A sztochasztika alapjai informatikusoknak

Kevei Péter

2024. április 15.

Tartalomjegyzék

1. Valószínűségi mező

1.1. Alapfogalmak

Véletlen (valószínűségi) kísérlet: lényegében azonos körülmények között tetszőlegesen sokszor megismételhető megfigyelés, melynek többféle kimenetele lehet, és a figyelembe vett körülmények nem határozzák meg egyértelműen a kimenetelt.

Ilyenre már sok példát láttunk: feldobunk egy érmét; dobunk egy kockával; 90 szám közül kihúzunk 5-öt; 26 tétel közül választunk 2-t; 8 csapatot véletlenszerűen párokba rendezünk, ...

A véletlen kísérlet lehetséges kimeneteleinek halmaza az **eseménytér**, jele Ω.

Az **esemény** olyan a kísérlettel kapcsolatban tett állítás, melynek igaz vagy hamis volta eldönthető a kísérlet lefolytatása után. Az **események halmaza** az Ω részhalmazainak egy olyan rendszere, mely σ-algebra.

Ha Ω véges halmaz (ami legtöbb példában teljesül), akkor az események halmaza Ω *hatványhalmaza*, azaz összes részhalmazának halmaza. A hatványhalmaz jele 2^Ω. Emlékeztetünk, hogy egy n elemű halmaznak pontosan 2^n db részhalmaza van. Valóban, minden elemre eldönthetem, hogy beleteszem-e a részhalmazba vagy sem. Így minden elemre 2 lehetőségem van, ez $2 \cdot 2 \cdot \ldots 2 = 2^n$ különböző részhalmaz. Tehát 2^Ω hatványhalmaz elemszáma $2^{|\Omega|}$.

Egy $\mathcal{A} \subset 2^\Omega$ halmazrendszert akkor nevezünk **σ-algebrának**, ha

- $\emptyset \in \mathcal{A}$;
- valahányszor $A \in \mathcal{A}$, mindannyiszor $A^c = \Omega \backslash A \in \mathcal{A}$ (azaz a halmazrendszer zárt a komplementerképzésre);
- valahányszor $A_1, A_2, \ldots \in \mathcal{A}$, mindannyiszor $\cup_{i=1}^\infty A_i \in \mathcal{A}$ (azaz a halmazrendszer zárt a megszámlálható unióképzésre).

1.1. Definíció. Egy $\mathbf{P} : \mathcal{A} \to [0,1]$ halmazfüggvény *valószínűségi mérték*, ha

- $\mathbf{P}(\Omega) = 1$, azaz a biztos esemény valószínűsége 1 (hát persze);
- ha az $A_1, A_2, \ldots \in \mathcal{A}$ halmazok (páronként) diszjunktak, akkor

$$\mathbf{P}(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mathbf{P}(A_i),$$

  azaz a halmazfüggvény σ-additív.

A fenti tulajdonságokkal rendelkező $(\Omega, \mathcal{A}, \mathbf{P})$ hármast **valószínűségi mezőnek** nevezzük.

Események jelölése: $A, B, C, A_1, A_2, \ldots$.

- $|A| = 1 \Leftrightarrow A = \{\omega\}$, $\omega \in \Omega$, elemi esemény;
- $\emptyset$ a lehetetlen esemény;
- Ω a biztos esemény;
- $A^c = \Omega - A$ az ellentett esemény;
- $A \cap B$ mindkét esemény bekövetkezik, azaz A *és* B;
- $A \cup B$ a két esemény közül legalább az egyik bekövetkezik, azaz A *vagy* B;
- $A \cap B = \emptyset$ a két esemény kizárja egymást;
- $A - B$ az A bekövetkezik de B nem;
- $A \subset B$ az A esemény maga után vonja B-t.

1.2. *Példa.* Két pénzérmét feldobunk. Ekkor az eseménytér

$$\Omega = \{(F,F),(F,I),(I,F),(I,I)\},$$

tehát felírom az első érmével mit dobtam, majd felírom, hogy a másodikkal mit dobtam.

Vegyük észre, hogy megkülönböztetem az érméket, hiszen két fejet csak egyféleképpen dobhatunk, míg egy fejet és egy írást kétféleképpen. Mindig különböztessük meg az érméket, kockákat!

Ekkor $|\Omega| = 2^2 = 4$ lehetséges kimenetel van. Az események halmaza

$$\begin{aligned} 2^\Omega = \Big\{ & \emptyset, \{(F,F)\}, \{(F,I)\}, \{(I,F)\}, \{(I,I)\}, \\ & \{(F,F),(F,I)\}, \{(F,F),(I,F)\}, \{(F,F),(I,I)\}, \\ & \{(F,I),(I,F)\}, \{(F,I),(I,I)\}, \{(I,F),(I,I)\}, \\ & \{(F,F),(F,I),(I,F)\}, \{(F,F),(F,I),(I,I)\}, \\ & \{(F,F),(I,F),(I,I)\}, \{(I,F),(F,I),(I,I)\}, \\ & \{(F,F),(F,I),(I,F),(I,I)\} \Big\}. \end{aligned}$$

Ez a hatványhalmaz, ilyet többet nem írunk ki. Ekkor $|2^\Omega| = 2^4 = 16$ az összes esemény száma.

Legyen $A_i = \{\text{az } i\text{-edik dobás fej}\}$, $i = 1, 2$. Ekkor

$$A_1 = \{(F,F),(F,I)\}, \quad A_2 = \{(F,F),(I,F)\}.$$

A pontosan egy fejet dobunk esemény

$$B = \{\text{pontosan 1 fej}\} = \{(F,I),(I,F)\} = (A_1 \cup A_2) \backslash (A_1 \cap A_2),$$

azaz az első fej *vagy* a második fej, *de* nem mindkettő fej. Az, hogy nem dobunk fejet

$$C = \{\text{egyik sem fej}\} = \{(I, I)\} = A_1^c \cap A_2^c,$$

azaz az első nem fej *és* a második nem fej.

1.3. Állítás (A valószínűség tulajdonságai)**.** *Legyen* $(\Omega, \mathcal{A}, \mathbf{P})$ *egy valószínűségi mező,* $A, B, A_1, A_2, \ldots \in \mathcal{A}$ *események.*

(i) Ha $A_i \cap A_j = \emptyset$*, minden* $i \neq j$ *párra, akkor*

$$\mathbf{P}(A_1 \cup \ldots \cup A_n) = \mathbf{P}(A_1) + \ldots + \mathbf{P}(A_n).$$

(ii) $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$.

(iii) $A \subset B \Rightarrow \mathbf{P}(B \backslash A) = \mathbf{P}(B) - \mathbf{P}(A)$*, és* $\mathbf{P}(A) \leq \mathbf{P}(B)$.

(iv) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(v) Szitaformula:

$$\mathbf{P}(A_1 \cup \ldots \cup A_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_k}).$$

(vi) $\mathbf{P}(A \cup B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

(vii) $\mathbf{P}(A_1 \cup \ldots \cup A_n) \leq \mathbf{P}(A_1) + \ldots + \mathbf{P}(A_n)$.

(viii) Ha $A_n \in \mathcal{A}$ *monoton növő halmazsorozat (azaz* $A_n \subset A_{n+1}$*), akkor* $\mathbf{P}(\cup_n A_n) = \lim_{n \to \infty} \mathbf{P}(A_n)$.

(ix) Ha $B_n \in \mathcal{A}$ *monoton csökkenő halmazsorozat (azaz* $B_n \supset B_{n+1}$*), akkor* $\mathbf{P}(\cap_n B_n) = \lim_{n \to \infty} \mathbf{P}(B_n)$.

Bizonyítás. (i) Legyen $A_{n+1} = A_{n+2} = \ldots = \emptyset$.

(ii) $1 = \mathbf{P}(\Omega) = \mathbf{P}(A \cup A^c) = \mathbf{P}(A) + \mathbf{P}(A^c)$.

(iii) $B = A \cup (B \backslash A)$,

$$\mathbf{P}(B) = \mathbf{P}(A) + \mathbf{P}(B \backslash A) \Rightarrow \mathbf{P}(B \backslash A) = \mathbf{P}(B) - \mathbf{P}(A) \geq 0.$$

(iv) $\mathbf{P}(A \cup B) = \mathbf{P}(A \cup (B - (A \cap B))) = \mathbf{P}(A) + \mathbf{P}(B - (A \cap B)) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$.

(v) Teljes indukcióval. Csak az $n = 3$ esetben bizonyítunk, az általános eset ugyanígy megy, csak macerásabb a jelölés. Az előző pont állítását

felhasználva, előbb az A_1, $A_2 \cup A_3$ eseményekre, majd az A_2, A_3 és $A_1 \cap A_2$, $A_1 \cap A_3$ eseményekre

$$\begin{aligned}
&\mathbf{P}(A_1 \cup A_2 \cup A_3) = \mathbf{P}(A_1 \cup (A_2 \cup A_3)) \\
&= \mathbf{P}(A_1) + \mathbf{P}(A_2 \cup A_3) - \mathbf{P}(A_1 \cap (A_2 \cup A_3)) \\
&= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \mathbf{P}(A_2 \cap A_3) - \mathbf{P}((A_1 \cap A_2) \cup (A_1 \cap A_3)) \\
&= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) - \mathbf{P}(A_2 \cap A_3) \\
&\qquad - (\mathbf{P}(A_1 \cap A_2) + \mathbf{P}(A_1 \cap A_3) - \mathbf{P}(A_1 \cap A_2 \cap A_1 \cap A_3)) \\
&= \mathbf{P}(A_1) + \mathbf{P}(A_2) + \mathbf{P}(A_3) \\
&\quad - \mathbf{P}(A_2 \cap A_3) - \mathbf{P}(A_1 \cap A_2) - \mathbf{P}(A_1 \cap A_3) + \mathbf{P}(A_1 \cap A_2 \cap A_3)
\end{aligned}$$

(vi) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B) \leq \mathbf{P}(A) + \mathbf{P}(B)$.

(vii) Teljes indukcióval.

□

Az $(\Omega, \mathcal{A}, \mathbf{P})$ hármast **valószínűségi mezőnek** nevezzük.

1.2. Klasszikus valószínűségi mező

Az $(\Omega, 2^\Omega, \mathbf{P})$ valószínűségi mező **klasszikus**, ha minden kimenetel egyformán valószínű, azaz $\mathbf{P}(\{\omega\}) = c$ minden $\omega \in \Omega$ esetén. Ekkor persze szükségképpen $c = 1/|\Omega|$. Tetszőleges A eseményre $\mathbf{P}(A) = \frac{|A|}{|\Omega|} = \frac{\text{kedvező}}{\text{összes}}$.

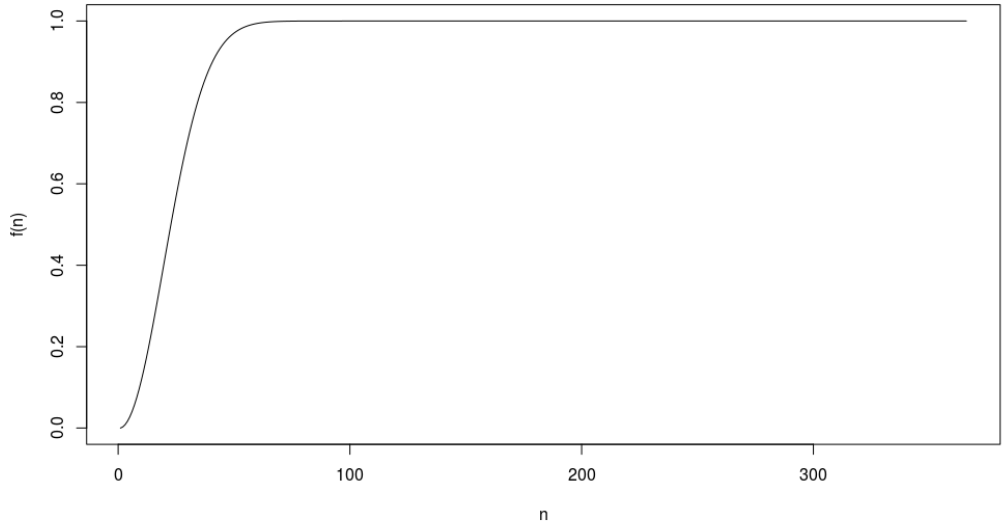
1.2.1. Születésnap probléma

Mekkora a valószínűsége annak, hogy n ember között van két olyan, akiknek ugyanazon a napon van a születésnapjuk?[1]

Jelölje $f(n)$ a keresett valószínűséget. Nyilván $f(1) = 0$, és $f(n) = 1$, ha $n \geq 366$, hiszen a skatulya-elv miatt biztosan van két ember, akik egy napon születtek.

Klasszikus valószínűségi mezőnk van, tehát eseteket számolunk. Az összes eset 365^n, hiszen minden ember 365 napon születhetett. Legyen $2 \leq n \leq 365$. Némi próbálgatás után rájöhetünk, hogy egyszerűbb a kedvezőtlen eseteket összeszámolni, azaz azokat az eseteket keressük, amikor mindenki más napon született. Valahogy (mondjuk ábécé sorrendben) sorba rakjuk az embereket.

[1]Ez egy matematika feladat, azaz hallgatólagosan feltesszük, hogy az év 365 napos (eltekintünk a február 29-ektől), és minden ember egymástól függetlenül egyforma valószínűséggel született az év bármely napján. Ezek viszonylag természetes feltevések, de azért nem mindig teljesülnek. Ha a társaságban vannak ikrek, akkor persze már nem teljesülnek a feltételek.



1. ábra. $f(n)$ valószínűségek n függvényében

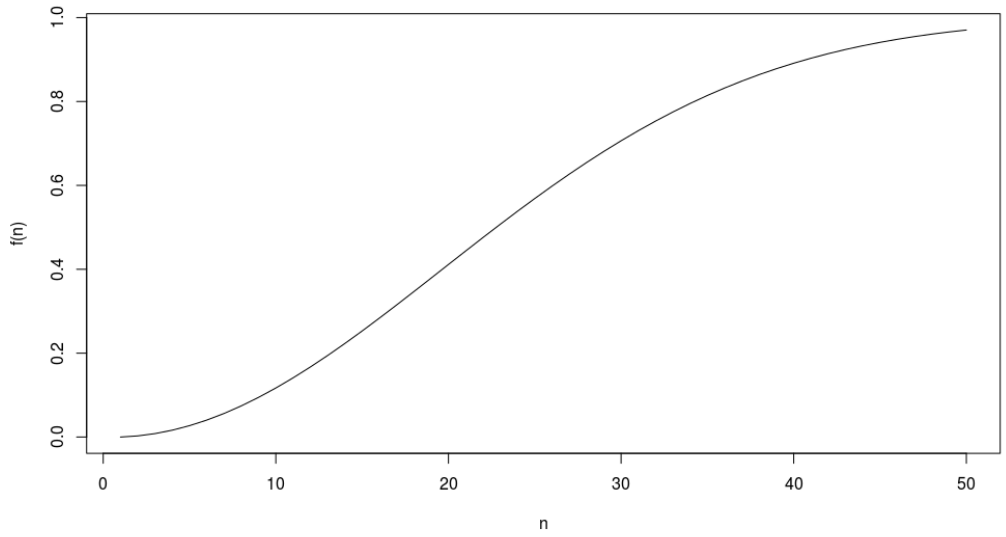
Ekkor az első ember 365 napon születhetett, a második nem születhetett azon a napon, amelyiken az első, ezért neki 364 lehetőség maradt. Hasonlóan, a harmadiknak már csak 363, stb. Végül a kedvezőtlen esetek számára $365 \cdot 364 \cdot \ldots (365 - n + 1)$ adódik. Így

$$\begin{aligned} f(n) &= \mathbf{P}\big(n \text{ ember között van 2, akiknek ugyanazon} \\ &\qquad \text{a napon van a születésnapja}\big) \\ &= 1 - \mathbf{P}(\text{ mindenkinek különböző napon van a születésnapja}) \\ &= 1 - \frac{365 \cdot 364 \cdot \ldots \cdot (365 - n + 1)}{365^n}. \end{aligned}$$

Világos, hogy $f(n)$ monoton nő, hiszen minél több ember van, annál nagyobb a közös születésnap esélye. Némileg meglepő, hogy $n = 23$ esetén a keresett valószínűség már $0,5$-nél nagyobb. Pontosabban

$$f(22) = 0,4757 < \frac{1}{2} < 0,5073 = f(23).$$

Az 1. ábrán azt is látjuk, hogy ez a valószínűség nagyon gyorsan tart 1-hez, $f(50) = 0,97$, azaz 50 ember között már nagyon valószínű hogy van két azonos születésnap. A 2. ábrán ugyanezeket a valószínűségeket látjuk, csak az érdekes tartományra koncentrálva.



2. ábra. $f(n)$ valószínűségek n függvényében, $n \leq 50$

Hogy jobban meglepődjünk, gondoljuk meg, hogy mekkora legyen n értéke ahhoz, hogy annak a valószínűsége, hogy valaki május 4-én született 0,5-nél nagyobb legyen? Megint a komplementer esemény valószínűségét számolva, n ember esetén annak a valószínűsége, hogy egyikük sem május 4-én született

$$\left(\frac{364}{365}\right)^n.$$

Ez pontosan akkor lesz 0,5-nél kisebb, ha

$$n \geq \frac{\log 2}{\log \frac{365}{364}} \approx 253.$$

Az előzőeket általánosítva, tegyük fel, hogy egy kísérletnek N lehetséges kimenetele van, és minden kimenetel egyformán valószínű. Legalább hányszor kell elvégezni a kísérletet ahhoz, hogy $0,5$-nél nagyobb legyen annak a valószínűsége, hogy kétszer van ismétlődés?

A születésnap probléma pontosan ez a feladat $N = 365$ választással. Ha pedig $N = 6$, akkor arra kapunk választ, hogy hányszor kell dobni egy dobókockával, hogy $0,5$-nél nagyobb eséllyel legyen ismétlődés. Ha n-szer ismétlünk, $2 \leq n \leq N$, akkor annak az $f(n)$ valószínűsége, hogy van ismétlés

$$f(n) = 1 - \frac{N(N-1)\dots(N-n+1)}{N^n}.$$

Ezt kicsit alakíthatjuk

$$1 - f(n) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\ldots\left(1 - \frac{n-1}{N}\right).$$

A kérdés, hogy a jobboldali szorzat mikor lesz először $0,5$-nél kisebb. Ha x kicsi, akkor $1 - x \approx e^{-x}$ (ez éppen a deriválás definíciója[2]), így a jobboldal nagyjából

$$\exp\left\{-\sum_{i=1}^{n-1}\frac{i}{N}\right\} = \exp\left\{-\frac{n(n-1)}{2N}\right\} \approx \exp\left\{-\frac{n^2}{2N}\right\}.$$

Ennek kell $0,5$-nek lenni, ahonnan logaritmust véve kapjuk, hogy

$$-\frac{n^2}{2N} = \ln\frac{1}{2},$$

azaz

$$n \approx \sqrt{2\ln 2}\sqrt{N} \approx 1.2\sqrt{N}.$$

Ez $N = 365$-re éppen $22,5$, tehát működik.

Pontosan ez a probléma lép fel hash függvények elemzése során. Ekkor egy inputot kódolunk véges hosszú 0-1 sorozatként. Az előzőek alapján, ha egy hash függvény kimenete m bites (ekkor 2^m a lehetséges 0-1 sorozatok száma), akkor $1.2 \cdot 2^{m/2}$ véletlenül választott input között $1/2$ valószínűséggel lesz kettő, melynek ugyanaz a hash értéke. A gyakorlatban általában olyan hash függvényt akarunk, melyre ütköző párokat nehéz találni, ezért m-et megfelelően nagyra kell választani. Bővebben, Buttyán, Vajda, *Kriptográfia és alkalmazásai*, Typotex, 2004.

1.2.2. A párosítási probléma

Veszünk n darab kártyát 1-től n-ig megszámozva. Összekeverjük, és véletlen sorrendben lerakjuk őket egy sorba. A k-adik helyen párosítás történik, ha a k-adik helyre a k sorszámú kártya kerül. (Tehát véletlen permutációk fixpontjait tekintjük.)

Arra keressük a választ, hogy mennyi a valószínűsége, hogy nem történik párosítás. Jelölje p_n ezt a valószínűséget.

Jelölje A_k azt az eseményt, hogy a k-adik helyen párosítás történik, $k = 1, 2, \ldots, n$. Ekkor az az esemény, hogy legalább egy párosítás történik éppen

[2]Legyen $g(x) = e^{-x}$. Ekkor $g'(0) = \lim_{x\to 0}\frac{g(x)-g(0)}{x} = \lim_{x\to 0}(e^{-x}-1)/x$. Ugyanakkor $g'(x) = -e^{-x}$, így $\lim_{x\to 0}(e^{-x}-1)/x = -1$, amit x-szel átszorozva kapjuk, hogy $e^{-x}-1 \approx -x$, vagy $1 - x \approx e^{-x}$.

$A_1 \cup A_2 \cup \ldots \cup A_n$. Ennek a valószínűségét a szitaformulával határozhatjuk meg. Eszerint

$$\mathbf{P}\left(A_1 \cup \ldots \cup A_n\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right).$$

Tetszőleges k és $1 \le i_1 < i_2 < \ldots < i_k \le n$ esetén az $A_{i_1} \cap \ldots \cap A_{i_k}$ esemény azt jelenti, hogy mind az A_{i_1}, mind az A_{i_2}, …, mind az A_{i_k} esemény bekövetkezik ($\cap$ az *és*), azaz az $i_1, \ldots, i_k$ elemek mind a helyükön maradnak, mind fixpontok. Ennek a valószínűsége,

$$\mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \frac{(n-k)!}{n!},$$

hiszen az összes eset továbbra is $n!$, míg a kedvező eseteknél az $i_1, \ldots, i_k$ elemek fixek, itt nincs választásom, és a maradék $n-k$ elemet pedig tetszőlegesen permutálhatom, erre $(n-k)!$ lehetőségem van. Vegyük észre, hogy a fenti valószínűség nem függ az $i_1, \ldots, i_k$ indexek értékétől, csak k-tól. Rögzített k-ra az indexeket $\binom{n}{k}$-féleképpen választhatom. Ezeket visszaírva

$$\begin{aligned}
\mathbf{P}\left(A_1 \cup \ldots \cup A_n\right) &= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right) \\
&= \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} \\
&= \sum_{k=1}^{n} \frac{(-1)^{k+1}}{k!}.
\end{aligned}$$

Ezek szerint

$$\begin{aligned}
p_n &= \mathbf{P}(\text{ nincs párosítás }) = \mathbf{P}(A_1^c \cap \ldots \cap A_n^c) \\
&= 1 - \mathbf{P}(A_1 \cup \ldots \cup A_n) = 1 - \left(- \sum_{k=1}^{n} \frac{(-1)^k}{k!} \right) \\
&= \sum_{k=0}^{n} \frac{(-1)^k}{k!}.
\end{aligned}$$

Tetszőleges x valós számra

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots,$$

ahonnan látjuk, hogy

$$\lim_{n\to\infty} p_n = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} = e^{-1} \approx 0,368.$$

Ezek után határozzuk meg azt a valószínűséget, hogy pontosan k darab párosítás történik. Vezessük be a

$$p_{n,k} = \mathbf{P}(n \text{ kártya van, és pontosan } k \text{ párosítás történik}),\ k = 0, 1, \ldots, n.$$

Nyilván $p_n = p_{n,0}$. Jelölje $N_{n,k}$ azon kimenetelek számát, amikor pontosan k párosítás történik n kártyával. Ezekkel a jelölésekkel $p_m = N_{m,0}/m!$ minden m természetes szám esetén. Könnyen meggondolható, hogy

$$\begin{aligned} p_{n,k} &= \frac{N_{n,k}}{n!} = \frac{\binom{n}{k} N_{n-k,0}}{n!} = \frac{\binom{n}{k}(n-k)! p_{n-k}}{n!} \\ &= \frac{p_{n-k}}{k!} = \frac{1}{k!} \sum_{j=0}^{n-k} \frac{(-1)^j}{j!}. \end{aligned}$$

Az utóbbi alakból rögtön látjuk, hogy

$$\lim_{n\to\infty} p_{n,k} = \frac{1}{k!} \sum_{j=0}^{\infty} \frac{(-1)^k}{k!} = \frac{e^{-1}}{k!}.$$

1.3. Geometriai valószínűségi mező

Akkor beszélünk geometriai valószínűségi mezőről, ha a kísérlettel kapcsolatos események egy geometriai alakzat részhalmazainak feleltethetők meg. Ekkor a lehetséges kimenetelek halmaza $\Omega = H \subset \mathbb{R}^n$, aminek a mértéke (hossza, területe, térfogata) pozitív és véges. Ekkor egy $A \subset H$ esemény valószínűsége arányos a halmaz mértékével, azaz

$$\mathbf{P}(A) = \frac{\lambda(A)}{\lambda(H)},$$

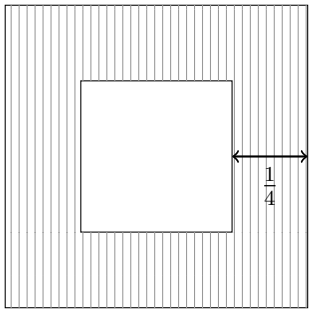
ahol λ az n-dimenziós térfogat (hossz, terület, térfogat). Ez persze nem klasszikus valószínűségi mező, hiszen a lehetséges kimenetelek halmaza végtelen (sőt, nagyon végtelen, kontinuum számosságú). Ugyanakkor a valószínűség definíciója hasonlít a klasszikus mező esetéhez, csak most a (kedvező esetek száma) / (összes esetek száma) formula helyett a

$$\mathbf{P}(A) = \frac{\text{kedvező terület}}{\text{összes terület}}$$

formula szerepel. A terület helyett lehet hossz, térfogat.

1.4. *Példa.* Egy négyzet belsejében egyenletes eloszlás szerint választunk egy pontot. Mennyi a valószínűsége, hogy a választott pont közelebb van valamelyik oldalhoz, mint 1/4?

A kísérlet egy geometriai valószínűségi mezőn írható le, ahol az eseménytér az egységnégyzet, az események az egységnégyzet Borel-halmazai (szép halmazai), és valószínűség pedig a terület, azaz $\Omega = [0,1]^2$, $\mathcal{A} = \mathcal{B}([0,1]^2)$, $\mathbf{P}(A) = \mathrm{ter}(A)$. A kedvező síkrész:



A kedvező terület, ami éppen a keresett valószínűség

$$\mathbf{P}(\text{közelebb van mint } 1/4) = 4 \cdot \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{4}.$$

2. Feltételes valószínűség és függetlenség

2.1. Feltételes valószínűség

Legyen $(\Omega, \mathcal{A}, \mathbf{P})$ egy valószínűségi mező, és ezen A, B események, és tegyük föl, hogy $\mathbf{P}(B) > 0$. Ekkor az A **esemény** B **eseményre vonatkozó feltételes valószínűsége**

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

Ha annyi információnk van a véletlen kísérletről, hogy a B esemény bekövetkezett, akkor az A esemény valószínűsége $\mathbf{P}(A|B)$.

2.1. Állítás. *Rögzítsünk egy tetszőleges B eseményt, melyre $\mathbf{P}(B) > 0$. Ekkor $\mathbf{P}_B(A) = \mathbf{P}(A|B)$ valószínűségi mérték $\mathcal{A}$-n.*

Bizonyítás. Világos, hogy tetszőleges A eseményre $\mathbf{P}_B(A) \geq 0$, és mivel $\mathbf{P}(A \cap B) \leq \mathbf{P}(B)$, így $\mathbf{P}_B(A) \leq 1$. Továbbá

$$\mathbf{P}_B(\Omega) = \frac{\mathbf{P}(\Omega \cap B)}{\mathbf{P}(B)} = \frac{\mathbf{P}(B)}{\mathbf{P}(B)} = 1.$$

Már csak az additivitás ellenőrzése maradt. Legyenek $A_1, A_2, \ldots \in \mathcal{A}$ diszjunktak. Ekkor a definíció szerint és a $\mathbf{P}$ valószínűségi mérték additivitása alapján

$$\begin{aligned}\mathbf{P}_B(\cup_{i=1}^{\infty} A_i) &= \mathbf{P}\left(\cup_{i=1}^{\infty} A_i | B\right) = \frac{\mathbf{P}\left((\cup_{i=1}^{\infty} A_i) \cap B\right)}{\mathbf{P}(B)} = \frac{\mathbf{P}\left(\cup_{i=1}^{\infty}(A_i \cap B)\right)}{\mathbf{P}(B)} \\ &= \frac{\sum_{i=1}^{\infty} \mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} = \sum_{i=1}^{\infty} \frac{\mathbf{P}(A_i \cap B)}{\mathbf{P}(B)} = \sum_{i=1}^{\infty} \mathbf{P}(A_i|B) \\ &= \sum_{i=1}^{\infty} \mathbf{P}_B(A_i),\end{aligned}$$

ami éppen a bizonyítandó egyenlőség. □

Ebből következik, hogy $\mathbf{P}_B$ halmazfüggvényre is teljesülnek a valószínűségi mérték tulajdonságai, melyeket a későbbiekben említés nélkül fölhasználunk.

2.2. Tétel (Szorzási szabály)**.** *Legyenek $A_1, A_2, \ldots, A_n$ tetszőleges olyan események, melyekre $\mathbf{P}(A_1 \cap \ldots \cap A_{n-1}) > 0$. Ekkor*

$$\mathbf{P}(A_1 \cap \ldots \cap A_n) = \mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1 \cap A_2)\ldots\mathbf{P}(A_n|A_1 \cap \ldots \cap A_{n-1}).$$

A bizonyítás előtt megjegyezzük a következőket:

1. A formulában szereplő összes feltétel valószínűsége pozitív, azaz minden jóldefiniált. Ez a $\mathbf{P}(A_1 \cap \ldots \cap A_{n-1}) > 0$ feltétel következménye.
2. Ha $\mathbf{P}(A_1 \cap \ldots \cap A_n) > 0$ is teljesül, akkor $n!$ darab különböző ilyen szabály van.
3. A szabályt az $n = 2$ esetben használjuk legtöbbször. Ekkor

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B|A) = \mathbf{P}(B)\mathbf{P}(A|B),$$

amennyiben A és B is pozitív valószínűségű esemény.

Bizonyítás. A feltételes valószínűség definíciója szerint

$$\begin{aligned}&\mathbf{P}(A_1)\mathbf{P}(A_2|A_1)\mathbf{P}(A_3|A_1 \cap A_2)\ldots\mathbf{P}(A_n|A_1 \cap \ldots \cap A_{n-1}) \\ &= \mathbf{P}(A_1)\frac{\mathbf{P}(A_1 \cap A_2)}{\mathbf{P}(A_1)}\frac{\mathbf{P}(A_1 \cap A_2 \cap A_3)}{\mathbf{P}(A_1 \cap A_2)}\cdots\frac{\mathbf{P}(A_1 \cap \ldots \cap A_{n-1} \cap A_n)}{\mathbf{P}(A_1 \cap \ldots \cap A_{n-1})} \\ &= \mathbf{P}(A_1 \cap \ldots \cap A_n),\end{aligned}$$

amint állítottuk. □

2.3. *Példa.* Egy dobozban 12 kék és 3 fehér golyó van. Visszatevés nélkül húzunk két golyót egymás után. Mi annak a valószínűsége, hogy mindkét golyó kék?

Jelölje K_i az az eseményt, hogy az i-ediknek kihúzott golyó kék. Ekkor a szorzási szabály szerint

$$\mathbf{P}(K_1 \cap K_2) = \mathbf{P}(K_1)\mathbf{P}(K_2|K_1) = \frac{12}{15} \cdot \frac{11}{14}.$$

Ugyanezt kapjuk a klasszikus valószínűségi mezőre vonatkozó kedvező/összes formulával is, mely szerint

$$\mathbf{P}(K_1 \cap K_2) = \frac{\binom{12}{2}}{\binom{15}{2}}.$$

A $B_1, B_2, \ldots$ események **teljes eseményrendszert** alkotnak, ha

- minden $i \neq j$ párra $B_i \cap B_j = \emptyset$;
- $\cup_{n=1}^{\infty} B_n = \Omega$.

2.4. Tétel (Teljes valószínűség tétele)**.** *Legyen $B_1, B_2, \ldots$ teljes eseményrendszer, melyre $\mathbf{P}(B_n) > 0$ minden n-re. Ekkor tetszőleges $A \in \mathcal{A}$ esemény esetén*

$$\mathbf{P}(A) = \sum_{n=1}^{\infty} \mathbf{P}(A|B_n)\mathbf{P}(B_n).$$

Bizonyítás. A feltételes valószínűség definíciója és a valószínűség additivitása alapján

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbf{P}(A|B_n)\mathbf{P}(B_n) &= \sum_{n=1}^{\infty} \frac{\mathbf{P}(A \cap B_n)}{\mathbf{P}(B_n)} \cdot \mathbf{P}(B_n) = \sum_{n=1}^{\infty} \mathbf{P}(A \cap B_n) \\ &= \mathbf{P}\left(\cup_{n=1}^{\infty}(A \cap B_n)\right) = \mathbf{P}\left(A \cap (\cup_{n=1}^{\infty} B_n)\right) \\ &= \mathbf{P}(A \cap \Omega) = \mathbf{P}(A).\end{aligned}$$

□

2.5. *Példa* (Monty Hall probléma). Egy játékos 3 csukott ajtó közül 1-et választhat. Kettő mögött kecske van, egy mögött autó, de nem tudja, hogy melyik hol van. A játékos autót szeretne nyerni. Miután egy ajtót kiválaszt, a játékvezető kinyittat egy ajtót, amit nem választott a játékos, és ami mögött kecske van. Ezután felajánlja a játékosnak, hogy cserélhet ajtót. Megéri-e váltani?[3]

[3]Feltesszük, hogy az autó 1/3 eséllyel van bármelyik ajtó mögött, továbbá ha a játékos azt az ajtót választotta, ami mögött az autó van, akkor a játékvezető $1/2 - 1/2$ valószínűséggel választja a másik két ajtó bármelyikét.

Tegyük fel, hogy a játékos az 1-es ajtót választotta, a játékvezető a 2-est nyittatta ki. Jelölje A_i azt az eseményt, hogy az autó az i-edik ajtó mögött van, B pedig azt, hogy a játékvezető a 2-es ajtót nyittatta ki. Ekkor

$$\mathbf{P}(A_1) = \mathbf{P}(A_2) = \mathbf{P}(A_3) = \frac{1}{3}.$$

Amire kíváncsiak vagyunk az a $\mathbf{P}(A_1|B)$ valószínűség. Definíció szerint

$$\mathbf{P}(A_1|B) = \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)}.$$

Továbbá

$$\mathbf{P}(A_1 \cap B) = \mathbf{P}(A_1)\mathbf{P}(B|A_1) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6},$$

hiszen ekkor a két kecske közül választhat a játékvezető. A teljes valószínűség tétele szerint

$$\begin{aligned}\mathbf{P}(B) &= \mathbf{P}(A_1)\mathbf{P}(B|A_1) + \mathbf{P}(A_2)\mathbf{P}(B|A_2) + \mathbf{P}(A_3)\mathbf{P}(B|A_3) \\ &= \frac{1}{3} \cdot \frac{1}{2} + \frac{1}{3} \cdot 0 + \frac{1}{3} \cdot 1 = \frac{1}{2}.\end{aligned}$$

Így

$$\mathbf{P}(A_1|B) = \frac{\mathbf{P}(A_1 \cap B)}{\mathbf{P}(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

Ezért

$$\mathbf{P}(A_3|B) = 1 - \mathbf{P}(A_1|B) = \frac{2}{3},$$

vagyis megéri váltani.[4]

2.6. *Példa* (Véletlenített válaszadás). Egy olyan kérdőívet készítünk, melyen kínos kérdések is szerepelnek. Ezekre az emberek még anonim kitöltés esetén sem szívesen válaszolnak őszintén. A következőt csináljuk. A kínos kérdésnél a kitöltő feldob egy (szabályos) pénzérmét. Ha fejet kap, akkor válaszoljon igennel, ha írást, akkor válaszoljon arra a kérdésre, hogy énekel-e a zuhany alatt. Mivel a pénzfeldobás eredményét csak a kitöltő látja, így nem tudhatjuk, hogy az igen válasz mire vonatkozott. A módszert alkalmazzák, eredetileg Warner javasolta 1965-ben[5]

[4]Ha váltunk, akkor pontosan akkor nyerünk autót, ha először kecskés ajtót választottunk, aminek a valószínűsége 2/3. Míg ha nem váltunk, akkor pontosan akkor nyerünk, ha elsőnek autót választottunk, aminek 1/3 az esélye.

[5]Warner, S. L. (1965). *Randomised response: a survey technique for eliminating evasive answer bias.*

Tegyük fel, hogy az emberek (ismeretlen!) p hányada énekel a zuhany alatt. Mennyi a valószínűsége, hogy a teszt kitöltése során valaki igennel válaszol? Az igen válaszok arányából hogy következtethetünk a valódi p arányra?

Jelölje I azt az eseményt, hogy a válasz igen, míg A azt az eseményt, hogy a kitöltő énekel a zuhany alatt. A feltétel szerint $\mathbf{P}(A) = p$, és ha véletlen kitöltő énekel a zuhany alatt, akkor a válasza mindenképpen igen, azaz $\mathbf{P}(I|A) = 1$, míg ha nem énekel, akkor a válasza csak akkor igen, ha az érmével fejet dob, azaz $\mathbf{P}(I|A^c) = \frac{1}{2}$. Így a teljes valószínűség tétele szerint

$$\mathbf{P}(I) = \mathbf{P}(A) \cdot \mathbf{P}(I|A) + \mathbf{P}(A^c) \cdot \mathbf{P}(I|A^c) = p + (1-p)\frac{1}{2} = \frac{1+p}{2}.$$

Ebből visszaszámolva p-t azt kapjuk, hogy ha az igenek arányára q értéket kapunk, akkor $p = 2q - 1$.

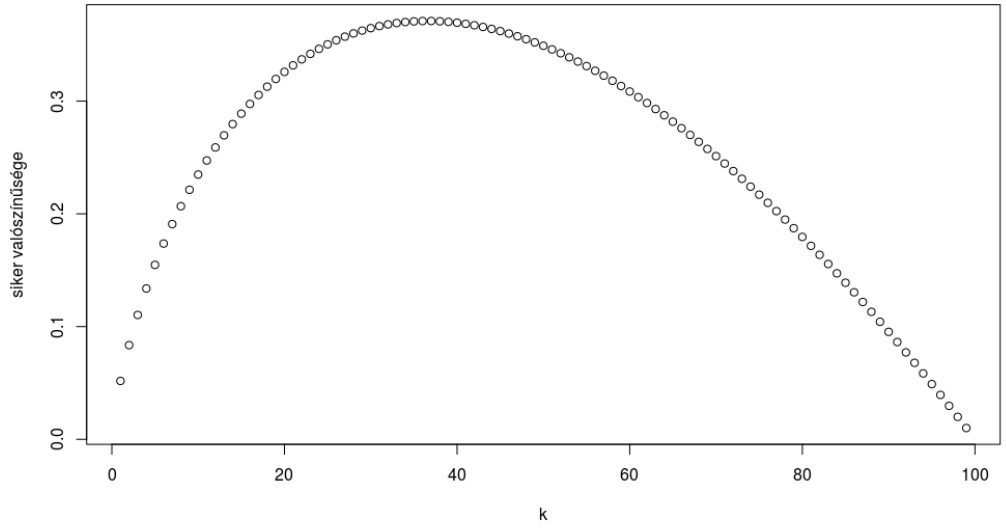
2.7. *Példa* (Szindbád). Szindbádnak jogában áll N háremhölgy közül egyet kiválasztania oly módon, hogy az előtte egyenként elvonuló hölgyek valamelyikére rámutat. Tegyük fel, hogy egyértelmű szigorúan monoton szépségi sorrendet tud felállítani, és a háremhölgyek bármely elvonulási sorrendje egyformán valószínű. Szindbád k hölgyet elenged, majd kiválasztja az elsőt, aki szebb az összes előtte elvonultnál. Mennyi a valószínűsége, hogy a legszebb hölgyet választja ki? Milyen k esetén lesz ez a valószínűség a legnagyobb, ha N elég nagy?

Jelölje A_i azt az eseményt, hogy a i-edik lány a legszebb, $i = 1, 2, \ldots, N$, és legyen B az az esemény, hogy Szindbád a legszebb lányt választja. Ekkor a teljes valószínűség tétele szerint

$$\mathbf{P}(B) = \sum_{i=1}^{N} \mathbf{P}(A_i)\, \mathbf{P}(B|A_i).$$

Világos, hogy $\mathbf{P}(A_i) = N^{-1}$ minden i-re, hiszen a legszebb lány egyforma valószínűséggel lehet az N hely bármelyikén (ezt feltettük). Ha $i \leq k$, akkor a választási szabály szerint Szindbád biztosan nem választotta a lányt, ezért $\mathbf{P}(B|A_i) = 0$ ha $i \leq k$.

Most kell egy fontos észrevétel. Ha $i > k$, akkor Szindbád pontosan akkor választja ki a legszebb háremhölgyet, ha az első $i-1$ lány közül a legszebb az első k-ban volt. Hát persze, hiszen ekkor az első $(i-1)$ közötti legszebbet a szabály szerint elengedte, és utána már csak az i-ediknek érkező szebb nála, aki történetesen a legszebb. Annak a valószínűsége, hogy az első $(i-1)$ közt a legszebb az első k közt van, éppen $k/(i-1)$, hiszen megint egyforma



3. ábra. A siker valószínűsége k függvényében, $N = 100$

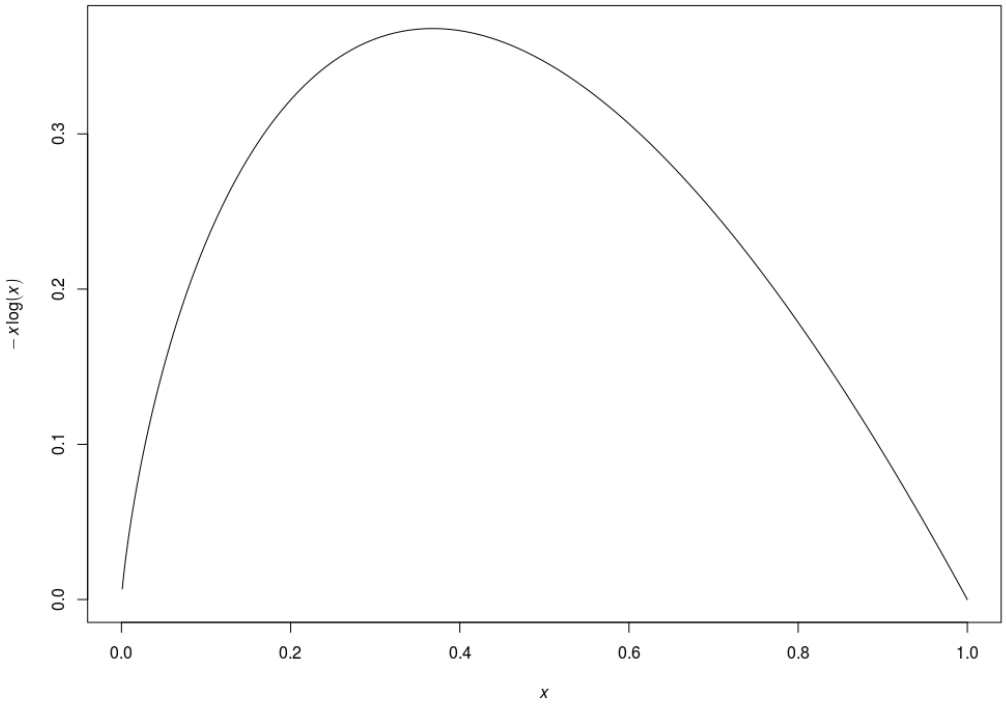
valószínűséggel lehet bármelyik a legszebb. Tehát $\mathbf{P}(B|A_i) = k/(i-1)$. Összegezve

$$\mathbf{P}(B) = \sum_{i=1}^{N} \mathbf{P}(A_i)\,\mathbf{P}(B|A_i) = \sum_{i=k+1}^{N} \frac{1}{N}\frac{k}{i-1}.$$

Ezzel a valószínűségszámítás részével megvagyunk a feladatnak. A 3. ábrán a legszebb lány kiválasztásának valószínűségét láthatjuk k függvényében, ha $N = 100$. A legjobb választás k-ra, azaz ami maximalizálja a legszebb lány választásának valószínűségét, $k = 37$, és ekkor $\mathbf{P}(B) = 0{,}37$.

Ha N nagy, akkor az optimális k értéket meghatározhatjuk közelítőleg. Ez nem is olyan nehéz. A harmonikus sor részletösszegének az aszimptotikája kell nevezetesen, hogy $1 + 2^{-1} + 3^{-1} + \ldots + n^{-1} \sim \ln n$. Ez kijön, ha a sorösszeget az $\int_1^n x^{-1}\mathrm{d}x = \ln x$ integrállal közelítjük. Ezek szerint $\sum_{i=k+1}^{N} \frac{1}{i-1} \sim \ln N - \ln k = \ln(N/k)$. Ezt visszahelyettesítve $\mathbf{P}(B) = \frac{k}{N}\ln\frac{N}{k}$. Ez egész barátságos, az $x = k/N \in (0,1)$ jelöléssel a $h(x) = x\ln x^{-1}$ függvény maximumhelyét kell megtalálnunk a $(0,1)$-en, lásd 4 ábra (vegyük észre, hogy ez pont olyan mint a 3 ábra). Ezt deriválva, $h'(x) = 0$ egyenletet megoldva látjuk, hogy $x = e^{-1}$ a maximumhely, ami éppen azt jelenti, hogy $k \approx N/e$.

Tehát ha N nagy akkor Szindbád optimális választása $k \approx N/e$. Ekkor a siker, azaz a legszebb lány kiválasztásának valószínűsége $\approx e^{-1} = 0,368$.



4. ábra. A $-x \ln x$ függvény a $[0, 1]$ intervallumon

Azaz, ha Szindbád ügyesen választ stratégiát, akkor 100 háremhölgy közül a legszebbet 37% eséllyel ki tudja választani. Azt kell tennie, hogy $100/e \approx 37$ lányt elenged, majd kiválasztja az elsőt, aki szebb az összes eddigi lánynál.

2.8. Tétel (Bayes-formula)**.** *Legyenek A és B olyan események, hogy $\mathbf{P}(A) > 0$, $\mathbf{P}(B) > 0$. Ekkor*

$$\mathbf{P}(B|A) = \frac{\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(A)}.$$

Bizonyítás. A definíció szerint

$$\frac{\mathbf{P}(A|B)\mathbf{P}(B)}{\mathbf{P}(A)} = \frac{\mathbf{P}(A \cap B)\mathbf{P}(B)}{\mathbf{P}(B)\mathbf{P}(A)} = \mathbf{P}(B|A).$$

□

2.9. Tétel (Bayes-tétel)**.** *Legyen $B_1, B_2, \ldots$ teljes eseményrendszer, melyre $\mathbf{P}(B_n) > 0$ minden n-re. Ekkor tetszőleges pozitív valószínűségű $A \in \mathcal{A}$ esemény esetén, tetszőleges k-ra*

$$\mathbf{P}(B_k|A) = \frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\sum_{n=1}^{\infty} \mathbf{P}(A|B_n)\mathbf{P}(B_n)}.$$

Bizonyítás. Előbb a teljes valószínűség tételét, majd a Bayes-formulát használva

$$\frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\sum_{n=1}^{\infty} \mathbf{P}(A|B_n)\mathbf{P}(B_n)} = \frac{\mathbf{P}(A|B_k)\mathbf{P}(B_k)}{\mathbf{P}(A)} = \mathbf{P}(B_k|A).$$

□

2.10. *Példa* (Doppingteszt). Kifejlesztenek egy új doppingtesztet, mely a dopingolók 99%-ánál pozitív eredményt ad, azonban a nem dopingoló sportolók 1%-nál is tévesen pozitív eredményt ad. Tegyük föl, hogy a sportolók 1%-a dopingol. Mennyi annak a valószínűsége, hogy egy véletlenül kiválasztott sportoló

(a) doppingtesztje pozitív?

(b) dopingolt, ha tudjuk, hogy a doppingtesztje pozitív?

Jelölje T azt az eseményt, hogy a teszt eredménye pozitív, és D azt az eseményt, hogy a sportoló dopingolt. Ekkor a feladat (a) része a $\mathbf{P}(T)$, a (b) része a $\mathbf{P}(D|T)$ valószínűséget kérdezi. A teljes valószínűség tételét alkalmazva a D, D^c eseményrendszerre kapjuk

$$\begin{aligned}\mathbf{P}(T) &= \mathbf{P}(D)\mathbf{P}(T|D) + \mathbf{P}(D^c)\mathbf{P}(T|D^c) \\ &= 0,01 \cdot 0,99 + 0,99 \cdot 0,01 = 0,0198.\end{aligned}$$

A Bayes-formula szerint

$$\mathbf{P}(D|T) = \frac{\mathbf{P}(T|D)\mathbf{P}(D)}{\mathbf{P}(T)} = \frac{0,99 \cdot 0,01}{0,0198} = \frac{1}{2}.$$

A feladat eredménye meglepő, hiszen egy látszólag jól működő teszt esetén, annak a valószínűsége, hogy egy sportoló tényleg doppingolt, feltéve, hogy a teszt eredménye pozitív, $1/2$. Világos, hogy ilyen tesztelés mellett nem vehetjük el senkitől az olimpiai aranyérmét. A hiba onnan jön, hogy ha 100 sportolóból 1 doppingol, akkor a teszt ezt az 1-et nagy valószínűséggel kimutatja, viszont a 99 becsületes sportoló közül is kb. egyet tévesen a doppingolók közé sorol. Így kb. két pozitív teszteredmény lesz, de a két sportoló közül csak az egyik doppingol.

2.2. Függetlenség

Két esemény függetlensége intuitívan azt jelenti, hogy bekövetkezéseik nem befolyásolják egymást. Másképpen, a B esemény bekövetkezése nem befolyásolja az A bekövetkezését, azaz $\mathbf{P}(A|B) = \mathbf{P}(A)$, ahonnan $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$.

2.11. Definíció. Az A és B események **függetlenek**, ha

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

A definícióból világos, hogy a függetlenség szimmetrikus. Továbbá, a biztos ill. a lehetetlen eseménytől minden esemény független.

2.12. *Példa.* Francia kártyapakliból véletlenszerűen húzunk egy lapot. Jelölje D azt az eseményt, hogy dámát húzunk, K pedig azt, hogy kőrt. Ekkor $D \cap K$ az az esemény, hogy a kőr dámát húztuk ki, így $\mathbf{P}(D \cap K) = 1/52$. Ugyanakkor $\mathbf{P}(D) = 4/52 = 1/13$ és $\mathbf{P}(K) = 13/52 = 1/4$, azaz a két esemény független.

2.13. Definíció. Az A, B, C események **függetlenek**, ha

$$\begin{aligned}
&\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B),\\
&\mathbf{P}(A \cap C) = \mathbf{P}(A)\mathbf{P}(C),\\
&\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C),\\
&\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C).
\end{aligned}$$

Továbbá, az A, B, C események **páronként függetlenek**, ha bármely kettő független, azaz a fenti első három egyenlőség fennáll.

A páronkénti függetlenségből nem következik a függetlenség.

2.14. Definíció. Az $A_1, A_2, \ldots, A_n$ események **függetlenek**, ha bármely $k \in \{2, 3, \ldots, n\}$ és $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ esetén

$$\mathbf{P}(A_{i_1} \cap \ldots \cap A_{i_k}) = \mathbf{P}(A_{i_1}) \ldots \mathbf{P}(A_{i_k}).$$

Végtelen sok esemény akkor független, ha közülük bármely véges sok független.

2.15. Állítás. *Ha az $A_1, \ldots, A_n$ események függetlenek, akkor tetszőleges $k \in \{1, 2, \ldots, n\}$ esetén az $\{A_1, \ldots, A_k\}$ eseményekből ill. az $\{A_{k+1}, \ldots, A_n\}$ eseményekből alkotott események függetlenek.*

Ezt nem bizonyítjuk. Az állítás szerint például ha A, B, C, D független események, akkor $A \cup B$ és $C \cap D$ is függetlenek.

2.16. *Példa* (Üzenetküldés zajos csatornán). Bináris, azaz $0 - 1$ jelsorozatot akarunk küldeni zajos csatornán keresztül. Tegyük fel, hogy a csatorna minden bitet egymástól függetlenül p valószínűséggel elront, azaz a 0-ból 1-est, az 1-ből 0-át csinál. Ő egy *bináris szimmetrikus csatorna.*

Ekkor annak a valószínűsége, hogy 5 bit hiba nélkül megérkezik, a függetlenség miatt

$$\mathbf{P}(\text{ 5 bit hiba nélkül átmegy }) = (1-p)^5.$$

Ez nem túl jó, ha $p = 0,1$, akkor ez a valószínűség $0,59$, míg $p = 0,2$ esetén csak $0,33$.

Csináljuk azt, hogy minden bitet háromszor küldünk el. Tehát a 0 helyett 000, az 1 helyett 111. Ha a vevő nem három egyforma értéket kap, akkor tudja, hogy hiba történt az átvitel során. Ekkor többségi szavazást tart a bitről, ha 001, 010, vagy 100 a bejövő jel, akkor azt 0-nak dekódolja, ha 110, 101, vagy 011, azt 1-nek. Ekkor 1 bitnek megfelelő részt pontosan akkor dekódol hibásan, ha *legalább kétszer* átfordult a bit, ennek valószínűsége

$$\begin{aligned} p' &= \mathbf{P}(\text{ hibásan dekódol }) \\ &= \mathbf{P}(\text{ pontosan 2 bit fordul át }) + \mathbf{P}(\text{ mindhárom átfordul }) \\ &= 3p^2(1-p) + p^3 = 3p^2 - 2p^3. \end{aligned}$$

Ez jóval kisebb, mint p. A háromszorozó kódolással annak a valószínűsége, hogy 5 bitnyi üzenetet helyesen dekódolunk, az előzőek szerint

$$\begin{aligned} &\mathbf{P}(\text{ 5 bitnyi üzenet háromszorozva hiba nélkül átmegy }) \\ &= (1-p')^5 = \left(1 - 3p^2 + 2p^3\right)^5. \end{aligned}$$

Így a hibavalószínűséget jelentősen csökkentettük, de azon az áron, hogy 1 bitnyi információ küldéséhez 3 bitet használtunk. Ennél lehet sokkal okosabban csinálni, lásd Shannon csatornakódolási tétele. Ez lesz Kódoláselmélet kurzuson mesterképzésben.

3. Diszkrét véletlen változók

3.1. Véletlen változók, eloszlásfüggvény, várható érték

3.1. Definíció. Tekintsünk egy $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőt. Az

$$\xi : \Omega \mapsto \mathbb{R}$$

függvényeket *véletlen változónak* nevezzük, ha a

$$\xi^{-1}((-\infty, a)) = \{\omega : \xi(\omega) < a\}$$

inverzkép $\mathcal{A}$-beli tetszőleges $a \in \mathbb{R}$ esetén.

Már sok példát láttunk véletlen változóra. Ilyen például a dobókockával dobott szám értéke, vagy ha három kockával dobunk, akkor a legkisebb dobott szám. Ilyen az ötöslottón kihúzott legnagyobb szám, vagy az egy szelvényen elért találatok száma. Véletlen változó az is, hogy a ropi hol törik el, vagy az egységnégyzetben egyenletesen választott pont milyen távol van a négyzet határától, stb.

3.2. Definíció. A ξ véletlen változó *eloszlásfüggvénye*[6] az

$$F(x) = \mathbf{P}(\xi < x) = \mathbf{P}(\{\omega : \xi(\omega) < x\}),\ x \in \mathbb{R},$$

függvény.

3.3. Tétel. *Legyen $F(x)$ egy ξ véletlen változó eloszlásfüggvénye. Ekkor*

(i) F monoton nemcsökkenő;

(ii) $\lim_{x\to\infty} F(x) = 1$ *és* $\lim_{x\to-\infty} F(x) = 0$*;*

(iii) F balról folytonos.

Bizonyítás. (i) Ha $x_1 < x_2$ akkor $\{\xi < x_1\} \subset \{\xi < x_2\}$ és így a mérték monotonitása miatt $F(x_1) = \mathbf{P}(\xi < x_1) \le \mathbf{P}(\xi < x_2) = F(x_2)$. □

[6]Szokás az eloszlásfüggvényt jobbról folytonosnak is definiálni, ekkor a definícióban $\le$ szerepel szigorúan $<$ helyett. Majd látjuk, hogy ez nem nagy különbség, a folytonos esetben a két definíció ugyanaz, míg a diszkrét esetben máshol vannak az üres meg a teli karikák az ugrásoknál.

Vegyük észre, hogy F monotonitásából következik az $F(x+)$ jobboldali határérték létezése is, azonban általában az $F(x) = F(x+)$ egyenlőség nem teljesül. Az előzőekhez hasonlóan látható, hogy $F(x+) = \mathbf{P}(\xi \leq x)$, továbbá

$$F(x+) = \mathbf{P}(\xi \leq x) = \mathbf{P}(\xi < x) + \mathbf{P}(\xi = x) = F(x) + \mathbf{P}(\xi = x).$$

Ez pedig éppen azt jelenti, hogy F pontosan akkor folytonos az x pontban, ha $\mathbf{P}(\xi = x) = 0$.

A definícióból adódik, hogy $a < b$ esetén $\mathbf{P}(a \leq \xi < b) = F(b) - F(a)$.

3.4. Definíció. Egy véletlen változó *diszkrét*, ha értékkészlete megszámlálható (azaz véges vagy megszámlálhatóan végtelen). Ha egy diszkrét véletlen változó lehetséges értékei $x_1, x_2, \ldots$, akkor $p_i = \mathbf{P}(\xi = x_i) > 0$ a változó eloszlása. A ξ diszkrét véletlen változó *várható értéke*

$$\mathbf{E}(\xi) = \sum_i x_i \mathbf{P}(\xi = x_i),$$

ha $\sum_i |x_i| \mathbf{P}(\xi = x_i) < \infty$. A ξ szórása

$$\mathbf{D}(\xi) = \sqrt{\mathbf{E}[(\xi - \mathbf{E}(\xi))^2]} = \sqrt{\mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2},$$

ahol a második momentum

$$\mathbf{E}(\xi^2) = \sum_i x_i^2 \mathbf{P}(\xi = x_i).$$

Ha (p_i) eloszlás, akkor $\sum_i p_i = 1$. Az eloszlásfüggvény $F(x) = \sum_{i:x_i<x} p_i$.

Egy kísérletet n-szer függetlenül ismétlünk és minden alkalommal megfigyeljük ξ értékét: $\xi_1, \xi_2, \ldots, \xi_n$. Ezen értékek átlagai egy számhoz tartanak, ez lesz $\mathbf{E}(\xi)$. A szórás a várható érték körüli ingadozás mérőszáma.

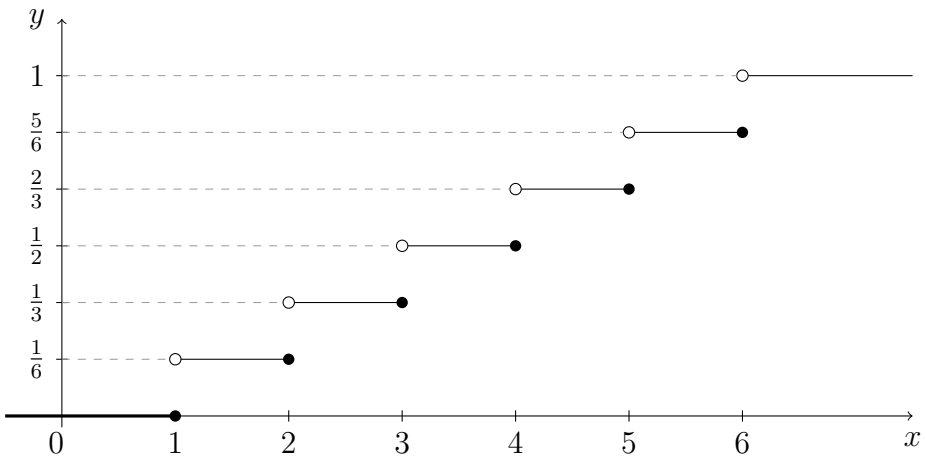
3.5. *Példa.* Dobókockával dobunk. Jelölje ξ a dobott értéket. Ekkor ξ lehetséges értékei: $1, 2, \ldots, 6$. Szabályos a kockánk, ezért minden értéket $1/6$ valószínűséggel dobunk, azaz $\mathbf{P}(\xi = k) = \frac{1}{6}$, $k = 1, 2, \ldots, 6$. Az eloszlásüggvény

$$F(x) = \mathbf{P}(\xi < x) = \begin{cases} 0, & x \leq 1, \\ \frac{\lceil x \rceil - 1}{6}, & 1 < x \leq 6, \\ 1, & x > 6. \end{cases}$$

Itt $\lceil x \rceil = \min\{n : n \geq x, n \in \mathbb{Z}\}$ jelöli x felső egészrészét.

A várható érték a definíció szerint

$$\mathbf{E}(\xi) = \sum_{k=1}^{6} k \cdot \mathbf{P}(\xi = k) = \frac{7}{2} = 3,5.$$



5. ábra. Dobott szám eloszlásfüggvénye

A szórásnégyzet $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2$, amiből a várható értéket már tudjuk. A *második momentum*

$$\mathbf{E}(\xi^2) = \sum_{k=1}^{6} k^2 \cdot \mathbf{P}(\xi^2 = k^2) = \sum_{k=1}^{6} k^2 \cdot \mathbf{P}(\xi = k) = \frac{1}{6} \sum_{k=1}^{6} k^2 = \frac{91}{6},$$

így $\mathbf{D}(\xi) = \sqrt{91 - (3,5)^2} \approx 1,7$.

3.2. Nevezetes diszkrét eloszlások

3.2.1. Binomiális eloszlás

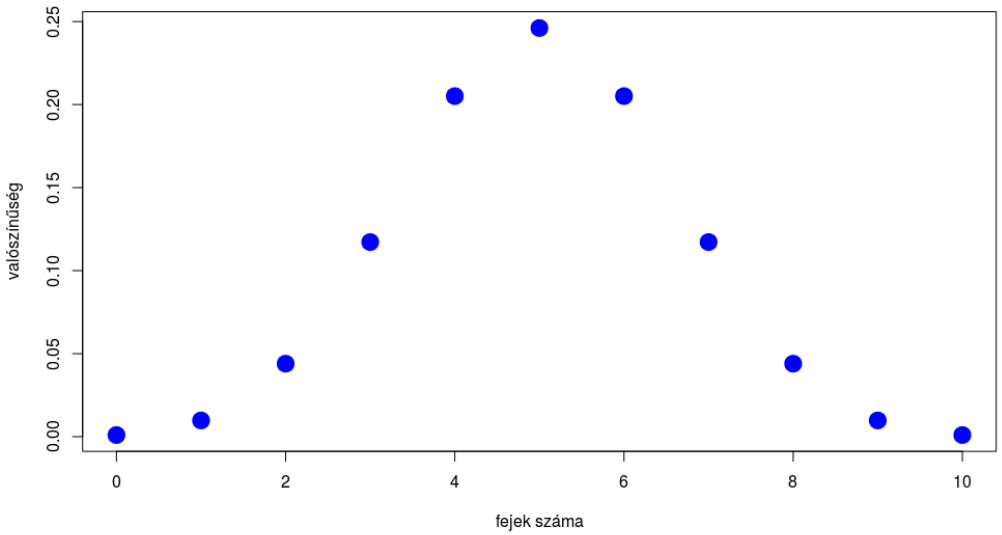
A ξ véletlen változó p *paraméterű Bernoulli-eloszlású*, $\xi \sim \text{Bernoulli}(p)$, $p \in [0,1]$, ha lehetséges értékei $0, 1$, és $\mathbf{P}(\xi = 1) = p = 1 - \mathbf{P}(\xi = 0)$. Várható értéke $\mathbf{E}(\xi) = p$, szórásnégyzete $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2 = p - p^2 = p(1-p)$.

Tipikus példa egy A esemény I_A indikátorváltozója.

A ξ véletlen változó (n, p) *paraméterű binomiális eloszlású*, $\xi \sim \text{Bin}(n, p)$, $n \in \{1, 2, \ldots\}$, $p \in [0, 1]$, ha lehetséges értékei $0, 1, \ldots, n$, és $\mathbf{P}(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$.

Ez tényleg eloszlás, hiszen a binomiális tétel szerint

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



6. ábra. Az $n = 10$, $p = 0.5$ paraméterű binomiális eloszlás valószínűségei

A 6 ábrán a $\mathbf{P}(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k}$ valószínűségeket látjuk $n = 10$ és $p = 0.5$ esetén. Észrevehetjük a haranggörbét, ami már előrevetíti a centrális határeloszlás-tételt. A 7 ábrán ebből az eloszlásból szimulálunk egy $N = 10$ és $N = 1000$ elemű mintát. Láthatjuk, hogy $N = 1000$ esetén a relatív gyakoriságok lényegében megegyeznek az elméleti valószínűségekkel, míg $N = 10$ esetén még nem látunk különösebb struktúrát.

A 8 ábrán egy nem szimmetrikus binomiális eloszlás valószínűségeit látjuk, $n = 10$ és $p = 0.7$ paraméterekkel.

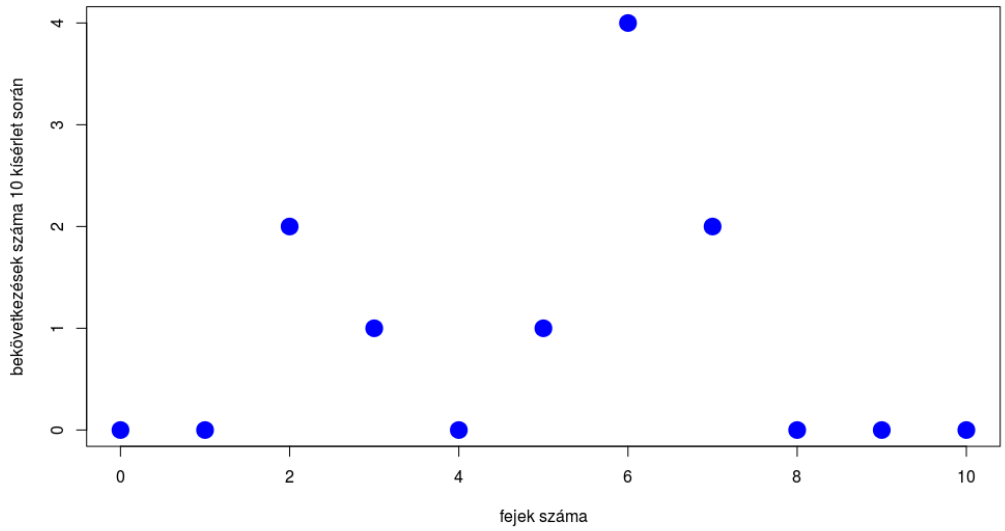
Vegyük észre, hogy a Bernoulli-eloszlás éppen az $(1, p)$ paraméterű binomiális eloszlás.

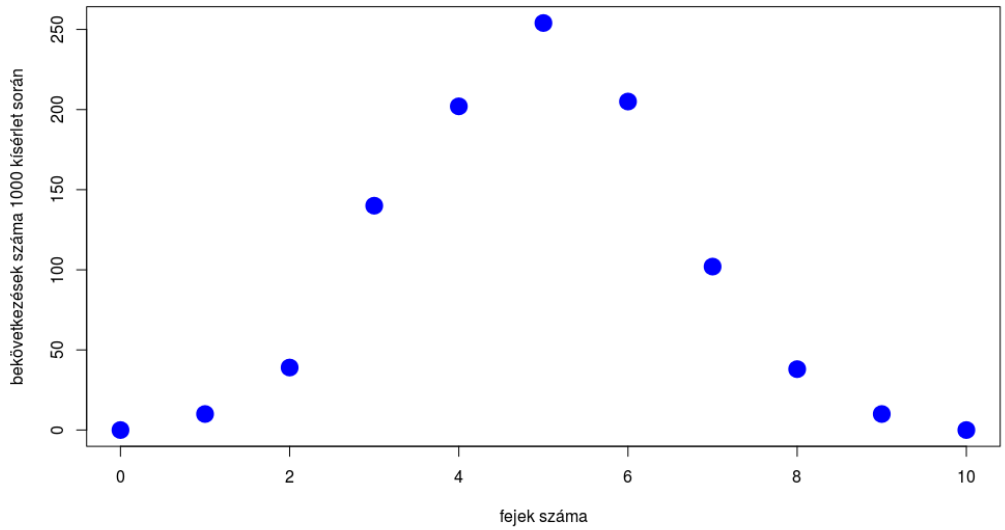
Várható értéke a definíció alapján, felhasználva a

$$k\binom{n}{k} = n\binom{n-1}{k-1}$$

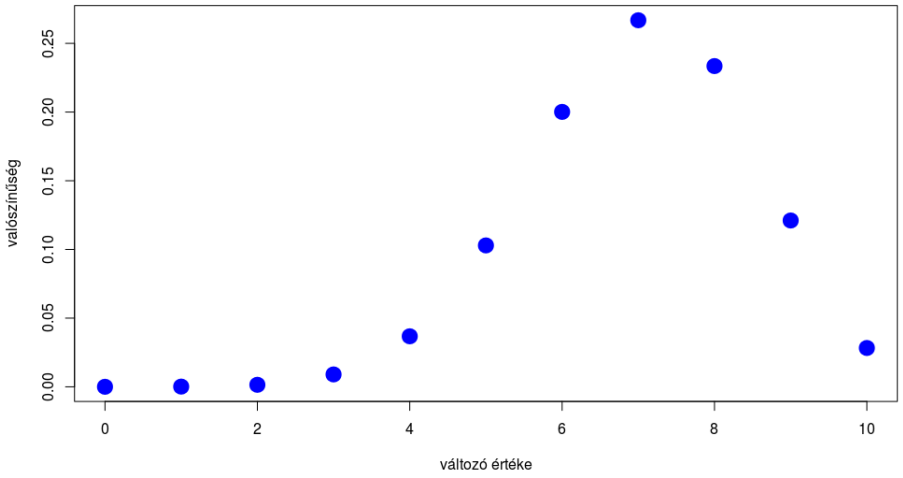
azonosságot,

$$\begin{aligned}\mathbf{E}(\xi) &= \sum_{k=0}^{n} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} n \cdot \binom{n-1}{k-1} p^k (1-p)^{n-k} \\ &= np \sum_{\ell=0}^{n-1} p^\ell (1-p)^{n-1-\ell} = np.\end{aligned}$$





7. ábra. Az $n = 10$, $p = 0.5$ paraméterű binomiális eloszlás szimulációja $N = 10$-szer, és $N = 1000$-szer



8. ábra. Az $n = 10$, $p = 0.7$ paraméterű binomiális eloszlás valószínűségei

A második momentum is hasonlóan kiszámolható, ugyanakkor az alábbi állítás segítségével egyszerűen adódik.

3.6. Állítás. *Legyenek $I_1, \ldots, I_n$ független, Bernoulli(p) eloszlású véletlen változók. Ekkor $\xi = \sum_{i=1}^{n} I_i$ binomiális eloszlású (n, p) paraméterekkel.*

Bizonyítás. Nyilván ξ lehetséges értékei $0, 1, \ldots, n$. Tetszőleges $0 \leq k \leq n$ esetén

$$\begin{aligned}
&\mathbf{P}(\xi = k) \\
&= \mathbf{P}\left(\exists\, 1 \leq i_1 < \ldots < i_k \leq n, \text{hogy } I_j = 1, \text{ ha } j \in \{i_1, \ldots, i_k\}, \text{ különben } 0\right) \\
&= \binom{n}{k} \mathbf{P}(I_1 = \ldots = I_k = 1, I_{k+1} = \ldots = I_n = 0) \\
&= \binom{n}{k} p^k (1-p)^{n-k},
\end{aligned}$$

ahol az utolsó egyenlőség a függetlenség miatt teljesül. Ez éppen azt jelenti, hogy ξ binomiális eloszlású, amint állítottuk. □

Ebből az előállításból gyorsan adódik a várható értékre és a szórásnégyzetre adott formula.

3.7. Következmény. *Legyen $\xi \sim Binom(n, p)$. Ekkor $\mathbf{E}(\xi) = np$, és $D(\xi) = \sqrt{np(1-p)}$.*

Bizonyítás. Az előző állítás szerint $\xi = I_1 + \ldots + I_n$, ahol $I_1, \ldots, I_n$ független p paraméterű Bernoullik. Így

$$\mathbf{E}(\xi) = \sum_{i=1}^{n} \mathbf{E}(I_i) = np.$$

Később belátjuk, hogy független változók összegének a szórásnégyzete az a szórásnégyzetek összege, így

$$\mathbf{D}^2(\xi) = \sum_{i=1}^{n} \mathbf{D}^2(I_i) = np(1-p).$$

□

Tipikus példa: egy p valószínűségű A esemény bekövetkezéseinek a számát vizsgáljuk n független kísérlet során. Ekkor, ha

$$I_j = \begin{cases} 1, & \text{ha a } j\text{-edik kísérletnél } A \text{ bekövetkezett,} \\ 0, & \text{különben,} \end{cases}$$

akkor $I_j \sim \text{Bernoulli}(p)$, és $\xi = \sum_{i=1}^{n} I_j \sim \text{Bin}(n, p)$.

3.2.2. Geometriai eloszlás

A ξ véletlen változó p *paraméterű geometriai eloszlású*, $\xi \sim \text{Geo}(p)$, ha a lehetséges értékek $1, 2, \ldots$ és

$$\mathbf{P}(\xi = k) = (1-p)^{k-1}p, \quad k = 1, 2, \ldots.$$

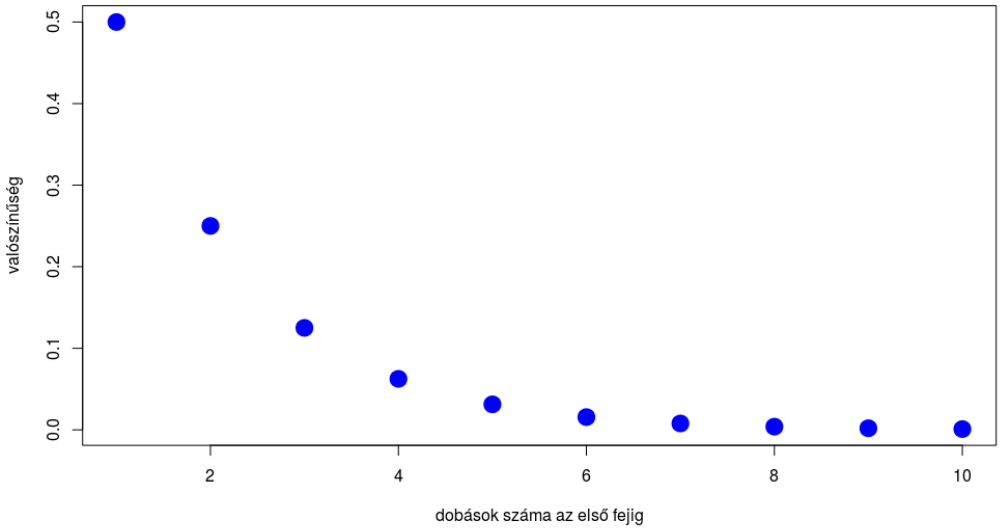
Ez tényleg eloszlás, hiszen

$$\sum_{k=1}^{\infty} p(1-p)^{k-1} = p\frac{1}{1-(1-p)} = 1.$$

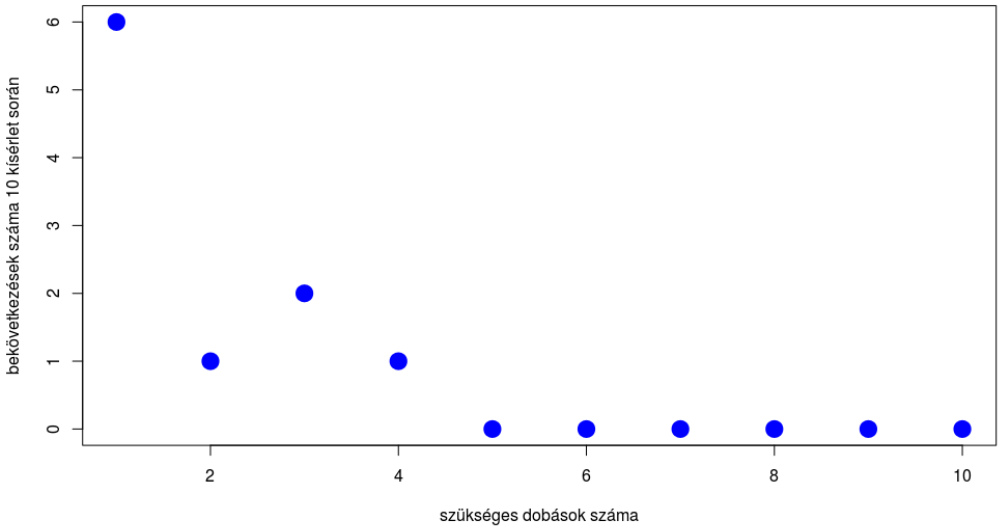
A 9 ábrán a $p = 0.5$ paraméterű geometriai eloszlás valószínűségeit látjuk. Ez a sorozat adja meg azokat a valószínűségeket, hogy egy szabályos érmét dobálva pontosan a k-adikra kapunk először fejet. A 10 ábrán ugyanebből az eloszlásból szimuláltunk 10-szer. Azaz 10-szer elvégeztük azt a kísérletet, hogy egy szabályos érmével az első fejig dobunk.

Megjegyzés. Most a geometriai/mértani sor kell. Volt, hogy

$$\frac{1}{1-x} = 1 + x + x^2 + \ldots = \sum_{n=0}^{\infty} x^n, \quad |x| < 1.$$



9. ábra. A $p = 0.5$ paraméterű geometriai eloszlás valószínűségei



10. ábra. A $p = 0.5$ paraméterű geometriai eloszlás szimulációja $N = 10$-szer

Sőt, még középiskolában volt mértani sorozat, aminek az összegképletére az adódott, hogy

$$1 + q + q^2 + \ldots + q^n = \frac{q^{n+1} - 1}{q - 1}.$$

Ezt úgy csináltuk, hogy beszoroztuk q-val a baloldalt, majd az eredményből kivontuk az eredeti összeget, és ekkor pont a jobb oldal nevezőjét kaptuk. Ennek a végtelen sornak a konvergenciasugara 1, azaz csak akkor értelmes a végtelen sor, ha $|x| < 1$. Sokat fogjuk használni, hogy konvergenciaintervallum belsejében a végtelen sor tagonként differenciálható akárhányszor (lásd Kalkulus 2, ha van).

Mivel

$$\begin{aligned}\sum_{k=1}^{\infty} kx^{k-1} &= \sum_{k=1}^{\infty} (x^k)' = \left(\sum_{k=1}^{\infty} x^k\right)' \\ &= \left(\frac{1}{1-x} - 1\right)' = \frac{1}{(1-x)^2},\end{aligned}$$

(itt most tagonként deriváltuk a sort, amit lehet az előző lábjegyzet szerint) ezért a várható érték

$$\mathbf{E}(\xi) = \sum_{k=1}^{\infty} kp(1-p)^{k-1} = \frac{1}{p}.$$

A második momentum hasonlóan számolható

$$\mathbf{E}(\xi^2) = \frac{2-p}{p^2},$$

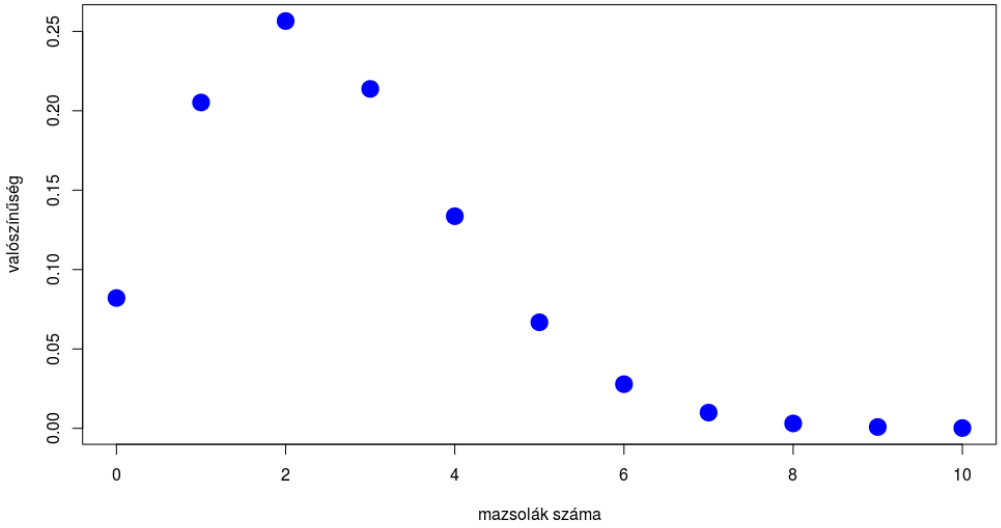
és így

$$\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2 = \frac{1-p}{p^2}.$$

Tipikus példa: addig ismétlünk egy kísérletet, amíg a vizsgált A esemény be nem következik.

A geometriai eloszlás a diszkrét örökifjú eloszlás, hiszen ha $k, \ell \in \mathbb{N}$, akkor

$$\begin{aligned}\mathbf{P}(\xi > k + \ell | \xi > k) &= \frac{\mathbf{P}(\xi > k + \ell)}{\mathbf{P}(\xi > k)} \\ &= \frac{q^{k+\ell}}{q^k} = q^\ell = \mathbf{P}(\xi > \ell).\end{aligned}$$



11. ábra. A $\lambda = 2.5$ paraméterű Poisson-eloszlás valószínűségei

3.2.3. Poisson-eloszlás

A ξ véletlen változó λ *paraméterű Poisson-eloszlású*, $\xi \sim \text{Poisson}(\lambda)$, $\lambda \geq 0$, ha ξ lehetséges értékei $0, 1, 2, \ldots$, és

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

Ez valóban eloszlás, hiszen

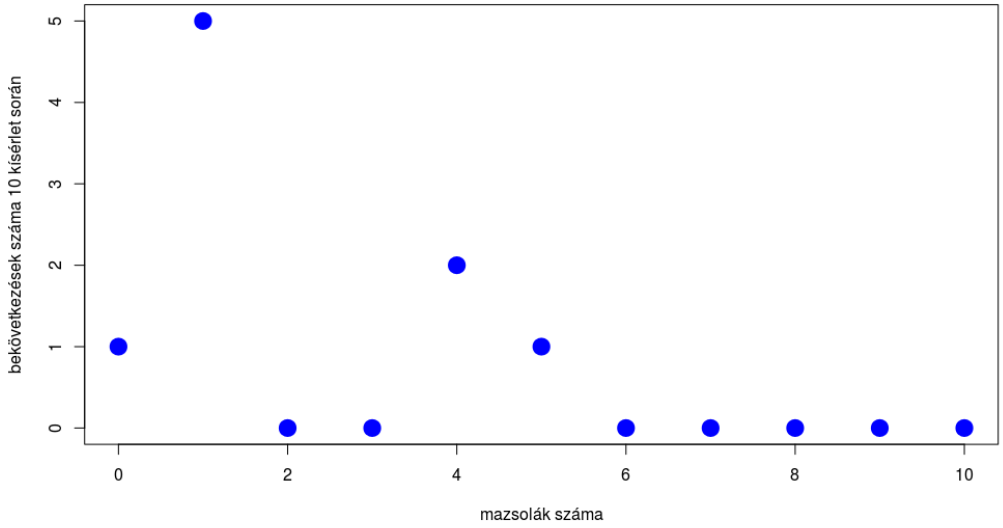
$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}.$$

A 11 ábrán a $\lambda = 2.5$ paraméterhez tartozó Poisson-eloszlás valószínűségeit látjuk. A 12 ábrán ilyen eloszlású véletlen változóból szimuláltunk 10-szer.

Megjegyzés. Sokszor használjuk, hogy az exponenciális függvény 0 pont körüli Taylor-sora (MacLaurin-sora)

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \ldots = \sum_{n=0}^{\infty} \frac{x^n}{n!}, \quad x \in \mathbb{R}.$$

A sor konvergenciasugara végtelen, a sorfejtés tetszőleges $x \in \mathbb{R}$ esetén fennáll. Ez lesz Kalkulus 2 kurzuson, akinek van ilyen.



12. ábra. A $\lambda = 2.5$ paraméterű Poisson-eloszlású véletlen változó szimulációja $N = 10$-szer

Várható értéke

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

Második momentuma hasonlóan számolható

$$\mathbf{E}(\xi^2) = \lambda^2 + \lambda,$$

így szórásnégyzete

$$\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2 = \lambda.$$

Poisson-eloszlás a binomiális eloszlás határeloszlásaként áll elő. Legyen $p = p_n = \lambda/n$, valamely $\lambda > 0$ számra. Ha $\xi_n \sim \mathrm{Bin}(n, p_n)$, akkor

$$\begin{aligned} \mathbf{P}(\xi_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Tartassuk n-et végtelenbe. A szorzat utolsó tényezőjében a nevező 1-hez tart, a kitevő konstans k, ezért az egész 1-hez tart. Az utolsó előtti tényező

1^∞ típusú határérték, innen rögtön beugrik az e. Kalkulusból volt, hogy tetszőleges $x \in \mathbb{R}$ esetén

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x.$$

Ezt $x = -\lambda$-val alkalmazva látjuk, hogy

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Végül a második tényezőben a nevezőben és a számlálóban is k tényezős szorzat van. Mivel $(n-i)/n \to 1$, tetszőleges $i = 1, 2, \ldots, k-1$ esetén (k rögzített!), így a második tényező határértéke 1. Összegezve

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

Azaz a Poisson-eloszlás jól közelíthető kis paraméterű binomiális eloszlással, ha n nagy. Ezek alapján azt látjuk, hogy akkor lép fel Poisson-eloszlás, ha egy kis valószínűségű eseményt sokszor „ismételünk”:

- téves telefonhívások száma;
- autóbalesetek száma;
- nyomdahubák száma egy oldalon;
- földrengések száma;
- csillagok száma egy adott térrészben;
- mazsolák száma a pudingban;
- programhibák száma egy kódban.

Az első példa Ladislaus Bortkiewicz (1868–1931) orosz közgazdásztól (statisztikus) származik: halálos lórúgások száma egy év alatt a porosz hadseregben (20 évig figyelt 14 lovas ezredet). 1898: A kis számok törvénye (Bortkiewicz-eloszlás).

4. Folytonos véletlen változók

4.1. Sűrűségfüggvény, várható érték

Egy véletlen változó értékkészlete nem feltétlenül megszámlálható. A ropi például bárhol eltörhet. Vagy gondolhatunk tetszőleges mérés eredményére, élettartamra, ... Ilyenkor a változó kontinuum sok értéket vehet fel, mindegyiket 0 valószínűséggel. Ez a mese, a definíció a következő.

4.1. Definíció. Egy ξ véletlen változó *folytonos eloszlású*, ha létezik egy nemnegatív f függvény, melyre

$$F(x) = \mathbf{P}(\xi < x) = \int_{-\infty}^{x} f(y)\mathrm{d}y, \quad x \in \mathbb{R}.$$

Az $f(x)$ függvény az ξ véletlen változó sűrűségfüggvénye. Egy ξ folytonos véletlen változó várható értéke

$$\mathbf{E}(\xi) = \int_{-\infty}^{\infty} yf(y)\mathrm{d}y,$$

ha $\int_{-\infty}^{\infty} |y|f(y)\mathrm{d}y < \infty$. A ξ szórása

$$\mathbf{D}(\xi) = \sqrt{\mathbf{E}[(\xi - \mathbf{E}(\xi))^2]} = \sqrt{\mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2},$$

ahol a második momentum

$$\mathbf{E}(\xi^2) = \int_{-\infty}^{\infty} x^2 f(x)\mathrm{d}x.$$

A definícióból világos, hogy $\mathbf{P}(\xi \in (a,b)) = \mathbf{P}(\xi \in (a,b]) = \int_a^b f(y)\mathrm{d}y$, $-\infty \le a \le b \le \infty$. Speciálisan

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = \mathbf{P}(\xi \in \mathbb{R}) = 1, \text{ és } \mathbf{P}(\xi = x) = \int_x^x f(y)\mathrm{d}y = 0.$$

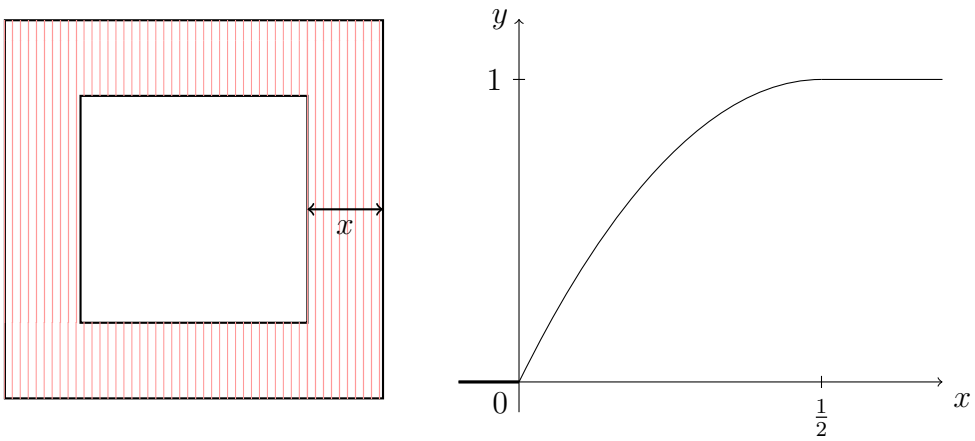
4.2. *Példa.* Egységnégyzetben választunk egyenletes eloszlás szerint egy pontot. Adjuk meg a pont a négyzet határától vett távolságának eloszlását!

Jelölje ξ a távolságot. Ekkor geometriai valószínűségi mezőn vagyunk, $\Omega = [0,1]^2$, $\mathcal{A} = \mathcal{B}([0,1]^2)$, és $\mathbf{P}(A) = |A|$, ahol $|\cdot|$ a terület. Könnyen látható, hogy $\xi : \Omega \to \mathbb{R}$, $(u,v) \mapsto \min\{u, v, 1-u, 1-v\}$, így

$$\begin{aligned}\{\xi < x\} &= \{\omega : \xi(\omega) < x\} \\ &= \begin{cases} \emptyset, & x < 0, \\ \{(u,v) : \min\{u,v,1-u,1-v\} \le x\}, & 0 \le x \le 1/2, \\ [0,1]^2, & x \ge 1/2. \end{cases}\end{aligned}$$

Azaz

$$\begin{aligned}F(x) = \mathbf{P}(\xi < x) &= \mathbf{P}(\{\omega\ : \xi(\omega) < x\}) \\ &= \begin{cases} 0, & \text{ha } x < 0, \\ 4x(1-x), & \text{ha } 0 \le x \le 1/2, \\ 1, & \text{ha } x \ge 1/2. \end{cases}\end{aligned}$$



13. ábra. A jó terület és az eloszlásfüggvény

A sűrűségfüggvény az eloszlásfüggvény deriváltja, azaz

$$f(x) = F'(x) = \begin{cases} 4 - 8x, & x \in (0, 1/2), \\ 0, & \text{különben.} \end{cases}$$

Így, a várható értéke

$$\mathbf{E}(\xi) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_{0}^{1/2} x(4 - 8x) \mathrm{d}x = \frac{1}{2} - \frac{1}{3} = \frac{1}{6}.$$

A szóráshoz még a második momentum kell, ami

$$\mathbf{E}(\xi^2) = \int_{-\infty}^{\infty} x^2 f(x) \mathrm{d}x = \int_{0}^{1/2} x^2(4 - 8x) \mathrm{d}x = \frac{1}{6} - \frac{1}{8} = \frac{1}{24}.$$

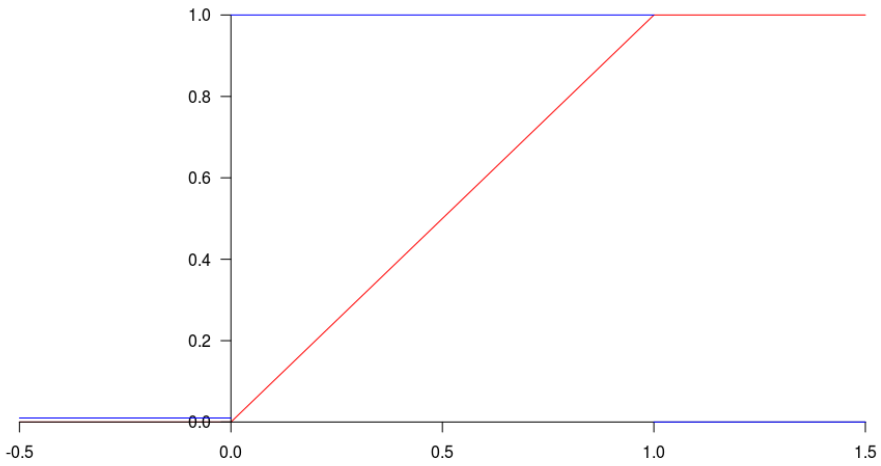
Tehát $\mathbf{D}(\xi) = \sqrt{\frac{1}{24} - \frac{1}{36}} \approx 0,118$.

4.2. Nevezetes folytonos eloszlások

4.2.1. Egyenletes eloszlás

A ξ véletlen változó *egyenletes eloszlású az* (a, b) *intervallumon*, $a < b$, jelben $\xi \sim$ Egyenletes(a, b), ha sűrűségfüggvénye

$$f(y) = \begin{cases} \frac{1}{b-a}, & \text{ha } y \in (a, b), \\ 0, & \text{különben.} \end{cases}$$



14. ábra. Az $a = 0$, $b = 1$ paraméterű egyenletes eloszlás sűrűség- és eloszlásfüggvénye

Ez tényleg sűrűségfüggvény, hiszen $f \geq 0$, és $\int_{-\infty}^{\infty} f(y)\mathrm{d}y = 1$.

Vegyük észre, hogy ez ugyanaz a definíció, mint korábban, a geometriai valószínűségi mezőnél. Valóban, ha $(c, d) \subset (a, b)$ egy tetszőleges részintervallum, akkor

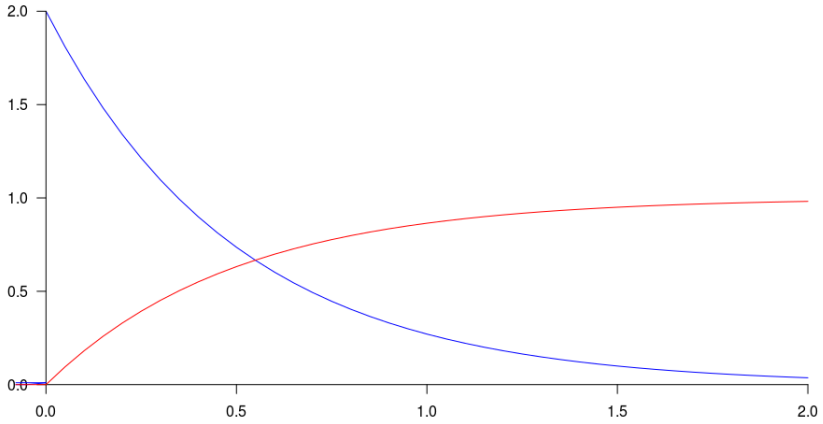
$$\mathbf{P}(\xi \in (c, d)) = \int_c^d f(y)\mathrm{d}y = \frac{d-c}{b-a}.$$

Eloszlásfüggvénye

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y = \begin{cases} 0, & \text{ha } x \leq a, \\ \frac{x-a}{b-a}, & \text{ha } x \in [a, b], \\ 1, & \text{ha } x \geq b. \end{cases}$$

Momentumai, $k \geq 1$

$$\begin{aligned} \mathbf{E}(\xi^k) &= \int_{-\infty}^{\infty} y^k f(y)\mathrm{d}y \\ &= \int_a^b y^k \frac{1}{b-a}\mathrm{d}y \\ &= \frac{b^{k+1} - a^{k+1}}{(k+1)\,(b-a)}. \end{aligned}$$



15. ábra. A $\lambda = 2$ paraméterű exponenciális eloszlás sűrűség- és eloszlásfüggvénye

Speciálisan

$$\mathbf{E}(\xi) = \frac{a+b}{2},\ \mathbf{D}^2(\xi) = \frac{(b-a)^2}{12}.$$

4.2.2. Exponenciális eloszlás

A ξ véletlen változó λ-*paraméterű exponenciális eloszlású*, $\xi \sim \mathrm{Exp}(\lambda)$, $\lambda > 0$, ha sűrűségfüggvénye

$$f(y) = \begin{cases} \lambda e^{-\lambda y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

Ez tényleg sűrűségfüggvény. A megfelelő eloszlásfüggvény

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(y)\mathrm{d}y \\ &= \begin{cases} 1 - e^{-\lambda x}, & x \geq 0, \\ 0, & x \leq 0. \end{cases} \end{aligned}$$

Momentumai

$$\mathbf{E}(\xi^k) = \int_{-\infty}^{\infty} y^k f(y)\mathrm{d}y = \int_0^{\infty} y^k \lambda e^{-\lambda y}\mathrm{d}y$$
$$= \lambda^{-k} \int_0^{\infty} z^k e^{-z}\mathrm{d}z = \lambda^{-k}\Gamma(k+1) = \frac{k!}{\lambda^k}.$$

Itt fölhasználtuk, hogy a

$$\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1}e^{-y}\mathrm{d}y,\ \alpha > 0,$$

Gamma-függvényre teljesül, hogy $\Gamma(k) = (k-1)!$, azaz a függvény a faktoriális folytonos kiterjesztése. Ez az azonosság következik a

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$$

azonosságból, ami parciális integrálással könnyen adódik. Valóban, az e^{-y} függvényt integrálva és a y^α függvényt deriválva kapjuk

$$\Gamma(\alpha+1) = \int_0^{\infty} y^\alpha e^{-y}\mathrm{d}y$$
$$= \left[-e^{-y}y^\alpha\right]_{y=0}^{\infty} - \int_0^{\infty}(-e^{-y})\alpha y^{\alpha-1}\mathrm{d}y$$
$$= 0 + \alpha\Gamma(\alpha).$$

Itt felhasználtuk, hogy $\lim_{y\to\infty} e^{-y}y^\alpha = 0$ tetszőleges $\alpha > 0$ esetén, azaz az exponenciális függvény minden hatványfüggvénynél gyorsabban tart végtelenben.

Ezek szerint

$$\mathbf{E}(\xi) = \frac{1}{\lambda},\ \mathbf{D}^2(\xi) = \frac{1}{\lambda^2}.$$

Az exponenciális eloszlás karakterizálja az ún. *örökifjú tulajdonság, vagy emlékezet nélküliség.* Ez azt jelenti, hogy tetszőleges $x, y > 0$ esetén

$$\mathbf{P}(\xi \geq x+y | \xi \geq x) = \mathbf{P}(\xi \geq y). \tag{1}$$

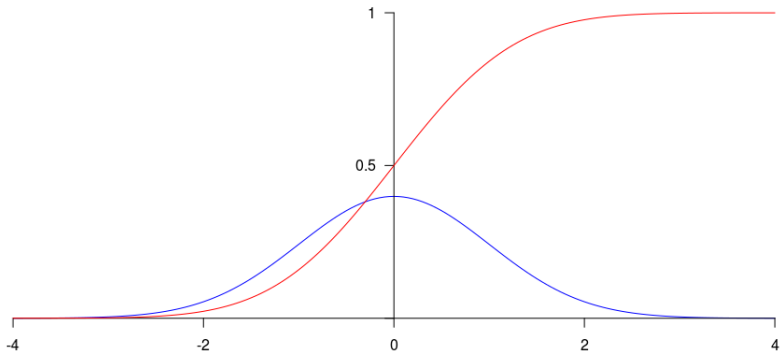
Ez valóban azt jelenti, hogy az eloszlás nem öregszik.

Ha $\xi \sim \mathrm{Exp}(\lambda)$, akkor ez teljesül, hiszen

$$\mathbf{P}(\xi \geq x+y | \xi \geq x) = \frac{\mathbf{P}(\xi \geq x+y)}{\mathbf{P}(\xi \geq x)}$$
$$= e^{-\lambda y} = \mathbf{P}(\xi \geq y),$$

ami éppen (1). A fordított irány, logaritmust véve, a Cauchy-féle függvényegyenlet megoldásából következik.

Tipikus példák: telefonhívás hossza, várakozási idő, alkatrészek élettartama, üvegpohár élethossza.



16. ábra. A standard normális eloszlás sűrűség- és eloszlásfüggvénye

4.2.3. Normális eloszlás

A ξ véletlen változó *normális eloszlású* μ *és* σ^2 *paraméterekkel*, $\xi \sim \mathrm{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, ha sűrűségfüggvénye

$$f_{\mu,\sigma}(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

A $\mu = 0$ és $\sigma = 1$ paraméterekhez tartozó eloszlást *standard normális eloszlásnak* nevezzük. A normális eloszlást nevezik Gauss-eloszlásnak is.

Könnyen látható, hogy $f_{\mu,\sigma}$ függvény μ-re szimmetrikus, azaz $f_{\mu,\sigma}(\mu+y) = f_{\mu,\sigma}(\mu - y)$, $y \in \mathbb{R}$, μ-ben van a maximuma, és $\mu \pm \sigma$ inflexiós pontok.

Ahhoz, hogy belássuk, hogy $f_{\mu,\sigma}$ sűrűségfüggvény, az alábbi lemmát bizonyítás nélkül felhasználjuk.

4.2.1. Lemma. *Az* $\int_{-\infty}^{\infty} e^{-t^2/2}\,\mathrm{d}t$ *integrál létezik mint improprius Riemann-integrál és értéke* $\sqrt{2\pi}$.

Az $f_{\mu,\sigma}$ függvény nemnegatív. A $t = (y-\mu)/\sigma$ helyettesítéssel

$$\begin{aligned}\int_{-\infty}^{\infty} f_{\mu,\sigma}(y)\mathrm{d}y &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\mathrm{d}y \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\mathrm{d}t = 1,\end{aligned}$$

ahol az utolsó egyenlőségnél a 4.2.1 Lemmát használtuk. Azaz $f_{\mu,\sigma}$ valóban sűrűség.

A várható érték

$$\begin{aligned}\mathbf{E}(\xi) &= \int_{-\infty}^{\infty} y f_{\mu,\sigma}(y)\mathrm{d}y \\ &= \frac{\sigma}{\sqrt{2\pi}}\left(\int_{-\infty}^{\infty}\frac{y-\mu}{\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\frac{1}{\sigma}\mathrm{d}y + \int_{-\infty}^{\infty}\frac{\mu}{\sigma}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\frac{1}{\sigma}\mathrm{d}y\right) = \mu,\end{aligned}$$

a szórásnégyzet pedig

$$\begin{aligned}\mathbf{D}^2(\xi) &= \mathbf{E}((\xi-\mu)^2) \\ &= \int_{-\infty}^{\infty}(y-\mu)^2 f_{\mu,\sigma}(y)\mathrm{d}y \\ &= \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(y-\mu)\cdot\frac{y-\mu}{\sigma^2}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\mathrm{d}y \\ &= \left[-\frac{\sigma}{\sqrt{2\pi}}(y-\mu)e^{-\frac{(y-\mu)^2}{2\sigma^2}}\right]_{-\infty}^{\infty} + \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-\frac{(y-\mu)^2}{2\sigma^2}}\mathrm{d}y = \sigma^2.\end{aligned}$$

Tehát a definícióban szereplő két paraméter az a várható érték és a szórásnégyzet.

Az ξ eloszlásfüggvénye a következőképpen számolható:

$$\begin{aligned}F(x) = \mathbf{P}(\xi \le x) &= \int_{-\infty}^{x} f_{\mu,\sigma}(y)\mathrm{d}y \\ &= \int_{-\infty}^{(x-\mu)/\sigma}\frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\mathrm{d}y = \Phi((x-\mu)/\sigma),\end{aligned} \tag{2}$$

ahol

$$\Phi(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{y^2}{2}}\mathrm{d}y$$

a standard normális eloszlás eloszlásfüggvénye. Ebből a számolásból világos, hogy elég a Φ függvény értékeit ismerni, és ebből tetszőleges paraméterű normális eloszlás eloszlásfüggvénye számolható.

Ugyancsak (2) egyszerű következménye az alábbi állítás.

4.3. Állítás. *Ha $\xi \sim N(\mu,\sigma^2)$, akkor $(\xi-\mu)/\sigma \sim N(0,1)$.*

Sőt, ez kicsit általánosabban is igaz: ha $\xi \sim \mathrm{N}(\mu,\sigma^2)$, és a, b valós állandók, $a \neq 0$, akkor $a\xi + b \sim \mathrm{N}(a\mu+b, a^2\sigma^2)$.

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7703	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

17. ábra. Standard normális eloszlástáblázat

A normális eloszlás nagyon erősen koncentrálódik a várható értéke körül. Valóban, ha $\xi \sim \mathrm{N}(\mu, \sigma^2)$ és $Z \sim \mathrm{N}(0,1)$, akkor

$$\mathbf{P}(|\xi - \mu| \le \lambda\sigma) = \mathbf{P}(|Z| \le \lambda) = \Phi(\lambda) - \Phi(-\lambda)$$
$$= 2\Phi(\lambda) - 1 = \begin{cases} 0,6827, & \lambda = 1, \\ 0,9545, & \lambda = 2, \\ 0,9973, & \lambda = 3, \\ 0,9999, & \lambda = 4. \end{cases}$$

5. Várható érték

5.1. Várható érték tulajdonságai, szórás, momentumok

Először felidézzük a várható érték definícióját diszkrét és folytonos esetben.

5.1. Definíció. Ha ξ diszkrét véletlen változó $x_1, x_2, \ldots$ lehetséges értékekkel, akkor az ξ *várható értéke*

$$\mathbf{E}(\xi) = \sum_i x_i \mathbf{P}(\xi = x_i),$$

ha $\sum_i |x_i| \mathbf{P}(\xi = x_i) < \infty$.

Ha ξ folytonos véletlen változó $f(x)$ sűrűségfüggvénnyel, akkor az ξ várható értéke

$$\mathbf{E}(\xi) = \int_{-\infty}^{\infty} y f(y) \mathrm{d}y,$$

ha $\int_{-\infty}^{\infty} |y| f(y) \mathrm{d}y < \infty$.

5.2. Állítás. *A következőkben a, b valós konstansok, $\xi, \eta, \xi_1, \ldots, \xi_n$ véletlen változók.*

(i) Tetszőleges $g : \mathbb{R} \to \mathbb{R}$ függvényre

$$\mathbf{E}(g(\xi)) = \sum_{i=1}^{\infty} g(x_i) \mathbf{P}(\xi = x_i), \text{ ill. } \mathbf{E}(g(\xi)) = \int_{-\infty}^{\infty} g(y) f(y) \mathrm{d}y.$$

Tetszőleges $h : \mathbb{R}^2 \to \mathbb{R}$ függvény és ξ, η diszkrét véletlen változók esetén

$$\mathbf{E}(h(\xi, \eta)) = \sum_{i,j} h(x_i, y_j) \mathbf{P}(\xi = x_i, \eta = y_j).$$

(ii) A várható érték lineáris, azaz tetszőleges $a, b \in \mathbb{R}$ állandókra

$$\mathbf{E}(a\xi + b) = a\mathbf{E}(\xi) + b.$$

(iii) *Ha* $a \leq \xi \leq b$, *akkor* $a \leq \mathbf{E}(\xi) \leq b$ *tetszőleges* $a, b \in \mathbb{R}$ *számok esetén.*

(iv) $\mathbf{E}(\xi + \eta) = \mathbf{E}(\xi) + \mathbf{E}(\eta)$.

(v) *Ha* $\xi_1, \xi_2, \ldots, \xi_n$ *véletlen változók, akkor*

$$\mathbf{E}\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathbf{E}(\xi_i).$$

Bizonyítás. (i) Csak diszkrét esetben bizonyítunk, és csak az első állítást. Mivel ξ diszkrét, ezért $g(\xi)$ is diszkrét, és lehetséges értékei $g(x_1), g(x_2), \ldots$. Így a várható érték definíciója szerint

$$\mathbf{E}(g(\xi)) = \sum_i g(x_i)\mathbf{P}(\xi = x_i).$$

(ii) Az előző állítást $g(x) = ax + b$ függvénnyel felírva

$$\mathbf{E}(a\xi + b) = \sum_i (ax_i + b)\mathbf{P}(\xi = x_i) = a\mathbf{E}(\xi) + b.$$

Folytonosra ugyanígy.

(iii)

$$a = a\sum_i \mathbf{P}(\xi = x_i) \leq \sum_i x_i \mathbf{P}(\xi = x_i) = \mathbf{E}(\xi) \leq \sum_i b\mathbf{P}(\xi = x_i) = b.$$

Folytonosra ugyanígy.

(iv) Csak diszkrét esetben bizonyítunk. Az (i) pont szerint $h(x, y) = x+y$ függvénnyel

$$\begin{aligned}
\mathbf{E}(\xi + \eta) &= \sum_i \sum_j (x_i + y_j)\mathbf{P}(\xi = x_i, \eta = y_j) \\
&= \sum_i x_i \sum_j \mathbf{P}(\xi = x_i, \eta = y_j) + \sum_j y_j \sum_i \mathbf{P}(\xi = x_i, \eta = y_j) \quad \text{tvt} \\
&= \sum_i x_i \mathbf{P}(\xi = x_i) + \sum_j y_j \mathbf{P}(\eta = y_j) \\
&= \mathbf{E}(\xi) + \mathbf{E}(\eta).
\end{aligned}$$

(v) Következik (iv)-ből teljes indukcióval. □

5.3. *Példa.* Csodaország munka törvénykönyve szerint egy cég minden munkása fizetett szabadságot kap azokon a napokon, amikor legalább az egyiküknek születésnapja van. Ezen napok kivételével azonban az év minden napján

mindenkinek dolgoznia kell. Minden munkás 1 TV-készüléket készít egy nap alatt. Hány alkalmazottat vegyen fel a cégtulajdonos, ha azt akarja, hogy a gyártott TV-készülékek számának a várható értéke maximális legyen?

Legyen n az alkalmazottak száma. Jelölje ξ_n az egy évben gyártott TV-k számát, és legyen η_i az i-edik napon gyártott TV-k száma, $i = 1, 2, \ldots, 365$. Világos, hogy

$$\xi_n = \eta_1 + \eta_2 + \ldots + \eta_{365}.$$

Másrészt η_i-k azonos eloszlásúak, lehetséges értékeik n vagy 0, és

$$\mathbf{P}(\eta_i = n) = \mathbf{P}(\text{nincs születésnap az } i\text{-edik napon}) = \left(\frac{364}{365}\right)^n.$$

Tehát

$$\mathbf{E}(\xi_n) = 365\, n \left(\frac{364}{365}\right)^n.$$

Egyszerű számolással kapjuk, hogy

$$\frac{\mathbf{E}(\xi_n)}{\mathbf{E}(\xi_{n+1})} \leq 1 \;\Leftrightarrow\; n \leq 364,$$

azaz a maximum az $n = 364$ és $n = 365$ helyeken vétetik fel.

5.4. *Példa* (Kupongyűjtő probléma.). Egy N különböző elemből álló sokaságból visszatevéses mintát veszünk. Jelölje S_r azt a véletlen számot, ahány elemet kellett húznunk, hogy kapjunk r különböző elemet. Határozzuk meg S_r várható értékét, majd adjunk $\mathbf{E}(S_N)$-re kezelhető aszimptotikus egyenlőséget!

Vezessük be az $\xi_k = S_{k+1} - S_k$ változót, $S_0 = 0$. Ekkor ξ_k geometriai eloszlású, ahol a siker valószínűsége $p_k = (N - k))/N$. Ezért $\mathbf{E}(\xi_k) = \frac{1}{p_k} = \frac{N}{N-k}$. Visszaírva S_r-be

$$\begin{aligned}\mathbf{E}(S_r) &= \mathbf{E}(\xi_0 + \xi_1 + \ldots + \xi_{r-1}) \\ &= \mathbf{E}(\xi_0) + \mathbf{E}(\xi_1) + \ldots + \mathbf{E}(\xi_{r-1} \\ &= \frac{N}{N} + \frac{N}{N-1} + \ldots + \frac{N}{N-r+1}.\end{aligned}$$

Ha teljes mintát akarunk, azaz $r = N$, akkor

$$\mathbf{E}(S_N) = N\left(\frac{1}{N} + \frac{1}{N-1} + \ldots + \frac{1}{1}\right) \sim N \ln N,$$

amint $N \to \infty$.

5.5. Definíció. Az ξ véletlen változó k-adik momentuma $\mathbf{E}(\xi^k)$, és k-adik centrális momentuma $\mathbf{E}[(\xi - \mathbf{E}\xi)^k]$, $k = 1, 2, \ldots$. Az 5.2 Állítás szerint

$$\mathbf{E}(\xi^k) = \begin{cases} \sum_i x_i^k \mathbf{P}(\xi = x_i), & \text{ha } \xi \text{ diszkrét,} \\ \int_{-\infty}^{\infty} x^k f(x) \mathrm{d}x, & \text{ha } \xi \text{ folytonos.} \end{cases}$$

5.6. Definíció. Az ξ véletlen változó szórása $\mathbf{D}(\xi) = \sqrt{\mathbf{E}(\xi - \mathbf{E}(\xi))^2}$.

A szórás annak a mérőszáma, hogy a változó mennyire tér el a várható értékétől. Mivel $\mathbf{E}(\xi - \mathbf{E}(\xi)) = 0$, $\mathbf{E}|\xi - \mathbf{E}(\xi)|$ pedig nehezen kezelhető (nem differenciálható az $|\cdot|$ függvény), ezért ez a legegyszerűbb ilyen.

5.7. Állítás. *Tetszőleges ξ véletlen változó és a, b valós számok esetén*

(i) $\mathbf{D}^2(\xi) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2$*;*

(ii) $\mathbf{D}^2(a\xi + b) = a^2\mathbf{D}^2(\xi)$*;*

(iii) $\mathbf{D}(\xi) = 0$ *akkor és csak akkor, ha* $\xi = \mathbf{E}(\xi)$*, azaz ξ konstans véletlen változó.*

Bizonyítás. A definíció alkalmazása. □

5.2. Huffman-kód

Legyen $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ véges halmaz, a *forrásábécé*. Ekkor tehát x_1, x_2, $\ldots, x_n$ a betűk. Minden betűt egy véges hosszú $0-1$ sorozattal kódolunk. Formálisan, ha η^* jelöli a kódszavak halmazát, azaz a véges hosszú $0-1$ sorozatokat, akkor a kód egy $f : \mathcal{X} \to \eta^*$ függvény. Az f-hez tartozó lehetséges kódszavak $f(x_1), f(x_2), \ldots, f(x_n)$.

Az olyan kódolások érdekelnek, melyek egyértelműen dekódolhatók. Az f kód *prefix*, ha a lehetséges kódszavak közül egyik sem folytatása a másiknak. Világos, hogy egy prefix kód egyértelműen dekódolható. Jelölje $x \in \mathcal{X}$ esetén $|f(x)|$ a kódszó hosszát.

5.8. *Példa.* Legyen $\mathcal{X} = \{a, b, c\}$, és legyen $f_1(a) = 0$, $f_1(b) = 01$, $f_1(c) = 011$. Ekkor f_1 nem prefix kód, de könnyen látható, hogy egyértelműen dekódolható. Az $f_2(a) = 01$, $f_2(b) = 00$, $f_2(c) = 1$, kód prefix.

Világos, hogy ha n nagy, akkor hosszú kódszavak is kellenek. Olyan kódolást keresünk, amiben a hosszú kódszavak ritkán jönnek elő. Legyen X egy véletlen betű, és eloszlása $\mathbf{P}(X = x_k) = p_k$, $k = 1, 2, \ldots, n$. (A forrásábécét helyettesíthetjük számokkal, és akkor a definíció szerinti véletlen változót kapunk.) Tehát p_k a x_k betű gyakorisága az adott nyelvben.

Adott f kód esetén egy hosszú szövegben az egy karakterre eső átlagos kódszóhossz az

$$\mathbf{E}(|f(X)|) = \sum_{k=1}^{n} p_k |f(x_k)|$$

várható érték. Ezt szeretnénk minimalizálni a prefix kódok körében.

A betűk átrendezhetők, ezért feltehető, hogy $p_1 \geq p_2 \geq \ldots \geq p_n$. Ha az f prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \ldots \leq |f(x_n)|.$$

Hát persze, hiszen ha nem így lenne, akkor felcserélhetnénk két kódszót, és kisebb várható értéket kapnánk;.

(ii) A két legkisebb valószínűséghez tartozó kód hossza egyenlő. Hát persze, hiszen ha $|f(x_{n-1})| < |f(x_n)|$ teljesülne, akkor $f(x_n)$-ből eldobhatnánk az utolsó bitet.

(iii) $f(x_{n-1})$ és $f(x_n)$ csak az utolsó bitben térnek el. Valóban, az előző pont alapján látjuk, hogy van olyan i, hogy $f(x_i)$ és $f(x_n)$ csak az utolsó bitben térnek el. Na de ekkor, $|f(x_i)| = |f(x_{n-1})| = |f(x_n)|$, és így az i-edik és $(n-1)$-edik kódszót felcserélhetem.

A fenti (i)–(iii) tulajdonságok segítségével megmutatható a következő.

5.9. Tétel. *Tegyük fel, hogy az*

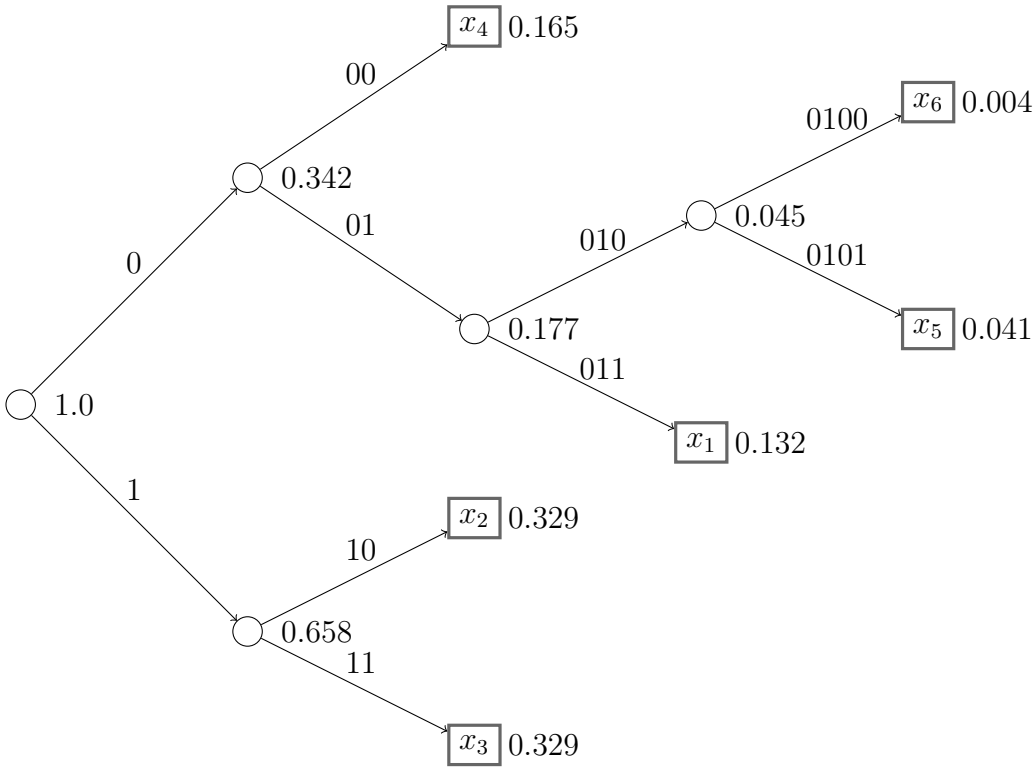
$$\mathcal{X}' = \{x_1, \ldots, x_{n-2}, y_{n-1}\}$$

$(n-1)$ *elemű forrásábécé és* $p_1, \ldots, p_{n-2}, p_{n-1}+p_n$ *eloszlás esetén* g *egy optimális prefix kód. Ekkor az eredeti problémához tartozó optimális prefix kódot kapunk, ha az* x_{n-1}*, ill.* x_n *kódját úgy választjuk, hogy a* $g(y_{n-1})$ *kódszót kiegészítjük 0-val, ill. 1-gyel, a többi kódszót változatlanul hagyjuk.*

Az előző tétel alapján indukcióval készíthetünk egy optimális prefix kódot. Minden lépésben vonjuk össze a két legkisebb valószínűségű betűt, a megfelelő valószínűségeket összeadva. Ezt addig csináljuk, amíg két valószínűségünk marad. Majd az összevonások megfordításával visszafelé haladva a megfelelő kódszó kétféle kiegészítésével optimális kódot kapunk. Ez a *Huffman-kód.*

5.10. *Példa.* Legyen $n = 6$, $\mathcal{X} = \{x_1, \ldots, x_6\}$, és $p_1 = 0.132$, $p_2 = 0.329$, $p_3 = 0.329$, $p_4 = 0.165$, $p_5 = 0.041$, $p_6 = 0.004$. Ekkor 4 összevonás kell, ezek sorrendje:

(i) $x_5, x_6 \to x_{56}$, $p_{56} = 0.045$;



18. ábra. Huffman-kód készítése

(ii) x_1, x_{56}, $p_{156} = 0.177$;

(iii) x_{156}, x_4, $p_{1564} = 0.342$;

(iv) x_2, x_3, $p_{23} = 0.658$.

Így az optimális kód:

x_1	x_2	x_3	x_4	x_5	x_6
011	10	11	00	0101	0100

Az így kapott optimális kód esetén a várható kódhossz

$$\begin{aligned}\mathbf{E}(f(X)) = {} & 0.132 \cdot 3 + 0.329 \cdot 2 + 0.329 \cdot 2 \\ & + 0.165 \cdot 2 + 0.041 \cdot 4 + 0.004 \cdot 4 = 2.22.\end{aligned}$$

Vegyük észre, hogy fix kódhossz esetén 3 bit kell a kódoláshoz.

Megmutatható, hogy az optimális várható kódhosszra teljesül, hogy

$$\sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} \leq \mathbf{E}(|f(X)|) < \sum_{k=1}^{n} p_k \log_2 \frac{1}{p_k} + 1.$$

A fenti kifejezésben megjelenő mennyiség a (p_k) valószínűségeloszlás *entrópiája*. Minél véletlenebb a sorozat, azaz minél közelebb van az egyenleteshez az eloszlás, annál nagyobb az entrópia, és annál nehezebb gazdaságosan kódolni.

A Huffman-kódot használják pl. JPEG és MP3 kódolásnál.

6. Véletlen változók függősége

6.1. Véletlen vektorváltozók

Egy kísérletnél sokszor több a kísérlet eredményét leíró adatra vagyunk kíváncsiak. Például testtömeg, testmagasság, vérnyomás, pulzus, ...

6.1. Definíció. Az $\xi = (\xi_1, \dots, \xi_n) : \Omega \to \mathbb{R}^n$ függvény véletlen vektorváltozó, ha minden komponense véletlen változó. Az ξ eloszlásfüggvénye

$$F(x_1, \dots, x_n) = \mathbf{P}(\xi_1 < x_1, \dots, \xi_n < x_n).$$

Az $(\xi_1, \dots, \xi_n)$ véletlen vektorváltozó diszkrét, ha értékkészlete megszámlálható.

Az ξ_i, $i = 1, 2, \dots, n$, változók eloszlását, peremeloszlásnak, vagy marginális eloszlásnak nevezzük.

6.2. *Példa* (Trinomiális eloszlás). Egy szabályos dobókockával n-szer dobunk. Jelölje ξ a hatosok, η egyesek számát!

Mind a hatosok, mind az egyesek száma *binomiális eloszlású* véletlen változó $(n, 1/6)$ paraméterrel. Eloszlásuk

$$\mathbf{P}(\xi = k) = \mathbf{P}(\eta = k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}, \quad k = 0, 1, \dots, n.$$

Így,

$$\mathbf{E}(\xi) = \frac{n}{6}, \quad \mathbf{D}^2(\xi) = n\frac{5}{36}.$$

A (ξ, η) véletlen vektorváltozó lehetséges értékei olyan (k, ℓ) párok, melyre $0 \le k, \ell \le n$, és $k + \ell \le n$, hiszen külön-külön 0 és n közötti egész értéket vehetnek fel, és az összegük legfeljebb n, a dobások száma. A megfelelő valószínűségek

$$\mathbf{P}(\xi = k, \eta = \ell) = \binom{n}{k}\binom{n-k}{\ell} \left(\frac{1}{6}\right)^{k+\ell} \left(\frac{4}{6}\right)^{n-k-\ell},$$

hiszen a k db hatost és ℓ db egyest $\binom{n}{k}\binom{n-k}{\ell}$ féleképpen választhatom ki, ezeken a helyeken $1/6$ valószínűséggel dobom azt, amit kell, a maradék $n -$

$k-\ell$ helyen pedig 4 lehetőségem van, hiszen nem lehet sem egyes, sem hatos. Ez a trinomiális eloszlás. A példára még később visszatérünk a kovariancia kiszámolásánál.

Legyenek $\xi_1, \dots, \xi_n$ az $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőn értelmezett véletlen változók.

6.3. Definíció. Az $\xi_1, \dots, \xi_n$ függetlenek, ha minden $x_1, \dots, x_n \in \mathbb{R}$ esetén

$$\mathbf{P}(\xi_1 < x_1, \dots, \xi_n < x_n) = \mathbf{P}(\xi_1 < x_1) \dots \mathbf{P}(\xi_n < x_n)$$

teljesül. Vagyis az együttes eloszlásfüggvény az egyes eloszlásfüggvények szorzata.

Diszkrét esetben ez a karakterizáció tovább egyszerűsíthető.

6.4. Állítás. *Legyenek $\xi_1, \dots, \xi_n$ diszkrét véletlen változók úgy, hogy ξ_i lehetséges értékei $x_1^{(i)}, x_2^{(i)}, \dots$, $i = 1, 2, \dots, n$. Ekkor $\xi_1, \dots, \xi_n$ pontosan akkor függetlenek, ha*

$$\mathbf{P}(\xi_1 = x_{i_1}^{(1)}, \dots, \xi_n = x_{i_n}^{(n)}) = \mathbf{P}(\xi_1 = x_{i_1}^{(1)}) \dots \mathbf{P}(\xi_n = x_{i_n}^{(n)})$$

teljesül tetszőleges $i_i, \dots, i_n$ indexekre.

6.5. *Példa* (Kockadobás). Egy szabályos dobókockával kétszer dobunk. Jelölje ξ az első, η a második dobás eredményét. Ekkor tetszőleges $k, \ell \in \{1, 2, \dots, 6\}$ esetén

$$\mathbf{P}(\xi = k, \eta = \ell) = \frac{1}{36} = \frac{1}{6} \cdot \frac{1}{6} = \mathbf{P}(\xi = k) \cdot \mathbf{P}(\eta = \ell),$$

azaz ξ és η függetlenek.

Ha a $\xi_1, \xi_2, \dots, \xi_n$ véletlen változók függetlenek, akkor a definíció szerint a $\{\xi_1 < x_1\}$, $\{\xi_2 < x_2\}, \dots$, $\{\xi_n < x_n\}$ események függetlenek, tetszőleges $x_1, x_2, \dots, x_n$ valós számokra. Ennél több is igaz. A változók pontosan akkor függetlenek, ha tetszőleges $B_1, \dots, B_n$ szép halmazokra (mondjuk intervallumokra) a $\{\xi_1 \in B_1\}$, …, $\{\xi_n \in B_n\}$ események függetlenek. Azaz, például a $\{\xi_1 > 3\}$, $\{\xi_2 = 5\}$ események is függetlenek.

6.6. Állítás. *Legyenek ξ, η független diszkrét véletlen változók. Ekkor*

$$\mathbf{E}\left(g_1(\xi) g_2(\eta)\right) = \mathbf{E}\left(g_1(\xi)\right) \mathbf{E}\left(g_2(\eta)\right).$$

Speciálisan, ha ξ és η függetlenek, akkor $\mathbf{E}(\xi\eta) = \mathbf{E}(\xi)\mathbf{E}(\eta)$.

Bizonyítás. Az 5.2 Állítás (i) pontja és a függetlenség szerint

$$\begin{aligned}\mathbf{E}(g_1(\xi)g_2(\eta)) &= \sum_i \sum_j g_1(x_i)g_2(y_j)\mathbf{P}(\xi = x_i, \eta = y_j) \\ &= \sum_i \sum_j g_1(x_i)g_2(y_j)\mathbf{P}(\xi = x_i)\mathbf{P}(\eta = y_j) \quad \text{függetlenség} \\ &= \sum_i g_1(x_i)\mathbf{P}(\xi = x_i) \sum_j g_2(y_j)\mathbf{P}(\eta = y_j) \\ &= \mathbf{E}(g_1(\xi))\mathbf{E}(g_2(\eta)).\end{aligned}$$

□

6.2. Kovariancia, korreláció

Véletlen változók függőségének mérőszámai a kovariancia és a korreláció.

6.7. Definíció. Az ξ és η véletlen változók *kovarianciája*

$$\mathbf{Cov}(\xi, \eta) = \mathbf{E}[(\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))],$$

korrelációja

$$\rho(\xi, \eta) = \frac{\mathbf{Cov}(\xi, \eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}.$$

A kovariancia egyszerű tulajdonságai:

6.8. Állítás. *Tetszőleges $\xi, \xi_1, \ldots, \xi_n, \eta, \eta_1, \ldots, \eta_m$ véletlen változók és a, b valós számok esetén igazak az alábbiak.*

(i) $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi)$;

(ii) $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi)$;

(iii) $\mathbf{Cov}(\xi, \eta) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta)$;

(iv) $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta)$;

(v) $\mathbf{Cov}\left(\sum_{i=1}^n \xi_i, \sum_{j=1}^m \eta_j\right) = \sum_{i=1}^n \sum_{j=1}^m \mathbf{Cov}(\xi_i, \eta_j)$;

(vi) ha ξ és η függetlenek, akkor $\mathbf{Cov}(\xi, \eta) = 0$.

Bizonyítás. (i) és (ii) a definíció azonnali következménye.

(iii) A zárójelet felbontva, a várható érték tulajdonságai alapján

$$\begin{aligned}\mathbf{Cov}(\xi, \eta) &= \mathbf{E}\left((\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))\right) \\ &= \mathbf{E}\left(\xi\eta - \xi\mathbf{E}(\eta) - \mathbf{E}(\xi)\eta + \mathbf{E}(\xi)\mathbf{E}(\eta)\right) \\ &= \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta).\end{aligned}$$

(iv) Definíció szerint

$$\begin{aligned}&\mathbf{Cov}(a(\xi+c),b(\xi+d))\\&=\mathbf{E}\left[(a(\xi+c)-\mathbf{E}(a(\xi+c)))(b(\eta+d)-\mathbf{E}(b(\eta+d)))\right]\\&=ab\mathbf{E}((\xi-\mathbf{E}(\xi))(\eta-\mathbf{E}(\eta)))=ab\mathbf{Cov}(\xi,\eta).\end{aligned}$$

(v) A várható érték linearitásából

$$\begin{aligned}\mathbf{Cov}\left(\sum_{i=1}^{n}\xi_i,\sum_{j=1}^{m}\eta_j\right)&=\mathbf{E}\left(\sum_{i=1}^{n}(\xi_i-\mathbf{E}(\xi_i))\sum_{j=1}^{m}(\eta_j-\mathbf{E}(\eta_j))\right)\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{E}((\xi_i-\mathbf{E}(\xi_i))(\eta_j-\mathbf{E}(\eta_j)))\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}\mathbf{Cov}(\xi_i,\eta_j).\end{aligned}$$

(vi) A függetlenségből következik, hogy $\mathbf{E}(\xi\eta)=\mathbf{E}(\xi)\mathbf{E}(\eta)$, így (iii) szerint $\mathbf{Cov}(\xi,\eta)=0$. □

6.9. Állítás. *(i) Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi,\eta)|\le\mathbf{D}(\xi)\mathbf{D}(\eta).$$

Innen adódik, hogy $\rho(\xi,\eta)\in[-1,1]$*;*

(ii) ha $\rho(\xi,\eta)=1$*, akkor*

$$\xi=\mathbf{E}(\xi)+\frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta-\mathbf{E}(\eta));$$

(iii) ha $\rho(\xi,\eta)=-1$*, akkor*

$$\xi=\mathbf{E}(\xi)-\frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta-\mathbf{E}(\eta)).$$

Bizonyítás. (i): Tekintsük az $U+tV$ véletlen változót, ahol t egy valós szám. Mivel $\mathbf{E}[(U+tV)^2]\ge 0$, ezért a

$$p(t)=\mathbf{E}[(U+tV)^2]=t^2\mathbf{E}(V^2)+2t\mathbf{E}(UV)+\mathbf{E}(U^2) \tag{3}$$

t-ben másodfokú polinom diszkriminánsa nempozitív. Azaz

$$4[\mathbf{E}(UV)]^2\le 4\mathbf{E}(U^2)\mathbf{E}(V^2), \tag{4}$$

amiből következik, hogy

$$|\mathbf{E}(UV)| \le \sqrt{\mathbf{E}(U^2)\mathbf{E}(V^2)}.$$

Ezt az egyenlőtlenséget az $U = \xi - \mathbf{E}(\xi)$ és $V = \eta - \mathbf{E}(\eta)$ változókra felírva kapjuk az állítást.

(ii) és (iii): Ha $|\rho(\xi, \eta)| = 1$, akkor a (4) egyenlőtlenség $U = \xi - \mathbf{E}(\xi)$ és $V = \eta - \mathbf{E}(\eta)$ változókra egyenlőség, azaz a másodfokú p polinom diszkriminánsa 0. Ezek szerint

$$t_0 = -\frac{\mathbf{E}[(\xi - \mathbf{E}\xi)(\eta - \mathbf{E}\eta)]}{\mathbf{E}[(\eta - \mathbf{E}\eta)^2]} = -\rho(\xi, \eta)\frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}$$

zérushely, vagyis

$$\xi - \mathbf{E}(\xi) + t_0(\eta - \mathbf{E}(\eta)) = 0,$$

ami éppen a bizonyítandó. □

Megjegyzés. Korreláció jelentése. Ha $\rho(\xi, \eta) = 0$, akkor ξ és η *korrelálatlanok*. A 6.8 Állítás (v) pontja szerint a függetlenségből következik a korrelálatlanság. Fordítva ez nem igaz, könnyű ellenpéldát gyártani. Mindenesetre, minél kisebb a korreláció annál gyengébb a két változó közötti függés. A 6.9 Állításból pedig azt látjuk, hogy minél közelebb van $|\rho(\xi, \eta)|$ értéke 1-hez, annál erősebb a változók közötti függés.

Ha a korreláció pozitív, akkor ha ξ nagy, akkor η is nagy, ha pedig negatív, akkor ha ξ nagy, akkor η kicsi, és fordítva. Ezek a megállapítások persze nem tehetők nagyon precízzé, ez a szemléletes jelentés. A korreláció jelentését a következő fejezetben, a lineáris regressziónál értjük meg jobban.

6.10. Állítás. *Legyenek $\xi_1, \xi_2, \ldots, \xi_n$ páronként független véletlen változók. Ekkor*

$$\mathbf{D}^2\left(\sum_{i=1}^{n} \xi_i\right) = \sum_{i=1}^{n} \mathbf{D}^2(\xi_i).$$

Bizonyítás. Egyszerű számolás. □

6.11. Állítás. *Legyenek ξ, η véletlen változók, melyek korrelációs együtthatója ρ. Ekkor*

$$\mathbf{D}^2(\xi + \eta) = \mathbf{D}^2(\xi) + 2\mathbf{D}(\xi)\mathbf{D}(\eta)\rho + \mathbf{D}^2(\eta).$$

Bizonyítás. A kovariancia tulajdonságai alapján

$$\begin{aligned}\mathbf{D}^2(\xi + \eta) &= \mathbf{Cov}(\xi + \eta, \xi + \eta) = \mathbf{Cov}(\xi, \xi) + 2\mathbf{Cov}(\xi, \eta) + \mathbf{Cov}(\eta, \eta) \\ &= \mathbf{D}^2(\xi) + 2\mathbf{D}(\xi)\mathbf{D}(\eta)\rho + \mathbf{D}^2(\eta).\end{aligned}$$

□

6.12. *Példa* (Trinomiális eloszlás (6.2 Példa folytatása)). Egy szabályos dobókockával n-szer dobunk. Jelölje ξ a hatosok, η egyesek számát!

A kovarianciát egy ügyes trükkel könnyen kiszámolhatjuk. A kovariancia tulajdonságai szerint

$$\mathbf{D}^2(\xi+\eta) = \mathbf{D}^2(\xi) + 2\mathbf{Cov}(\xi,\eta) + \mathbf{D}^2(\eta).$$

(Hát persze, a kovariancia olyan mint egy belső szorzat.) Innen ismerjük ξ és η szórását. Továbbá, $\xi+\eta$ az egyesek és hatosok összege, tehát ő binomiális eloszlást követ n és $2\cdot\frac{1}{6} = 1/3$ paraméterekkel. Így $\mathbf{D}^2(\xi+\eta) = n\frac{1}{3}\cdot\frac{2}{3}$. Rendezve adódik, hogy

$$\mathbf{Cov}(\xi,\eta) = -\frac{n}{36},$$

korrelációjuk pedig

$$\rho(\xi,\eta) = \frac{\mathbf{Cov}(\xi,\eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)} = \frac{-n\frac{1}{36}}{n\frac{5}{36}} = -\frac{1}{5}.$$

Látjuk, hogy a korreláció negatív, azaz ha sok hatost dobunk, akkor kevés egyest, és fordítva, ami teljesen természetes.

6.3. Lineáris regresszió

6.13. *Példa.* Három, külsőre egyforma érmével a fejdobás valószínűsége 1/4, 1/2, és 3/4. Véletlenszerűen választunk egy érmét, és azzal kétszer dobunk. Legyen η a fej valószínűsége a választott érmén, ξ a dobott fejek száma.

Világos, hogy ξ lehetséges értékei 0,1,2, míg η lehetséges értékei 1/4, 1/2, 3/4. A szorzási szabállyal kapjuk

$$\begin{aligned}\mathbf{P}\left(\eta=\frac{1}{4},\xi=0\right) &= \mathbf{P}\left(\eta=\frac{1}{4}\right)\cdot\mathbf{P}\left(\xi=0|\eta=\frac{1}{4}\right)\\ &= \frac{1}{3}\cdot\left(\frac{3}{4}\right)^2 = \frac{9}{48}.\end{aligned}$$

A többi eset hasonlóan számolható, pl.

$$\mathbf{P}\left(\eta=\frac{1}{2},\xi=1\right) = \frac{1}{3}\cdot 2\left(\frac{1}{2}\right)^2 = \frac{1}{6}.$$

Minden valószínűséget kiszámolva az alábbi táblázatot kapjuk, ami megadja a (ξ,η) együttes eloszlását:

$\eta\backslash\xi$	0	1	2	$\sum$
$\frac{1}{4}$	$\frac{9}{48}$	$\frac{6}{48}$	$\frac{1}{48}$	$\frac{1}{3}$
$\frac{1}{2}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
$\frac{3}{4}$	$\frac{1}{48}$	$\frac{6}{48}$	$\frac{9}{48}$	$\frac{1}{3}$
$\sum$	$\frac{14}{48}$	$\frac{20}{48}$	$\frac{14}{48}$	1

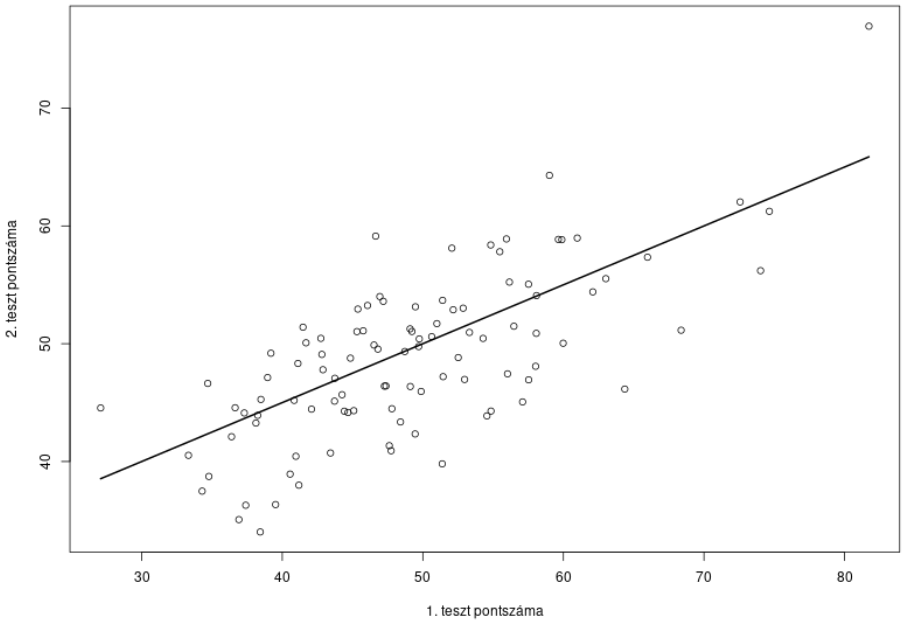
A táblázat utolsó oszlopában megjelenik η eloszlása, míg az utolsó sorban ξ eloszlása, ezek a peremeloszlások.

Ennél a kísérletnél η értékét nem tudom meghatározni, hiszen az érmék külsőre egyformák. Viszont világos, hogy ξ ismeretéből következtethetünk η értékére, hiszen minél nagyobb a fej valószínűsége, annál több fejet dobunk. Ez azt jelenti, hogy egy olyan (determinisztikus) g függvényt keresünk, melyre $g(\xi)$ közel van η-hoz. A közelség mérése függ az adott feladattól. A konkrét példában kereshetünk olyan g függvényt, melyre a $\mathbf{P}(g(\xi) = \eta)$ valószínűség maximális, azaz a lehető legnagyobb valószínűséggel adjuk meg a valódi η értéket. Egy másik lehetséges választás, hogy minimalizáljuk a $(g(\xi) - \eta)^2$ négyzetes hiba várható értékét. A regresszió ezt a problémát fogalmazza meg általános esetben.

6.14. *Példa.* A sztochasztika alapjai kurzus elején a hallgatók kitöltenek egy tesztet, mely az eddigi matematika tudásukat méri. A kurzus teljesítéséhez a félév végén is kitöltenek egy tesztet a kurzus anyagából. A korábbi évek tapasztalatai alapján feltehető, hogy az i-edik hallgató első teszten elért pontszáma ξ_i, a másodikon $\eta_i = \frac{\xi_i+\xi_i'}{2}$, ahol $\xi_1, \xi_2, \dots$ és $\xi_1', \xi_2', \dots$ független normális eloszlású változók $\mu = 50$ várható értékkel és $\sigma = 10$ szórással, $i = 1, 2, \dots, N$, ahol N a hallgatók száma. Ez azt fejezi ki, hogy a jobb alapokkal érkező diák (nagyobb ξ érték) valószínűleg sztochasztikából is sikeresebb lesz, de a függés nem determinisztikus, hiszen lehet, hogy az adott diáknak jobban megy vagy éppen kevésbé jól megy a sztochasztika, mint a többi matematika tárgy. A 19. ábrán $N = 100$ hallgató pontszámát látjuk. Minden pont egy hallgató pontszámaihoz tartozik, a pont első koordinátája az első teszten, a második koordinátája a második teszten elért pontszámot jelenti. Világos, hogy két pontszám között van kapcsolat. Általában magasabb első pontszámhoz, magasabb második pontszám tartozik. Az is látszik, hogy a függés nem determinisztikus.

Arra vagyunk kíváncsiak, hogy az ismert $\xi_1, \dots, \xi_N$ értékekből hogyan tudunk következtetni az $\eta_1, \dots, \eta_N$ értékekre. (Ez fontos lehet, ha előre kell tervezni, hogy hány gyakorlat kell majd a következő féléves *Alkalmazott statisztika* kurzuson.)

A feladatunk az, hogy a ξ értékéből adjunk előrejelzést η értékére. Világos, hogy ezt pontosan nem tudjuk megmondani, de valamit tudunk monda-



19. ábra. 100 hallgató pontszámának alakulása

ni. Ha ismerjük ξ értékét, akkor

$$\mathbf{E}[\eta|\xi] = \mathbf{E}\left[\frac{\xi+\xi'}{2}|\xi\right] = 25 + \frac{\xi}{2},$$

hiszen ξ értékét pontosan tudom, ξ' várható értéke pedig 50, és ξ független ξ'-től. Ez éppen a regressziós egyenes, amit behúztunk a 19. ábrán.

Számítsuk ki a két változó, ξ és η korrelációs együtthatóját! A kovariancia tulajdonságai szerint

$$\mathbf{Cov}(\xi,\eta) = \mathbf{Cov}\left(\xi, \frac{\xi+\xi'}{2}\right) = \frac{1}{2}\mathbf{Cov}(\xi,\xi) + \frac{1}{2}\mathbf{Cov}(\xi,\xi') = \frac{1}{2}\mathbf{D}^2(\xi).$$

A szórások $\mathbf{D}^2(\xi) = 100$, és a függetlenség miatt

$$\mathbf{D}^2(\eta) = \mathbf{D}^2\left(\frac{\xi+\xi'}{2}\right) = \frac{1}{4}\left(\mathbf{D}^2(\xi) + \mathbf{D}^2(\xi')\right) = 50,$$

ahonnan a korrelációs együttható

$$\rho(\xi,\eta) = \frac{\mathbf{Cov}(\xi,\eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)} = \frac{1}{2}\frac{\mathbf{D}^2(\xi)}{\mathbf{D}(\xi)\mathbf{D}(\eta)} = \frac{1}{\sqrt{2}}.$$

Tekintsük a (ξ,η) véletlen vektorváltozót. Az η változót tekintem *függő* változónak, ennek az értékére szeretnék következtetni a ξ *független változó* értékéből. Vagyis ismert ξ esetén szeretném megmondani η-t. Egy olyan determinisztikus g függvényt keresek, melyre az $\eta - g(\xi)$ véletlen változó kicsi valamilyen értelemben. Mivel ez így nehéz,[7] csak a lineáris függvények között keresünk. Tehát keressük azokat az a, b valós számokat, melyre a $\eta - (a\xi + b)$ változó kicsi. A kicsiséget négyzetes hibában mérve, keressük az

$$h(a,b) = \mathbf{E}\left[(\eta - (a\xi+b))^2\right]$$

függvény minimumhelyét, azaz a legjobb a, b választást. Mivel

$$\begin{aligned}
&\mathbf{E}\left[(\eta - (a\xi+b))^2\right] \\
&= \mathbf{E}\left[((\eta - a\xi) - \mathbf{E}(\eta - a\xi) + \mathbf{E}(\eta - a\xi) - b)^2\right] \\
&= \mathbf{E}\left[((\eta - a\xi) - \mathbf{E}(\eta - a\xi))^2\right] + \left[\mathbf{E}(\eta - a\xi) - b\right]^2 \\
&\quad + 2\mathbf{E}\left[((\eta - a\xi) - \mathbf{E}(\eta - a\xi))\right](\mathbf{E}(\eta - a\xi) - b) \\
&= \mathbf{E}\left[((\eta - a\xi) - \mathbf{E}(\eta - a\xi))^2\right] + \left[\mathbf{E}(\eta - a\xi) - b\right]^2
\end{aligned}$$

[7] A legjobb g az η változó ξ-re vett feltételes várható értéke lesz, amit konkrétan nehéz kiszámolni.

látjuk, hogy $b = \mathbf{E}(\eta) - a\mathbf{E}(\xi)$ választás adja b-ben a minimumhelyet. Tehát a $\mathbf{D}^2(\eta - a\xi)$ mennyiség minimuma kell a-ban. A szórásnégyzetre vonatkozó formulák szerint

$$\begin{aligned}\mathbf{D}^2(\eta - a\xi) &= \mathbf{Cov}(\eta - a\xi, \eta - a\xi) \\ &= \mathbf{D}^2(\eta) + a^2\mathbf{D}^2(\xi) - 2a\mathbf{Cov}(\eta, \xi).\end{aligned}$$

Ez a-ban másodfokú polinom, főegyütthatója pozitív, azaz egy felfelénéző parabola. Ennek minimumhelye

$$a = \frac{2\mathbf{Cov}(\eta, \xi)}{2\mathbf{D}^2(\xi)} = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}.$$

Ezek szerint a legjobb lineáris közelítést a

$$g(x) = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}(x - \mathbf{E}(\xi)) + \mathbf{E}(\eta)$$

függvény adja. Ő a *regressziós egyenes.*

Vegyük észre, hogy ha ξ és η korrelálatlanok, azaz $\mathbf{Cov}(\xi, \eta) = 0$, akkor a legjobb közelítés $\mathbf{E}(\eta)$, vagyis ξ semmi információt nem ad η értékéről.

7. Véletlen változók konvergenciája

7.1. Markov és Csebisev egyenlőtlenségei

7.1. Tétel (Markov-egyenlőtlenség)**.** *Legyen ξ egy véletlen változó $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőn, melynek véges a várható értéke. Ekkor tetszőleges pozitív c konstansra*

$$\mathbf{P}(|\xi| \geq c) \leq \frac{\mathbf{E}(|\xi|)}{c}.$$

Bizonyítás. Vegyük észre, hogy tetszőleges $c > 0$ esetén $c\mathbb{I}(|X| \geq c) \leq |X|$, ahonnan

$$c\,\mathbf{P}(|X| \geq c) \leq \mathbf{E}(|X|),$$

ami éppen az állítás. □

A Markov-egyenlőtlenség egyszerű alkalmazásával adódik a

7.2. Tétel (Csebisev-egyenlőtlenség)**.** *Legyen ξ egy véletlen változó $(\Omega, \mathcal{A}, \mathbf{P})$ valószínűségi mezőn, melynek véges a szórása. Ekkor tetszőleges pozitív c konstansra*

$$\mathbf{P}(|\xi - \mathbf{E}(\xi)| \geq c) \leq \frac{\mathbf{D}^2(\xi)}{c^2}.$$

Bizonyítás. A Markov-egyenlőtlenség szerint

$$\mathbf{P}(|\xi - \mathbf{E}(\xi)| \geq c) = \mathbf{P}((\xi - \mathbf{E}(\xi))^2 \geq c^2) \leq \frac{\mathbf{D}^2(\xi)}{c^2}.$$

□

7.2. Nagy számok gyenge törvénye

7.3. Tétel (Csebisev-féle nagy számok gyenge törvénye)**.** *Legyenek $\xi_1, \xi_2, \ldots$ páronként független, véges szórású véletlen változók, melyek közös várható értéke μ és szórásnégyzete σ^2. Ekkor tetszőleges $\varepsilon > 0$ esetén*

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{\xi_1 + \ldots + \xi_n}{n} - \mu\right| > \varepsilon\right) = 0.$$

Bizonyítás. A páronkénti függetlenség miatt

$$\mathbf{D}^2(\xi_1 + \ldots + \xi_n) = n\sigma^2.$$

A Csebisev-egyenlőtlenséget az $\xi = \xi_1 + \ldots + \xi_n$ változóra fölírva kapjuk, hogy

$$\mathbf{P}\left(\left|\frac{\xi_1 + \ldots + \xi_n}{n} - \mu\right| > \varepsilon\right) \leq \frac{\mathbf{D}^2(\xi_1 + \ldots + \xi_n)}{n^2\varepsilon^2} \leq \frac{\sigma^2}{n\varepsilon^2},$$

ami tart 0-hoz. □

A bizonyításból látjuk, hogy a páronkénti függetlenség helyett elég korrelálatlanságot feltenni.

Speciális esetként adódik a

7.4. Tétel (Bernoulli-féle nagy számok gyenge törvénye (1713))**.** *Jelölje S_n egy p valószínűségű A esemény bekövetkezéseinek a számát egy kísérlet n független ismétlése során. Ekkor tetszőleges $\varepsilon > 0$ esetén*

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\frac{S_n}{n} - p\right| > \varepsilon\right) = 0.$$

A tétel szerint a relatív gyakoriságok a fenti értelemben konvergálnak az igazi valószínűséghez. Mivel a valószínűség definícióját a relatív gyakoriságok tulajdonságai motiválták (additivitás), ezért a fenti tétel szerint a valószínűség tényleg az, amit akarunk.

7.3. Centrális határeloszlás-tétel

A nagy számok törvénye azt állítja, hogy független, azonos eloszlású véletlen változók átlagai közel vannak a várható értékhez. Az alábbiakban ezt a közelséget tesszük precízzé.

7.5. Tétel (Centrális határeloszlás-tétel)**.** *Legyenek $\xi, \xi_1, \xi_2, \dots$ független, azonos eloszlású véletlen változók közös $\mathbf{E}(\xi) = \mu$ várható értékkel, és véges $\mathbf{D}(\xi) = \sigma$ szórással. Ekkor tetszőleges $x \in \mathbb{R}$ esetén*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\sum_{i=1}^n (\xi_i - \mu)}{\sqrt{n}\sigma} < x\right) = \Phi(x),$$

ahol

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} \mathrm{d}y$$

a standard normális eloszlás eloszlásfüggvénye.

A tételt nem bizonyítjuk. A bizonyítás már komolyabb eszközökkel, a karakterisztikus függvények módszerével történik. Annyit megjegyzünk, hogy a tételben szereplő standardizált összeg várható értéke 0, míg szórása 1. Valóban

$$\mathbf{E}\left(\sum_{i=1}^n (\xi_i - \mu)\right) = \sum_{i=1}^n (\mathbf{E}(\xi_i) - \mu) = 0,$$

és a függetlenség miatt

$$\mathbf{D}^2(\sum_{i=1}^n (\xi_i - \mu)) = \sum_{i=1}^n \mathbf{D}^2(\xi_i) = n\sigma^2,$$

ezért

$$\mathbf{D}^2\left(\frac{\sum_{i=1}^n (\xi_i - \mu)}{\sigma\sqrt{n}}\right) = \frac{n\sigma^2}{n\sigma^2} = 1.$$

A CHT indikátorváltozókra vonatkozó speciális esete a de Moivre–Laplace-tétel.

7.6. Tétel (de Moivre–Laplace tétel)**.** *Jelölje S_n egy p valószínűségű A esemény bekövetkezéseinek a számát egy kísérlet n független ismétlése során. Ekkor tetszőleges $x \in \mathbb{R}$ esetén*

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

Valóban, korábban láttuk, hogy a p-paraméterű Bernoulli-eloszlás várható értéke p és szórása $\sqrt{p(1-p)}$. A speciális eset bizonyítása a binomiális együtthatók pontos aszimptotikájának meghatározásával történhet.

7.7. *Példa.* A kakucsretyegei polgármesterválasztáson két jelölt van: A és B. Kakucsretyege 40000 szavazója egymástól függetlenül, $1/2-1/2$ valószínűséggel szavaz a két jelölt egyikére. A feszült politikai helyzet miatt a szavazatok újraszámlálását rendelik el, ha a két jelöltre leadott szavazatok száma között legfeljebb 100 a különbség. Mi a valószínűsége, hogy újraszámlálásra kerül sor?

Ekkor tehát $n = 40000$, $p = \mathbf{P}(\text{A-ra szavaz valaki}) = 1/2$. Legyen S_n az A-ra szavazók száma, ekkor $n - S_n$ a B-re szavazók száma. A kérdés $\mathbf{P}(|S_n - (n - S_n)| \le 100)$. A CHT-ban előforduló mennyiségek $np = 20000$ és $\sqrt{np(1-p)} = 100$. Így a CHT szerint

$$\begin{aligned}\mathbf{P}(|S_n - (n - S_n)| \le 100) &= \mathbf{P}(-100 \le 2S_n - n \le 100)\\ &= \mathbf{P}\left(-0,5 \le \frac{S_n - np}{\sqrt{npq}} \le 0,5\right)\\ &\approx \Phi(0,5) - \Phi(-0,5)\\ &= 2\Phi(0,5) - 1 \approx 0,38.\end{aligned}$$

7.8. *Példa.* Budapesten meg akarják állapítani a dohányosok p arányát. Ehhez kiválasztanak n egyént úgy, hogy minden választásnál mindenki ugyanakkora valószínűséggel kerül kiválasztásra, és csak ezek közt nézik meg a dohányosok S számát. Legalább mekkora legyen az n, hogy a kapott $p' = S/n$ arány legalább 0,95 valószínűséggel legfeljebb 0,005 hibával közelítse a valódi p arányt, akármi is $p \in (0,1)$?

Ez már érdekesebb feladat, ugyanis semmi nincs megadva. Világos, hogy a feladat nagyon fontos, ugyanis a közvéleménykutatásokhoz pontosan ilyen típusú kérdést kell feltenni. Van valami ismeretlen valószínűség p, ami azt mutatja meg, hogy az emberek ilyen aránya szavazna az A párt jelöltjére. Nem tudjuk mi a p, de erről szeretnénk valamit mondani. Hány embert kell megkérdezni, hogy valami okosat mondhassunk?

Jelölje S a dohányzók számát a megkérdezettek között. Ekkor S binomiális eloszlású véletlen változó n (meghatározandó, de ismert) és p (ismeretlen) paraméterekkel. Világos, hogy az ismeretlen p értékre az S/n becslést adjuk. Elég összetett a kérdés, kicsit el kell rajta gondolkodni. A becslés hibája $|S/n - p|$. Azt akarjuk, hogy ez nagy valószínűséggel (0,95) kicsi legyen (0,005-nél kisebb), azaz olyan n értéket keresünk, amire

$$\mathbf{P}\left(\left|\frac{S}{n} - p\right| < 0,005\right) \ge 0,95.$$

(Annak a valószínűsége, hogy a hiba 0,005-nél kisebb, legalább 0,95.) A de Moivre–Laplace-tétel szerint

$$\mathbf{P}\left(a \le \frac{S-np}{\sqrt{np(1-p)}} \le b\right) \approx \Phi(b) - \Phi(a).$$

Így

$$\begin{aligned}
\mathbf{P}\left(\left|\frac{S}{n} - p\right| < 0{,}005\right) &= \mathbf{P}\left(\left|\frac{S-np}{n}\right| < 0{,}005\right) \\
&= \mathbf{P}\left(\left|\frac{S-np}{\sqrt{np(1-p)}}\right| < 0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) \\
&\approx \Phi\left(0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) - \Phi\left(-0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) \\
&= 2\Phi\left(0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1.
\end{aligned}$$

Annyit használtunk, hogy $|x| \le a$ pontosan akkor, ha $-a < x < a$, és hogy $\Phi(-x) = 1 - \Phi(x)$. Ezek szerint az kell, hogy

$$2\Phi\left(0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) - 1 \ge 0{,}95,$$

azaz

$$\Phi\left(0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) \ge 0{,}975.$$

A táblázatból kikeresve azt kapjuk, hogy

$$0{,}005\frac{\sqrt{n}}{\sqrt{p(1-p)}} \ge 1{,}96.$$

Átrendezve

$$n \ge 392^2 \cdot p(1-p). \tag{5}$$

Nem ismerjük p értékét. Úgy kell n-et választani, hogy a fenti egyenlőtlenség minden p-re igaz legyen. Tehát válasszuk p-t úgy, hogy a jobb oldal maximális legyen. Ez $p = 1/2$-nél van, értéke $1/4$. Tehát, ha

$$n \ge 396^2\frac{1}{4} = 38416,$$

akkor (5) teljesül minden $p \in [0,1]$ esetén.

7.9. *Példa.* A sztochasztika alapjai kurzus 700 hallgatójának mindegyike bemegy az első előadásra. Ezt követően minden hallgató minden további előadás előtt feldob egy szabályos pénzérmét. Ha fejet kap, akkor bemegy a következő előadásra, ha írást akkor nem, és utána már egyetlen előadásra sem megy be.

Véletlen Vince a 700 hallgató egyike. Mennyi a valószínűsége, hogy Vince az összes előadásra bemegy? Várhatóan hány előadáson vesz részt? Mennyi a valószínűsége, hogy a 13. (utolsó) előadáson lesz hallgató? Várhatóan hány hallgató lesz a 2., 3., utolsó előadáson? Mennyi a valószínűsége, hogy a negyedik előadáson 100-nál több hallgató vesz részt. És annak, hogy a tizenegyediken pontosan 2 hallgató vesz részt?

Vince egy szabályos érmét dobál legfeljebb 12-szer. Jelölje ξ annak a dobásnak a sorszámát, amikor először írást dob, ha csupa fejet, akkor legyen $\xi = 13$. Ekkor Vince pontosan ξ db előadáson vesz részt. (Ha pl. $\xi = 1$, akkor elsőre fejet dob, így a második előadásra már nem megy be.) Ha $k \leq 12$, akkor a $\xi = k$ esemény pontosan azt jelenti, hogy az első $k-1$ dobás fej, a k-adik írás, így

$$\mathbf{P}(\xi = k) = \frac{1}{2^k}, \quad k = 1, 2, \ldots, 12.$$

Ha $\xi = 13$, azaz minden előadásra bemegy, akkor mind a 12 dobás fej. Ennek a valószínűsége

$$\mathbf{P}(\text{Vince az összes előadásra bemegy}) = \mathbf{P}(\xi = 13) = \frac{1}{2^{12}} \approx 0.00024.$$

A látogatott előadások számának várható értéke $\mathbf{E}(\xi)$, ami definíció szerint

$$\mathbf{E}(\xi) = \sum_{k=1}^{12} \frac{k}{2^k} + \frac{13}{2^{12}} = 2 - \frac{1}{2^{12}} \approx 2.$$

Jelölje η_2 a 2. előadáson levő hallgatók számát, η_3 a 3. előadáson levő hallgatók számát, ... A 2. előadáson azok vesznek részt, akik elsőre fejet dobtak, tehát η_2 binomiális eloszlású $n = 700$ és $p_2 = 1/2$ paraméterekkel. Így

$$\mathbf{E}(\eta_2) = 700 \cdot \frac{1}{2} = 350.$$

Hasonlóan, a 3. előadáson résztvevők száma binomiális $n = 700$ és $p_3 = 1/4$ paraméterekkel, hiszen itt azok vannak, akik kétszer fejet dobtak. Így

$$\mathbf{E}(\eta_3) = 700 \cdot \frac{1}{4} = 175.$$

Míg az utolsó előadáson résztvevők száma binomiális $n = 700$, $p_{13} = 1/2^{12}$ paraméterrel, így

$$\mathbf{E}(\eta_{13}) = 700 \cdot \frac{1}{2^{12}} = 0.17.$$

Tehát az utolsó előadáson várhatóan 0.17 hallgató lesz.

A 4. előadáson résztvevők száma binomiális $n = 700$ és $p = 1/8$ paraméterekkel. Ekkor az n nagy és a p nem olyan kicsi, így normális közelítést használunk. A de Moivre–Laplace-tétel szerint

$$\begin{aligned}\mathbf{P}(\eta_4 > 100) &= \mathbf{P}\left(\frac{\eta_4 - 700 \cdot \frac{1}{8}}{\sqrt{700 \cdot \frac{1}{8} \cdot \frac{7}{8}}} > \frac{100 - 700 \cdot \frac{1}{8}}{\sqrt{700 \cdot \frac{1}{8} \cdot \frac{7}{8}}}\right) \\ &\approx \mathbf{P}\left(Z > \frac{10}{7}\right) \approx 1 - \Phi(1.43) \approx 0.077.\end{aligned}$$

A tizenegyedik előadáson résztvevők száma η_{11} binomiális eloszlású $n = 700$ és $p = 2^{-10}$ paraméterekkel. Ekkor az n nagy, viszont p nagyon kicsi, és $np = 700/1024 \approx 0.68$. Így nem normális, hanem Poisson-közelítést használunk. (A Poisson eloszlást éppen a binomiális határeloszlásaként definiáltuk.) Tehát az η_{11} változót egy $\lambda = \frac{700}{1024}$ paraméterű Poisson-eloszlású változóval közelíthetjük. Így

$$\mathbf{P}(\eta_{11} = 2) \approx \frac{\lambda^2}{2} e^{-\lambda} \approx 0.117946.$$

A pontos valószínűség, a binomiális eloszlás szerint

$$\mathbf{P}(\eta_{11} = 2) = \binom{700}{2} \left(2^{-10}\right)^2 \cdot \left(1 - 2^{-10}\right)^{698} \approx 0.117969.$$

8. Statisztikai alapfogalmak

Az $(\Omega, \mathcal{A}, \mathcal{P})$ hármast *statisztikai mezőnek nevezzük*, ha Ω tetszőleges halmaz, az eseménytér, $\mathcal{A}$ az események halmaza, $\mathcal{P}$ pedig valószínűségi mértékek halmaza.

A statisztikai problémák során a megfelelő valószínűségi mérték kiválasztása, illetve ennek valamilyen tulajdonságának meghatározása a feladat.

Független, azonos eloszlású véletlen változók egy $\xi, \xi_1, \xi_2, \ldots$ sorozatát *statisztikai mintának* nevezzük. Mivel statisztikai problémák során a közös eloszlást általában nem ismerjük, ezért a közös eloszlást *háttéreloszlásnak* is nevezik. Ha a minta véges, akkor n-elemű mintáról beszélünk. A minta egy adott realizációját $x_1, \ldots, x_n$ jelöli. A minta egy T függvényét *statisztikának* nevezzük.

8.1. Alapstatisztikák

Legyen $\xi_1, \xi_2, \ldots, \xi_n$ egy n-elemű minta. A várható érték, szórásnégyzet és kovariancia empirikus megfelelői az alábbiak. Ezek a megfelelő elméleti mennyiségek természetes becslései.

8.1. Definíció. Az

$$\overline{\xi}_n = \frac{1}{n}\sum_{i=1}^{n} \xi_i$$

a *mintaátlag.* Az *empirikus szórásnégyzet*

$$V_n(\xi) = V_n = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2.$$

Az alábbi egyszerű állítás megkönnyíti az empirikus szórásnégyzet számítását.

8.2. Tétel (Steiner-tétel)**.** *Tetszőleges $x_1, \ldots, x_n$ értékekre és c valós számra*

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + (\overline{x}_n - c)^2.$$

Bizonyítás. Egyszerű számolás. □

Ezek alapján a $c = 0$ választással

$$V_n = \frac{1}{n}\sum_{i=1}^{n}\xi_i^2 - (\overline{\xi}_n)^2.$$

Az $(\xi_1, \eta_1), \ldots, (\xi_n, \eta_n)$ minta *empirikus kovarianciája*

$$C_n(\xi, \eta) = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)(\eta_i - \overline{\eta}_n) = \frac{1}{n}\sum_{i=1}^{n}\xi_i\eta_i - \overline{\xi}_n\overline{\eta}_n.$$

Jelölés: A konkrét realizációból számolt értékeket a megfelelő kisbetűkkel jelöljük,[8] így konkrét $x_1, \ldots, x_n$ realizációhoz tartozó mintaátlag és empirikus szórásnégyzet

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad v_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_i)^2.$$

[8] Kivétel az empirikus eloszlásfüggvény.

8.3. Definíció. Az $\xi_1, \ldots, \xi_n$ minta *empirikus eloszlásfüggvénye*

$$F_n(x) = \frac{1}{n} \left| \{ i : \xi_i < x \} \right| .$$

Az empirikus eloszlásfüggvény indikátorváltozókkal is definiálható:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\xi_i < x).$$

A függetlenség miatt $\mathbb{I}(\xi_1 < x), \ldots, \mathbb{I}(\xi_n < x)$ független Bernoulli-eloszlású véletlen változók $p = F(x)$ paraméterrel, ahol F a közös elméleti eloszlásfüggvény. Innen következik, hogy $nF_n(x) \sim \mathrm{Binom}(n, F(x))$. Ezzel beláttuk az alábbit.

8.4. Állítás. *Legyen $\xi_1, \xi_2, \ldots$ az F háttéreloszlásból származó minta, és tekintsük ennek az F_n empirikus eloszlásfüggvényét. Ekkor*

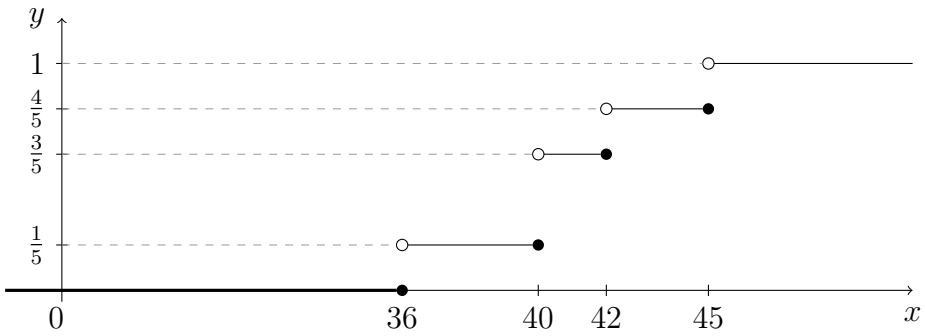
$$\mathbf{E}(F_n(x)) = F(x), \ \mathbf{D}^2(F_n(x)) = \frac{F(x)(1 - F(x))}{n}.$$

Továbbá, tetszőleges $x \in \mathbb{R}$ esetén, tetszőleges $\varepsilon > 0$ számra

$$\lim_{n \to \infty} \mathbf{P}(|F_n(x) - F(x)| > \varepsilon) = 0.$$

Az utolsó állítás a nagy számok Csebisev-féle gyenge törvénye.

8.5. *Példa.* Az $x_1 = 40$, $x_2 = 45$, $x_3 = 40$, $x_4 = 42$, $x_5 = 36$ minta empirikus eloszlásfüggvénye:



8.2. Torzítatlanság és konzisztencia

A továbbiakban $(\Omega, \mathcal{A}, \mathcal{P})$ egy statisztikai mező, ahol $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$. A feladat az ismeretlen θ paraméter, vagy annak valamely függvényének becslése egy minta alapján.

Bevezetjük a $\boldsymbol{\xi} = \boldsymbol{\xi}_n = (\xi_1, \ldots, \xi_n)$ jelölést.

8.6. Definíció. A $T(\boldsymbol{\xi}_n)$ statisztika $\psi(\theta)$ *torzítatlan* becslése, ha

$$\mathbf{E}_\theta(T(\boldsymbol{\xi}_n)) = \psi(\theta) \quad \text{minden } \theta \in \Theta \text{ esetén.}$$

A $T(\boldsymbol{\xi})$ statisztika $\psi(\theta)$ *aszimptotikusan torzítatlan* becslése, ha

$$\lim_{n\to\infty} \mathbf{E}_\theta(T(\boldsymbol{\xi}_n)) = \psi(\theta) \quad \text{minden } \theta \in \Theta \text{ esetén.}$$

A mintaátlag definíciójából és a várható érték linearitásából azonnal adódik a következő.

8.7. Állítás. *A mintaátlag torzítatlan becslése a várható értéknek, feltéve hogy az létezik.*

Vegyük észre, hogy a minta tetszőleges konvex kombinációja torzítatlan becslése lesz a várható értéknek. Így speciálisan egyetlen mintaelem ξ_1 is torzítatlan becslés. Tehát a torzítatlanság önmagában nem sokat mond.

8.8. Állítás. *Legyen $\xi, \xi_1, \xi_2, \ldots$ független, azonos eloszlású minta, egy olyan eloszlásból, melyre minden $\theta \in \Theta$ esetén $\mathbf{D}^2_\theta(\xi) < \infty$. Ekkor az empirikus szórásnégyzet nem torzítatlan, csak aszimptotikusan torzítatlan becslése a szórásnégyzetnek. A korrigált empirikus szórásnégyzet*

$$V_n^*(\xi) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2 = \frac{n}{n-1}V_n(\xi)$$

torzítatlan becslése a szórásnégyzetnek.

Bizonyítás. A Steiner-tétel és a szórás definíciója alapján

$$\begin{aligned}\mathbf{E}_\theta(V_n) &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_\theta(\xi_i^2) - \mathbf{E}_\theta\left((\overline{\xi}_n)^2\right) \\ &= \frac{n-1}{n}\left(\mathbf{E}_\theta(\xi^2) - (\mathbf{E}_\theta(\xi))^2\right) = \frac{n-1}{n}\mathbf{D}^2_\theta(\xi),\end{aligned}$$

amiből mindkét állítás adódik. □

Hasonló számolással adódik, hogy az empirikus kovariancia várható értéke

$$\mathbf{E}(C_n(\xi,\eta)) = \frac{n-1}{n}\mathbf{Cov}(\xi,\eta),$$

azaz az empirikus kovariancia nem torzítatlan, csak aszimptotikusan torzítatlan becslése a kovarianciának.

8.9. Definíció. A $T(\boldsymbol{\xi}_n)$ statisztika *gyengén konzisztens becslése* $\psi(\theta)$-nak, ha tetszőleges $\theta \in \Theta$ és $\varepsilon > 0$ esetén

$$\lim_{n\to\infty} \mathbf{P}_\theta \left(|T(\boldsymbol{\xi}_n) - \psi(\theta)| > \varepsilon\right) = 0.$$

A nagy számok Csebisev-féle gyenge törvénye éppen a mintaátlag gyenge konzisztenciáját jelenti.

8.10. Állítás. *Ha a háttéreloszlás szórásnégyzete létezik, akkor a mintaátlag gyengén konzisztens becslése a várható értéknek.*

Erősebb momentumfeltételek mellett némi számolással igazolható, hogy az empirikus szórásnégyzet és empirikus kovariancia konzisztenciája.

8.11. Tétel. *Ha* $\mathbf{E}_\theta(\xi^4) < \infty$, *minden* $\theta \in \Theta$ *esetén, akkor mind az empirikus szórásnégyzet, mind a korrigált empirikus szórásnégyzet gyengén konzisztens becslése a szórásnégyzetnek.*

8.12. Tétel. *Legyen* $(\xi, \eta), (\xi_1, \eta_1), (\xi_2, \eta_2), \ldots$ *olyan statisztikai minta, melyre* $\mathbf{E}_\theta(\xi^4) < \infty$, $\mathbf{E}_\theta(\eta^4) < \infty$, *minden* $\theta \in \Theta$ *esetén. Ekkor a*

$$C_n = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)(\eta_i - \overline{\eta}_n)$$

empirikus kovariancia gyengén konzisztens becslése a kovarianciának.

8.3. Maximum likelihood módszer

A maximum likelihood módszer, vagy a legnagyobb valószínűség elve, statisztikában nagyon gyakran használt paraméterbecslési eljárás. Tekintsünk egy $(\Omega, \mathcal{A}, \mathcal{P})$ statisztikai mezőt, ahol $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$, ahol $\Theta \subset \mathbb{R}^k$ (általában egy-, néha kétdimenziós). Egy adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén azt a θ paramétert fogadjuk el, mely mellett a legnagyobb a valószínűsége az adott realizációnak.

A precíz általános definíció előtt nézzünk egy példát. Tegyük föl, hogy egy Bernoulli-eloszlásból veszünk mintát, ahol a paraméter $\theta \in [0, 1]$ ismeretlen, ezt akarjuk becsülni. Tehát $\xi_1, \ldots, \xi_n$ független Bernoulli(θ)-eloszlású véletlen változók. A függetlenség miatt

$$\mathbf{P}_\theta \left((\xi_1, \ldots, \xi_n) = (x_1, \ldots, x_n)\right) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

Innen látjuk, hogy csak az számít a mintából, hogy hányszor következett be a vizsgált esemény. Tehát az $(x_1, \ldots, x_n)$ realizációhoz tartozó valószínűség, ha θ a valódi paraméter

$$L_\theta(x_1, \ldots, x_n) = \theta^{s_n}(1-\theta)^{n-s_n},$$

ahol $s_n = \sum_{i=1}^n x_i$. Azt a θ értéket gondoljuk az igazi paraméternek, mely esetén az adott kimenetel a legvalószínűbb. Azaz a θ paraméter becslésére azt a $\widehat{\theta}$ értéket választjuk, melyre

$$L_{\hat{\theta}}(\mathbf{x}) = \sup_{\theta\in[0,1]} L_\theta(\mathbf{x}).$$

Rövidebben

$$\widehat{\theta} = \operatorname{argmax}\{\sup_{\theta\in[0,1]} L_\theta(\mathbf{x})\}.$$

Adott s, n értékek mellett keressük a

$$L_\theta = \theta^s(1-\theta)^{n-s}$$

függvény maximumát a $\theta \in [0,1]$ intervallumon. Lederiválva

$$\frac{\mathrm{d}}{\mathrm{d}\theta} L_\theta = \theta^{s-1}(1-\theta)^{n-s-1}(s-n\theta).$$

Látjuk, hogy a derivált pozitív a $[0, s/n)$ intervallumon, s/n helyen 0, és negatív az $(s/n, 1]$ intervallumon. Ezek szerint a

$$\widehat{\theta}(\mathbf{x}) = \frac{s_n}{n}$$

becslést kapjuk, ami éppen a relatív gyakoriság, vagy empirikus várható érték.

8.13. Definíció. Amennyiben $\mathbf{P}_\theta$ melletti háttéreloszlás diszkrét minden $\theta \in \Theta$ paraméterre, akkor a *likelihood-függvény*

$$L_\theta(\mathbf{x}) = L_\theta(x_1, \ldots, x_n) = \mathbf{P}_\theta(\xi = \mathbf{x}) = \prod_{j=1}^n \mathbf{P}_\theta(\xi = x_j).$$

Ha $\mathbf{P}_\theta$ melletti háttéreloszlás folytonos minden $\theta \in \Theta$ esetén és sűrűségfüggvénye f_θ, akkor a *likelihood-függvény*

$$L_\theta(\mathbf{x}) = L_\theta(x_1, \ldots, x_n) = \prod_{j=1}^n f_\theta(x_i).$$

A θ paraméter *maximum likelihood becslése* az $\mathbf{x}$ minta alapján

$$\widehat{\theta}(\mathbf{x}) = \operatorname{argmax}_{\theta\in\Theta} L_\theta(\mathbf{x}).$$

A szorzatalak miatt általában kényelmesebb a likelihood függvény logaritmusával számolni, amit log-likelihood függvénynek nevezünk, azaz

$$\ell_\theta(\mathbf{x}) = \log L_\theta(\mathbf{x}).$$

Mivel a logaritmus függvény szigorúan monoton nő, ezért L és ℓ maximumhelye megegyezik.

Bizonyos általános feltételek mellett a $\widehat{\theta}$ maximum likelihood becslés konzisztens, aszimptotikusan torzítatlan.

8.14. *Példa* (Hipergeometrikus eloszlásból vett minta ML becslése). Egy területen meg akarjuk becsülni az ott élő madarak ismeretlen N számát. Meggyűrűzünk M madarat, majd befogunk $n \leq M$ madarat, melyek közül m van meggyűrűzve. Tegyük föl, hogy $m \geq 1$, különben fogjunk még madarat. Megadjuk N ML becslését.

A likelihood függvény

$$L_N(m) = \mathbf{P}_N(\xi = m) = \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}},$$

ahol ξ jelöli a másodszorra befogottak közül meggyűrűzöttek számát. A feladat meghatározni, hogy ez milyen N esetén lesz maximális. Ez az N lesz a ML becslés. Egyszerű számolással

$$\frac{L_{N+1}(m)}{L_N(m)} = \frac{(N+1-m)(N+1-n)}{(N+1)(N+1-M-n+m)} > 1.$$

pontosan akkor teljesül, ha

$$N < \frac{nM}{m} - 1.$$

Ezek szerint az $(L_N(m))_{N \geq M}$ sorozat monoton nő $[nM/m]$-ig, ahol $[\cdot]$ az egészrészfüggvény. Ezek szerint N ML becslése

$$\widehat{N} = \left[\frac{nM}{m}\right].$$

Ez persze logikus is.

8.15. *Példa* (Poisson-eloszlás paraméterének ML becslése). Legyenek $\xi_1, \ldots, \xi_n$ független, Poisson(λ) eloszlású véletlen változók. Adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén a log-likelihood függvény

$$\ell_\lambda(\mathbf{x}) = \log \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \sum_{i=1}^{n} x_i \log \lambda - \sum_{i=1}^{n} \log x_i! - \lambda n.$$

A maximumhelyet deriválással határozhatjuk meg,

$$\frac{\mathrm{d}\ell_\lambda(\mathbf{x})}{\mathrm{d}\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n,$$

amiből a

$$\frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

likelihood egyenlet adódik. Az $\overline{x}$ zérushely valóban maximum, tehát

$$\widehat{\lambda} = \overline{x}_n.$$

Azaz az ML becslés éppen a mintaátlag. Láttuk, hogy ez torzítatlan, gyengén konzisztens becslés.

8.16. *Példa* (Exponenciális eloszlás). Legyenek $\xi_1, \ldots, \xi_n$ független, $\mathrm{Exp}(\lambda)$ eloszlású véletlen változók. Adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén a log-likelihood függvény

$$\ell_\lambda(\mathbf{x}) = \log \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Innen a

$$\frac{n}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

likelihood egyenlet adódik, aminek megoldása

$$\widehat{\lambda} = \frac{1}{\overline{x}_n}.$$

Ez valóban maximumhely.

8.17. *Példa* (Normális eloszlás). Legyenek $\xi_1, \ldots, \xi_n$ független, $\mathrm{N}(\mu, \sigma^2)$ eloszlású véletlen változók. Adott $\mathbf{x} = (x_1, \ldots, x_n)$ realizáció esetén a log-likelihood függvény

$$\begin{aligned}\ell_{(\mu,\sigma^2)}(\mathbf{x}) &= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= -\frac{n}{2}(\log(2\pi) + \log \sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.\end{aligned}$$

Vegyük észre, hogy tetszőleges σ^2 esetén log-likelihood függvény

$$\widehat{\mu} = \overline{x}_n$$

helyen lesz maximális. Azaz a maximumhely μ-ben nem függ σ-tól. Ezt visszahelyettesítve, a

$$-\frac{n}{2}(\log(2\pi)+\log\sigma^2)-\frac{n}{2\sigma^2}v_n$$

függvény maximumát keressük σ^2 függvényeként. Lederiválva

$$\frac{\mathrm{d}\ell_{(\overline{x}_n,\sigma^2)}}{\mathrm{d}\sigma^2}=-\frac{n}{2}\left(\frac{1}{\sigma^2}-\frac{v_n}{(\sigma^2)^2}\right).$$

Innen a

$$\widehat{\sigma}^2=v_n$$

becslést kapjuk. Könnyen látható, hogy ez valóban maximum hely. Tehát, a ML becslés

$$(\widehat{\mu},\widehat{\sigma}^2)=(\overline{x}_n,v_n).$$

Láttuk, hogy $\overline{\xi}_n$ torzítatlan, konzisztens, míg V_n aszimptotikusan torzítatlan konzisztens becslés.[9]

8.18. *Példa* (Egyenletes eloszlás). Legyenek $\xi_1,\dots,\xi_n$ független, az (a,b)-n egyenletes eloszlású véletlen változók. Adott $\mathbf{x}$ realizáció esetén jelölje $x_{\min}$ és $x_{\max}$ a legkisebb és legnagyobb mintaelemet. Ekkor

$$L_{(a,b)}(\mathbf{x})=\left(\frac{1}{b-a}\right)^n\mathbb{I}(a\le x_{\min},x_{\max}\le b).$$

Ez akkor maximális, amikor az indikátor 1, és $b-a$ a lehető legkisebb. Tehát a

$$(\widehat{a},\widehat{b})=(x_{\min},x_{\max})$$

lesz az (a,b) ML becslése. Könnyen látható, hogy ez aszimptotikusan torzítatlan, konzisztens becslés.

8.4. Momentumok módszere

A módszert általában több paraméter együttes becslésére használják. Tegyük fel, hogy $\theta=(\theta_1,\dots,\theta_k)$, azaz $k\ge 1$ paraméterünk van. Válasszunk k darab momentumot, általában az első k-t, amelyek egyértelműen meghatározzák a $\theta=(\theta_1,\dots,\theta_k)$ paramétert. Vezessük be a

$$m_j=\mathbf{E}_\theta(\xi^j)=g_j(\theta_1,\dots,\theta_k),\quad j=1,2,\dots,k,$$

[9]Emlékeztetünk, hogy $\overline{\xi}_n$, V_n véletlen mennyiségek, míg $\overline{x}_n$, v_n ezek egy konkrét realizációhoz tartozó értékei.

jelölést. Az, hogy az első k momentum egyértelműen meghatározza a paramétereket, azt jelenti, hogy vannak olyan $h_1, \dots, h_k$ függvények, hogy

$$h_i(m_1, \dots, m_k) = \theta_i, \quad i = 1, 2, \dots, k.$$

8.19. Definíció. A $\theta = (\theta_1, \dots, \theta_k)$ *momentum becslése* a

$$\widehat{\theta}_i = h_i(\widehat{m}_1, \dots, \widehat{m}_k), \quad i = 1, 2, \dots, k,$$

statisztika, ahol $\widehat{m}_i$ az empirikus i-edik momentum, azaz

$$\widehat{m}_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i, \quad i = 1, 2, \dots, k.$$

A nagy számok gyenge törvénye szerint $\lim_{n\to\infty} \widehat{m}_i = m_i$. Innen pedig egyszerűen adódik, hogy $\widehat{\theta}_i$ konzisztens becslése θ_i-nek minden i-re.

A Poisson-, exponenciális és normális eloszlás esetén a paraméter(ek) momentum becslése megegyezik a ML becsléssel.

8.20. *Példa* (Poisson-eloszlás). A Poisson-eloszlás paramétere megegyezik a várható értékkel, azaz $m_1 = \mathbf{E}_\lambda(\xi) = \lambda$, azaz $h_1(x) = x$, vagyis a λ paraméter momentum becslése

$$\widehat{\lambda} = \widehat{m}_1 = \overline{x}_n,$$

éppen a mintaátlag.

8.21. *Példa* (Exponenciális eloszlás). Az exponenciális eloszlás esetén azaz $m_1 = \mathbf{E}_\lambda(\xi) = \frac{1}{\lambda}$, azaz $h_1(x) = x^{-1}$, vagyis a λ paraméter momentum becslése

$$\widehat{\lambda} = \frac{1}{\widehat{m}_1} = \frac{1}{\overline{x}_n},$$

éppen a ML becslés.

8.22. *Példa* (Normális eloszlás). Ha $\xi \sim \mathrm{N}(\mu, \sigma^2)$, akkor $m_1 = \mathbf{E}_{(\mu,\sigma^2)}(\xi) = \mu$ és $m_2 = \mathbf{E}_{(\mu,\sigma^2)}(\xi^2) = \sigma^2 + \mu^2$. Tehát

$$h_1(m_1, m_2) = m_1, \quad h_2(m_1, m_2) = m_2 - m_1^2.$$

Az empirikus momentumokat behelyettesítve

$$\widehat{\mu} = \widehat{m}_1 = \overline{x}_n$$
$$\widehat{\sigma^2} = \widehat{m}_2 - (\overline{x}_n)^2 = v_n.$$

Azaz a mintaátlag és az empirikus szórásnégyzet a megfelelő becslések.

Az egyenletes eloszlás momentum becslése nem egyezik meg a ML becsléssel.

8.23. *Példa* (Egyenletes eloszlás). Ha $\xi \sim$ Egyenletes(a,b), akkor $m_1 = \mathbf{E}_{(a,b)}(\xi) = \frac{a+b}{2}$ és $m_2 = \mathbf{E}_{(a,b)}(\xi^2) = \frac{a^2+ab+b^2}{3}$. Rövid számolás után kapjuk, hogy

$$a = m_1 - \sqrt{3(m_2 - m_1^2)}, \quad b = m_1 + \sqrt{3(m_2 - m_1^2)}.$$

Mivel $\widehat{m}_1 = \overline{x}_n$ és $\widehat{m}_2 - (\overline{x}_n)^2 = v_n$, így a

$$\widehat{a} = \overline{x}_n - \sqrt{3v_n}, \quad \widehat{b} = \overline{x}_n + \sqrt{3v_n}$$

momentum becslést kapjuk.

8.5. Lineáris regresszió

A következő modellt vizsgáljuk. Egy x ismert bemenethez tartozik egy y ismert kimenet. A két változó kapcsolatát $y = f(x) +$ hiba formula adja meg, ahol a hiba valamilyen értelemben kicsi, szimmetrikus. Feltesszük, hogy ez véletlen, várható értéke 0. A hiba eredhet a mérés pontatlanságából (mérési hiba), hozzáadott zajból. Nem ismerjük az f függvényt, ezt akarjuk meghatározni. Ez a mesterséges intelligencia egyik alapfeladata, a *tanulás*.

A legegyszerűbb esetet vizsgáljuk, amikor $f(x) = ax+b$ lineáris függvény, ahol $a, b \in \mathbb{R}$ nem ismertek. Ekkor az a, b értékek meghatározása, becslése a feladat. Adott az (x_i, y_i), $i = 1, 2, \ldots, n$ minta. Nem várunk pontos lineáris illeszkedést, nem lesz olyan a, b, hogy $y_i = ax_i + b$ minden $i = 1, 2, \ldots, n$ esetén teljesüljön. Négyzetes hibára minimalizálunk, azaz keressük azt az $(\hat{a}, \hat{b})$ párt, melyre

$$h(a,b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 \tag{6}$$

négyzetes hiba minimális. Azaz egy kétváltozós függvény minimumhelyét keressük[10]. Ez most egyszerű struktúrájú, hiszen a-ban, b-ben másodfokú. A $z_i = y_i - ax_i$ jelöléssel

$$\begin{aligned} h(a,b) &= \sum (z_i - b)^2 = \sum z_i^2 - 2b \sum z_i + b^2 n \\ &= n(b - \overline{z}_n)^2 + \sum z_i^2 - n(\overline{z}_n)^2 \end{aligned}$$

[10] Általános esetben kétváltozós függvény szélsőértékét úgy keressük, hogy megnézzük, hogy a parciális deriváltak hol 0-k, aztán vizsgáljuk a második deriváltakból álló mátrixot.

adódik. A 2. és 3. tagot kifejtve, felhasználva, hogy[11] $\overline{z}_n = \overline{y}_n - a\overline{x}_n$ és

$$\frac{1}{n}\sum x_i^2 - (\overline{x}_n)^2 = v_n,$$
$$\frac{1}{n}\sum x_i y_i - \overline{x}_n\overline{y}_n = c_n$$

kapjuk, hogy

$$\begin{aligned}
&\sum z_i^2 - n(\overline{z}_n)^2 \\
&= a^2\left(\sum x_i^2 - n(\overline{x}_n)^2\right) - 2a\left(\sum x_i y_i - n\overline{x}_n\overline{y}_n\right) + \sum y_i^2 - n(\overline{y}_n)^2 \\
&= nv_n\left(a - \frac{c_n}{v_n}\right)^2 - n\frac{c_n^2}{v_n} + \sum y_i^2 - n(\overline{y}_n)^2.
\end{aligned}$$

Tehát $h(a,b)$ kifejezést négyzetek összegére bontottuk

$$h(a,b) = n(b-\overline{z}_n)^2 + nv_n\left(a - \frac{c_n}{v_n}\right)^2 + \text{ konstans},$$

ahol a konstans nem függ a, b értékétől. Tehát $h(a,b)$ pontosan akkor minimális, ha a négyzetek 0-k, azaz

$$b = \overline{z}_n = \overline{y}_n - a\overline{x}_n.$$

és

$$a = \frac{c_n}{v_n} = \frac{\sum_{i=1}^n (y_i - \overline{y}_n)(x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}.$$

Ez a *legkisebb négyzetek módszere.*

Vegyük észre, hogy pontosan a korábban 6.3 alfejezetben megkapott elméleti értékek empirikus változatait kaptuk.

9. Konfidenciaintervallumok és próbák

Eddig az ismeretlen paramétert egyetlen számmal becsültük. Most egy egész intervallumot szeretnénk megadni, amibe a paraméter nagy valószínűséggel beleesik.

9.1. Definíció. A $(T_1(\xi), T_2(\xi))$ statisztikapárral definiált intervallum *legalább* $1-\varepsilon$ *megbízhatósági szintű konfidenciaintervallum a* $\psi(\theta)$ *paraméterre*, ha

$$\mathbf{P}_\theta(T_1(\xi) < \psi(\theta) < T_2(\xi)) \geq 1-\varepsilon, \quad \forall \theta \in \Theta,$$

[11] Lásd az alapstatisztikák definíciója, és Steiner-tétel.

ahol $\varepsilon > 0$ előre adott kicsi szám.

Amennyiben a fönti $\geq$ helyett $=$ szerepel, akkor $(T_1(\xi), T_2(\xi))$ *pontosan* $1-\varepsilon$ *megbízhatósági szintű konfidenciaintervallum.*

9.1. Konfidenciaintervallum normális eloszlás várható értékére ismert szórás esetén

Legyenek $\xi_1, \ldots, \xi_n \sim \mathrm{N}(\mu, \sigma_0^2)$ független véletlen változók, ahol μ az ismeretlen paraméter, σ_0^2 ismert. Megadunk μ-re pontosan $1-\varepsilon$ megbízhatósági szintű konfidenciaintervallumot.

Mivel $\overline{\xi}$ torzítatlan, erősen konzisztens becslés μ-re, ezért az intervallumot $(\overline{\xi} - r_\varepsilon, \overline{\xi} + r_\varepsilon)$ alakban keressük.

Szükségünk lesz az alábbi állításra, melyet nem bizonyítunk.

9.2. Tétel. *Ha ξ, η független normális eloszlású véletlen változók akkor $\xi + \eta$ is normális eloszlású.*

Ebből indukcióval az is következik, hogy független normálisok összege normális, és persze a várható értékek összeadódnak (azok mindig), és a szórásnégyzetek is összeadódnak (hiszen függetlenek az összeadandók). Ezért $\overline{\xi} \sim \mathrm{N}(\mu, \sigma_0^2/n)$. Tehát

$$\begin{aligned}
\mathbf{P}_\mu(\overline{\xi} - r_\varepsilon < \mu < \overline{\xi} + r_\varepsilon) &= \mathbf{P}_\mu\left(-\frac{r_\varepsilon}{\sigma_0/\sqrt{n}} < \frac{\overline{\xi}-\mu}{\sigma_0/\sqrt{n}} < \frac{r_\varepsilon}{\sigma_0/\sqrt{n}}\right) \\
&= \Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) - \Phi\left(-\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) \\
&= 2\Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) - 1 = 1-\varepsilon,
\end{aligned}$$

azaz

$$\Phi\left(\frac{\sqrt{n}r_\varepsilon}{\sigma_0}\right) = 1 - \frac{\varepsilon}{2}.$$

Tehát az

$$u_{\varepsilon/2} = \Phi^{-1}(1-\varepsilon/2)$$

jelöléssel, ahol Φ^{-1} a Φ függvény inverze, a keresett $1-\varepsilon$ megbízhatósági szintű konfidenciaintervallum

$$\left(\overline{\xi} - \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}, \overline{\xi} + \frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right).$$

Vegyük észre, hogy a megbízhatósági szint növelésével, azaz ε csökkenésével, az intervallum hossza nő, a mintaelemszám növelésével pedig csökken.

9.2. Konfidenciaintervallum normális eloszlás várható értékére ismeretlen szórás esetén

A feladat ugyanaz, mint az előbb, csak most nem mondja meg senki a szórást. Tehát a szórást is becsülni kell. A fenti számolásnál az volt a kulcs, hogy

$$\frac{\sum_{i=1}^n(\xi_i-\mu)}{\sqrt{n}\sigma} \quad \text{standard normális.}$$

Ismeretlen σ esetén ezt becsüljük a korrigált empirikus szórással és kapjuk a

$$\frac{\sum_{i=1}^n(\xi_i-\mu)}{\sqrt{nV_n^*}} \tag{7}$$

statisztikát, ahol

$$V_n^* = \frac{1}{n-1}\sum_{i=1}^n(\xi_i-\overline{\xi}_n)^2$$

a korrigált empirikus szórásnégyzet. A kapott statisztika Student-eloszlású, ami a következő.

9.3. Definíció. Legyenek η és $\eta_1,\dots,\eta_n$ független standard normálisok. Ekkor a

$$T_n = \frac{\eta}{\sqrt{(\eta_1^2+\ldots+\eta_n^2)/n}} \tag{8}$$

változó n szabadsági fokú Student-eloszlású (vagy t-eloszlású) véletlen változó. Eloszlásfüggvénye $\Phi_n(x) = \mathbf{P}(T_n \le x)$.

A Φ_n értékei is táblázatba vannak szedve.

Vegyük észre, hogy a (8) nevezőjében 1 várható értékű független azonos eloszlású véletlen változók átlaga szerepel. Ha n tart a végtelenbe, akkor a nagy számok törvénye szerint ez a hányados a várható értékhez, azaz 1-hez konvergál. Vagyis, ha n nagy, akkor a (8) hányados nevezője 1, így maga a hányados egy standard normális eloszláshoz van közel. A Student-eloszlás táblázatából láthatjuk, hogy az eloszlásfüggvénye nagy n esetén majdnem megegyezik a standard normális eloszlásfüggvénnyel.

Most is kell egy eredmény, amit nem bizonyítunk.

9.4. Tétel. *Legyenek $\xi_1,\dots,\xi_n$ független normális eloszlású véletlen változók μ várható értékkel és σ szórással. Ekkor az $\overline{\xi}_n$ mintaátlag és az S_n^2 empirikus szórásnégyzet függetlenek. Továbbá a*

$$\sqrt{n}\frac{\overline{\xi}_n-\mu}{\sqrt{V_n^*}}$$

hányados $n-1$ szabadsági fokú Student-eloszlású.

Ezen az állításon azért meg kell lepődni egy kicsit, nevezetesen a függetlenségen. Az az állítás, hogy az ugyanabból a mintából számolt mintaátlag és empirikus szórásnégyzet függetlenek.

Ezek után a konfidenciaintervallum meghatározása pontosan úgy megy mint korábban, csak nem Φ-ből hanem Φ_{n-1}-ből számolunk kritikus értéket. Tehát keressük a konfidenciaintervallumot $[\overline{\xi}_n - r_\varepsilon, \overline{\xi}_n + r_\varepsilon]$ alakban. Ekkor a definíció szerint

$$\begin{aligned}
\mathbf{P}_\mu(\overline{\xi} - r_\varepsilon < \mu < \overline{\xi} + r_\varepsilon) &= \mathbf{P}_\mu\left(-\frac{r_\varepsilon}{\sqrt{V_n^*}/\sqrt{n}} < \frac{\overline{\xi} - \mu}{\sqrt{V_n^*}/\sqrt{n}} < \frac{r_\varepsilon}{\sqrt{V_n^*}/\sqrt{n}}\right) \\
&= \Phi_{n-1}\left(\frac{\sqrt{n}r_\varepsilon}{\sqrt{V_n^*}}\right) - \Phi_{n-1}\left(-\frac{\sqrt{n}r_\varepsilon}{\sqrt{V_n^*}}\right) \\
&= 2\Phi_{n-1}\left(\frac{\sqrt{n}r_\varepsilon}{\sqrt{V_n^*}}\right) - 1 \\
&= 1 - \varepsilon,
\end{aligned}$$

azaz

$$\Phi_{n-1}\left(\frac{\sqrt{n}r_\varepsilon}{\sqrt{V_n^*}}\right) = 1 - \frac{\varepsilon}{2}.$$

Itt még annyit felhasználtunk, hogy a Student-eloszlás is szimmetrikus, azaz $\Phi_{n-1}(-x) = 1 - \Phi_{n-1}(x)$. Tehát az

$$t_{\varepsilon/2} = \Phi_{n-1}^{-1}(1 - \varepsilon/2)$$

jelöléssel, ahol Φ_{n-1}^{-1} a Φ_{n-1} függvény inverze, a keresett $1-\varepsilon$ megbízhatósági szintű konfidenciaintervallum

$$\left(\overline{\xi} - \frac{t_{\varepsilon/2}\sqrt{V_n^*}}{\sqrt{n}}, \overline{\xi} + \frac{t_{\varepsilon/2}\sqrt{V_n^*}}{\sqrt{n}}\right).$$

9.3. Konfidenciaintervallum normális eloszlások várható értékének különbségére

Sokszor találkozunk olyan problémával, ahol különböző adatokat akarunk összehasonlítani. Tegyük fel, hogy azt akarjuk eldönteni, hogy valamilyen vérnyomáscsökkentő hatásos-e. Ekkor van egy $\xi_1, \ldots, \xi_n$ és egy $\eta_1, \ldots, \eta_m$ mintánk, ahol az ξ-ek az adott gyógyszert szedők (szisztolés) vérnyomása, míg a η-ok a kontrollcsoportban a vérnyomásértékek. Azt szeretnénk eldönteni, hogy az ξ-ek várható értéke kisebb-e, azaz hatásos-e a gyógyszer.

Legyenek tehát $\xi_1, \ldots, \xi_n$ független, normális eloszlású véletlen változók μ_1 várható értékkel és σ szórással, míg $\eta_1, \ldots, \eta_m$ független, normális eloszlású véletlen változók, melyek az ξ-ektől is függetlenek, és a várható értékük μ_2, szórásuk σ. *Tehát feltesszük, hogy a szórások egyenlőek.* A $\mu_1 - \mu_2$ különbségre konstruálunk konfidenciaintervallumot. Nyilván μ_1 becslése $\overline{\xi}_n$, míg μ_2 becslése $\overline{\eta}_m$, tehát a $\delta = \mu_1 - \mu_2$ különbséget a

$$\widehat{\delta} = \overline{\xi}_n - \overline{\eta}_m$$

mennyiséggel becsüljük. Ennek a szórása a függetlenség miatt

$$\mathbf{D}^2(\widehat{\delta}) = \mathbf{D}^2(\overline{\xi}_n) + \mathbf{D}^2(\overline{\eta}_m) = \frac{1}{n}\sigma^2 + \frac{1}{m}\sigma^2 = \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right).$$

A 9.2 Tétel szerint $\overline{\xi}_n$ és $\overline{\eta}_m$ is normális eloszlásúak, és mivel függetlenek ugyanezen tétel szerint a különbségük is normális eloszlású (na jó itt használtuk, hogy normális eloszlású változó -1-szerese is normális, de ezt láttuk korábban). Tehát $\widehat{\delta}$ normális eloszlású, várható értéke $\mu_1 - \mu_2$, szórásnégyzete pedig $\sigma^2(n^{-1} + m^{-1})$. Tehát

$$\frac{\overline{\xi}_n - \overline{\eta}_m - (\mu_1 - \mu_2)}{\sigma\sqrt{n^{-1} + m^{-1}}}$$

standard normális. Ebből pedig legyárthatjuk a konfidenciaintervallumot a korábban látottak szerint. A konfidenciaintervallumot a szimmetria miatt $(\widehat{\delta} - r_\varepsilon, \widehat{\delta} + r_\varepsilon)$ alakban keresve

$$\mathbf{P}\left(\delta \in [\widehat{\delta} - r_\varepsilon, \widehat{\delta} + r_\varepsilon]\right) = \mathbf{P}\left(\frac{|\widehat{\delta} - \delta|}{\sigma\sqrt{n^{-1} + m^{-1}}} \leq \frac{r_\varepsilon}{\sigma\sqrt{n^{-1} + m^{-1}}}\right)$$
$$= 2\Phi\left(\frac{r_\varepsilon}{\sigma\sqrt{n^{-1} + m^{-1}}}\right) - 1 = 1 - \varepsilon,$$

ahonnan kapjuk, hogy

$$r_\varepsilon = \sigma\sqrt{n^{-1} + m^{-1}}\Phi^{-1}(1 - \varepsilon/2).$$

Tehát a konfidenciaintervallum

$$\left(\overline{\xi}_n - \overline{\eta}_m - \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}\Phi^{-1}\Big(1 - \frac{\varepsilon}{2}\Big), \overline{\xi}_n - \overline{\eta}_m + \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}\Phi^{-1}\Big(1 - \frac{\varepsilon}{2}\Big)\right).$$

Ez persze csak akkor működik, ha ismerjük a szórást. Ha nem ismerjük azt is becsüljük. Mivel az elméleti szórások megegyeznek, ezért a szórásnégyzet becslése

$$V^*_{n,m} = \frac{1}{n + m - 2}\left[\sum_{i=1}^{n}(\xi_i - \overline{\xi}_n)^2 + \sum_{j=1}^{m}(\eta_j - \overline{\eta}_m)^2\right].$$

Ez torzítatlan becslése a szórásnégyzetnek, hiszen láttuk korábban a 8.8 Állításnál, hogy

$$\mathbf{E}\sum_{i=1}^{n}(\xi_i-\overline{\xi}_n)^2=(n-1)\sigma^2,\quad \text{és } \mathbf{E}\sum_{j=1}^{m}(\eta_j-\overline{\eta}_m)^2=(m-1)\sigma^2.$$

A 9.4 Tételből következik, hogy

$$\frac{\overline{\xi}_n-\overline{\eta}_m-(\mu_1-\mu_2)}{\sqrt{D_{n,m}}\sqrt{n^{-1}+m^{-1}}}$$

Student-eloszlású $n+m-2$ szabadsági fokkal. Ezért ugyanúgy számolhatunk konfidenciaintervallumot, mint ismert szórásnál, csak a Student-eloszlásból kell kritikus értéket számolni. Azaz a konfidenciaintervallum

$$\left(\overline{\xi}_n-\overline{\eta}_m-\Delta\,\Phi^{-1}_{n+m-2}\Big(1-\frac{\varepsilon}{2}\Big),\overline{\xi}_n-\overline{\eta}_m+\Delta\,\Phi^{-1}_{n+m-2}\Big(1-\frac{\varepsilon}{2}\Big)\right),$$

ahol

$$\Delta=\sqrt{D_{n,m}}\sqrt{\frac{1}{n}+\frac{1}{m}}.$$

9.4. u-próba

Legyen ξ normális eloszlású véletlen változó ismert σ_0 szórással, és ismeretlen μ várható értékkel. Azt szeretnénk eldönti egy $(\xi_1,\dots,\xi_n)$ független minta alapján, hogy igaz-e hogy a várható érték μ_0.

A nullhipotézis, amit igaznak teszünk föl az az, hogy a várható érték valóban μ_0, míg az ellenhipotézis, vagy alternatív hipotézis az az, hogy ez nem teljesül. Azaz

$$H_0:\ \mu=\mu_0\quad \text{vs. } H_1:\ \mu\neq\mu_0. \tag{9}$$

Az

$$u=\frac{\overline{\xi}_n-\mu_0}{\sigma_0}\sqrt{n} \tag{10}$$

statisztika standard normális eloszlású a H_0 fennállása esetén. Nagyon fontos megjegyezni, hogy ha H_1 áll fenn, akkor u *nem* standard normális. Legyen $\varepsilon>0$ rögzített szignifikanciaszint. Az $u_{\varepsilon/2}=\Phi^{-1}(1-\varepsilon/2)$ jelöléssel,

$$\mathbf{P}_{\mu_0}\left(\mu_0\in\left(\overline{\xi}_n-\frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}},\overline{\xi}_n+\frac{u_{\varepsilon/2}\sigma_0}{\sqrt{n}}\right)\right)=\mathbf{P}_{\mu_0}(|u|\le u_{\varepsilon/2})=1-\varepsilon.$$

u-próba. Ezek alapján a következőképpen járunk el:

1. A minta alapján kiszámoljuk az u próbastatisztikát (lásd (10)).
2. Rögzített $\varepsilon > 0$ szignifikanciaszinthez meghatározzuk $u_{\varepsilon/2}$ értéket.
3. Ha $|u| \le u_{\varepsilon/2}$ akkor elfogadjuk H_0-t, különben elvetjük.

Valójában $1-\varepsilon$ megbízhatósági szintű konfidenciaintervallumot szerkesztünk a várható értékre, és akkor *fogadjuk el / vetjük el* a nullhipotézist, ha a μ_0 érték *beleesik / nem esik bele* a konfidenciaintervallumba. Ez minden próba esetén így van.

Ekkor kétféleképpen követhetünk el hibát: *Elsőfajú hiba* esetén H_0 fennáll, mégis elutasítjuk, *másodfajú hiba* esetén pedig H_0 nem áll fenn, mégis elfogadjuk.

Elsőfajú hibát akkor követünk el, ha a valódi paraméter μ_0, és $|u| \ge u_{\varepsilon/2}$, ennek valószínűsége $\mathbf{P}(|u| \ge u_{\varepsilon/2}) = \varepsilon$, azaz ez éppen az előre megadott ε szignifikanciaszint.

Másodfajú hibát akkor követünk el, ha a valódi paraméter $\mu \neq \mu_0$, azaz H_1 áll fönn, és $|u| \le u_{\varepsilon/2}$. Ennek valószínűsége $\mathbf{P}_\mu(|u| \le u_{\varepsilon/2})$ függ a valódi paraméter értékétől. A próba *erőfüggvénye*

$$\beta_n(\mu, \varepsilon) = 1 - \mathbf{P}_\mu(|u| \le u_{\varepsilon/2}) = \mathbf{P}_\mu(|u| \ge u_{\varepsilon/2}). \tag{11}$$

Legyen $\Delta_n = \frac{\mu-\mu_0}{\sigma_0}\sqrt{n}$. Ha H_1 áll fenn, és a várható érték valódi értéke μ, akkor $\sqrt{n}(\overline{\xi}_n - \mu)/\sigma_0$ változó lesz standard normális, ezért

$$\begin{aligned}
\beta_n(\mu, \varepsilon) &= \mathbf{P}_\mu(|u| \ge u_{\varepsilon/2}) \\
&= \mathbf{P}_\mu\left(\left|\frac{\overline{\xi}_n - \mu}{\sigma_0}\sqrt{n} + \Delta_n\right| \ge u_{\varepsilon/2}\right) \\
&= 1 - \Phi(u_{\varepsilon/2} - \Delta_n) + \Phi(-u_{\varepsilon/2} - \Delta_n) \\
&= 2 - \left[\Phi(u_{\varepsilon/2} - \Delta_n) + \Phi(u_{\varepsilon/2} + \Delta_n)\right].
\end{aligned}$$

Innen látjuk, hogy

$$\lim_{n\to\infty} \beta_n(\mu, \varepsilon) = 1, \quad \text{tetszőleges } \mu \neq \mu_0 \text{ esetén.}$$

Ez utóbbi állítás próba konzisztenciáját jelenti.

9.5. *Példa.* Azt szeretnénk tesztelni, hogy az 1 kg-os cukor tényleg 1kg-e? Persze nem várjuk, hogy minden egyes csomagban halálpontosan 1 kg cukor legyen. Azt tesszük fel, hogy egy csomag tömege normális eloszlást követ μ várható értékkel (ezt nem ismerjük), és $\sigma_0 = 0,05$ ismert szórással.

A nullhipotézisünk az, hogy valóban 1 kg a várható érték, $\mu = 1$, az ellenhipotézis pedig $\mu \neq 1$, azaz

$$H_0: \ \mu = 1 \quad \text{vs. } H_1: \ \mu \neq 1. \tag{12}$$

Egy 25 elemű minta alapján a mintaátlagra $\overline{x}_{25} = 0,98$ adódik. Ez kisebb, mint 1, de az eltérés adódhat a véletlenből is, nem feltétlenül csaltak. A próbastatisztika értéke

$$u = \frac{\overline{x}_{25} - 1}{0,05}\sqrt{25} = -2.$$

Válasszuk a szignifikanciaszintet 0,05-nek, azaz $\varepsilon = 0,05$. Ekkor a standard normális eloszlástáblázatából kiolvassuk, hogy $u_{\varepsilon/2} = \Phi^{-1}(0,975) = 1,96$. Mivel $|u| = 2 > 1,96$, ezért elvetjük a nullhipotézist, azaz úgy döntünk, hogy nem 1 a várható érték.

Ugyanakkor a gyártó a következőképpen védekezhet. Annak az esélye, hogy elvetjük a nullhipotézist annak ellenére, hogy az igaz, 0,05. Ez nem is olyan kicsi, 100 esetből 5-ször bekövetkezik. Lehet, hogy most is egy kis valószínűségű esemény következett be. Nézzük meg, hogy $\varepsilon = 0,01$ szignifikanciaszinten is elvetjük-e a nullhipotézist. Ekkor $u_{\varepsilon/2} = \Phi^{-1}(0,995) = 2,57$. Mivel $|u| = 2 < 2,57$, ezért a nullhipotézist elfogadjuk.

A feladat jellegétől függően, sokszor ún. egyoldali próbát használunk. Ebben a példában igazából akkor éri sérelem a vásárlót, ha a valódi várható érték kisebb, mint 1. Azaz, a

$$H_0 : \ \mu = 1 \quad \text{vs. } H_1 : \ \mu < 1, \tag{13}$$

egyoldali hipotézisvizsgálatot kell elvégezni. H_0-t akkor fogadjuk el, ha $u \geq -u_\varepsilon = \Phi^{-1}(1-\varepsilon)$. Ekkor az elsőfajú hiba (azaz, hogy H_0 fennáll, mégis elvetjük) valószínűsége

$$\mathbf{P}_{\mu_0}(u \leq -u_\varepsilon) = 1 - \Phi(u_\varepsilon) = \varepsilon.$$

Ebben az esetben $\varepsilon = 0,01$ szignifikanciaszinthez tartozó kritikus érték $u_\varepsilon = \Phi^{-1}(0,99) = 2,33$. Azaz ilyen szignifikanciaszinten ebben az esetben is felmentjük a gyártót. (Bár nagyon gyanús.)

10. Függő véletlen változók

10.1. Feltételes függetlenség

A Bayes-tétel egy tipikus alkalmazása a következő. Tudjuk, hogy bizonyos megbetegedésre milyen tünetek jellemzőek. A tünetekből szeretnénk a betegségre következtetni. A pollenallergia és a Covid-19 között szeretnénk dönteni. Tegyük fel, hogy allergia esetén az emberek 30%-a lázasodik be, míg Covid-19 esetén 80%-a. Tegyük fel továbbá, hogy az emberek 20%-a allergiás, és 10%-a covidos, és nem lehet mindkettő.

Ekkor, ha egy beteg lázas, akkor annak a valószínűsége, hogy allergiás

$$\begin{aligned}\mathbf{P}(\text{allergia}|\text{láz}) &= \frac{\mathbf{P}(\text{allergia}) \cdot \mathbf{P}(\text{láz}|\text{allergia})}{\mathbf{P}(\text{láz})} \\ &= \frac{\mathbf{P}(\text{allergia}) \cdot \mathbf{P}(\text{láz}|\text{allergia})}{\mathbf{P}(\text{allergia}) \cdot \mathbf{P}(\text{láz}|\text{allergia}) + \mathbf{P}(\text{covid}) \cdot \mathbf{P}(\text{láz}|\text{covid})} \\ &= \frac{0,2 \cdot 0,3}{0,2 \cdot 0,3 + 0,1 \cdot 0,8} = 0,43.\end{aligned}$$

Ugyanígy

$$\begin{aligned}\mathbf{P}(\text{covid}|\text{láz}) &= \frac{\mathbf{P}(\text{covid}) \cdot \mathbf{P}(\text{láz}|\text{covid})}{\mathbf{P}(\text{láz})} \\ &= \frac{0,1 \cdot 0,8}{0,2 \cdot 0,3 + 0,1 \cdot 0,8} = 0,57.\end{aligned}$$

Ez persze egy nagyon leegyszerűsített modell, de a módszert látjuk.

Most több tünetet is vizsgálunk együtt. Az alábbi táblázatban az látjuk, hogy adott betegség esetén a tünet milyen valószínűséggel jelentkezik:

	Láz	Orrfolyás	Köhögés
Allergia	0,3	0,7	0,3
Covid	0,8	0,2	0,7

Azaz, ha allergiásak vagyunk, 0,3 valószínűséggel van lázunk, formulával $\mathbf{P}(\text{láz}|\text{allergia}) = 0,3$. Tudjuk, hogy az emberek 20%-a allergiás, 10%-a covidos, és nem mindkettő.

Tegyük fel, hogy a tünetek egymástól függetlenek adott megbetegedés esetén.[12] Azaz

$$\mathbf{P}(\text{láz és orrfolyás}|\text{allergia}) = \mathbf{P}(\text{láz}|\text{allergia}) \cdot \mathbf{P}(\text{orrfolyás}|\text{allergia}).$$

Tegyük fel, hogy egy beteg lázas és köhög, de nem folyik az orra. Ekkor mennyi a valószínűsége, hogy covidos? Ekkor

$$\begin{aligned}&\mathbf{P}(\text{covid}|\text{láz, köhögés, nincs orrfolyás}) \\ &= \frac{\mathbf{P}(\text{covid}) \cdot \mathbf{P}(\text{láz, köhögés, nincs orrfolyás}|\text{covid})}{\mathbf{P}(\text{láz, köhögés, nincs orrfolyás})}.\end{aligned}$$

[12] Azaz a láz, orrfolyás, köhögés események az *allergiára feltételesen függetlenek egymástól.*

A feltételes függetlenség szerint

$$\begin{aligned}&\mathbf{P}(\text{láz, köhögés, nincs orrfolyás}|\text{covid})\\&= \mathbf{P}(\text{láz}|\text{covid}) \cdot \mathbf{P}(\text{köhögés}|\text{covid}) \cdot \mathbf{P}(\text{nincs orrfolyás}|\text{covid})\\&= 0,8 \cdot 0,7 \cdot 0,8 = 0,448.\end{aligned}$$

A teljes valószínűség tétele szerint

$$\begin{aligned}&\mathbf{P}(\text{láz, köhögés, nincs orrfolyás})\\&= \mathbf{P}(\text{covid}) \cdot \mathbf{P}(\text{láz, köhögés, nincs orrfolyás}|\text{covid})\\&\quad + \mathbf{P}(\text{allergia}) \cdot \mathbf{P}(\text{láz, köhögés, nincs orrfolyás}|\text{allergia})\\&= 0,1 \cdot 0,448 + 0,2 \cdot 0,3 \cdot 0,3 \cdot 0,3 = 0,0502.\end{aligned}$$

A kapott eredményt visszaírva

$$\mathbf{P}(\text{covid}|\text{láz, köhögés, nincs orrfolyás}) = \frac{0,0448}{0,0502} = 0,89.$$

Tehát ekkor már 89% a covid esélye.

Mesterséges intelligencia nyelven ez a Bayes-tanulás. Később tanultok Bayesi-hálókról Mesterséges intelligencia kurzuson.

10.2. Beszúró rendezés elemzése

Rendezési feladatoknál egy n elemű halmaz elemeit kell sorbarendezni nagyság szerint. Azaz kapunk egy $A = (a_1, a_2, \ldots, a_n)$ valós számokból álló vektort. Meg kell adnunk A elemeinek sorbarendezését növekvő sorrendben, azaz az elemek azt a permutációját, melyre $a'_1 \leq a'_2 \leq \ldots \leq a'_n$.

Az egyik legegyszerűbb rendezés a beszúró rendezés azt csinálja, hogy a már rendezett i elem közé a megfelelő helyre beszúrja az a_{i+1} elemet. Ezt megcsinálja $i = 1$-től $i = (n-1)$-ig. Könnyű látni, hogy az algoritmus működik. Arra vagyunk kíváncsiak, hogy milyen gyors.

Legrosszabb esetben a sorozat csökkenő sorrendben van. Ekkor a_{i+1} kisebb, mint az előtte már elrendezett i elem mindegyike, ezért i összehasonlítást kell végezni. Összesen ekkor

$$1 + 2 + \ldots + (n-1) = \frac{n(n-1)}{2} \sim \frac{n^2}{2}$$

összehasonlítást végzünk. A legjobb esetben a sorozat növekvő sorrendbe van rendezve. Ekkor a_{i+1} nagyobb, mint az előtte már elrendezett i elem mindegyike, ezért elég csak a legnagyobbal összehasonlítani. Így minden i-re

1 összehasonlítást végzünk, összesen $(n-1)$-et. Ez két szélsőséges eset. Az átlagos esetet egy véletlen bemeneten a összehasonlítások számának várható értéke adja.

Tegyük fel, hogy az A elemeinek $n!$ lehetséges permutációja egyformán valószínű[13] Jelölje ξ az összes szükséges összehasonlítások számát, és η_i az i-edik lépésben végzett összehasonlítások számát. Nyilván

$$\xi = \eta_1 + \eta_2 + \ldots + \eta_{n-1}.$$

Ekkor minden i-re η_i lehetséges értékei $1, 2, \ldots, i$, hiszen legalább 1 összehasonlítást kell végeznünk, legfeljebb pedig az összes eddig sorbarendezett elemmel kell összehasonlítani, amiből éppen i van. Vegyük észre, hogy ha a_{i+1} a j-edik legnagyobb elem az $a_1, a_2, \ldots, a_{i+1}$ elemek között, akkor pontosan j összehasonlítást kell elvégezni, ha $j = 1, 2, \ldots, i$. Valóban, ha $j = 1$ akkor a_{i+1} a legnagyobb az $a_1, \ldots, a_i, a_{i+1}$ elemek között, így csak az első i közül a legnagyobbal kell összehasonlítani, hogy ez kiderüljön. Ha $j = 2$, akkor a első és a második legnagyobbal, ... Ha a_{i+1} a legkisebb $j = i+1$ (vagy a második legkisebb $j = i$), akkor mindenkivel össze kell hasonlítani. Mivel a permutáció véletlen, ezért a_{i+1} egyforma eséllyel lehet bárhol a sorrendben az $a_1, \ldots, a_i, a_{i+1}$ elemek között, tehát

$$\mathbf{P}(\eta_i = j) = \begin{cases} \frac{1}{i+1}, & j = 1, 2, \ldots, i-1, \\ \frac{2}{i+1}, & j = i. \end{cases}$$

A várható értékre

$$\begin{aligned} \mathbf{E}(\eta_i) &= \sum_{j=1}^{i} j\mathbf{P}(\eta_i = j) = \frac{1}{i+1}\sum_{j=1}^{i-1} j + \frac{2}{i+1}i \\ &= \frac{i(i-1)+4i}{2(i+1)} = \frac{i^2+3i}{2(i+1)} = \frac{i}{2} + 1 - \frac{1}{i+1}. \end{aligned}$$

Ez nagyjából $i/2$, ha i nagy. Ez persze intuitívan világos volt, hiszen kb. az elemek felével kell összehasonlítani az új elemet. Innen kapjuk

$$\mathbf{E}(\xi) = \mathbf{E}\left(\sum_{i=1}^{n-1}\eta_i\right) = \sum_{i=1}^{n-1}\mathbf{E}(\eta_i) = \sum_{i=1}^{n-1}\left(\frac{i}{2} + 1 - \frac{1}{i+1}\right) \sim \frac{n^2}{4}.$$

Összegezve, véletlen bemeneten csak egy $1/2$ szorzót nyerünk a legrosszabb esethez képest.

[13]Ha az nem feltétlen igaz, akkor éppen permutálhatjuk az elemeket egy véletlen permutációval, hogy ez teljesüljön. A *véletlenített algoritmusok* csinálnak ilyet.

Nem túl nehéz megmutatni, hogy $\eta_1, \eta_2, \ldots, \eta_{n-1}$ független véletlen változók[14]. Emiatt ξ szórását is egyszerűen számolhatjuk a

$$\mathbf{D}^2(\xi) = \mathbf{D}^2\left(\sum_{i=1}^{n-1} \eta_i\right) = \sum_{i=1}^{n-1} \mathbf{D}^2(\eta_i)$$

formulával. Mivel $\mathbf{D}^2(\eta_i) = \mathbf{E}(\eta_i^2) - (\mathbf{E}(\eta_i))^2$, így

$$\begin{aligned}\mathbf{E}(\eta_i^2) &= \sum_{j=1}^{i} j^2 \mathbf{P}(\eta_i = j) = \frac{1}{i+1}\sum_{j=1}^{i} j^2 + \frac{i^2}{i+1} \\ &= \frac{1}{i+1}\frac{i(i+1)(2i+1)}{6} + \frac{i^2}{i+1} \sim \frac{i^2}{3},\end{aligned}$$

ha i nagy. Így

$$\mathbf{D}^2(\xi) \sim \sum_{i=1}^{n-1} \frac{i^2}{3} \sim \frac{n^3}{9},$$

tehát $\mathbf{D}(\xi) \sim n^{3/2}/3$.

10.3. PageRank algoritmus

Tekintsük a világháló a 20. ábrán szereplő egyszerűsített modelljét. Minden csúcs egy weboldalt jelent, az irányított élek pedig azt, hogy az oldalon hova mutatnak linkek. Tehát az 1-es oldalon két link szerepel, egy a 2-es, egy a 3-as oldalra. A 2-es oldalon is két link van, az 1-esre és az 5-ösre, ...

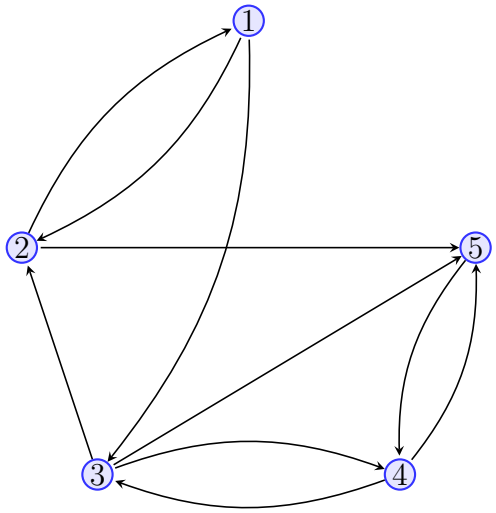
Tegyük fel, hogy egy *véletlen szörfös* minden weboldalon egyforma valószínűséggel kattint az ott szereplő linkek bármelyikére. Tehát, ha csak 1 link van, akkor biztosan arra, ha 2, akkor $1/2 - 1/2$ valószínűséggel valamelyikre, ... Legyen az 1-es a kezdőoldal, azaz innen indul a véletlen szörfös, és jelölje ξ_i, $i = 1, 2, \ldots$, azt az oldalt, ahova az i. lépés (kattintás) után jut. Ekkor

$$\mathbf{P}(\xi_1 = 2) = \mathbf{P}(\xi_1 = 3) = \frac{1}{2},$$

hiszen az 1-esen két link van, a 2-re és a 3-ra. Két lépés után már nehezebb a helyzet. Lehetünk az 1, 5, 2, 4 oldalakon. A megfelelő valószínűségeket a teljes valószínűség tételének segítségével számolhatjuk ki. Valóban, az 1-be csak a 2-ből léphet, így

$$\mathbf{P}(\xi_2 = 1) = \mathbf{P}(\xi_2 = 1|\xi_1 = 2) \cdot \mathbf{P}(\xi_1 = 2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

[14]Arról van szó, hogy a relatív rangok sorozata, azaz a_i helye az első i elem között, független véletlen változók.



20. ábra. Világháló egy modellje

Ez csak a szorzási szabály volt. Viszont az 5-be léphet a 2-ből és a 3-ból is, ezért

$$\begin{aligned}\mathbf{P}(\xi_2 = 5) &= \mathbf{P}(\xi_2 = 5|\xi_1 = 2) \cdot \mathbf{P}(\xi_1 = 2) + \mathbf{P}(\xi_2 = 5|\xi_1 = 3) \cdot \mathbf{P}(\xi_1 = 3) \\ &= \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{12}.\end{aligned}$$

Hasonlóan,

$$\begin{aligned}\mathbf{P}(\xi_2 = 2) &= \mathbf{P}(\xi_2 = 2|\xi_1 = 3) \cdot \mathbf{P}(\xi_1 = 3) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}, \\ \mathbf{P}(\xi_2 = 4) &= \mathbf{P}(\xi_2 = 4|\xi_1 = 3) \cdot \mathbf{P}(\xi_1 = 3) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}.\end{aligned}$$

Ellenőrizzük, hogy

$$\mathbf{P}(\xi_2 = 1) + \mathbf{P}(\xi_2 = 2) + \mathbf{P}(\xi_2 = 4) + \mathbf{P}(\xi_2 = 5) = 1,$$

mindenesetre megnyugtató.

Internetes keresőprogramok az 90-es évek közepe óta vannak, az első ilyenek a WebCrawler, go.com, Yahoo, AltaVista. Viszont nem működtek túl jól. 1997 novemberében a 4 legnépszerűbb keresőprogram közül csak egy találta meg saját magát. Sergey Brin és Larry Page 1998-ban olyan algoritmust kerestek, mely értelmes módon, fontossági sorrendbe rendezi a keresési találatokat. Olyan objektív rangsort keresünk, mely a világháló szerkezetéből adódik.

Rendeljünk minden oldalhoz egy nemnegatív mennyiséget, a fontosságot. Feltehető, hogy a fontosságok összege 1. Tehát az i oldal fontossága $\pi_i \geq 0$, és $\sum_{i=1}^{5} \pi_i = 1$. Egy oldal akkor fontos, ha sokan hivatkoznak rá, és az is számít, hogy milyen fontos oldalak hivatkoznak rá. Innen jön az ötlet, hogy legyen egy oldal fontossága a rá mutató oldalak súlyozott fontosságainak összege. Azaz,

$$\begin{aligned}
\pi_1 &= \frac{1}{2}\pi_2 \quad \text{(1-be csak a 2-ből fut él, és onnan 2 él indul)} \\
\pi_2 &= \frac{1}{2}\pi_1 + \frac{1}{3}\pi_3 \quad \text{(2-be az 1-ből és a 3-ból fut él)} \\
\pi_3 &= \frac{1}{2}\pi_1 + \frac{1}{2}\pi_4 \\
\pi_4 &= \frac{1}{3}\pi_3 + 1 \cdot \pi_5 \\
\pi_5 &= \frac{1}{2}\pi_3 + \frac{1}{3}\pi_3 + \frac{1}{2}\pi_4.
\end{aligned}$$

Tehát van 5 egyenletünk és 5 ismeretlenünk. Sőt, még azt is tudjuk, hogy $\sum_{i=1}^{5} \pi_i = 1$. Ez a 6 egyenlet lineárisan függő (azaz van egy olyan egyenlet, ami a másik 5-ből következik), és így kapunk egyértelmű megoldást. A fenti egyenletrendszer mátrix alakban így néz ki:

$$\begin{pmatrix}
0 & \frac{1}{2} & 0 & 0 & 0 \\
\frac{1}{2} & 0 & \frac{1}{3} & 0 & 0 \\
\frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\
0 & 0 & \frac{1}{3} & 0 & 1 \\
0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0
\end{pmatrix}
\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{pmatrix}
=
\begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \\ \pi_5 \end{pmatrix}.$$

Tehát éppen a fenti mátrix 1 sajátértékéhez tartozó sajátvektort[15] keressük. Ez a 25 milliárd dolláros sajátvektor[16]. 2022-ben nagyjából 2 milliárd ($2 \cdot 10^9$) weboldal van, azaz egy ennyi egyenletből álló rendszert kellene megoldani. Ez azért sok.

Példánkban az egyenletrendszert megoldva $\pi_1 = 0,045$, $\pi_2 = 0,091$, $\pi_3 = 0,205$, $\pi_4 = 0,364$, $\pi_5 = 0,295$ adódik.

Jelölje a fenti mátrixot P. Vegyük észre, hogy P első oszlopában éppen

[15]Egy $A \in \mathbb{R}^{d \times d}$ mátrixnak a $\lambda \in \mathbb{R}$ szám és $\mathbf{x} \in \mathbb{R}^d$ oszlopvektor sajátérték–sajátvektor párja, ha $A\mathbf{x} = \lambda\mathbf{x}$. Ekkor a mátrixhoz tartozó lineáris transzformáció az $\mathbf{x}$ vektor annak egy számszorosába viszi.

[16]Kurt Bryan, Tanya Leise: The $25,000,000,000 eigenvector. The linear algebra behind Google

a $\mathbf{P}(\xi_1 = i)$, $i = 1, \ldots, 5$, valószínűségek szerepelnek. Ha kicsit számolunk,

$$P^2 = \begin{pmatrix} \frac{1}{4} & 0 & \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{1}{4} & 0 & \frac{1}{6} & 0 \\ 0 & \frac{1}{4} & \frac{1}{6} & 0 & \frac{1}{2} \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} & \frac{2}{3} & 0 \\ \frac{5}{12} & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{2} \end{pmatrix}$$

adódik. Az első oszlop éppen a $\mathbf{P}(\xi_2 = i)$, $i = 1, \ldots, 5$ értékek. Ez általában is igaz. Megmutatható, hogy $\lim_{n\to\infty} P^n$ létezik, és a határérték mátrix minden oszlopa megegyezik $(\pi_1, \ldots, \pi_5)^\top$ sajátvektorral. Igazából nem is kell mátrixot hatványozni, elég csak a $\xi_3, \xi_4, \ldots$ változókhoz tartozó valószínűségeket számolni, ahogy kezdtük. Brin és Page eredeti cikkükben[17] 50 iterációt javasolnak. Így működik a google PageRank algoritmusa.

A (ξ_n) sorozat egy Markov-lánc, aminek stacionárius eloszlása (π_i). Ez azt jelenti, hogy a fejezet elején ismertetett véletlen szörfös, ha sokat kattintgat, akkor a teljes idő π_1 részét tölti az 1-es oldalon, π_2 részét a 2-es oldalon, ... Ha még számolunk egy kicsit és meghatározzuk ξ_3, ξ_4, ξ_5 lehetséges értékeit és a hozzájuk tartozó valószínűségeket, akkor azt látjuk, hogy ez egyre közelebb van a (π_i) értékekhez.

Markov-láncok lesznek Sztochasztikus modellek kurzuson a mesterképzésben.

[17] Sergey Brin, Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine, 1998