EXTRÉM STATISZTIKÁK ÉS FIZIKAI ALKALMAZÁSAIK

Speciálkollégium
Fizikus MSc. és PhD. kurzus, 2014. II. félév
3-6. előadás

Tartalomjegyzék

1. Bevezető

Ezeken a diákon a valószínűségszámítás néhány elemét idézzük fel, ezzel az extrém érték statisztikák elméletéhez szükséges matematikai hátteret alapozzuk meg.

A gyakorló feladatok általában könnyűek, nem szükséges megoldásukat beadni, de azokat a vizsgán kérdezhetem. A házi feladatok megoldása beadható, ez a vizsga részleges teljesítésének számít.

A diákon hibákat felfedező hallgatók a vizsgán kedvezőbb elbírálásra számíthatnak.

Felhasznált irodalom: [1] Stuart Coles: An Introduction to Statistical Modeling of Extreme Values, Springer Series in Statistics, 2001 (alkalmazott statisztikai megközelítés); [2] Rényi Alfréd: Valószínűségszámítás, Tankönyvkiadó (matematika tankönyv számos példával); [3] Filip Lindskog: The Mathematics and Fundamental Ideas of Extreme Value Theory, http://www.math.ethz.ch/∼embrechts/RM/evtnotes.pdf (matematikai kurzus jegyzet, rendkívül lényegretörő).

2. Emlékeztető: A valószínűségszámítás és a statisztikus inferencia alapjai

2.1. Sűrűség-, és eloszlásfüggvények egy folytonos változó esetén

2.1.1. Sűrűségfüggvény

$P(x)$: Valószínűség sűrűség (probability density function, PDF) – fizikai szövegben gyakran eloszlásnak (distribution) nevezik. Az $[a, b]$ intervallum valószínűsége (mértéke)

$$\text{Prob}(a \leq x \leq b) = \int_a^b \mathrm{d}x\, P(x) \tag{2.1}$$

Diszkrét érték véges súllyal: ha x_0 valószínűsége $0 < p_0 \leq 1$, akkor a PDF-ben fellép $p_0\delta(x - x_0)$.

Ha pl. az a-ban delta-csúcs áll, akkor a (2.1) integrálás alsó határa $a - 0$.

Ha a végpontokban a PDF síma, akkor mindegy, hogy az intervallum zárt vagy nyílt.

Az eloszlás tartója, ahol a PDF nem zérus.

Változócsere: ha ismerjük x PDF-jét, akkor annak valamely $y(x)$ függvényének PDF-jét is kifejezhetjük. Legyen $y = y(x)$ invertálható, $x = x(y)$, ahonnan

$$P_y(y)|\mathrm{d}y| = P_x(x)|\mathrm{d}x| \quad \Rightarrow \quad P_y(y) = P_x(x(y)) \left| \frac{\mathrm{d}x(y)}{\mathrm{d}y} \right|. \tag{2.2}$$

Formálisan így is eljárhatunk

$$P_y(y) = \int \mathrm{d}x\, \delta(y - y(x)) P_x(x) \quad \Rightarrow \quad P_y(y) = P_x(x(y)) \left| \frac{\mathrm{d}x(y)}{\mathrm{d}y} \right|. \tag{2.3}$$

A "Dirac deltás" kifejezés az általános receptje annak, hogyan írjuk fel leszármaztatott változók eloszlását. Statisztikus fizikai számításokban gyakran alkalmazzuk.

2.1. Gyakorló feladat. *Terjesszük ki a fenti formulát, ha $y = y(x)$ nem invertálható, de szakaszonként az!*

Általában nem fogjuk jelölni az indexet, $P(x)$ az x változó PDF-je.

Várható értékek: $\langle f(x) \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, P(x) f(x)$, norma $\langle 1 \rangle = 1$.

2.1.2. Integrált eloszlásfüggvény

Integrated probability distribution function (IPDF), fizikai szövegekben integrált vagy kumulatív eloszlásnak hívják, matematikai szövegben eloszlásfüggvény, szokásosan $F(x)$-szel jelölik. Nálunk jelölése

$$M(x) = \int_{-\infty}^{x} \mathrm{d}y\, P(y), \tag{2.4}$$

a $[-\infty, x]$ intervallum valószínűsége (mértéke). Változócsere (egy-egy értékű) $M_y(y) = M_x(x(y))$. Nyilván

$$\mathrm{Prob}(a \le x \le b) = M(b) - M(a). \tag{2.5}$$

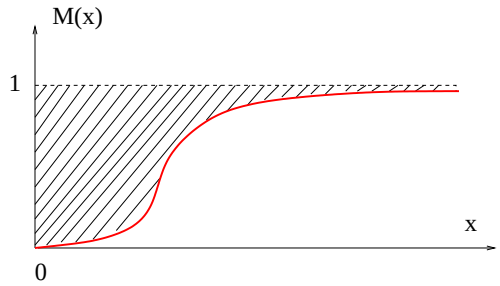
Ha a PDF delta-csúcsot tartalmaz, annak helyén az IPDF ugrik.

A várható érték számítható $M(x)$-szel is. Legyen a tartó a pozitív féltengely és $f(x)$

deriválható (a felírt integrálok végességét feltesszük)

$$\langle f(x)\rangle = \lim_{X\to\infty}\left\{\int_0^X \mathrm{d}x\, M'(x)f(x) = M(x)f(x)|_0^X - \int_0^X \mathrm{d}x\, M(x)f'(x)\right\}$$
$$= \lim_{X\to\infty}\left\{f(X) - \int_0^X \mathrm{d}x\, M(x)f'(x)\right\}$$
$$= \lim_{X\to\infty}\left\{\int_0^X \mathrm{d}x\, f'(x) + f(0) - \int_0^X \mathrm{d}x\, M(x)f'(x)\right\} = f(0) + \int_0^\infty \mathrm{d}x\,(1-M(x))f'(x). \tag{2.6}$$

Pl. az x várható értéke $\langle x\rangle = \int_0^\infty \mathrm{d}x\,(1-M(x))$. Noha a levezetéshez használtuk, hogy $M(x)$ deriválható, a végső képlet általánosabban igaz.



1. ábra. Az integrált eloszlás grafikai jelentése. Ha a tartó a pozitív félegyenes, akkor a jelölt terület, ha véges, az x várható értéke.

2.2. Gyakorló feladat. *Írjuk fel a várható értéket az IPDF-fel, ha az eloszlás tartója a valós tengely!*

2.2. Momentumok és kumulánsok

2.2.1. Generátor (karakterisztikus) függvények

Momentum generátor (a PDF Fourier-transzformáltja)

$$\Phi(z) = \left\langle e^{izx} \right\rangle, \qquad \left.\frac{\mathrm{d}^n \Phi(z)}{\mathrm{d}(iz)^n}\right|_{z=0} = \langle x^n \rangle = m_n \tag{2.7}$$

az n. momentum, ha ilyen létezik. Nyilván $\Phi(0) = 1 \Rightarrow m_0 = 1$. Ha Taylor-sorba fejthető, akkor

$$\Phi(z) = \sum_{n=0}^{\infty} \frac{m_n}{n!} (iz)^n. \tag{2.8}$$

Ha az $n \leq n_0$ momentumok léteznek, de az $n_0 + 1$-ik divergál, akkor a generátor z^{n_0} rendig a részletösszeggel közelíthető.

Kumuláns generátor

$$\Psi(z) = \ln \Phi(z) \underset{\substack{\uparrow \\ \text{ha Taylor}}}{=} \sum_{n=1}^{\infty} \frac{c_n}{n!} (iz)^n \tag{2.9}$$

Vegyük észre, hogy $\Psi(0) = 0$.

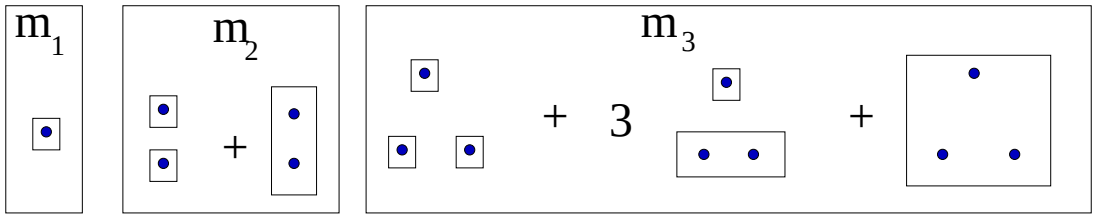
2.2.2. A momentumok és a kumulánsok közötti relációk

A generátorfüggvények Taylor-sorainak együtthatói között összefüggések találhatók. Az első három indexre könnyen beláthatjuk, hogy

$$\Phi(z) = e^{\Psi(z)} \quad \Rightarrow \quad m_1 = c_1, \quad m_2 = c_2 + c_1^2, \quad m_3 = c_3 + 3c_2c_1 + c_1^3, \ldots \tag{2.10}$$

$$\Psi(z) = \ln \Phi(z) \quad \Rightarrow \quad c_1 = m_1, \quad c_2 = m_2 - m_1^2, \quad c_3 = m_3 - 3m_2m_1 + 2m_1^3, \tag{2.11}$$

Magasabb indexekre mindez folytatható, de némi fáradsággal az általános reláció is felállítható. Ezt formula helyett grafikusan szemléltetjük a 2. ábrával.



2. ábra. A momentumok és kumulánsok relációjának grafikus illusztrálása. Az n pontból álló halmaz segítségével az m_n momentumot kifejezzük a c_k, $k = 1, \dots n$ kumulánsokkal. A k pontot tartalmazó keret c_k-t jelöli, az n pont egy partíciója pedig az itt fellépő c_k-k szorzatának feleltetendő meg. Az n. momentum az n pont összes lehetséges partíciója összegeként áll elő. A kombinatorikai együtthatók azt adják meg, hogy megkülönböztethető pontok esetén hányféleképp kapható ugyanolyan partíció. Harmadrendig (2.10) adódik.

2.1. Házi feladat. *Fejezzük ki az n-edik momentumot a kumulánsokkal és igazoljuk a 2. ábrán bemutatott receptet. (30%)*

2.2.3. Különleges momentumok

A centrális momentumok $\bar{m}_n = \langle (x - \langle x \rangle)^n \rangle$. Nyilván $\bar{m}_1 = 0$, továbbá könnyen látható, hogy $n = 2, 3$ mellett ezek éppen a c_n kumulánsok.

2.3. Gyakorló feladat. *Állítsuk elő a $\bar{m}_4$ és a c_4 mennyiségeket a momentumokkal. Ennek révén beláttuk, hogy általában különböznek.*

Matematikai szövegekben szokásos jelölés: átlag $\mathrm{E}(X) = \langle x \rangle = c_1$, variancia $\mathrm{Var}(X) = \langle (x - \langle x \rangle)^2 \rangle = c_2$. Az $x - \langle x \rangle$ átlaga zérus, a hozzá tartozó kumuláns generátorból hiányzik az $n = 1$ tag, egyébként azonos az x-ével.

Szórás: $c_2^{1/2}$, relatív szórás: $\frac{c_2^{1/2}}{c_1}$, ferdeség, skewness: $\beta_1 = \frac{c_3}{c_2^{3/2}}$, lapultság, kurtosis: $\beta_2 = \frac{c_4}{c_2^2} + 3 = \frac{\langle (x - \langle x \rangle)^4 \rangle}{c_2^2}$.

2.2.4. Nemnegatív változók generátor függvénye

Ha az x véletlen változó nemnegatív, akkor a Fourier transzformációval kapott momentum generátor ($\Phi(z)$) helyett célszerű a valós Laplace transzformáltat használni

$$G(s) = \left\langle e^{-sx} \right\rangle . \tag{2.12}$$

Formálisan $G(s) = \Phi(is)$, néhol csak ezt hívják karakterisztikus függvénynek. Nyilván

$$m_k = \left\langle x^k \right\rangle = (-1)^n G^{[n]}(0). \tag{2.13}$$

A kumuláns generátor $H(s) = \ln G(s) = \Psi(is)$

$$c_k = (-1)^n H^{[n]}(0). \tag{2.14}$$

2.2.5. Összeg eloszlása

Vizsgáljuk $x = x_1 + x_2$ eloszlását.

$$P(x) = \int \mathrm{d}x_1 \mathrm{d}x_2 \, \delta(x - x_1 - x_2) P_1(x_1) P_2(x_2) = \int \mathrm{d}y P_1(x-y) P_2(y) = (P_1 * P_2)(x), \tag{2.15}$$

ez a konvolúció. Vegyük észre, hogy $P_1 * P_2 = P_2 * P_1$.

2.4. Gyakorló feladat. *Mutassuk meg, hogy a momentum generátorok szorzódnak,* $\Phi(z) = \Phi_1(z)\Phi_2(z)$.

A kumuláns generátorok ebből következően $\Psi(z) = \Psi_1(z) + \Psi_2(z)$.

2.3. Diszkrét változók — az említés szintjén

Diszkrét értékkészletű esemény (k) és valószínűsége

$$k \sim p_k, \qquad \langle f_k \rangle = \sum_k f_k p_k. \tag{2.16}$$

PDF-ként előállítva Dirac delták sorát kapjuk

$$P(x) = \sum_k p_k \delta(x-k). \tag{2.17}$$

Generátor pl.

$$G(s) = \sum_k p_k e^{-sk}, \qquad (-1)^n G^{[n]}(0) = \langle k^n \rangle . \tag{2.18}$$

Kockavetés generátora: $G(s) = \frac{1}{6}(e^{-s} + e^{-2s} + \cdots + e^{-6s}) = \frac{e^{-s}}{6}\frac{1-e^{-6s}}{1-e^{-s}}, \qquad \langle k \rangle = 7/2.$

2.4. Néhány nevezetes eloszlás

2.4.1. Binomiális

n független bináris esemény, pl. p valószínűséggel „fej” és $1-p$-vel „írás”, esetén annak

a valószínűsége, hogy k alkalommal kapunk „fej”-et

$$p_k = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \sum_{k=0}^{n} p_k = 1. \tag{2.19}$$

Momentum generátor

$$G(s) \;=\; \sum_{k=0}^{n} \binom{n}{k} e^{-sk}\, p^k (1-p)^{n-k} = (1-p(1-e^{-s}))^n, \tag{2.20}$$

ahonnan a kumuláns generátor

$$H(s) \;=\; \ln G(s) = n \ln(1-p(1-e^{-s})). \tag{2.21}$$

A k várható értéke

$$-H'(s) = n\frac{pe^{-s}}{1-p(1-e^{-s})} \quad \Rightarrow \quad c_1 = \langle k \rangle = -H'(0) = np. \tag{2.22}$$

A második kumuláns

$$H''(s) = n\frac{p(1-p)e^{-s}}{(1-p(1-e^{-s}))^2} \quad \Rightarrow \quad c_2 = \langle k^2 \rangle - \langle k \rangle^2 = H''(0) = np(1-p). \tag{2.23}$$

Vegyük észre, hogy nagy n mellett az átlag $c_1 = O(n)$, az akörüli ingadozások $\sqrt{c_2} =$

$O(\sqrt{n})$, a relatív szórás $\frac{\sqrt{c_2}}{c_1} = O(n^{-\frac{1}{2}})$, azaz az eloszlás az átlag skáláján kicsúcsosodik.

2.4.2. Poisson

A binomiálisból az $n \to \infty, \quad p \to 0, \quad np \to \lambda$ átmenettel kapjuk

$$\binom{n}{k} p^k (1-p)^{n-k} \approx \frac{n^k}{k!} p^k (1-p)^n \approx \frac{\lambda^k}{k!} e^{-\lambda} = p_k. \tag{2.24}$$

Pl. hosszú felezési idejű ($p \to 0$) elem nagyszámú ($n \to \infty$) atomjának bomlását detektáljuk meghatározott idő alatt. Általában egymástól függetlenül bekövetkező események számának eloszlása adott időn belül Poisson.

2.5. Gyakorló feladat. *Állítsuk elő a $G(s) = \left\langle e^{-sk} \right\rangle$ generátort, s mutassuk meg, hogy ugyanezt kapjuk a binomiális eloszlás megfelelő limeszéből is. Igazoljuk, hogy a Poisson eloszlás mindegyik kumulánsa egyaránt $c_j \equiv \lambda$!*

2.4.3. Gamma

A formula azonos a Poissonéval, csakhogy a paraméter és a valószínűségi változó szerepe felcserélődik. Az egész k helyett megengedjük a valós $a > 0$ paramétert, $x > 0$ pedig a valószínűségi változó:

$$P(x; a) = \frac{x^{a-1}}{\Gamma(a)} e^{-x}. \tag{2.25}$$

A fenti sűrűségfüggvény normált! A függvénytáblázatokban szokásos alak x helyett annak lineáris transzformáltját tartalmazza, ez két további paramétert enged meg. A karakterisztikus függvény

$$G(s) = \left\langle e^{-sx} \right\rangle = (\Gamma(a))^{-1} \int_0^\infty \mathrm{d}x\, e^{-x(1+s)} x^{a-1} = (1+s)^{-a}, \tag{2.26}$$

ahonnan a kumuláns generátor $H(s) = -a \ln(1+s)$ és a kumulánsok $c_n = a(n-1)!$. Vegyük észre a következő tulajdonságot: ha x_1 és x_2 rendre a_1 és a_2 paraméterű Gamma eloszlásúak, akkor $x_3 = x_1 + x_2$ az $a_3 = a_1 + a_2$ paraméterű Gamma eloszlásnak tesz eleget.

2.6. Gyakorló feladat. *Számítsuk ki a Gamma eloszlás m_n momentumait!*

2.4.4. Gauss (normál)

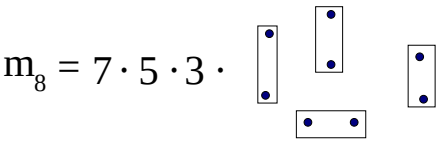
A másodiknál nagyobb kumulánsai zérusok, azaz $\Psi(z)$ másodfokú, ezért $\Phi(z)$ Gauss-függvény, s ennek $P(x)$ Fourier-transzformáltja is Gauss. Minden magasabb momentum kifejezhető az első kettővel.

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) = N(x;m,\sigma), \tag{2.27}$$

$$\Psi(z) = imz - \frac{1}{2}\sigma^2 z^2, \tag{2.28}$$

$$M(x;m,\sigma) = \int_{-\infty}^{x} \mathrm{d}y\, N(y;m,\sigma) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x-m}{\sigma\sqrt{2}}\right)+1\right]. \tag{2.29}$$

Az $m=0,\ \sigma=1$ eset a sztenderd normál eloszlás.



3. ábra. Ha minden kumuláns zérus kivéve c_2, akkor a 2. ábrán illusztrált eljárás alapján a momentumok $m_{2n} = \langle x^{2n}\rangle = (2n-1)!!\, c_2^n$.

2.7. Gyakorló feladat. *Mutassuk meg, hogy a binomiális eloszlásból, tetszőleges $0 < p < 1$ mellett, az $n,k \to \infty$ limeszben kapjuk az $m = np$, $\sigma^2 = np(1-p)$ Gauss-függvényt (Moivre & Laplace). (1) Használjuk a Stirling formulát és tegyük fel, hogy $z = k - np = O(\sqrt{n})$! E feltétel jogosságát vizsgáljuk utólag. (2) Használjuk a kumuláns generátort, amely nagy n mellett közelíthető a minimuma körüli értékeivel.*

2.4.5. χ^2 (khi négyzet)

Ha $x_1,\ldots,x_k$ (sztenderd) normál eloszlású, akkor $z = x_1^2 + \cdots + x_k^2$ eloszlása a χ_k^2. Ez a statisztikus modellezés szempontjából fontos, a hibák négyzetösszegének eloszlását adja, ha az egyes hibák gaussiak.

$k = 1, \quad z = x^2,\ z \geq 0$:

$$M_1(z) = \operatorname{erf}\left(\sqrt{\frac{z}{2}}\right) \tag{2.30}$$

$$P_1(z) = \frac{1}{\sqrt{2\pi z}} \exp\left(-\frac{z}{2}\right), \tag{2.31}$$

a karakterisztikus függvény pedig

$$G_1(s) = \left\langle e^{-sz} \right\rangle = \int_0^\infty P_1(z) e^{-sz} dz = (1+2s)^{-1/2}. \tag{2.32}$$

$k > 1$: A $k = 1$ eloszlás önmagával vett konvolúciójával kapható. A generátorfüggvény

$$G_k(s) = G_1^k(s) = (1+2s)^{-k/2}. \tag{2.33}$$

2.8. Gyakorló feladat. *Mutassuk meg, hogy a χ^2_k eloszlás azonos a $\frac{k}{2}$ paraméterű Gamma eloszlással az $\frac{x}{2}$ változóban.*

$k \to \infty$: később látjuk, hogy Gauss-hoz tart a megfelelő skálán.

2.4.6. Gausshoz tartó eloszlások

Különböző eloszlások tarthatnak a gaussihoz, midőn valamely paraméter limeszhez tart, s ezzel egyidejűleg megfelelő lineáris változócserét végzünk. A kumuláns sorfejtésből könnyen észrevehető, ha az eloszlás a gaussihoz tart.

Binomiális: $n \to \infty$. Vezessük be az

$$x = \frac{k - \langle k \rangle}{\sqrt{n}} = \frac{k}{\sqrt{n}} - p\sqrt{n} \tag{2.34}$$

változót, s menjen $n, k \to \infty$. Az x kumuláns generátora

$$\begin{aligned} \Psi_x(q) &= \ln \left\langle e^{iqx} \right\rangle = -iqp\sqrt{n} + \ln \left\langle e^{iqk/\sqrt{n}} \right\rangle = -iqp\sqrt{n} + H_k(-iq/\sqrt{n}) \\ &= -iqp\sqrt{n} + n \ln\left[1 - p\left(1 - e^{iq/\sqrt{n}}\right)\right] = -p(1-p)\frac{q^2}{2} + O(n^{-\frac{1}{2}}). \end{aligned} \tag{2.35}$$

Itt $H_k(s)$ a k binomiális eloszlású változó kumuláns generátora, amelyet (2.21) ad meg. Tehát a határeloszlás az $N(x; 0, \sqrt{p(1-p)})$ normál.

2.9. Gyakorló feladat. *Mutassuk meg, ha eredetileg az*

$$x = \frac{k - c_1}{\sqrt{c_2}} \tag{2.36}$$

új változót vezettük volna be, akkor a sztenderd normál lenne a határeloszlás.

Gamma: $a \to \infty$. Tudjuk, hogy

$$H_x(s) = -a \ln(1 + s) = a \sum_{n=1}^{\infty} \frac{(-s)^n}{n}. \tag{2.37}$$

Bevezetve az

$$y = \frac{x - c_1}{\sqrt{c_2}} = \frac{x - a}{\sqrt{a}} \tag{2.38}$$

új változót, nyerjük

$$\Psi_y(q) = H_x(-iq/\sqrt{c_2}) - iqc_1/\sqrt{c_2} = a \sum_{n=2}^{\infty} \frac{1}{n} \left(\frac{iq}{\sqrt{a}} \right)^n. \tag{2.39}$$

A nagy a limeszben tehát a sztenderd normál adódik:

$$\Psi_y(q) = -\frac{q^2}{2}. \tag{2.40}$$

Chi-négyzet (χ_k^2): $k \to \infty$. Miután a χ_k^2 az $a = k/2$ paraméterű Gamma eloszlással egyenlő, a nagy k limeszben az

$$y = \frac{x - c_1}{\sqrt{c_2}} \tag{2.41}$$

változóban a sztenderd normál eloszlást kapjuk.

2.5. Szimultán és feltételes eloszlások

Szimultán (joint, multivariate) eloszlás: $P(x_1, \ldots, x_n) = P(\boldsymbol{x})$.

Független változók: $P(\boldsymbol{x}) = \prod_{i=1}^{n} P_i(x_i)$.

Független, azonos eloszlású (independent, identically distributed, i.i.d.) változók: $P_i(x) \equiv P(x)$.

Várható értékek: $\langle f(\boldsymbol{x})\rangle = \int \mathrm{d}^n x\, f(\boldsymbol{x})P(\boldsymbol{x})$.

Kovariancia: $\sigma_{ij} = \langle x_i\, x_j\rangle - \langle x_i\rangle \langle x_j\rangle$.

Redukált eloszlás (n-ről k változóra): $P(x_1,\ldots,x_k) = \int \mathrm{d}x_{k+1}\ldots\mathrm{d}x_n\, P(\boldsymbol{x})$.

Egy változóra a marginális eloszlás: $P(x_k) = \int \mathrm{d}x_1\ldots\mathrm{d}x_{k-1}\mathrm{d}x_{k+1}\ldots\mathrm{d}x_n\, P(\boldsymbol{x})$.

Feltételes eloszlás (x_1 eloszlása, ha x_2 adott):

$$P(x_1|x_2) = \frac{P(x_1,x_2)}{\int \mathrm{d}x_1\, P(x_1,x_2)} = \frac{P(x_1,x_2)}{P(x_2)}. \tag{2.42}$$

A $P(x_1|x_2)$-ben x_1 valószínűségi változó, x_2 pedig paraméter, ezért az előbbiben normált $\int P(x_1|x_2)\mathrm{d}x_1 = 1$.

Minden eloszlás valójában feltételes, legfeljebb természetesnek tekintjük s ezért nem mondjuk ki a feltételt.

2.6. Határeloszlások

Eddig is láttunk határeloszlásokat: a Poisson és a Gauss a binomiálisból bizonyos limeszekben adódtak. Továbbá a bevezetőben bemutatott extrém érték eloszlások is

határesetként adódtak megfelelő skálázással.

Határozzuk meg az $X = \sum_{i=1}^n x_i$ összeg eloszlását, ha az x_i-k i.i.d. változók, és 3. kumulánsuk véges. A közös kumuláns generátor tehát harmadrendig

$$\psi(z) = ic_1 z - \frac{1}{2}c_2 z^2 - \frac{1}{6}ic_3 z^3 + \ldots. \tag{2.43}$$

Korábban láttuk, hogy az összeg kumuláns generátora az egyes tagok kumuláns generátorainak összege (ld. 2.2.5. fejezet)

$$\ln\left\langle e^{iXz}\right\rangle = \Psi(z) = n\psi(z) = nc_1 iz - \frac{n}{2}c_2 z^2 - i\frac{n}{6}c_3 z^3 + \ldots. \tag{2.44}$$

Láthatóan $\langle X\rangle = nc_1 = n\langle x\rangle$, $\mathrm{Var}(X) = nc_2 = n\mathrm{Var}(x)$. Vezessük be az

$$Y = \frac{X - nc_1}{\sqrt{nc_2}} \tag{2.45}$$

új változót, ennek generátorát úgy kapjuk, hogy Ψ-ből a z-vel arányos tagot elhagyjuk

és áttérünk a $w = z\sqrt{nc_2}$ változóra. Ekkor az $n \to \infty$ limeszben

$$\ln \left\langle e^{iYw} \right\rangle = \Psi(w) = -\frac{1}{2}w^2 - \frac{i}{6}\frac{nc_3}{(nc_2)^{3/2}}w^3 + \ldots \quad \to \quad -\frac{1}{2}w^2, \tag{2.46}$$

azaz a sztenderd normál eloszlást kaptuk.

A fentiekben azt mutattuk meg, hogy az X változó "ferdeség"-e zérushoz tart.

A levezetéshez valójában elegendő azt feltenni, hogy a kumuláns generátor másodrendig sorbafejthető, s a maradék magasabb rendű.

Tájékoztatásul a matematikai irodalomból idézzük annak szükséges és elegendő feltételét, hogy azonos eloszlású, független változók összege határesetben gaussi legyen: a $P(x)$ PDF-re teljesüljön, hogy $x \to \infty$ mellett

$$\frac{x^2 \left(\int_{-\infty}^{-x} + \int_{x}^{\infty} \right) P(y)\mathrm{d}y}{\int_{-x}^{x} y^2\, P(y)\mathrm{d}y} \to 0. \tag{2.47}$$

2.2. Házi feladat. *(i) Mutassuk meg, hogy ha a második momentum véges, akkor (2.47) fennáll. (ii) Ha a PDF nagy y mellett y^{-3} szerint cseng le, akkor nincs második momentum, de (2.47) teljesül. Határozzuk meg a kumuláns generátor vezető tagját (az egyszerűség kedvéért tegyük, fel, hogy az első kumuláns zérus), s próbáljuk meg a határeloszlást a fenti levezetéssel analóg módon kiszámítani. (15%)*

Különböző eloszlású, független változók összege: a j. eloszlás kumulánsai legyenek $c_{j,n}$, ekkor

$$\langle X \rangle = \sum_{j=1}^{n} c_{j,1}, \;\; \mathrm{Var}(X) = \sum_{j=1}^{n} c_{j,2}, \tag{2.48}$$

s a magasabb kumulánsokat is hasonlóan kapjuk (ha léteznek). A Ψ-ből „elhagyjuk" az átlagot és skálázunk $w = z\sqrt{\mathrm{Var}(X)}$ szerint. A kumuláns generátor

$$\Psi(w) = -\frac{1}{2}w^2 - \frac{i}{6}\frac{\sum_j c_{j,3}}{(\sum_j c_{j,2})^{3/2}}w^3 + \frac{1}{24}\frac{\sum_j c_{j,4}}{(\sum_j c_{j,2})^2}w^4 + \ldots. \tag{2.49}$$

Innen a sztenderd normál eloszlást kapjuk, ha az $n \to \infty$ limeszben a kvadratikusnál magasabb tagok zérushoz tartanak. Ljapunov-feltétel: a második kumulánsok átlaga pozitív és a harmadiké nem divergál ("túl erős", elégséges de nem szükséges),

$$\frac{1}{n}\sum_j c_{j,2} \to C_2 > 0, \quad \text{és} \quad \frac{1}{n}\sum_j c_{j,3} \to C_3, \quad |C_3| < \infty. \tag{2.50}$$

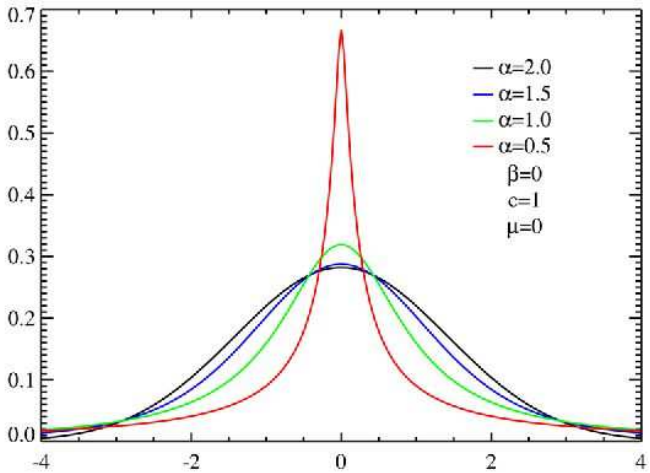
A centrális határeloszlás tétele (central limit theorem, CLT) hagyományosan annak a rigorózus megfogalmazása, hogy bizonyos feltételek mellett nagyszámú véletlen változó összege a limitben gaussi eloszlású.

Általánosítás x^{-3} hatványnál lassabban lecsengő PDF-ekre. Ha a kumuláns generátor $\psi_j(z) = c_{j,1} iz - c_{j,\alpha}|z|^\alpha + \ldots$, ahol $0 < \alpha < 2$, rögzített, akkor a határeloszlás generátorát a

$$\Psi(w) \approx -|w|^\alpha \tag{2.51}$$

alakra skálázzuk. Ez a forma invariáns a konvolúcióval szemben, és a szimmetrikus Lévy eloszlást definiálja

$$P_\alpha(x) = \int \frac{\mathrm{d}w}{2\pi} \cos(xw)\, e^{-|w|^\alpha} \tag{2.52}$$



4. ábra. Szimmetrikus Lévy-sűrűségek.

Gyakorlati szempontból azt mondhatjuk, ha a sűrűségfüggvény nagy $|x|$-re az $|x|^{-\alpha-1}$ szerint cseng le, akkor $0 < \alpha < 2$ esetén a határeloszlás az α indexű Lévy-függvény. Ha a lecsengés különböző kitevőkkel megy pozitív ill. negatív x-re, akkor a kisebbik számít.

Az $\alpha = 1$ mellett a határeloszlás Cauchy-féle

$$P(x) = [\pi(1 + x^2)]^{-1}. \tag{2.53}$$

A képlet a Lorentz-görbe.

Ha $\alpha \geq 2$, vagy a lecsengés minden hatványnál gyorsabb, beleértve azt az esetet is, amikor az eloszlás tartója korlátos, a határeloszlás gaussi.

2.6.1. Stabil eloszlások

A centrális határeloszlásokat önmagukkal konvolválva az eredeti határeloszlást kapjuk vissza, lineáris változócsere erejéig. A követelmény általános alakja, hogy minden m, σ párhoz található olyan m', σ', melyre

$$P(x) * \frac{1}{\sigma} P\left(\frac{x-m}{\sigma}\right) = \frac{1}{\sigma'} P\left(\frac{x-m'}{\sigma'}\right). \tag{2.54}$$

Pl. Gauss eloszlásra $m' = m, \sigma' = \sqrt{1+\sigma^2}$, míg a Cauchy eloszlásra $m' = m, \sigma' = 1+\sigma$. Tájékoztatásul közöljük, hogy a kumuláns generátorok nyelvén megadható a fenti feltétel általános megoldása

$$\Psi(z) = iaz - b|z|^\alpha \left(1 + ic\,\mathrm{sgn}(z)\, f_\alpha(z)\right), \tag{2.55}$$

ahol $0 < \alpha \leq 2$, $b > 0$ és $-1 \leq c \leq 1$, továbbá

$$f_\alpha(z) = \begin{cases} \tan\dfrac{\pi\alpha}{2}, & \text{ha} \quad \alpha \neq 1 \\ \dfrac{2}{\pi}\log|z|, & \text{ha} \quad \alpha = 1 \end{cases} \tag{2.56}$$

A szimmetrikus Lévy határeloszlást a $c = 0$ esetben kapjuk.

Vegyük észre, hogy centrális határeloszlásnak ki kell elégítenie a (2.54) követelményt, de innen még nem következik, hogy (2.54) minden megoldása más eloszlások határa is egyben.

2.7. A statisztikus modellezés elemei

Adatsorok alapján azok eloszlását szeretnénk meghatározni. Valamilyen előzetes, prior feltételezésre tehetünk, pl. sejtjük az eloszlásfüggvény típusát, s ennek paramétereit számítjuk. Az is lehetséges, hogy előfeltevéssel nem élünk, s az eloszlás momentumait/kumulánsait próbáljuk becsülni.

2.7.1. Empirikus átlagok

Tegyük fel, hogy az $\{x_i\}_{i=1}^n$ adatok iid véletlen számok valamely eloszlás szerint. Ennek kumulánsai c_i, az első kettő szokásosan jelölve $c_1 = m,\ c_2 = \sigma^2$. Vezessük be az egzakt m, σ^2 paraméterek becslése céljából az empirikus $m_{\rm emp}$ átlagot és $\sigma^2_{\rm emp}$ varianciát

$$m \approx m_{\rm emp} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{2.57}$$

$$\sigma^2 \approx \sigma^2_{\rm emp} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_{\rm emp})^2, \quad (n > 1). \tag{2.58}$$

Az empirikus paramétereket azért definiáltuk ily módon, mert átlagaik az egzakt értékeket adják. Hangsúlyozzuk, hogy itt általános eloszlást engedtünk meg, azaz nem csak gaussi lehet például, de természetesen Gauss-eloszlásra is fennállnak a fenti definíciók.

2.10. Gyakorló feladat. *Mutassuk meg, hogy* $\langle m_{\rm emp}\rangle = m$, $\langle \sigma^2_{\rm emp}\rangle = \sigma^2$. *A második egyenlőség indokolja az empirikus szórás definiáló relációjában található* $n-1$ *osztót.* Vegyük észre, hogy a fenti empirikus mennyiségek akkor is végesek, ha bármelyikük egzakt értéke végtelen. Ebben semmiféle rejtély nincs, ismételten előállítva n darab új

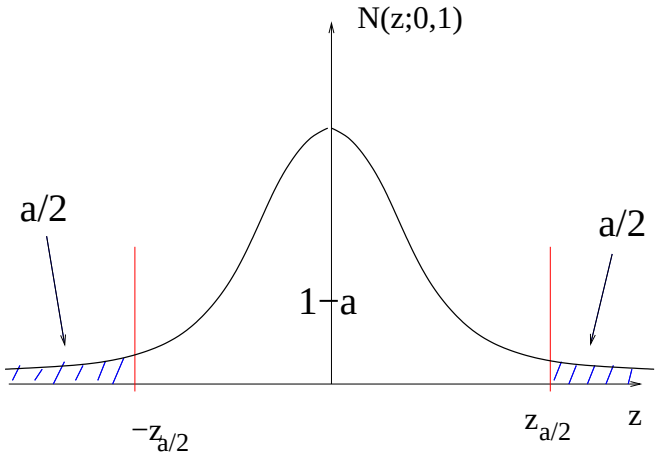
x_i értéket, az empirikus paraméter átlaga az ismétlésekre véve az egzakthoz fog tartani. Ha netán emez végtelen, akkor az empirikus paraméter empirikus átlaga is divergál nagyszámú ismétlés esetén.

2.3. Házi feladat. *Határozzuk meg az empirikus* $m_{\rm emp}$ *és* $\sigma^2_{\rm emp}$ *mennyiségek szórását az átlagaik körül az egzakt* c_i *kumulánsokkal kifejezve! (15%)*

2.4. Házi feladat. *Írjuk fel az empirikus harmadik kumulánst, melynek átlaga éppen az egzakt* c_3*! (10%) Javasoljunk empirikus formulát a ferdeségre. Ennek átlaga az egzakt ferdeség? (5%) Adjunk formulát az empirikus lapultságra. (15%)* Mekkora az empirikus harmadik kumuláns szórása? (15%)

2.7.2. Konfidencia intervallum

Az empirikus paraméterek csupán becslések. Hogyan jellemezhetjük a becslések jóságát?



5. ábra. A sztenderd normál eloszlás $1-a$ konfidenciaintervalluma. Az egyes tartományok súlyai $a/2, 1-a, a/2$.

Konfidencia intervallum: olyan tartomány, amelybe a jósolt paraméter adott, mondjuk $1-a$ valószínűséggel esik. Például, sztenderd normál eloszlás esetén az

$$a/2 = \int_{-\infty}^{-z_{a/2}} \mathrm{d}z\, N(z) \equiv M(-z_{a/2}) \tag{2.59}$$

transzcendens egyenlet megoldása adja az $1-a$ (azaz $100(1-a)\%$-hoz tartozó), átlagra centrált konfidencia intervallumát: $(-z_{a/2}, z_{a/2})$.

Elnevezés: A z_p az $1-p$ valószínűséghez tartozó kvantilis, azaz $1-p = M(z_p)$. A p kvantilise $-z_p$.

2.7.3. Egyszerű példa: Gauss eloszlás paramétereinek meghatározása

Ha az iid $\{x_i\}_{i=1}^n$ adatsort normál eloszlás generálta, akkor adjuk meg a konfidencia intervallumokat a fentebb bevezetett empirikus paraméterek körül.

Tudjuk, hogy (minden n-re) az m_{emp} normál eloszlású m, σ^2/n kumulánsokkal, ezért az alábbi z változó sztenderd normál eloszlású

$$z = \frac{m_{\mathrm{emp}} - m}{\sigma/\sqrt{n}} \;\sim\; N(z;0,1) \equiv N(z). \tag{2.60}$$

Ennek alapján empirikus konfidencia intervallumot adhatunk m-re. A Gauss eloszlású empirikus átlagra $1-a$ valószínűséggel teljesül

$$m - z_{a/2}\sigma/\sqrt{n} \le m_{\mathrm{emp}} \le m + z_{a/2}\sigma/\sqrt{n}. \tag{2.61}$$

Innen az σ^2_{emp} empirikus szórást beírva kapjuk ismert m_{emp} esetén az m-re vonatkozó becslést

$$m_{\text{emp}} - z_{a/2}\sigma_{\text{emp}}/\sqrt{n} \leq m \leq m_{\text{emp}} + z_{a/2}\sigma_{\text{emp}}/\sqrt{n}. \tag{2.62}$$

Felhívjuk a figyelmet arra, hogy ehhez a konfidenciaintervallumhoz csupán közelítőleg tartozik az $1-a$ valószínűség. A valódi valószínűség ennél kisebb, hiszen az empirikus szórás is becslésből származott.

2.5. Házi feladat. *Becsüljük meg a szórás $1-a$ konfidenciaintervallumát nagy n mellett! Használjuk az σ^2_{emp}-re vonatkozóan a centrális határeloszlás tételét. Vegyük észre, hogy a (2.61)-ben szereplő egzakt szórás helyére (2.62)-ben az empirikust írtuk. Mennyire csökken ezáltal a konfidenciaintervallumhoz tartozó valószínűség $1-a$ értékhez képest? (10-10%)*

2.7.4. Hipotézis valószínűsége – Bayes elve

Okozza az $\mathcal{A}$ (jelenség, törvény, fizikai mennyiség, paraméter érték) a $\mathcal{B}$ mért adatokat. Tegyük fel $P(\mathcal{B}|\mathcal{A})$-t ismertnek, azaz ha ismerjük az okot, akkor az adatok eloszlását ki tudjuk számítani. Most azonban a fordított helyzettel állunk szemben: ismerjük $\mathcal{B}$-t,

ennek alapján mekkora valószínűséggel tehetjük fel mögötte az $\mathcal{A}$ okot? Bayes formulája

$$P(\mathcal{A}|\mathcal{B}) = \frac{P(\mathcal{A},\mathcal{B})}{P(\mathcal{B})} = \frac{P(\mathcal{B}|\mathcal{A})P(\mathcal{A})}{\sum_{\mathcal{A}} P(\mathcal{B}|\mathcal{A})P(\mathcal{A})} \tag{2.63}$$

matematikai trivialitás, de értelme szerint a statisztikus inferencia alaprelációja. Ha ismert a $\mathcal{B}$ következmény, a formula megadja az azt kiváltó lehetséges $\mathcal{A}$ hipotézisekhez rendelhető valószínűségeket.

"Input": előzetes, prior, az adatok ismerete nélküli $P(\mathcal{A})$ valószínűség. Ez az eljárás érzékeny eleme, a prior valószínűség gyakran konvenció kérdése.

2.6. Házi feladat. *Valamely részecske tömegére a $P_{\rm emp}(m_{\rm emp})$ empirikus eloszlást mérik, ahol $m_{\rm emp}$ a kísérleti (esetleg más paramétereken keresztül visszaszámolt) tömeg. A szokásos eljárás az, hogy a $P_{\rm emp}(m)$ eloszlást tekintik egyben a valódi m tömegre vonatkozó hipotézis valószínűségének. Lehetséges lenne, hogy itt prior valószínűségeloszlás nélkül kapunk eredményt? (10%)*

2.7.5. Példák

1. Bináris kísérlet.

Kísérlet kimenetele lehet p valószínűséggel „1” és $1-p$ -vel „0”, azonban nem ismerjük p-t. Független kísérletek sorozatából nyert adatsor, pl. 00101110101110011000 10 alapján kívánjuk p-t megbecsülni. Végezzünk n kísérletet, melyekben k alkalommal kapunk 1-et: az $\mathcal{A}$ ok most p, a $\mathcal{B}$ mért adat pedig k. Adott k esetén milyen valószínűséggel állíthatjuk, hogy p volt a kísérletsorozatban az „1” bekövetkezésének valószínűsége?

A k binomiális eloszlást követ

$$P(\mathcal{B}|\mathcal{A}) = P(k|p) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{2.64}$$

s tegyük fel, hogy a prior $P(p) \equiv 1,\ (0 \le p \le 1)$. A keresett valószínűség

$$P(\mathcal{A}|\mathcal{B}) = P(p|k) = \frac{P(k|p)}{\int_0^1 \mathrm{d}p\, P(k|p)}. \tag{2.65}$$

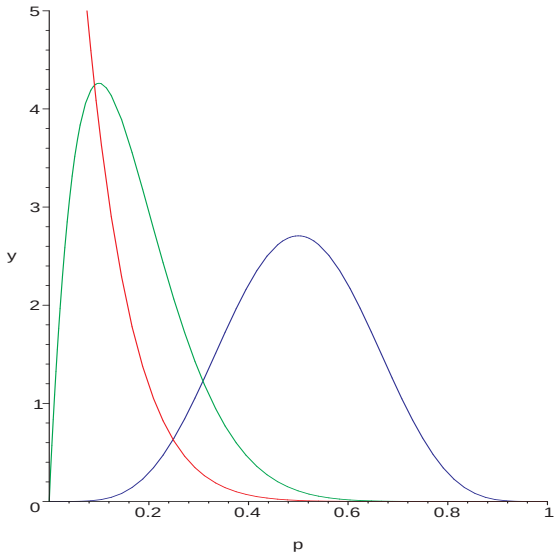
Ez a binomiális eloszlás p-re normálva. Számítsuk ki a normálási faktort

$$\int_0^1 \mathrm{d}p\, p^k (1-p)^{n-k} = B(k+1, n-k+1) = \frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \frac{k!(n-k)!}{(n+1)!} \tag{2.66}$$

(itt B neve béta-függvény), ahonnan a hipotézis valószínűsége

$$P(p|k) = (n+1)\binom{n}{k} p^k (1-p)^{n-k}. \tag{2.67}$$

Szimmeria:
$P(p|k) = P(1-p|n-k)$
Ábra: $n = 10$ mellett a $P(p|k)$ a $k = 0, 1, 5$ esetekben.



Az átlag és a szórás

$$\langle p \rangle = \frac{k+1}{n+2}, \qquad \Delta^2 p = \frac{(k+1)(n-k+1)}{(n+2)^2(n+3)}. \tag{2.68}$$

Vegyük észre, hogy noha $\frac{k}{n}$ átlaga p, az átlagos p az adatok alapján nem $\frac{k}{n}$ mért értéke! Ennek jelentősége az, hogy homogén 0 jelsorozat, azaz $k = 0$ esetén $\langle p \rangle = \frac{1}{n+2} > 0$. Ez lehet kicsi, de nem zárhatjuk ki, hogy a p pozitív volt. Homogén 1 jelek pedig

megengedik az egynél kisebb p-t is, $\langle p \rangle = \frac{n+1}{n+2} < 1$.

2.11. Gyakorló feladat. *Határozzuk meg a Δp-hez tartozó konfidencia szintet! Azaz milyen valószínűséggel esik a p az átlagtól egyszeres szórás távolságon belülre?*

2. Mekkora valószínűséggel kel fel holnap reggel a nap?

Tegyük fel, hogy eddig n-szer ismerten felkelt (ékírás ill. hieroglifák 5000 évesek, vagy használjuk a Föld életkorát, cca. 5 Gév). Legyen a napkelte valószínűsége p, ha ezt időben állandónak és napkeltéket egymástól független eseményeknek tekintjük, akkor az előzőeket a $k = n$ esetre alkalmazva

$$\begin{aligned} P(p|n) &= (n+1)p^n, \\ \langle p \rangle &= \frac{n+1}{n+2} \approx 1 - \frac{1}{n}, \quad \Delta p \approx \frac{1}{n}. \end{aligned} \tag{2.69}$$

A $p = 1$ közelítőleg egyszeres szórásra van az átlagtól.

2.7. Házi feladat. *Tegyük fel, hogy egységnyi idő alatt k radioaktív bomlást mérünk. Milyen valószínűséggel mondhatjuk, hogy a bomlási állandó λ? (10%)*

3. Hamis vagy nem hamis?

Egy dobókockáról annyit tudunk, hogy vagy hamis, ezesetben minden dobásra hatost

mutat (ólmozott, esetleg minden oldalán hatos áll), vagy "tiszta", s ekkor $1/6$ valószínűséggel mutatja bármelyik oldalát. Legyen n dobás eredménye hatos, milyen valószínűséggel állíthatjuk, hogy a kocka hamis ill. tiszta?

Rövid megoldás: a hamis kockával 6^n féleképpen kaphattuk a hatosokat, míg a tisztával egyféleképp. A hamis ill. tiszta esetek száma osztva az összes esettel adja

$$P(\text{hamis}) = \frac{6^n}{6^n + 1}, \qquad P(\text{tiszta}) = \frac{1}{6^n + 1}. \tag{2.70}$$

Látszólag nem használtunk prior valószínűségeket, hogyan lehetséges ez?

2.8. Házi feladat. *Oldjuk meg a problémát a Bayes formula alapján!(15%)*