{1}{\mu}\int_0^z x dF(x) = T(z), \quad z \geq 0,$$

ha F abszolút folytonos, akkor T is, és

$$t(z) = T'(z) = \frac{zf(z)}{\mu}.$$

A teljes élettartam aszimptotikus várható értéke nyilván $\frac{\mu^2+\sigma^2}{\mu}$.

Bizonyítás. ■