Tómács Tibor

Matematikai statisztika

ÚJ SZÉCHENYI TERV

Eszterházy Károly Főiskola
Matematikai és Informatikai Intézet

Tómács Tibor

Matematikai statisztika

Eger, 2012

Szerző:
Dr. Tómács Tibor
főiskolai docens
Eszterházy Károly Főiskola

Bíráló:
Dr. Sztrik János
egyetemi tanár
Debreceni Egyetem

Készült a TÁMOP-4.1.2-08/1/A-2009-0038 támogatásával.

Tartalomjegyzék

Előszó

Ez a tananyag az egri Eszterházy Károly Főiskola matematikai statisztika előadásaiból készült, melyet matematika tanárszakos és programtervező informatikus hallgatóknak szánunk.

Az összeállításnál nem volt cél a matematikai statisztika összes fontos ágának ismertetése, inkább arra törekedtünk, hogy a taglalt témakörök mindegyikére kellő idő jusson az egy féléves kurzus alatt.

A *Valószínűségszámítás* című fejezet nem kerül ismertetésre a kurzus idején. A célja azoknak a fontos fogalmaknak az összefoglalása, melyekre szükségünk lesz a matematikai statisztika megértéséhez. Ennek átismétlését az Olvasóra bízzuk. Ezen fejezet másik célja, hogy a valószínűségszámítás és a matematikai statisztika szóhasználatát és jelöléseit összehangoljuk. A jelöléseket külön is összegyűjtöttük.

A szükséges definíciókon, tételeken és bizonyításokon túl, elméleti számításokat igénylő feladatokat is megoldunk. Ezek gyakorlatilag olyan tételek, amelyeknek a bizonyításán érdemes önállóan is gondolkodni, mielőtt a megoldást elolvasnánk.

Ehhez a tananyaghoz kapcsolódik Tómács Tibor [17] jegyzete, amely a gyakorlati órák témáit dolgozza fel. Itt számítógéppel megoldható gyakorlatokat találunk. A statisztikában szokásos táblázatokat nem mellékeljük, mert az ezekben található értékeket szintén számítógéppel fogjuk kiszámolni.

Jelölések

Általános

$\mathbb{N}$	a pozitív egész számok halmaza
$\mathbb{R}$	a valós számok halmaza
$\mathbb{R}^n$	$\mathbb{R}$-nek önmagával vett n-szeres Descartes-szorzata
$\mathbb{R}_+$	a pozitív valós számok halmaza
(a, b)	rendezett elempár vagy nyílt intervallum
$\simeq$	közelítőleg egyenlő
$[x]$	az x valós szám egész része
f^{-1}	az f függvény inverze
$\lim_{x \to a+0} f(x)$	az f függvény a-beli jobb oldali határértéke
$A^\top$	az A mátrix transzponáltja
A^{-1}	az A mátrix inverze
$\det A$	az A mátrix determinánsa

Valószínűségszámítás

$(\Omega, \mathcal{F}, \mathrm{P})$	valószínűségi mező
$\mathrm{P}(A)$	az A esemény valószínűsége
$\mathrm{E}\,\xi$	ξ várható értéke
$\mathrm{E}(\xi \mid \eta)$	feltételes várható érték
$\mathrm{E}(\xi \mid \eta = y)$	feltételes várható érték
$\mathrm{D}\,\xi,\ \mathrm{D}^2\,\xi$	ξ szórása illetve szórásnégyzete
$\mathrm{cov}(\xi, \eta)$	kovariancia
$\mathrm{corr}(\xi, \eta)$	korrelációs együttható
φ	a standard normális eloszlás sűrűségfüggvénye
Φ	a standard normális eloszlás eloszlásfüggvénye
Γ	Gamma-függvény
I_A	az A esemény indikátorváltozója
$\mathbf{Bin}(r; p)$	az r-edrendű p paraméterű binomiális eloszlású valószínűségi változók halmaza
$\mathbf{Exp}(\lambda)$	a λ paraméterű exponenciális eloszlású valószínűségi változók halmaza
$\mathbf{Norm}(m; \sigma)$	az m várható értékű és σ szórású normális eloszlású valószínűségi változók halmaza

$\mathbf{Norm}_d(m; A)$	az m és A paraméterű d-dimenziós normális eloszlású valószínűségi változók halmaza
$\mathbf{Gamma}(r; \lambda)$	az r-edrendű λ paraméterű gamma-eloszlású valószínűségi változók halmaza
$\mathbf{Khi}(s)$	az s szabadsági fokú khi-négyzet eloszlású valószínűségi változók halmaza
$\mathbf{t}(s)$	az s szabadsági fokú t-eloszlású valószínűségi változók halmaza
$\mathbf{F}(s_1; s_2)$	az s_1 és s_2 szabadsági fokú F-eloszlású valószínűségi változók halmaza
$F \sim \mathbf{V}$	Ha ξ valószínűségi változó, és $\mathbf{V}$ a ξ-vel azonos eloszlású valószínűségi változók halmaza, akkor ez azt jelöli, hogy F a $\mathbf{V}$-beli valószínűségi változók közös eloszlásfüggvénye. Például $\Phi \sim \mathbf{Norm}(0; 1)$.

Matematikai statisztika

$(\Omega, \mathcal{F}, \mathcal{P})$	statisztikai mező
F_n^*	tapasztalati eloszlásfüggvény
$\overline{\xi}$	a ξ-re vonatkozó minta átlaga (mintaátlag)
$S_n,\ S_n^2$	tapasztalati szórás illetve szórásnégyzet
$S_{\xi,n},\ S_{\xi,n}^2$	ξ-re vonatkozó tapasztalati szórás illetve szórásnégyzet
$S_n^*,\ S_n^{*2}$	korrigált tapasztalati szórás illetve szórásnégyzet
$S_{\xi,n}^*,\ S_{\xi,n}^{*2}$	ξ-re vonatkozó korrigált tapasztalati szórás illetve szórásnégyzet
$\xi_1^*, \ldots, \xi_n^*$	rendezett minta
$\mathrm{Cov}_n(\xi, \eta)$	tapasztalati kovariancia
$\mathrm{Corr}_n(\xi, \eta)$	tapasztalati korrelációs együttható
Θ	paramétertér
P_ϑ	a ϑ paraméterhez tartozó valószínűség
E_ϑ	a ϑ paraméterhez tartozó várható érték
$\mathrm{D}_\vartheta,\ \mathrm{D}_\vartheta^2$	a ϑ paraméterhez tartozó szórás illetve szórásnégyzet
$f_\vartheta,\ F_\vartheta$	a ϑ paraméterhez tartozó sűrűség- illetve eloszlásfüggvény
I_n	Fisher-féle információmennyiség
l_n	likelihood függvény
L_n	loglikelihood függvény
$\widehat{\vartheta}$	a ϑ paraméter becslése
$H_0,\ H_1$	nullhipotézis, ellenhipotézis
$\mathcal{P}_{H_0},\ \mathcal{P}_{H_1}$	H_0 illetve H_1 esetén lehetséges valószínűségek halmaza

1. Valószínűségszámítás

Ennek a fejezetnek a célja, hogy átismételjük a valószínűségszámítás azon fogalmait és jelöléseit, amelyek szükségesek a matematikai statisztikához. Az itt kimondott állításokat és tételeket nem bizonyítjuk, feltételezzük, hogy ezek már ismertek a korábban tanultak alapján.

1.1. Valószínűségi mező

1.1.1. Véletlen esemény

Egy véletlen kimenetelű kísérlet matematikai modellezésekor azt tekintjük eseménynek, amelyről egyértelműen eldönthető a kísérlet elvégzése után, hogy bekövetkezett-e vagy sem. Így az, hogy egy esemény bekövetkezett, logikai ítélet. Ebből a logika és a halmazelmélet ismert kapcsolata alapján az eseményeket halmazokkal modellezhetjük.

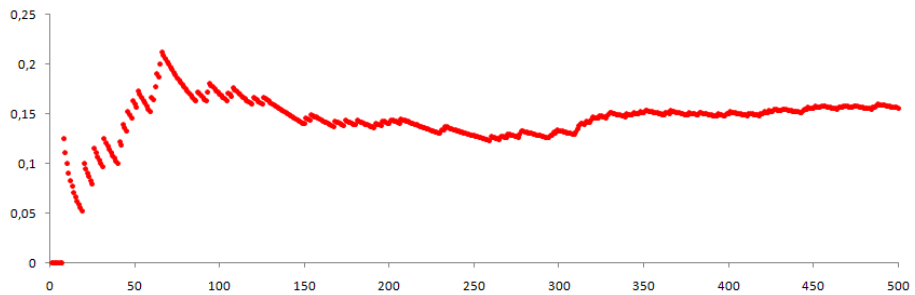
Ha egy kísérletben az A és B halmazok eseményeket modelleznek, akkor az $A \cup B$ bekövetkezése azt jelenti, hogy A és B közül legalább az egyik bekövetkezik. Erről egyértelműen eldönthető a kísérlet elvégzése után, hogy bekövetkezett-e, ezért ez is eseményt modellez. Másrészt, ha A esemény, akkor az A ellenkezője is az. Jelöljük ezt $\overline{A}$-val. Az $A \cup \overline{A}$ biztosan bekövetkezik, ezért ezt *biztos eseménynek* nevezzük és Ω-val jelöljük. Ebből látható, hogy $\overline{A}$ az A-nak Ω-ra vonatkozó komplementere, továbbá minden esemény az Ω egy részhalmaza. Az adott kísérletre vonatkozó események rendszerét jelöljük $\mathcal{F}$-fel, mely tehát az Ω hatványhalmazának egy részhalmaza.

Ahhoz, hogy az eseményeket megfelelően tudjuk modellezni, nem elég véges sok esemény uniójáról feltételezni, hogy az is esemény. Megszámlálhatóan végtelen sok esemény uniójának is eseménynek kell lennie. Tehát a következő definíciót mondhatjuk ki:

1.1. Definíció. Legyen Ω egy nem üres halmaz és $\mathcal{F}$ részhalmaza az Ω hatványhalmazának. Tegyük fel, hogy teljesülnek a következők:
(1) $\Omega \in \mathcal{F}$;
(2) Ha $A \in \mathcal{F}$, akkor $\overline{A} \in \mathcal{F}$, ahol $\overline{A} = \Omega \setminus A$;
(3) Ha $A_i \in \mathcal{F}$ $(i \in \mathbb{N})$, akkor $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
Ekkor $\mathcal{F}$-fet σ-algebrának, elemeit eseményeknek, illetve Ω-t biztos eseménynek nevezzük. A mértékelméletben az $(\Omega, \mathcal{F})$ rendezett párost *mérhető térnek* nevezzük. Ha $A, B \in \mathcal{F}$ és $A \subset B$, akkor azt mondjuk, hogy B teljesül az A-n.

1.1.2. Valószínűség

A modellalkotás következő lépéséhez szükség van egy tapasztalati törvényre az eseményekkel kapcsolatosan, melyet *Jacob Bernoulli* (1654–1705) svájci matematikus publikált. Egy dobókockát dobott fel többször egymásután. A hatos dobások számának és az összes dobások számának arányát, azaz a hatos dobás *relatív gyakoriságát* ábrázolta a dobások számának függvényében:



Bernoulli azt tapasztalta, hogy a hatos dobás relatív gyakorisága a dobások számának növelésével egyre kisebb mértékben ingadozik $\frac{1}{6}$ körül. Más véletlen kimenetelű kísérlet eseményeire is hasonló a tapasztalat, azaz a kísérletek számának növelésével a figyelt esemény bekövetkezésének relatív gyakorisága egyre kisebb mértékben ingadozik egy konstans körül. Ezt a konstanst a figyelt esemény *valószínűségének* fogjuk nevezni.

A továbbiakban $\mathrm{P}(A)$ jelölje az A esemény bekövetkezésének valószínűségét. Könnyen látható, hogy $\mathrm{P}(A) \geqslant 0$ minden esetben, a biztos esemény valószínűsége 1, illetve egyszerre be nem következő események uniójának valószínűsége az események valószínűségeinek összege.

Mindezeket a következő definícióban foglaljuk össze:

1.2. Definíció. Legyen $(\Omega, \mathcal{F})$ mérhető tér és $\mathrm{P}\colon \mathbb{R} \to [0, \infty)$ olyan függvény, melyre teljesülnek a következők:
(1) $\mathrm{P}(\Omega) = 1$;
(2) $\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$, ha $A_i \in \mathcal{F}$ páronként diszjunktak.
Ekkor a P függvényt *valószínűségnek*, a $\mathrm{P}(A)$ számot az A *esemény valószínűségének*, illetve az $(\Omega, \mathcal{F}, \mathrm{P})$ rendezett hármast *valószínűségi mezőnek* nevezzük. Ha egy $A \in \mathcal{F}$ esetén $\mathrm{P}(A) = 1$ teljesül, akkor azt mondjuk, hogy A *majdnem biztosan teljesül.*

Ha $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mező, akkor belátható, hogy $\mathrm{P}(\emptyset) = 0$, így mértékelméleti értelemben a valószínűségi mező *véges mértéktér.*

A valószínűségi mező tehát egy véletlen kimenetelű kísérletet modellez. De a matematikai statisztikában egy ilyen kísérletet többször is el kell végezni egymástól

függetlenül. Ezen független kísérleteket egyetlen valószínűségi mezőben le tudjuk írni az alábbiak szerint.

1.3. Definíció. Legyen az $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mező,

$$\Omega_n := \Omega \times \cdots \times \Omega, \quad (n\text{-szeres Descartes-szorzat}),$$

$\mathcal{F}_n$ a legszűkebb σ-algebra, mely tartalmazza az

$$\left\{ A_1 \times \cdots \times A_n : A_i \in \mathcal{F} \ (i = 1, \ldots, n) \right\}$$

halmazt, továbbá legyen $\mathrm{P}_n \colon \mathcal{F}_n \to \mathbb{R}$ olyan valószínűség, melyre minden $A_i \in \in \mathcal{F}$ $(i = 1, \ldots, n)$ esetén

$$\mathrm{P}_n(A_1 \times \cdots \times A_n) = \mathrm{P}(A_1) \cdots \mathrm{P}(A_n)$$

teljesül. (Ilyen valószínűség a Caratheodory-féle kiterjesztési tétel miatt egyértelműen létezik.) Ekkor az $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$-t *független kísérletek valószínűségi mezőjének* nevezzük.

Tehát $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ az $(\Omega, \mathcal{F}, \mathrm{P})$ kísérlet n-szeri független elvégzését modellezi.

1.2. Valószínűségi változó

Egy eseményt a gyakorlatban legtöbbször a következőképpen szoktunk megadni: Egy függvénnyel az Ω minden eleméhez hozzárendelünk egy valós számot, majd megadunk egy $I \subset \mathbb{R}$ intervallumot. Tekintsük az Ω azon elemeit, melyekhez ez a függvény I-beli értéket rendel. Az ilyen elemekből álló halmaz jelentse a vizsgálandó eseményt. Ehhez viszont az kell, hogy ez a halmaz valóban esemény legyen. Az olyan függvényt, mely minden I intervallumból eseményt származtat az előbbi módon, *valószínűségi változónak* nevezzük.

Bizonyítható, hogy elég csak az $I = (-\infty, x)$ alakú intervallumok esetén feltételezni, hogy az előbb megadott halmaz eleme $\mathcal{F}$-nek, ebből már következik minden más intervallum esetén is. Összefoglalva, kimondhatjuk tehát a következő definíciót:

1.4. Definíció. Legyen $(\Omega, \mathcal{F})$ mérhető tér és $\xi : \Omega \to \mathbb{R}$ olyan függvény, melyre teljesül, hogy $\{ \omega \in \Omega : \xi(\omega) < x \} \in \mathcal{F}$ minden $x \in \mathbb{R}$ esetén. Ekkor a ξ függvényt *valószínűségi változónak* nevezzük.

A továbbiakban az $\{ \omega \in \Omega : \xi(\omega) < x \}$ halmazt a mértékelméletből megszokottak szerint $\Omega(\xi < x)$ vagy rövidebben $\xi < x$ módon fogjuk jelölni. Az ilyen alakú

halmazokat ξ *nívóhalmazainak* is szokás nevezni. Hasonló jelölést alkalmazunk „$<$” helyett más relációk esetén is. A valószínűségi változó ekvivalens a mértékelméletbeli *mérhető függvény* fogalmával.

1.3. Eloszlás- és sűrűségfüggvény

A valószínűségi változó jellemzésére általános esetben jól használható az úgynevezett eloszlásfüggvény:

1.5. Definíció. Legyen $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mező és $\xi\colon \Omega \to \mathbb{R}$ egy valószínűségi változó. Ekkor a ξ *eloszlásfüggvénye*

$$F\colon \mathbb{R} \to \mathbb{R}, \quad F(x) := \mathrm{P}(\xi < x).$$

1.6. Tétel. *Legyen F egy tetszőleges valószínűségi változó eloszlásfüggvénye. Ekkor teljesülnek a következők:*
(a) *F monoton növekvő;*
(b) *F minden pontban balról folytonos;*
(c) $\lim_{x\to\infty} F(x) = 1$*;*
(d) $\lim_{x\to-\infty} F(x) = 0$.

1.7. Tétel. *Ha egy tetszőleges $F\colon \mathbb{R} \to \mathbb{R}$ függvényre teljesülnek az* (a)–(d) *tulajdonságok, akkor létezik olyan valószínűségi változó, melynek F az eloszlásfüggvénye.*

Ezen két tétel alapján jogos a következő elnevezés:

1.8. Definíció. Az $F\colon \mathbb{R} \to \mathbb{R}$ függvényt *eloszlásfüggvénynek* nevezzük, ha teljesülnek rá az (a)–(d) tulajdonságok.

1.9. Tétel. *Ha F a ξ valószínűségi változó eloszlásfüggvénye, akkor teljesülnek a következők:*
(1) $\mathrm{P}(a \leqslant \xi < b) = F(b) - F(a)$ *minden* $a, b \in \mathbb{R},\ a < b$ *esetén;*
(2) $\lim_{x\to a+0} F(x) = F(a) + \mathrm{P}(\xi = a)$ *minden* $a \in \mathbb{R}$ *esetén;*
(3) $\mathrm{P}(\xi = a) = 0$ *pontosan akkor, ha F az $a \in \mathbb{R}$ pontban folytonos.*

Ha ξ *diszkrét valószínűségi változó*, azaz ha R_ξ (ξ értékkészlete) megszámlálható, akkor az előző tétel (2) pontja alapján a ξ eloszlásfüggvénye egyértelműen meghatározott a $\mathrm{P}(\xi = k),\ k \in \mathrm{R}_\xi$ értékekkel. A $k \mapsto \mathrm{P}(\xi = k),\ k \in \mathrm{R}_\xi$ hozzárendelést ξ *eloszlásának* nevezzük.

Az eloszlás elnevezés más jelentésben is előfordul: Két tetszőleges (nem feltétlenül diszkrét) valószínűségi változót *azonos eloszlásúnak* nevezzük, ha az eloszlásfüggvényeik megegyeznek.

Gyakorlati szempontból a diszkrét valószínűségi változók mellett az úgynevezett abszolút folytonos valószínűségi változók osztálya is nagyon fontos.

1.10. Definíció. A ξ valószínűségi változót *abszolút folytonosnak* nevezzük, ha létezik olyan $f\colon \mathbb{R} \to [0,\infty)$ függvény, melyre

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$$

teljesül minden $x \in \mathbb{R}$ esetén, ahol F a ξ eloszlásfüggvénye. Ekkor f-fet a ξ *sűrűségfüggvényének* nevezzük.

1.11. Tétel. *Ha a ξ abszolút folytonos valószínűségi változó eloszlásfüggvénye F és sűrűségfüggvénye f, akkor F folytonos (következésképpen $\mathrm{P}(\xi = x) = 0,\ \forall x \in \mathbb{R}$) és Lebesgue-mérték szerint majdnem mindenütt differenciálható – nevezetesen, ahol f folytonos –, továbbá a differenciálható pontokban $F'(x) = f(x)$.*

1.12. Tétel. *Ha a ξ abszolút folytonos valószínűségi változó sűrűségfüggvénye f, akkor*
(1) $\mathrm{P}(a < \xi < b) = \int_a^b f(x)\,\mathrm{d}x$ *minden* $a, b \in \mathbb{R},\ a < b$ *esetén;*
(2) $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

1.13. Tétel. *Ha $f\colon \mathbb{R} \to [0,\infty)$ és $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$, akkor van olyan abszolút folytonos valószínűségi változó, melynek f a sűrűségfüggvénye.*

Ezen két tétel alapján jogos a következő elnevezés:

1.14. Definíció. Az $f\colon \mathbb{R} \to [0,\infty)$ függvényt *sűrűségfüggvénynek* nevezzük, ha $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$.

1.4. Várható érték, szórásnégyzet

A valószínűségi változók fontos paramétere a valószínűség szerinti integrálja.

1.15. Definíció. Legyen $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mező és $\xi\colon \Omega \to \mathbb{R}$ egy valószínűségi változó. Ha az $\int \xi\,\mathrm{dP}$ integrál létezik akkor azt $\mathrm{E}\,\xi$ módon jelöljük, és ξ *várható értékének* nevezzük. Ha ez az integrál nem létezik, akkor azt mondjuk, hogy ξ-nek nem létezik várható értéke.

Ha két valószínűségi változó eloszlása megegyezik, és valamelyiknek létezik a várható értéke, akkor a másiknak is létezik, továbbá a két várható érték megegyezik. Tehát a várható érték valójában az eloszlásfüggvénytől függ.

A várható érték előbbi értelmezése szerint lehet $+\infty$ illetve $-\infty$ is. Ha a valószínűségszámítást mértékelméleti alapok nélkül tárgyalják, akkor általában feltételezik a várható érték végességét, és csak diszkrét illetve abszolút folytonos eseteket tárgyalják. A következő tétel rávilágít a várható érték gyakorlati jelentőségére.

1.16. Tétel. *Ha a ξ valószínűségi változó értékkészlete $\{ x_1, \dots, x_n \}$, akkor $\mathrm{E}\,\xi = \sum_{i=1}^{n} x_i \,\mathrm{P}(\xi = x_i)$.*

Tehát a várható érték a ξ lehetséges értékeinek az eloszlás szerinti súlyozott átlagát jelenti. A későbbiekben tárgyalt Kolmogorov-féle nagy számok erős törvénye mutatja, hogy bizonyos feltételekkel egy kísérletsorozatban egy ξ valószínűségi változó értékeinek számtani közepe várhatóan (pontosabban 1 valószínűséggel) $\mathrm{E}\,\xi$-hez konvergál.

1.17. Tétel. *Legyen $\{ x_i \in \mathbb{R} : i \in \mathbb{N} \}$ a ξ valószínűségi változó értékkészlete. ξ-nek pontosan akkor véges a várható értéke, ha*

$$\sum_{i=1}^{\infty} |x_i| \,\mathrm{P}(\xi = x_i) < \infty,$$

továbbá ekkor

$$\mathrm{E}\,\xi = \sum_{i=1}^{\infty} x_i \,\mathrm{P}(\xi = x_i).$$

1.18. Tétel. *Legyen ξ abszolút folytonos valószínűségi változó, melynek f a sűrűségfüggvénye. A ξ-nek pontosan akkor véges a várható értéke, ha*

$$\int_{-\infty}^{\infty} |x| f(x) \,\mathrm{d}x < \infty,$$

továbbá ekkor

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

1.19. Tétel. *Ha ξ-nek létezik várható értéke és $\xi = \eta$ majdnem biztosan teljesül, akkor η-nak is létezik a várható értéke, továbbá megegyezik a ξ várható értékével.*

1.20. Tétel. *Ha ξ és η véges várható értékkel rendelkező valószínűségi változók, akkor $a\xi + b\eta$ $(a, b \in \mathbb{R})$ is az, továbbá*

$$\mathrm{E}(a\xi + b\eta) = a\,\mathrm{E}\,\xi + b\,\mathrm{E}\,\eta.$$

1.21. Tétel (Jensen-egyenlőtlenség)**.** *Ha $I \subset \mathbb{R}$ nyílt intervallum, $\xi\colon \Omega \to I$ olyan valószínűségi változó, melyre $\mathrm{E}\,|\xi| < \infty$ teljesül, továbbá $g\colon I \to \mathbb{R}$ Borel-mérhető konvex függvény, akkor*

$$g(\mathrm{E}\,\xi) \leqslant \mathrm{E}\,g(\xi).$$

A valószínűségi változó értékeinek ingadozását az átlag – pontosabban a várható érték – körül, az úgynevezett szórásnégyzettel jellemezzük, amely nem más, mint az átlagtól való négyzetes eltérés átlaga.

1.22. Definíció. A ξ valószínűségi változó *szórásnégyzete* illetve *szórása*

$$\mathrm{D}^2\,\xi := \mathrm{E}(\xi - \mathrm{E}\,\xi)^2, \quad \mathrm{D}\,\xi = \sqrt{\mathrm{E}(\xi - \mathrm{E}\,\xi)^2}.$$

feltéve, hogy ezek a várható értékek léteznek.

1.23. Tétel. *Ha ξ-nek létezik a szórásnégyzete, akkor*
(1) $\mathrm{D}^2\,\xi = \mathrm{E}\,\xi^2 - \mathrm{E}^2\,\xi$*;*
(2) $\mathrm{D}(a\xi + b) = |a|\,\mathrm{D}\,\xi$*, ahol* $a, b \in \mathbb{R}$*.*

1.5. Valószínűségi vektorváltozók

1.24. Definíció. Legyenek $\xi_1, \ldots, \xi_d$ tetszőleges valószínűségi változók. Ekkor a $(\xi_1, \ldots, \xi_d)$ rendezett elem d-est (*d-dimenziós*) *valószínűségi vektorváltozónak* nevezzük.

1.25. Definíció. A $\xi := (\xi_1, \ldots, \xi_d)$ valószínűségi vektorváltozó *eloszlásfüggvénye*

$$F\colon \mathbb{R}^d \to \mathbb{R}, \quad F(x_1, \ldots, x_d) := \mathrm{P}(\xi_1 < x_1, \ldots, \xi_d < x_d).$$

ξ *abszolút folytonos*, ha létezik olyan $f\colon \mathbb{R}^d \to [0, \infty)$ függvény, melyre

$$F(x_1, \ldots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(t_1, \ldots, t_d)\,\mathrm{d}t_1 \cdots \mathrm{d}t_d$$

teljesül minden $x_1, \ldots, x_d \in \mathbb{R}$ esetén. Ekkor f-et a ξ *sűrűségfüggvényének* nevezzük.

1.26. Tétel. *Ha a $\xi := (\xi_1, \ldots, \xi_d)$ abszolút folytonos valószínűségi vektorváltozó sűrűségfüggvénye f, és $g\colon \mathbb{R}^d \to \mathbb{R}$ Borel-mérhető függvény, akkor*

$$\mathrm{E}\,g(\xi_1, \ldots, \xi_d) = \int_{\mathbb{R}^d} g(x_1, \ldots, x_d) f(x_1, \ldots, x_d)\,\mathrm{d}x_1 \cdots \mathrm{d}x_d$$

olyan értelemben, hogy a két oldal egyszerre létezik vagy nem létezik, és ha létezik, akkor egyenlőek.

1.6. Feltételes várható érték

A feltételes várható értéket az egyszerűség kedvéért csak két speciális esetben definiáljuk. Az általános definíciót lásd például Mogyoródi J., Somogyi Á. [11].

1.27. Definíció. Legyenek az $\eta, \xi_1, \dots, \xi_k$ diszkrét valószínűségi változók értékkészletei rendre $R_\eta, R_{\xi_1}, \dots, R_{\xi_k}$, tegyük fel, hogy $\mathrm{E}\,\eta$ véges, továbbá legyen

$$g\colon R_{\xi_1} \times \dots \times R_{\xi_k} \to \mathbb{R}, \quad g(x_1, \dots, x_k) := \sum_{y_i \in R_\eta} y_i \frac{\mathrm{P}(\eta = y_i, \xi_1 = x_1, \dots, \xi_k = x_k)}{\mathrm{P}(\xi_1 = x_1, \dots, \xi_k = x_k)}.$$

Ekkor a $g(\xi_1, \dots, \xi_k)$ valószínűségi változót η-nak $(\xi_1, \dots, \xi_k)$-ra vonatkozó *feltételes várható értékének* nevezzük, és $\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)$ módon jelöljük. A $g(x_1, \dots, x_k)$ $(x_i \in R_{\xi_i},\ i = 1, \dots, k)$ értéket $\mathrm{E}(\eta \mid \xi_1 = x_1, \dots, \xi_k = x_k)$ módon jelöljük.

1.28. Definíció. Legyen az $(\eta, \xi_1, \dots, \xi_k)$ abszolút folytonos valószínűségi vektorváltozó sűrűségfüggvénye f, a $(\xi_1, \dots, \xi_k)$ sűrűségfüggvénye h, tegyük fel, hogy $\mathrm{E}\,\eta$ véges, továbbá legyen

$$g\colon \mathbb{R}^k \to \mathbb{R}, \quad g(x_1, \dots, x_k) := \int_{-\infty}^{\infty} y \frac{f(y, x_1, \dots, x_k)}{h(x_1, \dots, x_k)}\,\mathrm{d}y.$$

Ekkor a $g(\xi_1, \dots, \xi_k)$ valószínűségi változót η-nak $(\xi_1, \dots, \xi_k)$-ra vonatkozó *feltételes várható értékének* nevezzük, és $\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)$ módon jelöljük. A $g(x_1, \dots, x_k)$ $(x_i \in R_{\xi_i},\ i = 1, \dots, k)$ értéket $\mathrm{E}(\eta \mid \xi_1 = x_1, \dots, \xi_k = x_k)$ módon jelöljük.

A feltételes várható értékre teljesülnek a következők:
$\mathrm{E}\,\eta = \mathrm{E}\big(\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)\big)$;
$\mathrm{E}(a\xi + b\eta \mid \xi_1, \dots, \xi_k) = a\,\mathrm{E}(\xi \mid \xi_1, \dots, \xi_k) + b\,\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan, minden $a, b \in \mathbb{R}$ esetén;
$\mathrm{E}\big(\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k) \mid \xi_1, \dots, \xi_k\big) = \mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan;
$\mathrm{E}(\xi\eta \mid \xi_1, \dots, \xi_k) = \xi\,\mathrm{E}(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan.

1.7. Független valószínűségi változók

Az A és B események függetlenek, ha $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$. Valószínűségi változók függetlenségét nívóhalmazaik függetlenségével definiáljuk.

1.29. Definíció. A $\xi_1, \ldots, \xi_n$ valószínűségi változókat *függetleneknek* nevezzük, ha

$$\mathrm{P}(\xi_1 < x_1, \ldots, \xi_n < x_n) = \prod_{k=1}^{n} \mathrm{P}(\xi_k < x_k)$$

minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén teljesül. A $\xi_1, \ldots, \xi_n$ valószínűségi változók *páronként függetlenek*, ha közülük bármely kettő független. Végtelen sok valószínűségi változót függetleneknek nevezzük, ha bármely véges részrendszere független.

Szükségünk lesz a valószínűségi vektorváltozók függetlenségének fogalmára is. Ehhez bevezetünk egy jelölést. Legyen $\xi = (\xi_1, \ldots, \xi_d)$ egy valószínűségi vektorváltozó és $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. Ekkor a $\xi < x$ esemény alatt azt értjük, hogy a $\xi_k < x_k$ események minden $k = 1, \ldots, d$ esetén teljesülnek.

1.30. Definíció. A $\zeta_1, \ldots, \zeta_n$ d-dimenziós valószínűségi vektorváltozókat *függetleneknek* nevezzük, ha minden $x_1, \ldots, x_n \in \mathbb{R}^d$ esetén

$$\mathrm{P}(\zeta_1 < x_1, \ldots, \zeta_n < x_n) = \prod_{k=1}^{n} \mathrm{P}(\zeta_k < x_k)$$

teljesül. A $\zeta_1, \ldots, \zeta_d$ valószínűségi vektorváltozók *páronként függetlenek*, ha közülük bármely kettő független. Végtelen sok valószínűségi vektorváltozót függetleneknek nevezzük, ha bármely véges részrendszere független.

1.31. Tétel. *A $\xi_1, \ldots, \xi_n$ diszkrét valószínűségi változók pontosan akkor függetlenek, ha*

$$\mathrm{P}(\xi_1 = x_1, \ldots, \xi_n = x_n) = \prod_{k=1}^{n} \mathrm{P}(\xi_k = x_k)$$

teljesül minden $x_1 \in \mathrm{R}_{\xi_1}, \ldots, x_n \in \mathrm{R}_{\xi_n}$ esetén.

1.32. Tétel. *Legyen $(\xi_1, \ldots, \xi_n)$ abszolút folytonos valószínűségi vektorváltozó. A $\xi_1, \ldots, \xi_n$ valószínűségi változók pontosan akkor függetlenek, ha*

$$f(x_1, \ldots, x_n) = \prod_{k=1}^{n} f_k(x_k)$$

teljesül minden $x_1, \ldots, x_n \in \mathbb{R}$ esetén, ahol f_k a ξ_k sűrűségfüggvénye, továbbá f a $(\xi_1, \ldots, \xi_n)$ sűrűségfüggvénye.

1.33. Tétel (Konvolúció)**.** *Ha ξ és η független abszolút folytonos valószínűségi változók f illetve g sűrűségfüggvénnyel, akkor $\xi + \eta$ is abszolút folytonos, továbbá a*

sűrűségfüggvénye $x \in \mathbb{R}$ helyen

$$h(x) = \int_{-\infty}^{\infty} f(t)g(x-t)\,\mathrm{d}t.$$

1.34. Tétel. *Ha ξ és η független abszolút folytonos valószínűségi változók f illetve g sűrűségfüggvénnyel, akkor $\xi\eta$ is abszolút folytonos, továbbá a sűrűségfüggvénye $x \in \mathbb{R}$ helyen*

$$h(x) = \int_{-\infty}^{\infty} g(t) f\left(\frac{x}{t}\right) \frac{1}{|t|}\,\mathrm{d}t.$$

1.35. Tétel. *Ha ξ és η független abszolút folytonos valószínűségi változók f illetve g sűrűségfüggvénnyel, akkor $\frac{\xi}{\eta}$ is abszolút folytonos, továbbá a sűrűségfüggvénye $x \in \mathbb{R}$ helyen*

$$h(x) = \int_{-\infty}^{\infty} |t| g(t) f(xt)\,\mathrm{d}t.$$

1.8. Kovariancia és korrelációs együttható

1.36. Definíció. A ξ és η valószínűségi változók *kovarianciája*

$$\operatorname{cov}(\xi,\eta) := \mathrm{E}\big((\xi - \mathrm{E}\,\xi)(\eta - \mathrm{E}\,\eta)\big),$$

feltéve, hogy ezek a várható értékek léteznek.

Könnyen belátható, hogy $\operatorname{cov}(\xi,\eta) = \mathrm{E}\,\xi\eta - \mathrm{E}\,\xi\,\mathrm{E}\,\eta$.

1.37. Tétel. *Ha a ξ és η független valószínűségi változóknak létezik a várható értékeik, akkor létezik a kovarianciájuk is és $\operatorname{cov}(\xi,\eta) = 0$, azaz $\mathrm{E}\,\xi\eta = \mathrm{E}\,\xi\,\mathrm{E}\,\eta$.*

1.38. Definíció. A $\xi_1, \ldots, \xi_n$ valószínűségi változókat *korrelálatlanoknak* nevezzük, ha $\operatorname{cov}(\xi_i, \xi_j) = 0$ minden $i, j \in \{1, \ldots, n\}$, $i \neq j$ esetén.

1.39. Tétel. *Ha a $\xi_1, \ldots, \xi_n$ valószínűségi változók esetén létezik $\operatorname{cov}(\xi_i, \xi_j)$ minden $i, j \in \{1, \ldots, n\}$ esetén, akkor $\sum_{i=1}^n \xi_i$-nek létezik a szórásnégyzete, továbbá*

$$\mathrm{D}^2\left(\sum_{i=1}^n \xi_i\right) = \sum_{i=1}^n \mathrm{D}^2\,\xi_i + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \operatorname{cov}(\xi_i, \xi_j).$$

1.40. Tétel. *Ha a $\xi_1, \ldots, \xi_n$ páronként független valószínűségi változóknak léteznek a szórásnégyzeteik, akkor a $\sum_{i=1}^n \xi_i$ valószínűségi változónak is van szórásnégyzete, továbbá $\mathrm{D}^2\left(\sum_{i=1}^n \xi_i\right) = \sum_{i=1}^n \mathrm{D}^2\,\xi_i$.*

1.41. Definíció. Ha ξ és η pozitív szórású valószínűségi változók, akkor a *korrelációs együtthatójuk*

$$\operatorname{corr}(\xi,\eta) := \frac{\operatorname{cov}(\xi,\eta)}{\mathrm{D}\,\xi\,\mathrm{D}\,\eta}.$$

1.42. Tétel. *Legyen ξ pozitív szórású valószínűségi változó, továbbá $\eta := a\xi + b$, ahol $a, b \in \mathbb{R},\ a \neq 0$. Ekkor létezik ξ és η korrelációs együtthatója, és*

$$\operatorname{corr}(\xi,\eta) = \begin{cases} 1, & \text{ha } a > 0, \\ -1, & \text{ha } a < 0. \end{cases}$$

1.43. Tétel. *Ha $|\operatorname{corr}(\xi,\eta)| = 1$, akkor léteznek olyan $a, b \in \mathbb{R},\ a \neq 0$ konstansok, melyekre $\mathrm{P}(\eta = a\xi + b) = 1$ teljesül.*

1.9. Nevezetes eloszlások

1.9.1. Diszkrét egyenletes eloszlás

1.44. Definíció. Legyen $\{ x_1, \ldots, x_r \}$ a ξ valószínűségi változó értékkészlete és

$$\mathrm{P}(\xi = x_i) = \frac{1}{r} \quad (i = 1, \ldots, r).$$

Ekkor ξ-t diszkrét egyenletes eloszlásúnak nevezzük az $\{ x_1, \ldots, x_r \}$ halmazon.

1.45. Tétel. $\mathrm{E}\,\xi = \frac{1}{r}\sum_{i=1}^{r} x_i$ *és* $\mathrm{D}^2\,\xi = \frac{1}{r}\sum_{i=1}^{r} x_i^2 - \left(\frac{1}{r}\sum_{i=1}^{r} x_i\right)^2$.

1.9.2. Karakterisztikus eloszlás

1.46. Definíció. Az A esemény *indikátorváltozójának* az

$$\mathrm{I}_A \colon \Omega \to \mathbb{R}, \quad \mathrm{I}_A(\omega) := \begin{cases} 1, & \text{ha } \omega \in A, \\ 0, & \text{ha } \omega \notin A, \end{cases}$$

valószínűségi változót nevezzük, továbbá az I_A-t $\mathrm{P}(A)$ paraméterű *karakterisztikus eloszlásúnak* nevezzük.

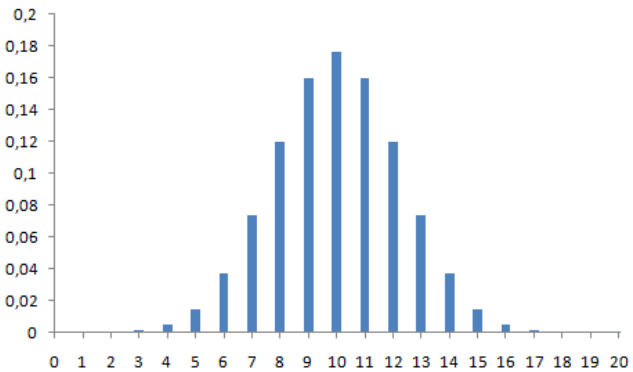
1.47. Tétel. $\mathrm{E}\,\mathrm{I}_A = \mathrm{P}(A)$ *és* $\mathrm{D}^2\,\mathrm{I}_A = \mathrm{P}(A)\big(1 - \mathrm{P}(A)\big)$.

1.9.3. Binomiális eloszlás

1.48. Definíció. Legyen $\{0, 1, \dots, r\}$ a ξ valószínűségi változó értékkészlete és $p \in$ $\in (0, 1)$. Ha minden $k \in \{0, 1, \dots, r\}$ esetén

$$\mathrm{P}(\xi = k) = \binom{r}{k} p^k (1-p)^{r-k},$$

akkor ξ-t r-edrendű p paraméterű *binomiális eloszlású* valószínűségi változónak nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $\mathbf{Bin}(r; p)$ módon jelöljük.



1.1. ábra. $r = 20$ rendű $p = 0{,}5$ paraméterű binomiális eloszlás vonaldiagramja

Egy tetszőleges A esemény gyakorisága r kísérlet után r-edrendű $\mathrm{P}(A)$ paraméterű binomiális eloszlású valószínűségi változó.

Az $r = 1$ rendű p paraméterű binomiális eloszlás megegyezik a p paraméterű karakterisztikus eloszlással, vagyis a p paraméterű karakterisztikus eloszlású valószínűségi változók halmaza $\mathbf{Bin}(1; p)$.

Másrészt r darab független p paraméterű karakterisztikus eloszlású valószínűségi változó összege r-edrendű p paraméterű binomiális eloszlású.

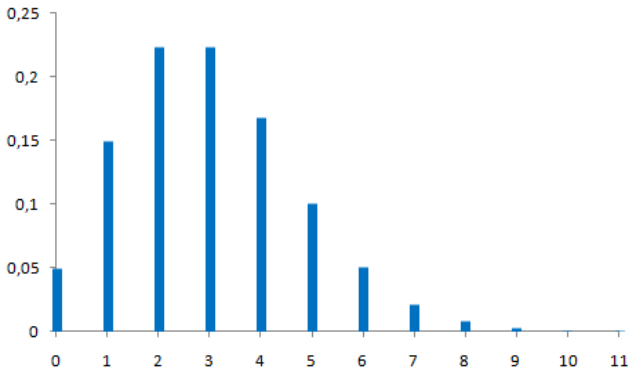
1.49. Tétel. *$\xi \in \mathbf{Bin}(r; p)$ esetén* $\mathrm{E}\,\xi = rp$ *és* $\mathrm{D}^2\,\xi = rp(1-p)$.

1.9.4. Poisson-eloszlás

1.50. Definíció. Legyen $\{0, 1, 2, \dots\}$ a ξ valószínűségi változó értékkészlete, $\lambda \in \mathbb{R}_+$ és

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

Ekkor ξ-t λ paraméterű *Poisson-eloszlású* valószínűségi változónak nevezzük.



1.2. ábra. $\lambda = 3$ paraméterű Poisson-eloszlás vonaldiagramja

1.51. Tétel. *Ha ξ egy $\lambda \in \mathbb{R}_+$ paraméterű Poisson-eloszlású valószínűségi változó, akkor* $\mathrm{E}\,\xi = \mathrm{D}^2\,\xi = \lambda$.

1.9.5. Egyenletes eloszlás

1.52. Definíció. Legyen ξ abszolút folytonos valószínűségi változó, $a, b \in \mathbb{R}$ és $a < b$. Ha ξ sűrűségfüggvénye

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} \frac{1}{b-a}, & \text{ha } a \leqslant x \leqslant b, \\ 0 & \text{egyébként,} \end{cases}$$

akkor ξ-t *egyenletes eloszlásúnak* nevezzük az $[a, b]$ intervallumon.

1.53. Tétel. *Ha ξ egyenletes eloszlású az $[a, b]$ intervallumon, akkor az eloszlásfüggvénye*

$$F\colon \mathbb{R} \to \mathbb{R}, \quad F(x) = \begin{cases} 0, & \textit{ha } x < a, \\ \frac{x-a}{b-a}, & \textit{ha } a \leqslant x \leqslant b, \\ 1, & \textit{ha } x > b, \end{cases}$$

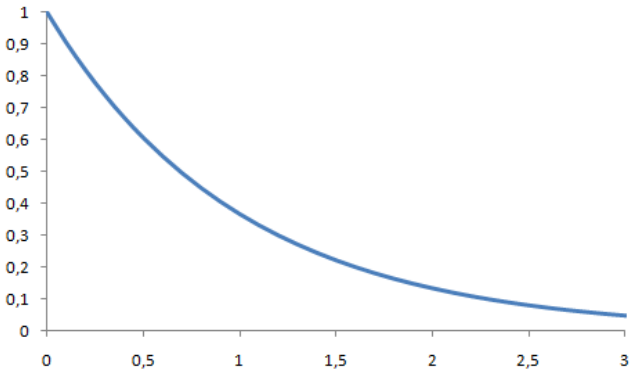
továbbá $\mathrm{E}\,\xi = \frac{a+b}{2}$ *és* $\mathrm{D}\,\xi = \frac{b-a}{\sqrt{12}}$.

1.9.6. Exponenciális eloszlás

1.54. Definíció. Legyen ξ abszolút folytonos valószínűségi változó, és $\lambda \in \mathbb{R}_+$. Ha ξ sűrűségfüggvénye

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \lambda e^{-\lambda x}, & \text{ha } x > 0, \end{cases}$$

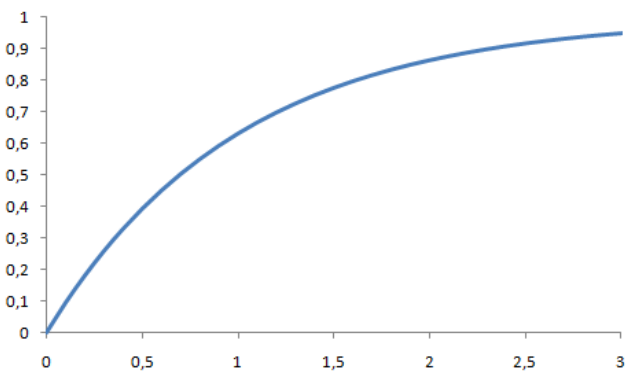
akkor ξ-t λ paraméterű *exponenciális eloszlású* valószínűségi változónak nevezzük. Az ilyen valószínűségi változók halmazát $\mathbf{Exp}(\lambda)$ módon jelöljük.



1.3. ábra. $\lambda = 1$ paraméterű exponenciális eloszlású valószínűségi változó sűrűségfüggvénye

1.55. Tétel. *$\xi \in \mathbf{Exp}(\lambda)$ esetén $\mathrm{E}\,\xi = \mathrm{D}\,\xi = \frac{1}{\lambda}$, továbbá ξ eloszlásfüggvénye*

$$F\colon \mathbb{R} \to \mathbb{R}, \quad F(x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ 1 - e^{-\lambda x}, & \text{ha } x > 0. \end{cases}$$



1.4. ábra. $\lambda = 1$ paraméterű exponenciális eloszlású valószínűségi változó eloszlásfüggvénye

1.56. Definíció. A ξ valószínűségi változót *örökifjú* tulajdonságúnak nevezzük, ha $\mathrm{P}(\xi \geqslant x + y) = \mathrm{P}(\xi \geqslant x)\,\mathrm{P}(\xi \geqslant y)$ minden $x, y \in \mathbb{R}_+$ esetén.

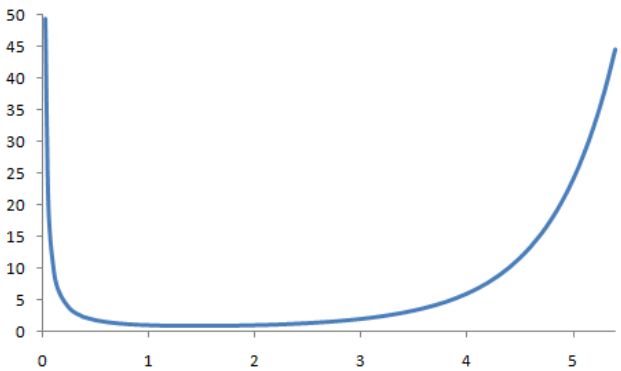
1.57. Tétel. *Egy abszolút folytonos valószínűségi változó pontosan akkor örökifjú tulajdonságú, ha exponenciális eloszlású.*

1.9.7. Gamma-eloszlás

A következőkben szükségünk lesz az úgynevezett *gamma-függvényre*:

$$\Gamma\colon \mathbb{R}_+ \to \mathbb{R}, \quad \Gamma(x) := \int_0^\infty u^{x-1}e^{-u}\,\mathrm{d}u.$$

$\Gamma(\frac{1}{2}) = \sqrt{\pi}$ illetve ha $n \in \mathbb{N}$, akkor $\Gamma(n) = (n-1)!$.



1.5. ábra. A gamma-függvény grafikonja

1.58. Definíció. Legyen $r, \lambda \in \mathbb{R}_+$ és a ξ valószínűségi változó sűrűségfüggvénye

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) := \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \frac{\lambda^r x^{r-1}}{\Gamma(r)} e^{-\lambda x}, & \text{ha } x > 0. \end{cases}$$

Ekkor ξ-t r-edrendű λ *paraméterű gamma-eloszlásúnak* nevezzük. Az ilyen valószínűségi változók halmazát $\mathbf{Gamma}(r;\lambda)$ módon jelöljük.

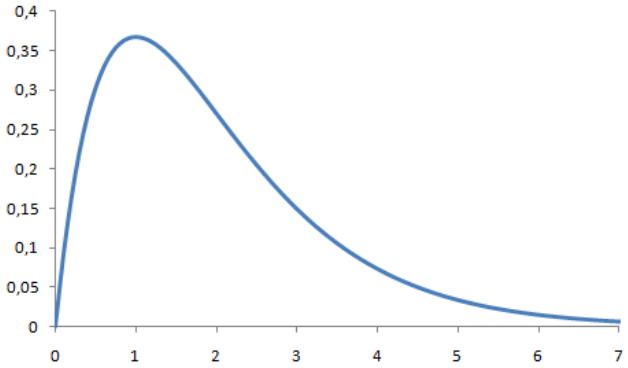
A definíció következménye, hogy $\mathbf{Exp}(\lambda) = \mathbf{Gamma}(1;\lambda)$.

1.59. Tétel. *$\xi \in \mathbf{Gamma}(r;\lambda)$ esetén $\mathrm{E}\,\xi = \frac{r}{\lambda}$ és $\mathrm{D}^2\,\xi = \frac{r}{\lambda^2}$.*

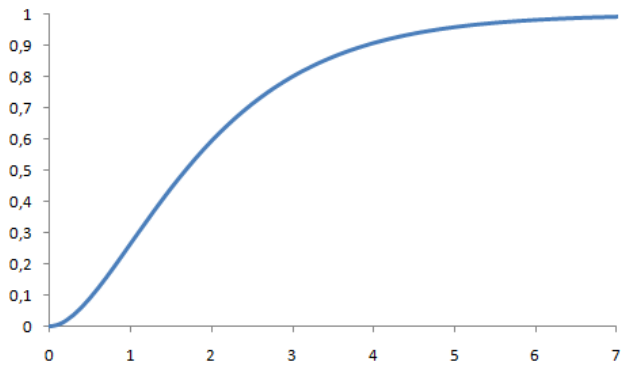
1.60. Tétel. *Ha $r \in \mathbb{N}$ és $\xi_1, \ldots, \xi_r$ azonos $\lambda > 0$ paraméterű exponenciális eloszlású független valószínűségi változók, akkor*

$$\xi_1 + \cdots + \xi_r \in \mathbf{Gamma}(r;\lambda).$$

1.61. Lemma. *Ha $r \geqslant 1$ és $\xi \in \mathbf{Gamma}(r;1)$ eloszlásfüggvénye F_r, akkor $0{,}5 < < F_r(r) < 0{,}7$.*



1.6. ábra. $r = 2$ rendű $\lambda = 1$ paraméterű gamma-eloszlású valószínűségi változó sűrűségfüggvénye



1.7. ábra. $r = 2$ rendű $\lambda = 1$ paraméterű gamma-eloszlású valószínűségi változó eloszlásfüggvénye

1.9.8. Normális eloszlás

1.62. Definíció. A ξ abszolút folytonos valószínűségi változót *standard normális eloszlásúnak* nevezzük, ha a sűrűségfüggvénye

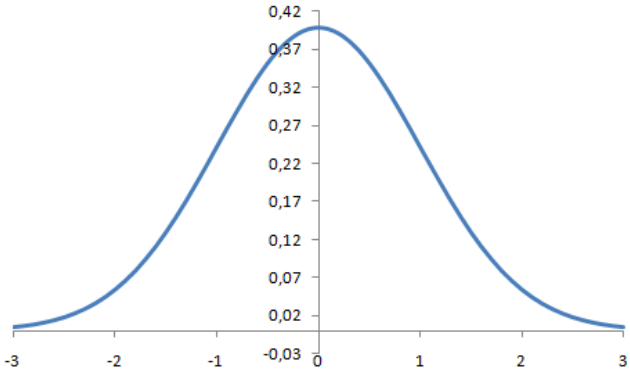
$$\varphi\colon \mathbb{R} \to \mathbb{R}, \quad \varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

A standard normális eloszlású valószínűségi változó eloszlásfüggvényét Φ-vel jelöljük, mely a sűrűségfüggvény definíciója szerint

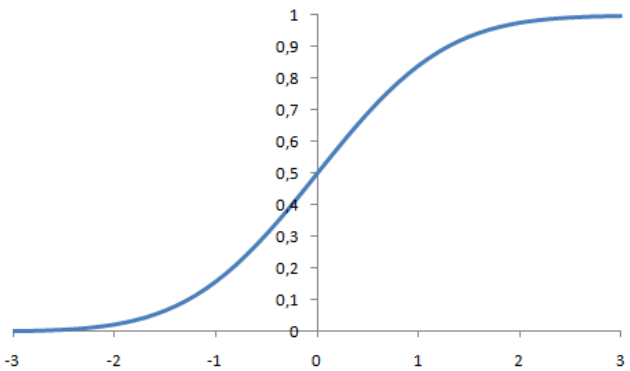
$$\Phi\colon \mathbb{R} \to \mathbb{R}, \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t.$$

Φ-re nincs zárt formula, közelítő értékeinek kiszámítására például a Taylor-sora használható:

$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^k (2k+1) k!} x^{2k+1}.$$



1.8. ábra. Standard normális eloszlású valószínűségi változó sűrűségfüggvénye



1.9. ábra. Standard normális eloszlású valószínűségi változó eloszlásfüggvénye

Megemlítjük még a $\Phi(x)$ egy egyszerű közelítő formuláját. Johnson és Kotz 1970-ben bizonyították (lásd [6]), hogy az

$$1 - 0{,}5(1 + ax + bx^2 + cx^3 + dx^4)^{-4}$$

kifejezéssel $x \geqslant 0$ esetén $2{,}5 \cdot 10^{-4}$-nél kisebb hibával közelíthető $\Phi(x)$, ahol

$$a = 0{,}196854,\ b = 0{,}115194,\ c = 0{,}000344,\ d = 0{,}019527.$$

Mivel φ páros függvény, ezért minden $x \in \mathbb{R}$ esetén $\Phi(-x) = 1 - \Phi(x)$.

1.63. Tétel. *Ha ξ standard normális eloszlású valószínűségi változó, akkor* $\mathrm{E}\,\xi = 0$ *és* $\mathrm{D}\,\xi = 1$.

1.64. Definíció. Legyen η standard normális eloszlású valószínűségi változó, $m \in \mathbb{R}$ és $\sigma \in \mathbb{R}_+$. Ekkor a $\sigma\eta + m$ valószínűségi változót m és σ paraméterű *normális eloszlásúnak* nevezzük. Az ilyen valószínűségi változók halmazát $\mathbf{Norm}(m; \sigma)$ módon jelöljük.

Definíció alapján a standard normális eloszlású valószínűségi változók halmaza $\mathbf{Norm}(0;1)$.

1.65. Tétel. *$\xi \in \mathbf{Norm}(m;\sigma)$ esetén $\mathrm{E}\,\xi = m$, $\mathrm{D}\,\xi = \sigma$, továbbá ξ eloszlásfüggvénye*

$$F\colon \mathbb{R} \to \mathbb{R}, \quad F(x) = \Phi\left(\frac{x-m}{\sigma}\right),$$

illetve sűrűségfüggvénye

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right).$$

1.66. Tétel. *Ha $\xi_1, \dots, \xi_n$ független, normális eloszlású valószínűségi változók, akkor $\xi_1 + \dots + \xi_n$ is normális eloszlású.*

1.67. Tétel. *Ha $\xi_1, \dots, \xi_n$ normális eloszlású valószínűségi változók és minden $i, j \in \{\,1, \dots, n\,\}$, $i \neq j$ esetén $\operatorname{cov}(\xi_i, \xi_j) = 0$, akkor $\xi_1, \dots, \xi_n$ függetlenek.*

1.68. Definíció. A ξ valószínűségi változó eloszlásának *ferdesége* illetve *lapultsága*

$$\frac{\mathrm{E}(\xi - \mathrm{E}\,\xi)^3}{\mathrm{D}^3\,\xi} \quad \text{illetve} \quad \frac{\mathrm{E}(\xi - \mathrm{E}\,\xi)^4}{\mathrm{D}^4\,\xi} - 3,$$

feltéve, hogy ezek a kifejezések léteznek.

1.69. Tétel. *Ha ξ normális eloszlású valószínűségi változó, akkor az eloszlásának ferdesége és lapultsága is* 0.

Ha $\xi \in \mathbf{Bin}(n;p)$, akkor $\frac{\xi - np}{\sqrt{np(1-p)}}$ közelítőleg standard normális eloszlású (lásd Moivre–Laplace-tétel). A közelítés akkor tekinthető megfelelően pontosnak, ha

$$\min\{\, np, n(1-p)\,\} \geqslant 10.$$

1.9.9. Többdimenziós normális eloszlás

1.70. Definíció. Legyenek $\eta_1, \dots, \eta_d$ független standard normális eloszlású valószínűségi változók. Ekkor az $(\eta_1, \dots, \eta_d)$ valószínűségi vektorváltozót *d-dimenziós standard normális eloszlásúnak* nevezzük.

1.71. Definíció. Ha $\eta = (\eta_1, \dots, \eta_d)$ d-dimenziós standard normális eloszlású valószínűségi vektorváltozó, A egy $d \times d$ típusú valós mátrix és $m = (m_1, \dots, m_d) \in \mathbb{R}^d$, akkor a

$$\xi := \eta A + m$$

valószínűségi vektorváltozót *d-dimenziós normális eloszlásúnak* nevezzük. A ξ-vel azonos eloszlású valószínűségi vektorváltozók halmazát $\mathbf{Norm}_d(m; A)$ módon jelöljük.

1.72. Tétel. *Ha* $\xi = (\xi_1, \dots, \xi_d) \in \mathbf{Norm}_d(m; A)$, *akkor*

$$m = (\mathrm{E}\,\xi_1, \dots, \mathrm{E}\,\xi_d),$$

$$D := A^\top A = \big(\mathrm{cov}(\xi_i, \xi_j)\big)_{d\times d},$$

továbbá ha $\det D \neq 0$, *akkor* ξ *sűrűségfüggvénye*

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad f(x) = \frac{1}{\sqrt{(2\pi)^d \det D}} \exp\left(-\tfrac{1}{2}(x-m)D^{-1}(x-m)^\top\right).$$

1.73. Tétel. *Legyen* $(\xi_1, \dots, \xi_d) \in \mathbf{Norm}_d(m; A)$. *Ekkor* $\xi_1, \dots, \xi_d$ *pontosan akkor korrelálatlanok, ha függetlenek.*

1.74. Tétel. *Ha* $(\xi_1, \dots, \xi_d) \in \mathbf{Norm}_d(m; A)$, *akkor létezik* $a_2, \dots, a_d \in \mathbb{R}$, *hogy* $\mathrm{E}(\xi_1 \mid \xi_2, \dots, \xi_d) = a_2\xi_2 + \dots + a_d\xi_d$.

1.9.10. Khi-négyzet eloszlás

1.75. Definíció. Legyenek $\xi_1, \dots, \xi_s$ független standard normális eloszlású valószínűségi változók. Ekkor a $\xi_1^2 + \dots + \xi_s^2$ valószínűségi változót s *szabadsági fokú khi-négyzet eloszlásúnak* nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $\mathbf{Khi}(s)$ módon jelöljük.

1.76. Tétel. *Ha* $\xi \in \mathbf{Khi}(s_1)$ *és* $\eta \in \mathbf{Khi}(s_2)$ *függetlenek, akkor*

$$\xi + \eta \in \mathbf{Khi}(s_1 + s_2).$$

1.77. Tétel. $\mathbf{Khi}(s) = \mathbf{Gamma}\left(\frac{s}{2}; \frac{1}{2}\right)$, *azaz* $\xi \in \mathbf{Khi}(s)$ *sűrűségfüggvénye*

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} 0, & \text{ha } x \leqslant 0, \\ \frac{2^{-\frac{s}{2}} x^{\frac{s}{2}-1}}{\Gamma\left(\frac{s}{2}\right)} e^{-\frac{x}{2}}, & \text{ha } x > 0. \end{cases}$$

1.78. Következmény. $\xi \in \mathbf{Khi}(s)$ *esetén* $\mathrm{E}\,\xi = s$ *és* $\mathrm{D}^2\,\xi = 2s$.

1.79. Tétel. *Legyen* $A_1, \dots, A_r$ *egy teljes eseményrendszer (azaz uniójuk a biztos esemény és páronként diszjunktak). Jelölje* ϱ_i *az* A_i *esemény gyakoriságát* n *kísérlet*

után. Tegyük fel, hogy $p_i := \mathrm{P}(A_i) > 0$ *minden* $i \in \{\,1, \dots, r\,\}$ *esetén. Ekkor*

$$\chi^2 := \sum_{i=1}^{r} \frac{(\varrho_i - np_i)^2}{np_i}$$

eloszlása $r-1$ *szabadsági fokú khi-négyzet eloszláshoz konvergál* $n \to \infty$ *esetén.*

A bizonyítás a karakterisztikus függvények elméletén és lineáris algebrán alapul (lásd például Fazekas I. [2, 161–162. oldal]). A gyakorlatban a tétel azt jelenti, hogy $F \sim \mathbf{Khi}(r-1)$ jelöléssel

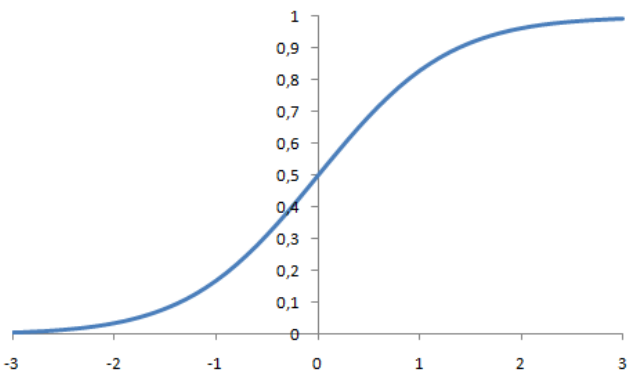
$$\mathrm{P}(\chi^2 < x) \simeq F(x).$$

A közelítés már jónak tekinthető, ha $\min\{\,\varrho_1, \dots, \varrho_r\,\} \geqslant 10$.

1.80. Lemma. *Ha a* $\xi \in \mathbf{Khi}(s)$ *valószínűségi változó eloszlásfüggvénye* F_s*, akkor* $0{,}5 < F_s(s) < 0{,}7$.

1.9.11. t-eloszlás

1.81. Definíció. Ha $\xi \in \mathbf{Norm}(0;1)$ és $\eta \in \mathbf{Khi}(s)$ függetlenek, akkor a $\xi\sqrt{\frac{s}{\eta}}$ valószínűségi változót s *szabadsági fokú t-eloszlásúnak* nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $\mathbf{t}(s)$ módon jelöljük.

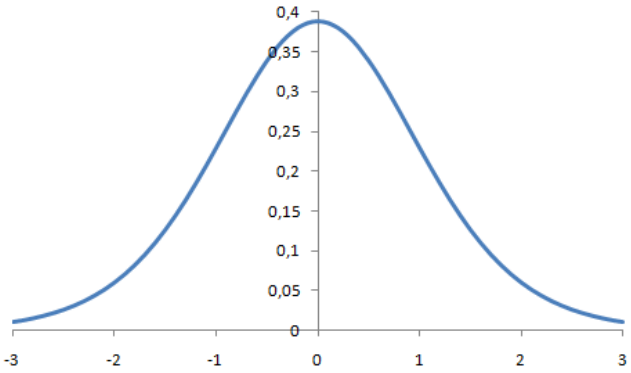


1.10. ábra. $s = 10$ szabadsági fokú t-eloszlású valószínűségi változó eloszlásfüggvénye

A t-eloszlás William Sealy Gosset (1876–1937) nevéhez köthető, aki Student álnéven publikált. Ezért a t-eloszlás Student-eloszlás néven is ismert.

1.82. Tétel. *Ha* $\xi \in \mathbf{t}(s)$*, akkor a sűrűségfüggvénye*

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \frac{\Gamma\left(\frac{s+1}{2}\right)}{\sqrt{s\pi}\,\Gamma\left(\frac{s}{2}\right)\left(1+\frac{x^2}{s}\right)^{\frac{s+1}{2}}}.$$



1.11. ábra. $s = 10$ szabadsági fokú t-eloszlású valószínűségi változó sűrűségfüggvénye

1.83. Következmény. *$f(-x) = f(x)$ és $F(-x) = 1 - F(x)$ minden $x \in \mathbb{R}$ esetén, ahol f illetve F a $\xi \in \mathbf{t}(s)$ sűrűség- illetve eloszlásfüggvénye.*

1.84. Tétel. *Ha $\xi \in \mathbf{t}(s)$, akkor $s \geqslant 2$ esetén $\mathrm{E}\,\xi = 0$, illetve $s \geqslant 3$ esetén $\mathrm{D}^2\,\xi = \frac{s}{s-2}$. Ezektől eltérő esetekben nem létezik ξ várható értéke illetve szórása.*

1.85. Tétel. *Ha $\xi_s \in \mathbf{t}(s)$ minden $s \in \mathbb{N}$ esetén, akkor $\lim\limits_{s\to\infty} \mathrm{P}(\xi_s < x) = \Phi(x)$ minden $x \in \mathbb{R}$-re, azaz a t-eloszlás konvergál a standard normális eloszláshoz, ha a szabadsági fok tart ∞-be.*

Gyakorlatilag $s \geqslant 50$ esetén a $\xi_s \in \mathbf{t}(s)$ eloszlásfüggvénye és Φ között elhanyagolhatóan kicsi a különbség.

1.9.12. Cauchy-eloszlás

1.86. Definíció. Egy valószínűségi változót *Cauchy-eloszlásúnak* nevezünk, ha a sűrűségfüggvénye

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) := \frac{1}{\pi(1+x^2)}.$$

1.87. Tétel. *Cauchy-eloszlású valószínűségi változó eloszlásfüggvénye*

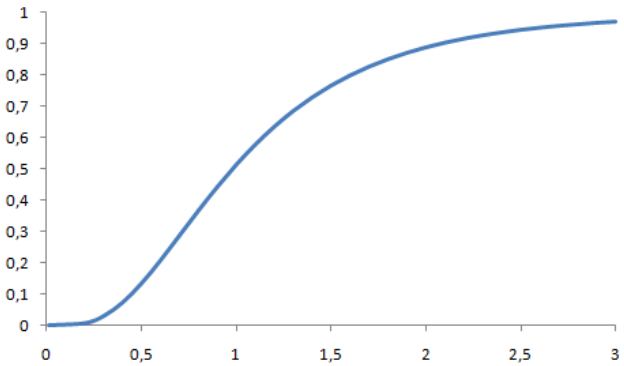
$$F\colon \mathbb{R} \to \mathbb{R}, \quad F(x) = \frac{1}{\pi}\operatorname{arctg} x + \frac{1}{2}.$$

1.88. Tétel. *A Cauchy-eloszlás megegyezik az 1 szabadsági fokú t-eloszlással.*

1.89. Következmény. *Cauchy-eloszlású valószínűségi változónak nem létezik várható értéke illetve szórása.*

1.9.13. F-eloszlás

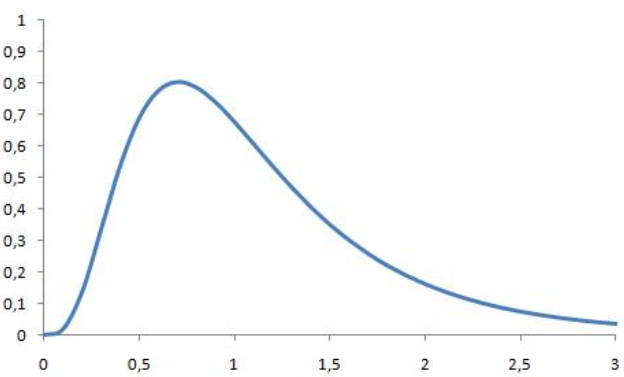
1.90. Definíció. Ha $\xi_1 \in \mathbf{Khi}(s_1)$ és $\xi_2 \in \mathbf{Khi}(s_2)$ függetlenek, akkor az $\frac{s_2\xi_1}{s_1\xi_2}$ valószínűségi változót *s_1 és s_2 szabadsági fokú F-eloszlásúnak* nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $\mathbf{F}(s_1; s_2)$ módon jelöljük.



1.12. ábra. $s_1 = 10$ és $s_2 = 15$ szabadsági fokú F-eloszlású valószínűségi változó eloszlásfüggvénye

1.91. Tétel. *Ha $\xi \in \mathbf{F}(s_1; s_2)$, akkor a sűrűségfüggvénye*

$$f\colon \mathbb{R} \to \mathbb{R}, \quad f(x) = \begin{cases} 0, & ha\ x \leqslant 0, \\ \frac{\Gamma\left(\frac{s_1+s_2}{2}\right)}{\Gamma\left(\frac{s_1}{2}\right)\Gamma\left(\frac{s_2}{2}\right)} \sqrt{\frac{s_1^{s_1} s_2^{s_2} x^{s_1-2}}{(s_1x+s_2)^{s_1+s_2}}}, & ha\ x > 0. \end{cases}$$



1.13. ábra. $s_1 = 10$ és $s_2 = 15$ szabadsági fokú F-eloszlású valószínűségi változó sűrűségfüggvénye

1.92. Tétel. *Ha $\xi \in \mathbf{F}(s_1; s_2)$, akkor $\frac{1}{\xi} \in \mathbf{F}(s_2; s_1)$.*

1.93. Tétel. *Ha $\xi \in \mathbf{F}(s_1; s_2)$, akkor $s_2 \geqslant 3$ esetén $\mathrm{E}\,\xi = \frac{s_2}{s_2-2}$ illetve $s_2 \geqslant 5$ esetén $\mathrm{D}^2\,\xi = \frac{2s_2^2(s_1+s_2-2)}{s_1(s_2-2)^2(s_2-4)}$.*

1.94. Tétel. *Ha $\xi \in \mathbf{t}(s)$, akkor $\xi^2 \in \mathbf{F}(1;s)$.*

1.95. Lemma. *Legyen $\xi \in \mathbf{F}(s_1;s_2)$ eloszlásfüggvénye F_{s_1,s_2}. Ekkor F_{s_1,s_2} az s_1 változóban monoton csökkenő, míg az s_2 változóban monoton növekvő, továbbá $0{,}3 < F_{s_1,1}(1) \leqslant F_{s_1,s_2}(1) \leqslant F_{1,s_2}(1) < 0{,}7$.*

1.10. Nagy számok törvényei

1.96. Tétel (Csebisev-egyenlőtlenség)**.** *Ha ξ véges szórással rendelkező valószínűségi változó, akkor minden $\varepsilon \in \mathbb{R}_+$ esetén*

$$\mathrm{P}\Big(|\xi - \mathrm{E}\,\xi| \geqslant \varepsilon\Big) \leqslant \frac{\mathrm{D}^2\,\xi}{\varepsilon^2}.$$

Speciálisan, ha ξ relatív gyakoriságot jelent, akkor kapjuk a következő fontos tételt.

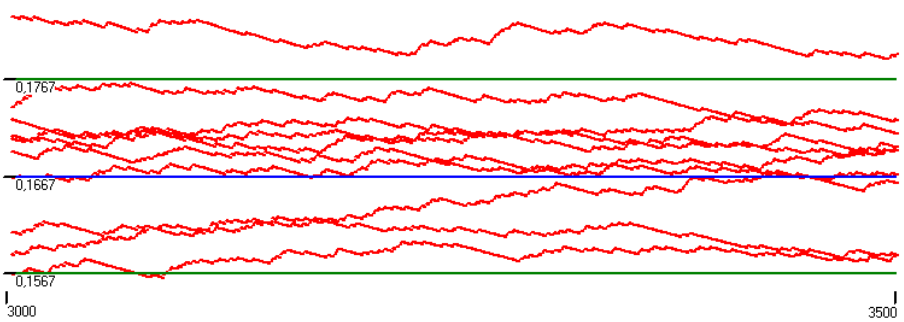
1.97. Tétel (Bernoulli-féle nagy számok törvénye)**.** *Legyen $\frac{\varrho_n}{n}$ az A esemény relatív gyakorisága n kísérlet után. Ekkor*

$$\mathrm{P}\left(\left|\frac{\varrho_n}{n} - \mathrm{P}(A)\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{P}(A)\,\mathrm{P}(\overline{A})}{n\varepsilon^2}$$

minden $\varepsilon \in \mathbb{R}_+$ esetén.

Tehát annak a valószínűsége, hogy az A esemény relatív gyakorisága $\mathrm{P}(A)$-nak az ε sugarú környezetén kívül legyen, az n növelésével egyre kisebb, határértékben 0. Ez pontosan ráillik a Bernoulli-féle tapasztalatra.

A következő ábrán a hatos dobás relatív gyakoriságát láthatjuk szabályos kockával 10 dobássorozat után, 3000-től 3500 dobásig.



A kék vonal jelzi a hatos dobás valószínűségét, míg a zöld vonalak annak $\varepsilon = 0{,}01$ sugarú környezetét. Az ábrán láthatjuk, hogy a 10 dobássorozatból 8 esetén

a relatív gyakoriság 0,01 pontossággal megközelítette a valószínűséget a 3000-től 3500-ig terjedő intervallumon.

A következő videóban az előző kísérletsorozatot vizsgáljuk többféle paraméterezéssel.

VIDEÓ

Az előző videóban használt program letölthető innen:

PROGRAM

A Bernoulli-féle nagy számok törvénye megfogalmazható valószínűségi változókkal is. Hajtsunk végre egy kísérletet n-szer egymástól függetlenül. Ha egy A esemény az i-edik kísérletben bekövetkezik, akkor a ξ_i valószínűségi változó értéke legyen 1, különben pedig 0. A $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók ekkor $\mathrm{P}(A)$ paraméterű karakterisztikus eloszlású páronként független valószínűségi változók, melyeknek a számtani közepe az A relatív gyakorisága, másrészt ekkor $\mathrm{E}\,\xi_1 = \mathrm{P}(A)$ és $\mathrm{D}^2\,\xi_1 = \mathrm{P}(A)\,\mathrm{P}(\overline{A})$. Így tehát bármely $\varepsilon \in \mathbb{R}_+$ esetén

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathrm{E}\,\xi_1\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{D}^2\,\xi_1}{n\varepsilon^2}.$$

Más eloszlású valószínűségi változók számtani közepe is hasonló tulajdonságot mutat.

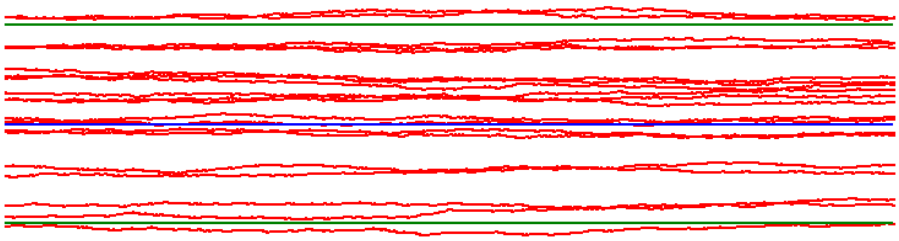
1.98. Tétel (Nagy számok gyenge törvénye)**.** *Legyenek $\xi_1, \xi_2, \dots, \xi_n$ véges várható értékű és szórású, azonos eloszlású, páronként független valószínűségi változók. Ekkor*

$$\mathrm{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mathrm{E}\,\xi_1\right| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{D}^2\,\xi_1}{n\varepsilon^2},$$

minden $\varepsilon \in \mathbb{R}_+$ esetén.

Tehát annak a valószínűsége, hogy a valószínűségi változók számtani közepe a várható érték ε sugarú környezetén kívül legyen, az n növelésével egyre kisebb, határértékben 0.

A következő ábrán n darab standard normális eloszlású páronként független valószínűségi változó számtani közepét láthatjuk n függvényében $n = 29\,500$-tól $n = 30\,000$-ig, 20 kísérletsorozat után.



A kék vonal jelzi a várható értéket (ez most 0), míg a zöld vonalak annak $\varepsilon = 0{,}01$ sugarú környezetét. Az ábrán láthatjuk, hogy a 20 kísérletsorozatból 17 esetén a számtani közép 0,01 pontossággal megközelítette a várható értéket a 29 500-tól 30 000-ig terjedő intervallumon.

A következő videóban az előző kísérletsorozatot vizsgáljuk többféle eloszlás esetén.

V I D E Ó

Két független standard normális eloszlású valószínűségi változó hányadosa Cauchy-eloszlású. Erről ismert, hogy nincs várható értéke. Így erre nem teljesül a nagy számok gyenge törvénye. Ezt szemlélteti a következő videó.

V I D E Ó

1.99. Tétel (Nagy számok Kolmogorov-féle erős törvénye)**.** *$\xi_1, \xi_2, \dots$ legyenek független, azonos eloszlású valószínűségi változók és* $\mathrm{E}\,|\xi_1| \in \mathbb{R}$*. Ekkor*

$$\mathrm{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \xi_i = \mathrm{E}\,\xi_1\right) = 1.$$

Ez a tétel az előzőnél erősebb állítást fogalmaz meg. *Etemadi* (1981) és *Petrov* (1987) eredményeiből kiderült, hogy a nagy számok Kolmogorov-féle erős törvényének állítása páronkénti függetlenség esetén is igaz marad.

1.11. Centrális határeloszlási tétel

A valószínűségszámításban és a matematikai statisztikában központi szerepe van a standard normális eloszlásnak. Ennek okát mutatja a következő tétel.

1.100. Tétel (Centrális határeloszlási tétel)**.** *Legyenek $\xi_1, \xi_2, \dots$ független, azonos eloszlású, pozitív véges szórású valószínűségi változók. Ekkor*

$$\eta_n := \frac{\sum_{i=1}^{n} \xi_i - \mathrm{E}\sum_{i=1}^{n} \xi_i}{\mathrm{D}\sum_{i=1}^{n} \xi_i}$$

határeloszlása standard normális, azaz

$$\lim_{n\to\infty} \mathrm{P}\left(\eta_n < x\right) = \Phi(x)$$

minden $x \in \mathbb{R}$ *esetén.*

Speciálisan, ha $\xi_1, \xi_2, \ldots$ függetlenek és p paraméterű karakterisztikus eloszlásúak, akkor $\sum_{i=1}^{n} \xi_i$ egy n-edrendű p paraméterű binomiális eloszlású valószínűségi változó. Ennek várható értéke np és szórásnégyzete $np(1-p)$. Erre alkalmazva a centrális határeloszlás tételét, kapjuk, hogy minden $x \in \mathbb{R}$ esetén

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{\sum_{i=1}^{n} \xi_i - np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

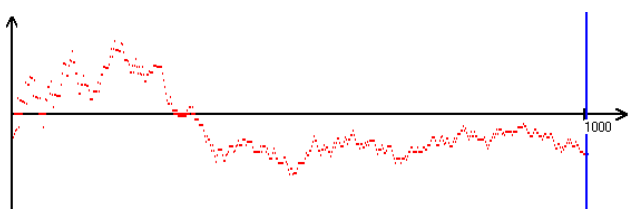
Ez az ún. *Moivre–Laplace-tétel*. Ez ekvivalens azzal, hogy $x \in \mathbb{R}$ és $\Delta x > 0$ esetén

$$\lim_{n\to\infty} \mathrm{P}\left(x \leqslant \frac{\sum_{i=1}^{n} \xi_i - np}{\sqrt{np(1-p)}} < x + \Delta x\right) = \frac{1}{\sqrt{2\pi}} \int_{x}^{x+\Delta x} e^{-\frac{t^2}{2}}\, \mathrm{d}t.$$

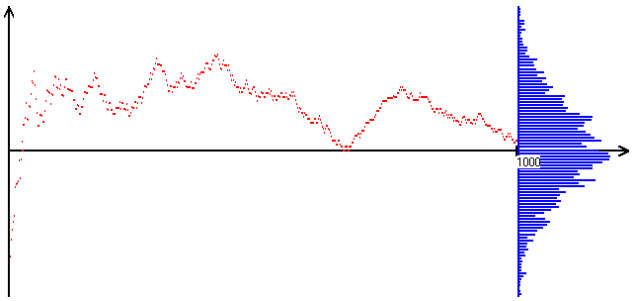
Így nagy n és kicsiny Δx esetén

$$\frac{1}{\Delta x}\, \mathrm{P}\left(x \leqslant \frac{\sum_{i=1}^{n} \xi_i - np}{\sqrt{np(1-p)}} < x + \Delta x\right) \simeq \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

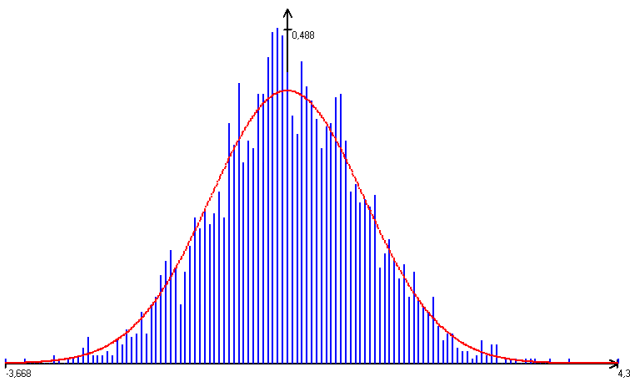
Legyen k_m egy p valószínűségű esemény gyakorisága m kísérlet után. Ábrázoljuk m függvényében a $\frac{k_m - mp}{\sqrt{mp(1-p)}}$ értékeket, ahol $m = 1, 2, \ldots, n$. A következő ábra ezt mutatja $p = 0{,}5$ és $n = 1000$ esetén.



A kísérletsorozatot megismételjük N-szer. A kék vonalon ábrázoljuk a becsapódások számát vonaldiagrammal. A következő ábrán ez látható $N = 3000$ esetén.



Végül a vonaldiagramot normáljuk N-nel és Δx-szel, mely már összehasonlítható a standard normális eloszlás sűrűségfüggvényével.



A következő videóban az előző kísérletsorozatot folyamatában vizsgáljuk.

VIDEÓ

2. A matematikai statisztika alapfogalmai

A valószínűségszámítás órákon tárgyalt feladatokban mindig szerepel valamilyen információ bizonyos típusú véletlen események valószínűségére vonatkozóan. Például:

- *Mi a valószínűsége annak, hogy két szabályos kockával dobva a kapott számok összege 7?*
  Itt a szabályosság azt jelenti, hogy a kocka bármely oldalára $\frac{1}{6}$ valószínűséggel eshet.

- *Egy boltban az átlagos várakozási idő 2 perc. Mi a valószínűsége, hogy 3 percen belül nem kerülünk sorra, ha a várakozási idő exponenciális eloszlású?*
  Itt az adott információk alapján $1 - e^{-\frac{x}{2}}$ annak a valószínűsége, hogy a várakozási idő kevesebb mint x perc.

Ha egy hasonló feladatban a megoldáshoz szükséges információk nem mindegyike ismert, akkor azokat nekünk kell tapasztalati úton meghatározni. A *matematikai statisztika* ilyen jellegű problémákkal foglalkozik.

A statisztikai feladatokban tehát az események rendszere, pontosabban az $(\Omega, \mathcal{F})$ mérhető tér adott, de a valószínűség nem.

Legyen $\mathcal{P}$ azon $\mathrm{P}\colon \mathcal{F} \to \mathbb{R}$ függvények halmaza, melyekre $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mező. Ekkor az $(\Omega, \mathcal{F}, \mathcal{P})$ rendezett hármast *statisztikai mezőnek* nevezzük. Az ideális az lenne, ha $\mathcal{P}$-ből ki tudnánk választani az igazi P-t. Sok esetben azonban erre nincs is szükség. Például ha az A és B események függetlenségét kell kimutatnunk, akkor csak azt kell megvizsgálni, hogy az igazi P-re teljesül-e az a tulajdonság, hogy $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$.

A statisztikai feladatokról azt is fontos tudnunk, hogy azok mindig megfogalmazhatók valószínűségi (vektor)változók segítségével. Ennek szemléltetésére tekintsük a következő példákat.

- *Döntsük el egy dobókockáról, hogy az cinkelt-e.* A probléma matematikai modellezésében legyen $\Omega = \{\,1, 2, 3, 4, 5, 6\,\}$, $\mathcal{F}$ az Ω hatványhalmaza és $\xi\colon \Omega \to \mathbb{R}$, $\xi(k) = k$. Ekkor azt kell kideríteni, hogy ξ diszkrét egyenletes eloszlású-e, azaz teljesül-e az igazi P-re, hogy minden $k = 1, 2, 3, 4, 5, 6$ esetén $\mathrm{P}(\xi = k) = \frac{1}{6}$.

- *Az emberek szem- és hajszíne független, vagy van közöttük genetikai kapcsolat?* A H halmaz elemei legyenek a haj lehetséges színei, illetve az S halmaz elemei a szem lehetséges színei. Legyen $\Omega := H \times S$ és $\mathcal{F}$ az Ω hatványhalmaza. Ekkor például a (barna, kék) $\in \Omega$ elemi esemény modellezze

azt, hogy a véletlenül kiválasztott személy barna hajú és kék szemű. Legyen $\xi\colon \Omega \to \mathbb{R},\ \xi(h,s) = 1,2,3,\dots$ aszerint, hogy h = szőke, barna, fekete, ... és $\eta\colon \Omega \to \mathbb{R},\ \eta(h,s) = 1,2,3,\dots$ aszerint, hogy s = kék, barna, zöld, ... Ekkor a $\zeta = (\xi,\eta)$ valószínűségi vektorváltozó eloszlását kell meghatározni, pontosabban az a kérdés, hogy az igazi P-re teljesül-e, hogy

$$\mathrm{P}(\xi = i, \eta = j) = \mathrm{P}(\xi = i)\,\mathrm{P}(\eta = j)$$

minden $i = 1,2,\dots$ és $j = 1,2,\dots$ esetén.

- *Két esemény közül döntsük el, hogy melyiknek nagyobb a valószínűsége.* Legyen a két esemény A és B. Ezen események indikátorváltozóira teljesülnek, hogy $\mathrm{E}\,\mathrm{I}_A = \mathrm{P}(A)$ és $\mathrm{E}\,\mathrm{I}_B = \mathrm{P}(B)$. Így tehát azt kell eldöntenünk, hogy a két esemény indikátorváltozói közöl melyiknek nagyobb a várható értéke.

2.1. Minta és mintarealizáció

A statisztikában tehát egy valószínűségi (vektor)változóra vonatkozólag kell információkat gyűjteni. Jelöljük ezt ξ-vel. Tegyük fel, hogy ξ az $(\Omega_0, \mathcal{F}_0, \mathrm{P}_0)$ valószínűségi mezőben van értelmezve, ahol P_0 a valódi (általunk nem ismert) valószínűséget jelenti. Az adatgyűjtésnek a statisztikában egyetlen módja van, a ξ-t meg kell figyelni (mérni) többször, egymástól függetlenül. Az i-edik megfigyelés eredményét jelölje ξ_i, amely egy véletlen érték, vagyis valószínűségi (vektor)változó. Mindez a következőképpen modellezhető.

Legyen $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ azon független kísérletek valószínűségi mezője, amely az $(\Omega_0, \mathcal{F}_0, \mathrm{P}_0)$ kísérlet n-szeri független elvégzését modellezi. Tegyük fel, hogy ξ d-dimenziós. Legyen

$$\xi_i\colon \Omega_n \to \mathbb{R}^d,\ \xi_i(\omega_1,\dots,\omega_n) := \xi(\omega_i)\ (i = 1,\dots,n).$$

Ekkor tetszőleges $x \in \mathbb{R}^d$ esetén

$$\begin{aligned}
&\mathrm{P}_n(\xi_i < x) = \mathrm{P}_n\big(\{(\omega_1,\dots,\omega_n) \in \Omega_n : \xi(\omega_i) < x\}\big) = \\
&= \mathrm{P}_n\big(\Omega_0 \times \cdots \times \Omega_0 \times \{\xi < x\} \times \Omega_0 \times \cdots \times \Omega_0\big) = \\
&= \mathrm{P}_0(\Omega_0) \cdots \mathrm{P}_0(\Omega_0)\,\mathrm{P}_0(\xi < x)\,\mathrm{P}_0(\Omega_0) \cdots \mathrm{P}_0(\Omega_0) = \mathrm{P}_0(\xi < x),
\end{aligned}$$

azaz ξ_i és ξ azonos eloszlású. Másrészt tetszőleges $x_1, \dots, x_n \in \mathbb{R}^d$ esetén

$$\begin{aligned}
&\mathrm{P}_n(\xi_1 < x_1, \dots, \xi_n < x_n) = \\
&= \mathrm{P}_n\big(\{(\omega_1, \dots, \omega_n) \in \Omega_n : \xi(\omega_1) < x_1, \dots, \xi(\omega_n) < x_n\}\big) = \\
&= \mathrm{P}_n\big((\xi < x_1) \times \dots \times (\xi < x_n)\big) = \prod_{i=1}^{n} \mathrm{P}_0(\xi < x_i) = \prod_{i=1}^{n} \mathrm{P}_n(\xi_i < x_i),
\end{aligned}$$

azaz a ξ_i valószínűségi változók függetlenek.

Összefoglalva tehát az n megfigyelés modellezhető $\xi_1, \dots, \xi_n$ független, ξ-vel azonos eloszlású valószínűségi (vektor)változókkal. Mivel valójában minket csak a ξ valódi eloszlása érdekel, matematikai értelemben nincs jelentősége, hogy a ξ és ξ_i-k különböző valószínűségi mezőben vannak értelmezve. Ezért megállapodunk abban, hogy a továbbiakban a $\xi, \xi_1, \xi_2, \dots$ valószínűségi változók ugyanazon $(\Omega, \mathcal{F}, \mathrm{P})$ valószínűségi mezőn értelmezettek, ahol P az általunk nem ismert valódi valószínűség.

2.1. Definíció. A ξ valószínűségi (vektor)változóra vonatkozó *n elemű minta* alatt a ξ-vel azonos eloszlású $\xi_1, \dots, \xi_n$ független, valószínűségi (vektor)változókat értünk. A ξ_k-t *k-adik mintaelemnek*, n-et pedig a *mintaelemek számának* nevezzük.

Természetesen, ha több valószínűségi (vektor)változóra is szükségünk van, akkor mindegyikre kell megfigyeléseket végezni, így több mintánk is lesz.

A gyakorlatban nem mintával dolgozunk, hanem konkrét értékekkel, melyek a mintaelemek lehetséges értékei.

2.2. Definíció. Ha $\xi_1, \dots, \xi_n$ a ξ valószínűségi (vektor)változóra vonatkozó minta és $\omega \in \Omega$, akkor a $\xi_1(\omega), \dots, \xi_n(\omega)$ értékeket ξ-re vonatkozó *mintarealizációnak* nevezzük. Az olyan $(x_1, \dots, x_n)$ elem n-esek halmazát, melyekre teljesül, hogy az x_i benne van a ξ értékkészletében $(i = 1, \dots, n)$, *mintatérnek* nevezzük.

Statisztikai feladatokban mintarealizáció alapján számolunk. Az így meghozott döntés nem biztos, hogy megfelel a valóságnak, csak annyit mondhatunk róla, hogy nem mond ellent a mintarealizációnak. Azaz az ilyen döntés hibás is lehet, így a válaszunkban azt is meg kell adni, hogy mi a valószínűsége ennek a hibának.

2.2. Tapasztalati eloszlásfüggvény

Ebben a részben feltételezzük, hogy egy ξ valószínűségi változó (tehát nem vektorváltozó) tulajdonságait kell megfigyelni. A legjobb az lenne, ha az F eloszlásfüggvényét

sikerülne meghatározni. Valójában – az előbb elmondottak miatt – F-fet meghatározni a mintarealizáció alapján nem tudjuk, de becsülni igen. Egy rögzített $x \in \mathbb{R}$ esetén $F(x) = \mathrm{P}(\xi < x)$. Tehát egy esemény valószínűségét kell megbecsülni. A valószínűség definícióját a relatív gyakoriság tulajdonságai sugallták, így az a sejtésünk, hogy egy esemény valószínűségét a relatív gyakoriságával lenne érdemes becsülni. A $\xi < x$ esemény relatív gyakorisága a ξ-re vonatkozó $\xi_1, \ldots, \xi_n$ minta alapján könnyen megadható indikátorváltozókkal: $\frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\xi_i < x}$. Itt $\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x}$ azon mintaelemek számát jelenti, melyek kisebbek x-nél. A későbbiekben látni fogjuk, hogy ez a becslés valóban megfelelő lesz számunkra.

2.3. Definíció. Legyen $\xi_1, \ldots, \xi_n$ egy ξ valószínűségi változóra vonatkozó minta. Ekkor az

$$x \mapsto F_n^*(x) := \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\xi_i < x} \quad (x \in \mathbb{R})$$

függvényt a ξ-re vonatkozó n elemű mintához tartozó *tapasztalati eloszlásfüggvényének* nevezzük.

Az $F_n^*(x)$ minden rögzített $x \in \mathbb{R}$ esetén egy valószínűségi változó. Ha a kísérletsorozatban az $\omega \in \Omega$ elemi esemény következett be, azaz a mintarealizáció $\xi_1(\omega), \ldots, \xi_n(\omega)$, akkor az

$$x \mapsto \big(F_n^*(x)\big)(\omega) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\xi_i(\omega) < x} \quad (x \in \mathbb{R})$$

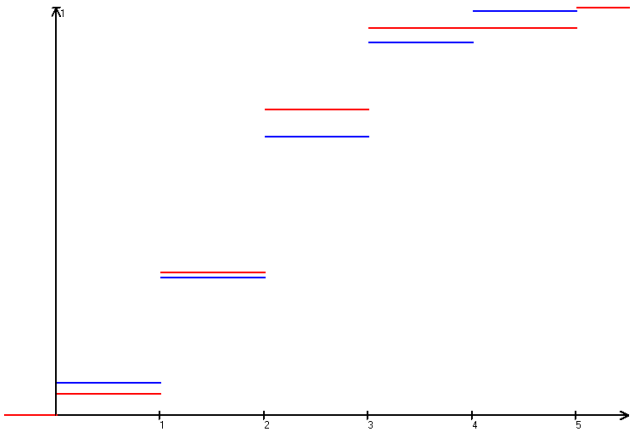
hozzárendelés egy valós függvény. Ezt a függvényt a *tapasztalati eloszlásfüggvény egy realizációjának* nevezzük, de a továbbiakban a rövidség kedvéért ezt is csak tapasztalati eloszlásfüggvényként emlegetjük és F_n^* módon jelöljük.

Példaként legyen ξ egy dobókockával dobott szám, és a mintarealizáció 3, 4, 5, 3, 6, 2, 3, 3, 5, 2. Ekkor

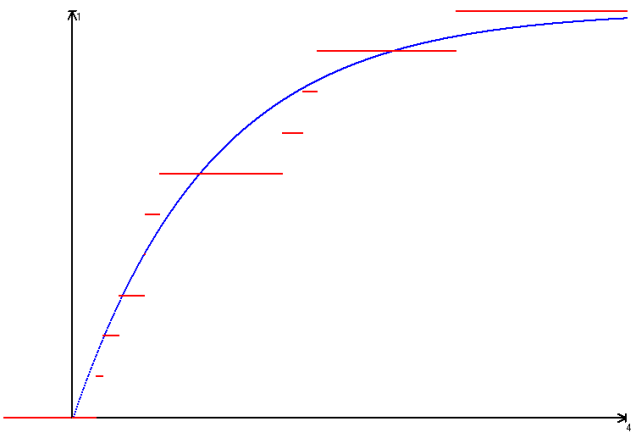
$$F_{10}^*(x) = \begin{cases} 0 & \text{ha } x \leqslant 2, \\ 0{,}2 & \text{ha } 2 < x \leqslant 3, \\ 0{,}6 & \text{ha } 3 < x \leqslant 4, \\ 0{,}7 & \text{ha } 4 < x \leqslant 5, \\ 0{,}9 & \text{ha } 5 < x \leqslant 6, \\ 1 & \text{ha } x > 6. \end{cases}$$

A következő ábrán egy $\mathbf{Bin}(5; 0{,}2)$-beli valószínűségi változóra vonatkozó 20 elemű

mintához tartozó tapasztalati eloszlásfüggvényt láthatunk.



A kék grafikon a valódi eloszlásfüggvényt jelenti, a piros a tapasztalatit. Vegyük észre, hogy a tapasztalati eloszlásfüggvény mindig lépcsős függvény, azaz az értékkészlete véges. Nevezetesen n elemű minta esetén az F_n^* maximálisan $n+1$ féle értéket vehet fel. Így felmerül a kérdés, hogy a lépcsős tapasztalati eloszlásfüggvény hogyan néz ki folytonos eloszlásfüggvényű valószínűségi változó esetén. A következő ábrán egy **Exp**(1)-beli valószínűségi változóra vonatkozó 10 elemű mintához tartozó tapasztalati eloszlásfüggvényt láthatunk.



A kék grafikon itt is a valódi eloszlásfüggvényt jelenti, a piros a tapasztalatit.

A tapasztalati eloszlásfüggvény megfelelő becslése-e a valódi eloszlásfüggvénynek? Az előző példákban, ahol a megfigyelések száma (n) viszonylag kevés, elég nagy eltéréseket láthatunk. De az n növelésével javul-e ez a helyzet? A következő Glivenkotól és Cantellitől származó tétel erről ad információt.

2.4. Tétel (A matematikai statisztika alaptétele)**.** *Legyen a ξ valószínűségi változó valódi eloszlásfüggvénye F és a ξ-re vonatkozó n elemű mintához tartozó tapasztalati*

eloszlásfüggvény F_n^*. Ekkor

$$\mathrm{P}\left(\lim_{n\to\infty}\sup_{x\in\mathbb{R}}|F_n^*(x)-F(x)|=0\right)=1,$$

azaz F_n^* egyenletesen konvergál $\mathbb{R}$-en F-hez majdnem biztosan.

Bizonyítás. Legyen $\varepsilon\in\mathbb{R}_+$ rögzített és $m\in\mathbb{N}$ olyan, hogy $\frac{1}{m}<\frac{\varepsilon}{2}$. Ha $k\in$ $\in\{1,\dots,m-1\}$, akkor az F balról való folytonossága miatt az $\{x\in\mathbb{R}:F(x)\leqslant\frac{k}{m}\}$ halmaznak létezik maximuma. Ezt a maximumot jelöljük x_k-val. Legyen továbbá $x_0:=-\infty$ és $x_m:=\infty$. Ekkor

$$\mathrm{P}(\xi<x_k)=F(x_k)\leqslant\frac{k}{m}\leqslant\lim_{x\to x_k+0}F(x)=\mathrm{P}(\xi\leqslant x_k)\quad(k=0,\dots,m).$$

Így

$$\mathrm{P}(\xi<x_k)\leqslant\frac{k-1}{m}+\frac{1}{m}\leqslant\mathrm{P}(\xi\leqslant x_{k-1})+\frac{1}{m}.$$

Jelentse A_k azt az eseményt, hogy $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i<x_k}=\mathrm{P}(\xi<x_k)$, illetve B_k azt, hogy $\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i\leqslant x_k}=\mathrm{P}(\xi\leqslant x_k)$. A nagy számok erős törvénye miatt $\mathrm{P}(A_k)=\mathrm{P}(B_k)=1$ $(k=0,\dots,m)$. Ebből

$$A:=\bigcap_{k=0}^m\bigcap_{l=0}^m(A_k\cap B_l)$$

jelöléssel $\mathrm{P}(A)=1$ teljesül. Emiatt létezik $N\in\mathbb{N}$, hogy minden $n>N$ egész szám és $k=0,\dots,m$ esetén az A-n teljesül, hogy

$$\left|\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i<x_k}-\mathrm{P}(\xi<x_k)\right|<\frac{\varepsilon}{2}\quad\text{és}\quad\left|\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i\leqslant x_k}-\mathrm{P}(\xi\leqslant x_k)\right|<\frac{\varepsilon}{2}.$$

Legyen $x\in\mathbb{R}$ rögzített. Ekkor létezik $t\in\{1,\dots,m\}$, hogy

$$x_{t-1}<x\leqslant x_t.$$

Mindezek alapján minden $n>N$ egész esetén az A-n teljesül, hogy

$$F(x)-F_n^*(x)=\mathrm{P}(\xi<x)-\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i<x}\leqslant$$

$$\leqslant\mathrm{P}(\xi<x_t)-\frac{1}{n}\sum_{i=1}^n\mathrm{I}_{\xi_i<x}\leqslant$$

$$\leqslant \frac{1}{m} + \mathrm{P}(\xi \leqslant x_{t-1}) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x} \leqslant$$
$$\leqslant \frac{1}{m} + \mathrm{P}(\xi \leqslant x_{t-1}) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i \leqslant x_{t-1}} < \frac{1}{m} + \frac{\varepsilon}{2} < \varepsilon.$$

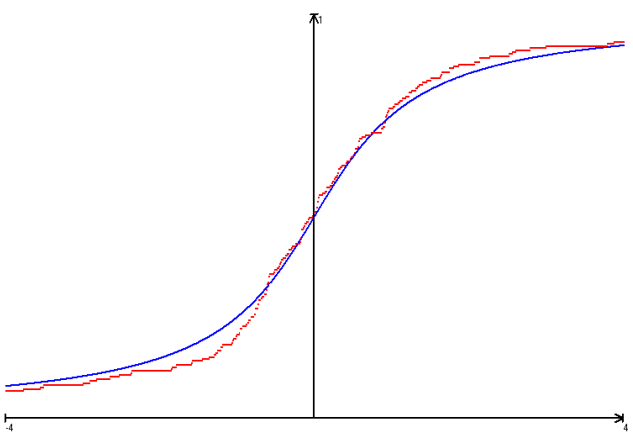
Hasonlóan teljesül minden $n > N$ egész esetén az A-n, hogy

$$F(x) - F_n^*(x) = \mathrm{P}(\xi < x) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x} \geqslant$$
$$\geqslant \mathrm{P}(\xi \leqslant x_{t-1}) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x} \geqslant$$
$$\geqslant -\frac{1}{m} + \mathrm{P}(\xi < x_t) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x} \geqslant$$
$$\geqslant -\frac{1}{m} + \mathrm{P}(\xi < x_t) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{\xi_i < x_t} > -\frac{1}{m} - \frac{\varepsilon}{2} > -\varepsilon.$$

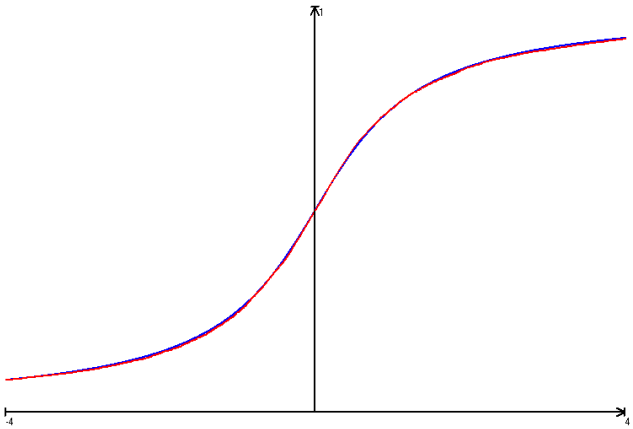
Így $|F(x) - F_n^*(x)| < \varepsilon$ teljesül az A-n, ha $n > N$. Ebből már következik a tétel.

Az előző tételben fontos az egyenletes konvergencia. Ugyanis ha csak pontonkénti lenne, akkor a számegyenes különböző helyein más és más sebességű lehetne. Így ebben az esetben a tapasztalati eloszlásfüggvény alakjából a valódira nem lehetne következtetni.

A következő két ábrán egy Cauchy-eloszlású valószínűségi változóra vonatkozó 200 illetve 10 000 elemű mintának a tapasztalati eloszlásfüggvényét látjuk. (Két független standard normális eloszlású valószínűségi változó hányadosát nevezzük Cauchy-eloszlásúnak.)



2.1. ábra. F_{200}^* grafikonja



2.2. ábra. $F^*_{10\,000}$ grafikonja

A kék grafikon a valódi eloszlásfüggvényt jelenti, míg a piros a tapasztalatit.

Látható, hogy 10 000-es mintaelemszám esetén már gyakorlatilag megegyezik a tapasztalati és a valódi eloszlásfüggvény. Az utóbbi ábrán úgy tűnhet, hogy a tapasztalati eloszlásfüggvény nem lépcsős. Természetesen ez nem igaz, pusztán arról van szó, hogy egy „lépcsőfok" hossza olyan kicsi, hogy az a rajz felbontása miatt csak egy pontnak látszik.

A következő videóban többféle eloszlással vizsgáljuk a tapasztalati eloszlásfüggvény konvergenciáját.

V I D E Ó

Az előző videóban használt program letölthető innen:

P R O G R A M

2.3. Tapasztalati eloszlás, sűrűséghisztogram

Tapasztalati eloszlásfüggvény helyett más lehetőség is van valószínűségi változók eloszlásának vizsgálatára.

Diszkrét valószínűségi változó esetén vizsgálhatjuk az úgynevezett *tapasztalati eloszlást* is, mely a valószínűségi változó egy lehetséges értékéhez hozzárendeli a kísérletsorozatbeli relatív gyakoriságát. Azaz, ha a ξ valószínűségi változó értékkészlete $\{ x_1, \ldots, x_k \}$ és a ξ-re vonatkozó minta $\xi_1, \ldots, \xi_n$, akkor a tapasztalati eloszlás az

$$x_t \mapsto r_t := \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\xi_i = x_t} \ (t = 1, \ldots, k)$$

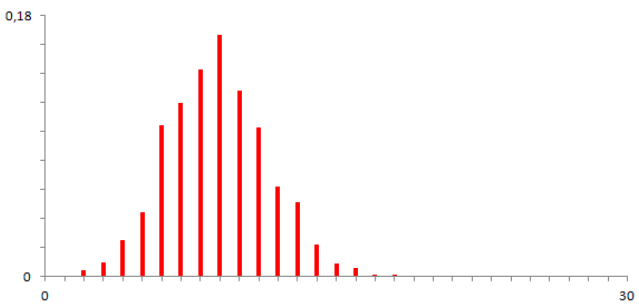
hozzárendelés. (Tehát nr_t a mintában az x_t-vel egyenlő elemek számát jelenti.)

Ha a kísérletsorozatban az $\omega \in \Omega$ elemi esemény következett be, azaz a mintarealizáció $\xi_1(\omega), \ldots, \xi_n(\omega)$, akkor az

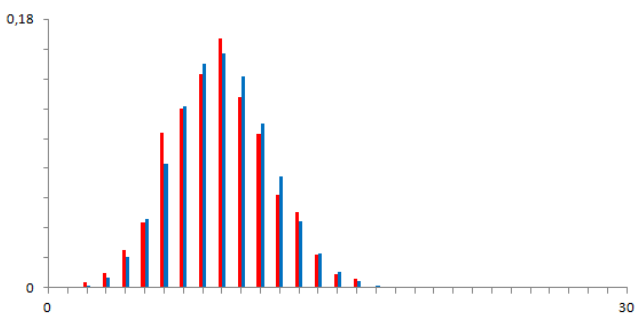
$$x_t \mapsto r_t(\omega) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\xi_i(\omega)=x_t} \ (t = 1, \ldots, k)$$

hozzárendelést a *tapasztalati eloszlás egy realizációjának* nevezzük, de a továbbiakban a rövidség kedvéért ezt is csak tapasztalati eloszlásként emlegetjük. Ezt célszerű *vonaldiagrammal* ábrázolni. Ez azt jelenti, hogy az $(x_t,0)$ koordinátájú pontot összekötjük az $\big(x_t, r_t(\omega)\big)$ ponttal minden t-re.

A következő képen egy **Bin**$(30; 0{,}3)$-beli valószínűségi változóra vonatkozó 1000 elemű mintarealizációból számolt tapasztalati eloszlást láthatunk vonaldiagrammal ábrázolva.



Ugyanezen az ábrán kékkel felrajzoljuk a valódi eloszlást is, mely jól mutatja a hasonlóságot.



Abszolút folytonos ξ valószínűségi változó esetén az ún. sűrűséghisztogram vizsgálata is célravezető lehet a tapasztalati eloszlásfüggvény mellett. Legyen $r \in \mathbb{N}$, $x_0, x_1, \ldots, x_r \in \mathbb{R}$ és $x_0 < x_1 < \cdots < x_r$. Tegyük fel, hogy a ξ-re vonatkozó $\xi_1(\omega), \ldots, \xi_n(\omega)$ mintarealizáció minden eleme benne van az (x_0, x_r) intervallumban. Minden $[x_{j-1}, x_j)$ intervallum fölé rajzoljunk egy y_j magasságú téglalapot úgy, hogy a téglalap területe a valódi f sűrűségfüggvény görbéje alatti területet becsülje

az $[x_{j-1}, x_j)$ intervallumon. Hasonlóan az eddigiekhez, egy esemény valószínűségét itt is az esemény relatív gyakoriságával becsüljük. Így tehát

$$\int_{x_{j-1}}^{x_j} f(x)\,\mathrm{d}x = \mathrm{P}(x_{j-1} \leqslant \xi < x_j) \simeq \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}_{x_{j-1}\leqslant\xi_i(\omega)<x_j} = y_j(x_j - x_{j-1}),$$

melyből

$$y_j = \frac{\sum_{i=1}^{n} \mathrm{I}_{x_{j-1}\leqslant\xi_i(\omega)<x_j}}{n(x_j - x_{j-1})} \quad (j = 1, \ldots, r).$$

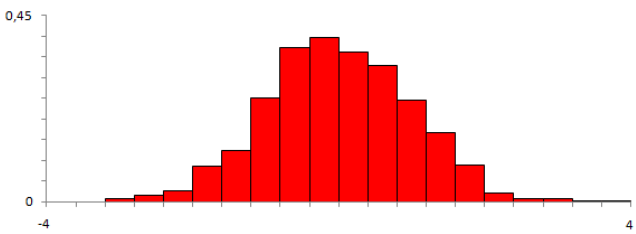
A kapott oszlopdiagramot *sűrűséghisztogramnak* nevezzük, amely tehát a valódi f sűrűségfüggvényt a j-edik részintervallumon az y_j konstanssal közelíti.

A sűrűséghisztogram megadása a mintarealizáció alapján nem egyértelmű, függ az osztópontok választásától. Az osztópontok felvételéhez csak annyi általános irányelv mondható, hogy függetlennek kell lennie a minta értékeitől.

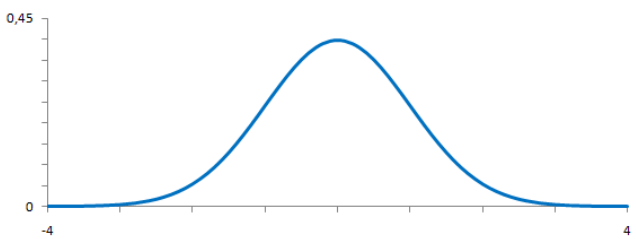
Az is fontos, hogy az osztópontok ne helyezkedjenek el túl sűrűn a mintarealizáció elemeihez képest, mert ekkor egy részintervallumba túl kevés mintaelem fog esni, s így nagyon pontatlan lesz a becslés. Azaz ebben az esetben a sűrűséghisztogramból nem lehet következtetni a valódi sűrűségfüggvény alakjára.

Másrészt, ha az osztópontok túl ritkák, azaz a részintervallumok száma kevés, akkor a sűrűségfüggvény becsült pontjainak száma túl kevés ahhoz, hogy a sűrűséghisztogramból következtetni lehessen a valódi sűrűségfüggvény alakjára.

A következő ábrán standard normális eloszlású 1000 elemű mintára vonatkozó sűrűséghisztogramot láthatunk $r = 20$, $x_0 = -4$, $x_{20} = 4$ választással, továbbá a részintervallumok egyenlő hosszúságúak.



Összehasonlításképpen a következő ábrán a standard normális eloszlás sűrűségfüggvényét láthatjuk a $[-4, 4]$ intervallumon.



2.4. Statisztikák

Tegyük fel, hogy egy ismeretlen eloszlású ξ valószínűségi változó várható értékét kell meghatározni. Mivel az eloszlást nem ismerjük, ezért a minta alapján kell becslést adni. A későbbiekben látni fogjuk, hogy bizonyos szempontból jó becslése a várható értéknek a ξ-re vonatkozó $\xi_1, \dots, \xi_n$ minta elemeinek a számtani közepe, azaz $\frac{1}{n}(\xi_1 + + \cdots + \xi_n)$. Általánosan fogalmazva itt egy olyan függvényt definiáltunk, amely egy valószínűségi változókból álló rendezett n-eshez egy valószínűségi változót rendel. Az ilyen függvényeket *statisztikának* nevezzük, és a következőkben kiemelt szerepük lesz.

2.5. Definíció. Legyen $\xi_1, \dots, \xi_n$ egy ξ valószínűségi változóra vonatkozó minta, továbbá

$$T\colon \mathbb{R}^n \to \mathbb{R}$$

olyan függvény, melyre $T(\xi_1, \dots, \xi_n)$ valószínűségi változó. Ekkor ezt a valószínűségi változót a minta egy *statisztikájának* nevezzük. Ha $\xi_1(\omega), \dots, \xi_n(\omega)$ egy a ξ-re vonatkozó mintarealizáció, akkor a $T\big(\xi_1(\omega), \dots, \xi_n(\omega)\big)$ számot az előbbi *statisztika egy realizációjának* nevezzük.

Ha T Borel-mérhető függvény, akkor $T(\xi_1, \dots, \xi_n)$ mérhető, azaz valószínűségi változó. Például $F_n^*(x)$ minden rögzített $x \in \mathbb{R}$ esetén statisztika.

2.6. Definíció. Legyen $\xi_1, \dots, \xi_n$ egy ξ valószínűségi változóra vonatkozó minta. A következő nevezetes statisztikákat definiáljuk:

mintaátlag	$\overline{\xi} := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \xi_i$
tapasztalati szórásnégyzet	$S_n^2 := \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2$
tapasztalati szórás	$S_n := \sqrt{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}$
korrigált tapasztalati szórásnégyzet	$S_n^{*2} := \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2$
korrigált tapasztalati szórás	$S_n^* := \sqrt{\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} (\xi_i - \overline{\xi})^2}$
k-adik tapasztalati momentum ($k \in \mathbb{N}$)	$\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \xi_i^k$

k-adik tapasztalati centrált momentum ($k \in \mathbb{N}$)	$\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^k$
tapasztalati ferdeség	$\frac{\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^3}{S_n^3}$
tapasztalati lapultság	$\frac{\frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^4}{S_n^4} - 3$

Ha több valószínűségi változót is vizsgálunk és hangsúlyozni szeretnénk, hogy a tapasztalati illetve korrigált tapasztalati szórás a ξ-re vonatkozik, akkor azokat $S_{\xi,n}$ illetve $S^*_{\xi,n}$ módon fogjuk jelölni.

2.7. Tétel (Steiner-formula)**.** *Bármely $c \in \mathbb{R}$ esetén*

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

Bizonyítás. Legyen $c \in \mathbb{R}$ tetszőlegesen rögzített. Ekkor

$$\begin{aligned} S_n^2 &= \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \overline{\xi})^2 = \frac{1}{n}\sum_{i=1}^{n}\big((\xi_i - c) - (\overline{\xi} - c)\big)^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(\xi_i - c)^2 - \frac{1}{n}\sum_{i=1}^{n}2(\overline{\xi} - c)(\xi_i - c) + \frac{1}{n}\sum_{i=1}^{n}(\overline{\xi} - c)^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(\xi_i - c)^2 - 2(\overline{\xi} - c)^2 + (\overline{\xi} - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - c)^2 - (\overline{\xi} - c)^2. \end{aligned}$$

2.8. Definíció. Legyen $\xi_1, \ldots, \xi_n$ egy ξ valószínűségi változóra vonatkozó minta, továbbá $(x_1, \ldots, x_n) \in \mathbb{R}^n$ esetén jelölje $r_1, \ldots, r_n$ az $1, \ldots, n$ számok egy olyan permutációját, melyre teljesül, hogy

$$x_{r_1} \leqslant x_{r_2} \leqslant \ldots \leqslant x_{r_n}.$$

Legyen

$$T_i \colon \mathbb{R}^n \to \mathbb{R}, \quad T_i(x_1, \ldots, x_n) := x_{r_i} \quad (i = 1, \ldots, n).$$

Ekkor a $\xi_i^* := T_i(\xi_1, \ldots, \xi_n)$ $(i = 1, \ldots, n)$ valószínűségi változókat *rendezett mintának* nevezzük. (Vegyük észre, hogy $\xi_1^* = \min\{\, \xi_1, \ldots, \xi_n \,\}$ és $\xi_n^* = \max\{\, \xi_1, \ldots, \xi_n \,\}$.)

A $\xi_n^* - \xi_1^*$ statisztikát *mintaterjedelemnek* nevezzük. A $\frac{\xi_1^* + \xi_n^*}{2}$ az úgynevezett *terjedelemközép.*

A *tapasztalati medián* legyen $\xi^*_{\frac{n+1}{2}}$, ha n páratlan, illetve $\frac{1}{2}\left(\xi^*_{\frac{n}{2}} + \xi^*_{\frac{n}{2}+1}\right)$, ha n páros.

Legyen $0 \leqslant t \leqslant 1$. A $100t\,\%$*-os tapasztalati kvantilis* legyen $\xi^*_{[nt]+1}$, ha $nt \notin \mathbb{N}$, illetve $t\xi^*_{nt} + (1-t)\xi^*_{nt+1}$, ha $nt \in \mathbb{N}$. (Vegyük észre, hogy az 50%-os tapasztalati kvantilis a tapasztalati mediánnal egyenlő.) A 25%-os tapasztalati kvantilist *tapasztalati alsó kvartilisnek*, illetve a 75%-os tapasztalati kvantilist *tapasztalati felső kvartilisnek* nevezzük.

A *tapasztalati módusz* a mintaelemek között a leggyakrabban előforduló. Ha több ilyen is van, akkor azok között a legkisebb.

2.9. *Megjegyzés.* Az előbbi T_i függvények Borel-mérhetőek, így a rendezett minta elemei statisztikák.

Ha a kísérletsorozatban az $\omega \in \Omega$ elemi esemény következett be, azaz a mintarealizáció $\xi_1(\omega), \ldots, \xi_n(\omega)$, akkor a $\overline{\xi}(\omega) = \frac{1}{n}\sum_{i=1}^n \xi_i(\omega)$ számot is mintaátlagnak nevezzük. Hasonlóan állapodunk meg minden nevezetes statisztika esetén. (Azaz például $S_n(\omega)$-t is tapasztalati szórásnak nevezzük.)

A következőben a statisztika fogalmát kiterjesztjük arra az esetre, amikor a minta elemei valószínűségi vektorváltozók.

2.10. Definíció. Legyen $\xi_1, \ldots, \xi_n$ egy d-dimenziós ξ valószínűségi vektorváltozóra vonatkozó minta, továbbá

$$T\colon (\mathbb{R}^d)^n \to \mathbb{R}$$

olyan függvény, melyre $T(\xi_1, \ldots, \xi_n)$ valószínűségi változó. Ekkor ezt a valószínűségi változót a minta egy *statisztikájának* nevezzük. Ha $\xi_1(\omega), \ldots, \xi_n(\omega)$ egy a ξ-re vonatkozó mintarealizáció, akkor a $T\big(\xi_1(\omega), \ldots, \xi_n(\omega)\big)$ számot az előbbi *statisztika egy realizációjának* nevezzük.

2.11. Definíció. Legyen $\xi = (\eta, \zeta)$ kétdimenziós valószínűségi vektorváltozó, továbbá a rávonatkozó minta $(\eta_1, \zeta_1), \ldots, (\eta_n, \zeta_n)$. Ennek a mintának a *tapasztalati kovarianciája*

$$\mathrm{Cov}_n(\eta, \zeta) := \frac{1}{n}\sum_{i=1}^n \eta_i\zeta_i - \frac{1}{n}\sum_{i=1}^n \eta_i \cdot \frac{1}{n}\sum_{i=1}^n \zeta_i,$$

illetve *tapasztalati korrelációs együtthatója*

$$\mathrm{Corr}_n(\eta, \zeta) := \frac{\mathrm{Cov}_n(\eta, \zeta)}{S_{\eta,n} \cdot S_{\zeta,n}}.$$

2.12. Definíció. Legyen $\xi_1, \ldots, \xi_n$ egy ξ valószínűségi (vektor)változóra vonatkozó minta. A $T(\xi_1, \ldots, \xi_n)$ statisztikát *szimmetrikusnak* nevezzük, ha az $1, \ldots, n$ számok

minden $i_1, \dots, i_n$ permutációja esetén

$$T(\xi_1, \dots, \xi_n) = T(\xi_{i_1}, \dots, \xi_{i_n}).$$

Vegyük észre, hogy az előzőekben definiált minden nevezetes statisztika szimmetrikus.

Még tovább általánosítható a statisztika fogalma, ha több valószínűségi vektorváltozóra vonatkozik.

2.13. Definíció. Legyen $\xi_1^{(i)}, \dots, \xi_{n_i}^{(i)}$ egy d_i-dimenziós $\xi^{(i)}$ valószínűségi vektorváltozóra vonatkozó minta $(i = 1, \dots, k)$, továbbá

$$T\colon (\mathbb{R}^{d_1})^{n_1} \times \cdots \times (\mathbb{R}^{d_k})^{n_k} \to \mathbb{R}$$

olyan függvény, melyre $T\left(\xi_1^{(1)}, \dots, \xi_{n_1}^{(1)}, \dots, \xi_1^{(k)}, \dots, \xi_{n_k}^{(k)}\right)$ valószínűségi változó. Ekkor ezt a valószínűségi változót az előbbi k darab minta egy *statisztikájának* nevezzük.

Ilyen statisztikákra példát, majd a hipotézisvizsgálatoknál látunk.

3. Pontbecslések

3.1. A pontbecslés feladata és jellemzői

Tegyük fel, hogy a vizsgált ξ valószínűségi változóról tudjuk, hogy egyenletes eloszlású az $[a, b]$ intervallumon, de az a és b paramétereket nem ismerjük. Ekkor a vizsgálandó statisztikai mező leszűkül az

$$(\Omega, \mathcal{F}, \mathcal{P}), \quad \mathcal{P} = \{ \mathrm{P}_\vartheta : \vartheta \in \Theta \}$$

mezőre, ahol $\Theta = \{ (a, b) \in \mathbb{R}^2 : a < b \}$ és P_ϑ olyan valószínűség az $(\Omega, \mathcal{F})$ téren, melyre $\mathrm{P}_\vartheta(\xi < x) = \frac{x-a}{b-a}$ teljesül minden $\vartheta = (a, b) \in \Theta$ és $a < x < b$ esetén.

A pontbecslés feladata ebben az esetben az a illetve b valódi értékének becslése. De nem mindig van szükség az összes ismeretlen paraméterre. Például előfordulhat, hogy csak a ξ várható értékére vagyunk kíváncsiak. Ekkor a fenti esetben az $\frac{a+b}{2}$ valódi értékét kell megbecsülni.

Az eljárás a ξ-re vonatkozó $\xi_1(\omega), \ldots, \xi_n(\omega)$ mintarealizáció alapján úgy fog történni, hogy bizonyos kritériumokat figyelembe véve megadunk egy statisztikát, melynek az ω helyen vett realizációja adja a becslést.

Most általánosítjuk az előzőeket. Legyen $v \in \mathbb{N}$, $\Theta \subset \mathbb{R}^v$ az úgynevezett *paramétertér*. Feltesszük, hogy $\Theta \neq \emptyset$. Jelöljön F_ϑ eloszlásfüggvényt minden $\vartheta = (\vartheta_1, \ldots, \vartheta_v) \in \Theta$ esetén. Feltesszük, hogy $\vartheta \neq \vartheta'$ esetén $F_\vartheta \neq F_{\vartheta'}$. Ez az úgynevezett *identifikálható* tulajdonság. Tegyük fel, hogy a vizsgált ξ valószínűségi változóról tudjuk, hogy az eloszlásfüggvénye az

$$\{ F_\vartheta : \vartheta = (\vartheta_1, \ldots, \vartheta_v) \in \Theta \}$$

halmaz (eloszláscsalád) eleme, de a $\vartheta_1, \ldots, \vartheta_v$ paraméterek valódi értékei ismeretlenek. Ekkor a vizsgált statisztikai mező leszűkül az

$$(\Omega, \mathcal{F}, \mathcal{P}), \quad \mathcal{P} = \{ \mathrm{P}_\vartheta : \vartheta \in \Theta \}$$

mezőre, ahol P_ϑ olyan valószínűség az $(\Omega, \mathcal{F})$ téren, melyre

$$\mathrm{P}_\vartheta(\xi < x) = F_\vartheta(x)$$

teljesül minden $x \in \mathbb{R}$ és $\vartheta \in \Theta$ esetén. A továbbiakban mindezt úgy fogalmazzuk meg, hogy legyen ξ a vizsgálandó valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{ \mathrm{P}_\vartheta : \vartheta \in$

$\in \Theta\,\}$ statisztikai mezőn.

Legyen $g\colon \Theta \to \mathbb{R}$ egy tetszőleges függvény. A *pontbecslés feladata* a $g(\vartheta)$ valódi értékének becslése egy statisztikával. Ezt a statisztikát és annak realizációját is a $g(\vartheta)$ *pontbecslésének* nevezzük.

Fontos kérdés, hogy milyen szempontok szerint válasszuk ki a pontbecslést megadó statisztikát. A következő természetesnek tűnő feltételeket adjuk:

- ingadozzon a $g(\vartheta)$ valódi értéke körül;
- szórása a lehető legkisebb legyen;
- a minta elemszámának végtelenbe divergálása esetén konvergáljon a $g(\vartheta)$ valódi értékéhez.

A következőkben ezeket a feltételeket fogalmazzuk meg pontosabban. Legyen $\xi_1, \xi_2, \dots$ az előbbi ξ valószínűségi változóra vonatkozó végtelen elemszámú minta (azaz $\xi_1, \xi_2, \dots$ független ξ-vel azonos eloszlású valószínűségi változók), továbbá jelölje E_ϑ, D_ϑ illetve cov_ϑ a P_ϑ-ból származtatott várható értéket, szórást illetve kovarianciát.

3.1. Definíció. A $T(\xi_1, \dots, \xi_n)$ statisztika $g(\vartheta)$ *torzítatlan becslése*, ha

$$\mathrm{E}_\vartheta\, T(\xi_1, \dots, \xi_n) = g(\vartheta)$$

minden $\vartheta \in \Theta$ esetén. Ha ez nem teljesül, akkor $T(\xi_1, \dots, \xi_n)$ a $g(\vartheta)$ *torzított becslése.*

3.2. Feladat. Bizonyítsa be, hogy $F_n^*(x)$ torzítatlan becslése $F(x)$-nek bármely $x \in$ $\in \mathbb{R}$ esetén, ahol F a ξ eloszlásfüggvénye és F_n^* a tapasztalati eloszlásfüggvény.

Bizonyítás. Az $nF_n^*(x)$ egy n-edrendű $p = F(x)$ paraméterű binomiális eloszlású valószínűségi változó. Így $\mathrm{E}_p\, F_n^*(x) = \frac{1}{n}\,\mathrm{E}_p\big(nF_n^*(x)\big) = \frac{1}{n}np = p = F(x)$.

3.3. Feladat. Legyen $\tau_k := T_k(\xi_1, \dots, \xi_n)$ torzítatlan becslése ϑ_k-nak minden $k =$ $= 1, \dots, r$ esetén, és $h\colon \mathbb{R}^r \to \mathbb{R}$ olyan függvény, melyre $h(\tau_1, \dots, \tau_r)$ valószínűségi változó. Bizonyítsa be, hogy $h(\tau_1, \dots, \tau_r)$ nem feltétlenül torzítatlan becslése $h(\vartheta)$-nak.

Bizonyítás. Legyen például ξ egy olyan esemény indikátorváltozója, melynek p valószínűségére $0 < p < 1$ teljesül. Könnyen látható, hogy $\mathrm{E}_p\,\overline{\xi} = p$, azaz $\overline{\xi}$ torzítatlan becslése p-nek. Másrészt $h\colon \mathbb{R} \to \mathbb{R},\ h(x) := x^2$ jelöléssel

$$\mathrm{E}_p\, h(\overline{\xi}) = \mathrm{E}_p\,\overline{\xi}^2 = \mathrm{D}_p^2\,\overline{\xi} + \mathrm{E}_p^2\,\overline{\xi} = \frac{1}{n^2} n\,\mathrm{D}_p^2\,\xi + \mathrm{E}_p^2\,\xi =$$

$$= \frac{1}{n}p(1-p) + p^2 \neq p^2 = h(p),$$

azaz $h(\overline{\xi})$ torzított becslése $h(p)$-nek.

3.4. Definíció. A $T_n(\xi_1, \dots, \xi_n)$ $(n \in \mathbb{N})$ statisztikasorozat $g(\vartheta)$ *aszimptotikusan torzítatlan becsléssorozata*, ha minden $\vartheta \in \Theta$ esetén teljesül, hogy

$$\lim_{n\to\infty} \mathrm{E}_\vartheta\, T_n(\xi_1, \dots, \xi_n) = g(\vartheta).$$

3.5. Definíció. Egy $T(\xi_1, \dots, \xi_n)$ statisztikát véges szórásúnak nevezünk, ha minden $\vartheta \in \Theta$ esetén $\mathrm{D}_\vartheta\, T(\xi_1, \dots, \xi_n) \in \mathbb{R}$.

3.6. Definíció. Legyenek $T_1(\xi_1, \dots, \xi_n)$ és $T_2(\xi_1, \dots, \xi_n)$ véges szórású torzítatlan becslései $g(\vartheta)$-nak. A $T_1(\xi_1, \dots, \xi_n)$ *hatásosabb* becslése $g(\vartheta)$-nak mint $T_2(\xi_1, \dots, \xi_n)$, ha minden $\vartheta \in \Theta$ esetén teljesül, hogy

$$\mathrm{D}_\vartheta\, T_1(\xi_1, \dots, \xi_n) \leqslant \mathrm{D}_\vartheta\, T_2(\xi_1, \dots, \xi_n).$$

3.7. Definíció. A $g(\vartheta)$ összes véges szórású torzítatlan becslése közül a leghatásosabbat a $g(\vartheta)$ *hatásos* becslésének nevezzük.

Nem biztos, hogy $g(\vartheta)$-nak létezik hatásos becslése, hiszen egy alulról korlátos számhalmaznak nem mindig van minimuma. De ha létezik hatásos becslés, akkor az majdnem biztosan egyértelmű. Ezt fogalmazza meg a következő tétel.

3.8. Tétel. *A hatásos becslés 1 valószínűséggel egyértelmű, azaz, ha $T_1(\xi_1, \dots, \xi_n)$ és $T_2(\xi_1, \dots, \xi_n)$ a $g(\vartheta)$-nak hatásos becslései, akkor minden $\vartheta \in \Theta$ esetén*

$$\mathrm{P}_\vartheta\big(T_1(\xi_1, \dots, \xi_n) = T_2(\xi_1, \dots, \xi_n)\big) = 1.$$

Bizonyítás. Legyen $\tau_1 := T_1(\xi_1, \dots, \xi_n)$, $\tau_2 := T_2(\xi_1, \dots, \xi_n)$, $\tau := \frac{\tau_1+\tau_2}{2}$ és $\vartheta \in \Theta$. Ekkor

$$\mathrm{E}_\vartheta\, \tau = \frac{1}{2}(\mathrm{E}_\vartheta\, \tau_1 + \mathrm{E}_\vartheta\, \tau_2) = \frac{1}{2}\big(g(\vartheta) + g(\vartheta)\big) = g(\vartheta),$$

azaz τ torzítatlan becslése $g(\vartheta)$-nak. Így τ_1 hatásossága miatt

$$\mathrm{D}^2_\vartheta\, \tau_1 \leqslant \mathrm{D}^2_\vartheta\, \tau = \mathrm{D}^2_\vartheta\, \frac{\tau_1 + \tau_2}{2} =$$
$$= \frac{1}{4}\big(\mathrm{D}^2_\vartheta\, \tau_1 + \mathrm{D}^2_\vartheta\, \tau_2 + 2\,\mathrm{cov}_\vartheta(\tau_1, \tau_2)\big) = \frac{1}{4}\big(2\,\mathrm{D}^2_\vartheta\, \tau_1 + 2\,\mathrm{cov}_\vartheta(\tau_1, \tau_2)\big).$$

Ebből kapjuk, hogy $0 \leqslant \mathrm{D}^2_\vartheta(\tau_1 - \tau_2) = 2\,\mathrm{D}^2_\vartheta\,\tau_1 - 2\,\mathrm{cov}_\vartheta(\tau_1, \tau_2) \leqslant 0$, azaz $\mathrm{D}^2_\vartheta(\tau_1 - \tau_2) = 0$. De ez csak úgy lehetséges, ha

$$\mathrm{P}_\vartheta\big(\tau_1 - \tau_2 = \mathrm{E}_\vartheta(\tau_1 - \tau_2)\big) = 1.$$

Ebből már következik az állítás, hiszen $\mathrm{E}_\vartheta(\tau_1 - \tau_2) = 0$.

3.9. Definíció. A $T_n(\xi_1, \ldots, \xi_n)$ $(n \in \mathbb{N})$ statisztikasorozat $g(\vartheta)$-nak *konzisztens becsléssorozata*, ha bármely $\varepsilon > 0$ és $\vartheta \in \Theta$ esetén

$$\lim_{n\to\infty} \mathrm{P}_\vartheta\big(|T_n(\xi_1, \ldots, \xi_n) - g(\vartheta)| \geqslant \varepsilon\big) = 0.$$

3.10. Feladat. Bizonyítsa be, hogy létezik nem konzisztens torzítatlan becsléssorozat.

Bizonyítás. Legyen $\xi \in \mathbf{Norm}(m; 1)$, ahol az $m \in \mathbb{R}$ paraméternek a valódi értéke ismeretlen. Ekkor ξ_n torzítatlan becsléssorozat, hiszen $\mathrm{E}_m\,\xi_n = m$, de $\varepsilon > 0$ esetén

$$\mathrm{P}_m\big(|\xi_n - m| \geqslant \varepsilon\big) = 1 - \mathrm{P}\big(|\xi_n - m| < \varepsilon\big) = 2 - 2\Phi(\varepsilon),$$

azaz $\lim\limits_{n\to\infty} \mathrm{P}_m\big(|\xi_n - m| \geqslant \varepsilon\big) \neq 0$. Így ξ_n nem konzisztens becsléssorozat.

A torzítatlan becsléssorozatok konzisztenciájához tudunk adni elégséges feltételt.

3.11. Tétel. *Ha $T_n(\xi_1, \ldots, \xi_n)$ torzítatlan becslése $g(\vartheta)$-nak minden $n \in \mathbb{N}$ esetén, és $\lim\limits_{n\to\infty} \mathrm{D}^2_\vartheta\, T_n(\xi_1, \ldots, \xi_n) = 0$ minden $\vartheta \in \Theta$ esetén, akkor $T_n(\xi_1, \ldots, \xi_n)$ a $g(\vartheta)$ konzisztens becsléssorozata.*

Bizonyítás. Legyen $\tau_n := T_n(\xi_1, \ldots, \xi_n)$, $\varepsilon > 0$ és $\vartheta \in \Theta$. Ekkor τ_n torzítatlansága, a Csebisev-egyenlőtlenség és $\lim\limits_{n\to\infty} \mathrm{D}^2_\vartheta\,\tau_n = 0$ miatt

$$\lim_{n\to\infty} \mathrm{P}_\vartheta\big(|\tau_n - g(\vartheta)| \geqslant \varepsilon\big) = \lim_{n\to\infty} \mathrm{P}_\vartheta\big(|\tau_n - \mathrm{E}_\vartheta\,\tau_n| \geqslant \varepsilon\big) \leqslant \lim_{n\to\infty} \frac{\mathrm{D}^2_\vartheta\,\tau_n}{\varepsilon^2} = 0.$$

Ebből már következik, hogy τ_n a $g(\vartheta)$ konzisztens becsléssorozata.

3.12. Definíció. A $T_n(\xi_1, \ldots, \xi_n)$ $(n \in \mathbb{N})$ statisztikasorozat $g(\vartheta)$-nak *erősen konzisztens becsléssorozata*, ha minden $\vartheta \in \Theta$ esetén

$$\mathrm{P}_\vartheta\left(\lim_{n\to\infty} T_n(\xi_1, \ldots, \xi_n) = g(\vartheta)\right) = 1.$$

3.13. *Megjegyzés.* Mivel a majdnem mindenütti konvergenciából következik a mértékben való konvergencia, ezért az erősen konzisztens becsléssorozat egyúttal konzisztens becsléssorozat is.

3.1.1. Várható érték becslése

Ebben az alszakaszban feltesszük, hogy $\mathrm{E}_\vartheta\,\xi \in \mathbb{R}$ minden $\vartheta \in \Theta$ esetén.

3.14. Feladat. Bizonyítsa be, hogy ha $c_1, \dots, c_n \in \mathbb{R}$ és $c_1 + \dots + c_n = 1$, akkor $\sum_{i=1}^n c_i\xi_i$ torzítatlan becslése ξ várható értékének.

Bizonyítás. $\mathrm{E}_\vartheta \sum_{i=1}^n c_i\xi_i = \sum_{i=1}^n c_i\,\mathrm{E}_\vartheta\,\xi_i = \sum_{i=1}^n c_i\,\mathrm{E}_\vartheta\,\xi = \mathrm{E}_\vartheta\,\xi \sum_{i=1}^n c_i = \mathrm{E}_\vartheta\,\xi$.

3.15. Feladat. Bizonyítsa be, hogy a mintaátlag torzítatlan becslése a várható értéknek.

Bizonyítás. Az előző következménye $c_i = \frac{1}{n}$ $(i = 1, \dots, n)$ választással.

3.16. Feladat. Bizonyítsa be, hogy ha ξ véges szórású, akkor a mintaátlag konzisztens becsléssorozata a várható értéknek.

Bizonyítás. Az állítás a nagy számok gyenge törvényével ekvivalens. De belátható a konzisztencia elégséges feltételének vizsgálatával is, hiszen

$$\lim_{n\to\infty} \mathrm{D}^2_\vartheta\,\overline{\xi} = \lim_{n\to\infty} \frac{1}{n}\,\mathrm{D}^2_\vartheta\,\xi = 0,$$

melyből következik az állítás.

3.17. Feladat. Bizonyítsa be, hogy ha $\mathrm{E}_\vartheta\,|\xi| \in \mathbb{R}$ minden $\vartheta \in \Theta$ esetén, akkor a mintaátlag erősen konzisztens becsléssorozata a várható értéknek.

Bizonyítás. Az állítás a Kolmogorov-féle nagy számok erős törvényével ekvivalens.

3.18. Feladat. Bizonyítsa be, hogy ha ξ véges szórású bármely $\vartheta \in \Theta$ esetén, akkor $\overline{\xi}$ hatásosabb becslése a várható értéknek, mint $\sum_{i=1}^n c_i\xi_i$, bármely $c_1, \dots, c_n \in \mathbb{R}$, $c_1 + + \dots + c_n = 1$ esetén.

Bizonyítás. $\mathrm{D}^2_\vartheta\,(\sum_{i=1}^n c_i\xi_i) = \sum_{i=1}^n c_i^2\,\mathrm{D}^2_\vartheta\,\xi = \mathrm{D}^2_\vartheta\,\xi \sum_{i=1}^n c_i^2 \geqslant \mathrm{D}^2_\vartheta\,\xi\,\frac{1}{n}(c_1 + \dots + c_n)^2 = = \frac{1}{n}\,\mathrm{D}^2_\vartheta\,\xi = \mathrm{D}^2_\vartheta\,\overline{\xi}$. Itt felhasználtuk a számtani és a *négyzetes közép* közötti relációt, mely szerint tetszőleges $a_1, \dots, a_n \in \mathbb{R}$ esetén $\frac{a_1+\dots+a_n}{n} \leqslant \sqrt{\frac{a_1^2+\dots+a_n^2}{n}}$. (Ez a Cauchy-egyenlőtlenségből következik.)

Tehát a várható értéknek a $\sum_{i=1}^{n} c_i\xi_i$ alakú, úgynevezett *lineáris becslések* között a leghatásosabb becslése a mintaátlag. Vajon az összes véges szórású torzítatlan becslés közül is ez a leghatásosabb, azaz hatásos? A következő feladat állítása erre ad általánosságban nemleges választ.

3.19. Feladat. Bizonyítsa be, hogy ha ξ egyenletes eloszlású a $[0, b]$ intervallumon ($b \in \mathbb{R}_+$), akkor a terjedelemközép hatásosabb becslése a várható értéknek a mintaátlagnál.

Bizonyítás. A bizonyítás terjedelmes, csak a fontosabb lépéseket közöljük. A minta legyen $\xi_1, \dots, \xi_n$. Először be kell látni, hogy a terjedelemközép a várható érték torzítatlan becslése, majd meg kell mutatni, hogy ennek szórása kisebb a mintaátlag szórásánál. Ehhez először a $\xi_1^*, \dots, \xi_n^*$ rendezett minta elemeinek eloszlását vizsgáljuk meg. Mivel $i \in \{\, 1, \dots, n \,\},\ 0 < x < b$, esetén

$$\begin{aligned}&\mathrm{P}_b(\xi_1 < x, \dots, \xi_i < x, \xi_{i+1} \geqslant x, \dots, \xi_n \geqslant x) = \\ &= \big(\mathrm{P}_b(\xi < x)\big)^i \big(\mathrm{P}_b(\xi \geqslant x)\big)^{n-i} = \left(\frac{x}{b}\right)^i \left(1 - \frac{x}{b}\right)^{n-i},\end{aligned}$$

ezért annak a valószínűsége, hogy $\xi_1, \dots, \xi_n$ közül pontosan i darab kisebb x-nél,

$$\binom{n}{i} \left(\frac{x}{b}\right)^i \left(1 - \frac{x}{b}\right)^{n-i},\ 0 < x < b.$$

A $\xi_k^* < x$ esemény azt jelenti, hogy pontosan k vagy pontosan $k+1$ vagy … pontosan n darab mintaelem kisebb x-nél. Így

$$\mathrm{P}_b(\xi_k^* < x) = \sum_{i=k}^{n} \binom{n}{i} \left(\frac{x}{b}\right)^i \left(1 - \frac{x}{b}\right)^{n-i},\ k \in \{\, 1, \dots, n \,\},\ 0 < x < b.$$

Ebből belátható, hogy ξ_k^* sűrűségfüggvénye x helyen

$$\frac{n}{b}\binom{n-1}{k-1} \left(\frac{x}{b}\right)^{k-1} \left(1 - \frac{x}{b}\right)^{n-k},\ k \in \{\, 1, \dots, n \,\},\ 0 < x < b.$$

Így $k \in \{\, 1, \dots, n \,\}$ esetén

$$\mathrm{E}_b\, \xi_k^* = \int_0^b x \frac{n}{b}\binom{n-1}{k-1} \left(\frac{x}{b}\right)^{k-1} \left(1 - \frac{x}{b}\right)^{n-k} \mathrm{d}x = \dots = \frac{kb}{n+1}.$$

Ebből $\mathrm{E}_b \frac{\xi_1^* + \xi_n^*}{2} = \frac{1}{2}\left(\frac{b}{n+1} + \frac{nb}{n+1}\right) = \frac{b}{2} = \mathrm{E}_b\, \xi$. Tehát a terjedelemközép a várható

érték torzítatlan becslése. Most rátérünk a szórás meghatározására. A korábbiak alapján

$$\mathrm{E}_b\,\xi_k^{*2} = \int_0^b x^2 \frac{n}{b}\binom{n-1}{k-1}\left(\frac{x}{b}\right)^{k-1}\left(1-\frac{x}{b}\right)^{n-k}\mathrm{d}x = \cdots = \frac{k(k+1)b^2}{(n+1)(n+2)}$$

teljesül minden $k \in \{\,1,\ldots,n\,\}$ esetén. Másrészt az előzőekhez hasonló gondolatmenettel ξ_k^* és ξ_l^* együttes sűrűségfüggvénye $1 \leqslant k < l \leqslant n$ esetén, az $(x,y) \in \mathbb{R}^2$ $(0 \leqslant \leqslant x < y \leqslant b)$ helyen

$$\frac{n!}{b^2(k-1)!(l-k-1)!(n-l)!}\left(\frac{x}{b}\right)^{k-1}\left(\frac{y}{b}-\frac{x}{b}\right)^{l-k-1}\left(1-\frac{y}{b}\right)^{n-l}.$$

Ebből bizonyítható, hogy

$$\mathrm{E}_b(\xi_k^*\xi_l^*) = \frac{k(l+1)b^2}{(n+1)(n+2)},\ 1 \leqslant k < l \leqslant n.$$

Így a szórásnégyzet:

$$\begin{aligned}
&\mathrm{D}_b^2\,\frac{\xi_1^*+\xi_n^*}{2} = \mathrm{E}_b\left(\frac{\xi_1^*+\xi_n^*}{2}\right)^2 - \mathrm{E}_b^2\,\frac{\xi_1^*+\xi_n^*}{2} = \\
&= \frac{1}{4}\,\mathrm{E}_b(\xi_1^*+\xi_n^*)^2 - \frac{b^2}{4} = \frac{1}{4}\,\mathrm{E}_b\,\xi_1^{*2} + \frac{1}{4}\,\mathrm{E}_b\,\xi_n^{*2} + \frac{1}{2}\,\mathrm{E}_b(\xi_1^*\xi_n^*) - \frac{b^2}{4} = \\
&= \frac{1}{4}\cdot\frac{2b^2}{(n+1)(n+2)} + \frac{1}{4}\cdot\frac{nb^2}{n+2} + \frac{1}{2}\cdot\frac{b^2}{n+2} - \frac{b^2}{4} = \frac{b^2}{2(n+1)(n+2)}.
\end{aligned}$$

Mivel $\mathrm{D}_b^2\,\overline{\xi} = \frac{1}{n}\,\mathrm{D}_b^2\,\xi = \frac{b^2}{12n}$, ezért az állítás ekvivalens az

$$\frac{1}{2(n+1)(n+2)} \leqslant \frac{1}{12n}$$

egyenlőtlenséggel. Könnyen látható, hogy ez minden $n \in \mathbb{N}$ esetén teljesül, és csak $n = 1$ illetve $n = 2$ esetén lehet egyenlőség. Az $n = 1$ illetve $n = 2$ esetén kapott egyenlőség nem meglepő, hiszen ekkor $\frac{\xi_1^*+\xi_n^*}{2} = \overline{\xi}$. Ezzel bizonyított az állítás.

Tehát van olyan eset, amikor a várható értéknek nem a mintaátlag a hatásos becslése. De vajon a mintaátlag sohasem lehet hatásos becslése a várható értéknek? A valószínűség becslése során látni fogjuk, hogy például karakterisztikus eloszlás esetén az.

3.1.2. Valószínűség becslése

3.20. Feladat. Bizonyítsa be, hogy egy esemény relatív gyakorisága torzítatlan becslése az esemény valószínűségének.

Bizonyítás. Legyen ξ a vizsgált esemény indikátorváltozója. Ekkor az esemény relatív gyakorisága $\overline{\xi}$-vel egyenlő, másrészt ξ várható értéke a vizsgált esemény valószínűsége. Így az állítás annak a speciális esete, hogy a mintaátlag torzítatlan becslése a várható értéknek.

3.21. Feladat. Bizonyítsa be, hogy egy esemény relatív gyakorisága erősen konzisztens becsléssorozata az esemény valószínűségének.

Bizonyítás. Az állítás annak a speciális esete, hogy a mintaátlag erősen konzisztens becsléssorozata a várható értéknek.

3.22. Feladat. Bizonyítsa be, hogy egy ismeretlen $0 < p < 1$ valószínűségű esemény relatív gyakorisága hatásos becslése p-nek. (Azaz $0 < p < 1$ paraméterű karakterisztikus eloszlású valószínűségi változóra vonatkozó mintából számolt mintaátlag hatásos becslése a várható értéknek.)

Bizonyítás. Legyen ξ a vizsgált esemény indikátorváltozója és $\xi_1, \dots, \xi_n$ egy ξ-re vonatkozó minta. Ekkor az esemény relatív gyakorisága $\overline{\xi}$, továbbá az eddigiek alapján $\overline{\xi}$ a p torzítatlan becslése. Legyen $T(\xi_1, \dots, \xi_n)$ tetszőleges torzítatlan becslése p-nek,

$$K := \{\, i = (i_1, \dots, i_n) : i_1, \dots, i_n \text{ az } 1, \dots, n \text{ permutációja} \,\}$$

és

$$S(\xi_1, \dots, \xi_n) := \frac{1}{n!} \sum_{i \in K} T(\xi_{i_1}, \dots, \xi_{i_n}).$$

Könnyen látható, hogy $S(\xi_1, \dots, \xi_n)$ szimmetrikus statisztika és torzítatlan becslése p-nek. Ha a $\xi_1(\omega), \dots, \xi_n(\omega)$ mintarealizációban pontosan k darab 1 van, akkor függetlenül attól, hogy pontosan melyek azok, a szimmetria miatt az $S\big(\xi_1(\omega), \dots, \xi_n(\omega)\big)$ értéke mindig ugyanaz. Ezt a közös értéket jelöljük S_k-val. Annak a valószínűsége, hogy a mintarealizációban pontosan k darab 1 van

$$\binom{n}{k} p^k (1-p)^{n-k} > 0.$$

Mindezekből a torzítatlanság miatt

$$0 = \mathrm{E}_p\big(S(\xi_1, \dots, \xi_n) - \overline{\xi}\big) = \sum_{k=0}^{n} \left(S_k - \frac{k}{n} \right) \binom{n}{k} p^k (1-p)^{n-k},$$

azaz

$$\sum_{k=0}^{n}\left(S_k-\frac{k}{n}\right)\binom{n}{k}\left(\frac{p}{1-p}\right)^k=0$$

minden $p\in(0,1)$ esetén. Ez pedig csak úgy lehetséges, ha $S_k=\frac{k}{n}$ minden $k=$ $=0,\dots,n$ esetén. Ebből az következik, hogy

$$S(\xi_1,\dots,\xi_n)=\overline{\xi}.$$

Így azt kell belátni, hogy $\mathrm{D}_p^2\,S(\xi_1,\dots,\xi_n)\leqslant\mathrm{D}_p^2\,T(\xi_1,\dots,\xi_n)$, amely azzal ekvivalens a torzítatlanság miatt, hogy $\mathrm{E}_p\,S^2(\xi_1,\dots,\xi_n)\leqslant\mathrm{E}_p\,T^2(\xi_1,\dots,\xi_n)$. Legyen

$$G_k:=\big\{\,x=(x_1,\dots,x_n):x_i\in\{0,1\},\ i=1,\dots,n,\ x_1+\dots+x_n=k\,\big\}.$$

Ekkor az előzőekhez hasonlóan látható, hogy

$$\begin{aligned}
\mathrm{E}_p\,S^2(\xi_1,\dots,\xi_n)&=\sum_{k=0}^{n}\sum_{x\in G_k}S^2(x)p^k(1-p)^{n-k}=\\
&=\sum_{k=0}^{n}\sum_{x\in G_k}\left(\frac{1}{n!}\sum_{i\in K}T(x_{i_1},\dots,x_{i_n})\right)^2p^k(1-p)^{n-k}=\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(\frac{1}{n!}\sum_{x\in G_k,i\in K}T(x_{i_1},\dots,x_{i_n})\right)^2p^k(1-p)^{n-k}=\\
&=\sum_{k=0}^{n}\binom{n}{k}\left(\frac{k!(n-k)!}{n!}\sum_{x\in G_k}T(x)\right)^2p^k(1-p)^{n-k}=\\
&=\sum_{k=0}^{n}\frac{1}{\binom{n}{k}}\left(\sum_{x\in G_k}T(x)\right)^2p^k(1-p)^{n-k}.
\end{aligned}$$

Másrészt

$$\mathrm{E}_p\,T^2(\xi_1,\dots,\xi_n)=\sum_{k=0}^{n}\sum_{x\in G_k}T^2(x)p^k(1-p)^{n-k},$$

így elég azt belátni, hogy

$$\frac{1}{\binom{n}{k}}\left(\sum_{x\in G_k}T(x)\right)^2\leqslant\sum_{x\in G_k}T^2(x).$$

Ez viszont teljesül a számtani és a négyzetes közép relációja miatt, hiszen G_k-nak $\binom{n}{k}$ darab eleme van.

3.1.3. Szórásnégyzet becslése

Ebben az alszakaszban feltesszük, hogy $\mathrm{D}_\vartheta\,\xi \in \mathbb{R}$ minden $\vartheta \in \Theta$ esetén.

3.23. Feladat. Bizonyítsa be, hogy a tapasztalati szórásnégyzet torzított becslése a szórásnégyzetnek.

Bizonyítás. A Steiner-formula és $\mathrm{E}_\vartheta\,\xi^2 = \mathrm{D}^2_\vartheta\,\xi + \mathrm{E}^2_\vartheta\,\xi$ miatt

$$\begin{aligned}
\mathrm{E}_\vartheta\,S_n^2 &= \mathrm{E}_\vartheta\left(\frac{1}{n}\sum_{i=1}^n \xi_i^2 - \overline{\xi}^2\right) = \frac{1}{n}\sum_{i=1}^n \mathrm{E}_\vartheta\,\xi_i^2 - \mathrm{E}_\vartheta\,\overline{\xi}^2 = \\
&= \frac{1}{n}\sum_{i=1}^n \mathrm{E}_\vartheta\,\xi^2 - \mathrm{D}^2_\vartheta\,\overline{\xi} - \mathrm{E}^2_\vartheta\,\overline{\xi} = \mathrm{E}_\vartheta\,\xi^2 - \mathrm{D}^2_\vartheta\,\overline{\xi} - \mathrm{E}^2_\vartheta\,\overline{\xi} = \\
&= \mathrm{D}^2_\vartheta\,\xi + \mathrm{E}^2_\vartheta\,\xi - \mathrm{D}^2_\vartheta\,\overline{\xi} - \mathrm{E}^2_\vartheta\,\overline{\xi} = \mathrm{D}^2_\vartheta\,\xi + \mathrm{E}^2_\vartheta\,\xi - \mathrm{D}^2_\vartheta\,\overline{\xi} - \mathrm{E}^2_\vartheta\,\xi = \\
&= \mathrm{D}^2_\vartheta\,\xi - \mathrm{D}^2_\vartheta\,\overline{\xi} = \mathrm{D}^2_\vartheta\,\xi - \frac{1}{n^2}\sum_{i=1}^n \mathrm{D}^2_\vartheta\,\xi_i = \mathrm{D}^2_\vartheta\,\xi - \frac{1}{n^2}\sum_{i=1}^n \mathrm{D}^2_\vartheta\,\xi = \\
&= \mathrm{D}^2_\vartheta\,\xi - \frac{1}{n}\,\mathrm{D}^2_\vartheta\,\xi = \frac{n-1}{n}\,\mathrm{D}^2_\vartheta\,\xi \neq \mathrm{D}^2_\vartheta\,\xi.
\end{aligned}$$

3.24. Feladat. Bizonyítsa be, hogy a tapasztalati szórásnégyzet aszimptotikusan torzítatlan becsléssorozata a szórásnégyzetnek.

Bizonyítás. Láttuk, hogy $\mathrm{E}_\vartheta\,S_n^2 = \frac{n-1}{n}\,\mathrm{D}^2_\vartheta\,\xi$, így $\lim\limits_{n\to\infty} \mathrm{E}_\vartheta\,S_n^2 = \mathrm{D}^2_\vartheta\,\xi$.

3.25. Feladat. Bizonyítsa be, hogy a tapasztalati szórásnégyzet erősen konzisztens becsléssorozata a szórásnégyzetnek.

Bizonyítás. A Kolmogorov-féle nagy számok törvénye miatt

$$\mathrm{P}_\vartheta\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \xi_i^2 = \mathrm{E}_\vartheta\,\xi^2\right) = 1 \quad \text{és} \quad \mathrm{P}_\vartheta\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \xi_i = \mathrm{E}_\vartheta\,\xi\right) = 1.$$

Így a Steiner-formulából kapjuk az állítást.

3.26. Feladat. Bizonyítsa be, hogy a korrigált tapasztalati szórásnégyzet torzítatlan becslése a szórásnégyzetnek.

Bizonyítás. Láttuk, hogy $\mathrm{E}_\vartheta\,S_n^2 = \frac{n-1}{n}\,\mathrm{D}^2_\vartheta\,\xi$, így $\mathrm{E}_\vartheta\,S_n^{*2} = \mathrm{E}_\vartheta\,\frac{n}{n-1}S_n^2 = \mathrm{D}^2_\vartheta\,\xi$.

3.27. Feladat. Bizonyítsa be, hogy a korrigált tapasztalati szórásnégyzet erősen konzisztens becsléssorozata a szórásnégyzetnek.

Bizonyítás. Az állítás a tapasztalati szórásnégyzet erős konzisztenciájából következik, hiszen $S_n^{*2} = \frac{n}{n-1}S_n^2$.

3.2. Információs határ

Legyen ξ egy ismeretlen $0 < p < 1$ paraméterű karakterisztikus eloszlású valószínűségi változó, továbbá a rávonatkozó minta $\xi_1, \dots, \xi_n$. Korábban bizonyítottuk, hogy $\overline{\xi}$ hatásos becslése p-nek. Mivel $\mathrm{D}_p^2 \overline{\xi} = \frac{1}{n} \mathrm{D}_p^2 \xi = \frac{p(1-p)}{n}$, ezért azt kapjuk, hogy a p összes véges szórású torzítatlan becslésének szórása nagyobb vagy egyenlő, mint $\frac{p(1-p)}{n}$.

Általánosságban, ha $g(\vartheta)$ összes véges szórású $T(\xi_1, \dots, \xi_n)$ torzítatlan becslésének szórása nagyobb vagy egyenlő, mint egy T-től független érték, akkor ezt *információs határnak* nevezzük.

Ennek a szakasznak a célja az információs határ meghatározása azzal a feltevéssel, hogy ξ abszolút folytonos vagy diszkrét, illetve $\Theta \subset \mathbb{R}$, azaz csak egy paraméter ismeretlen ($v = 1$). Feltesszük még, hogy Θ nyílt halmaz. Amennyiben ξ abszolút folytonos, akkor f_ϑ jelölje ξ-nek a P_ϑ-ból származó sűrűségfüggvényét. A ξ-re vonatkozó minta legyen $\xi_1, \dots, \xi_n$, továbbá a ξ értékkészlete legyen $\mathfrak{X}$, azaz a mintatér $\mathfrak{X}^n$.

3.28. Definíció. A $\xi_1, \dots, \xi_n$ minta *likelihood függvénye*

$$l_n \colon \mathfrak{X}^n \times \Theta \to \mathbb{R}, \quad l_n(x_1, \dots, x_n, \vartheta) := \begin{cases} \prod_{i=1}^n f_\vartheta(x_i), & \text{ha } \xi \text{ absz. folyt.,} \\ \prod_{i=1}^n \mathrm{P}_\vartheta(\xi_i = x_i), & \text{ha } \xi \text{ diszkrét.} \end{cases}$$

A $\xi_1, \dots, \xi_n$ minta *loglikelihood függvénye* $L_n := \ln l_n$.

3.29. Definíció. A $\xi_1, \dots, \xi_n$ minta *Fisher-féle információmennyisége*

$$I_n \colon \Theta \to \mathbb{R}, \quad I_n(\vartheta) := \mathrm{E}_\vartheta \left(\frac{\partial}{\partial \vartheta} L_n(\xi_1, \dots, \xi_n, \vartheta) \right)^2,$$

feltéve, hogy ez a függvény értelmezhető. Ellenkező esetben azt mondjuk, hogy a Fisher-féle információmennyiség nem létezik.

3.30. Definíció. Legyen $T \colon \mathbb{R}^n \to \mathbb{R}$ egy tetszőleges függvény. Azt mondjuk, hogy *Tl_n-re teljesül a bederiválási feltétel*, ha

$$\frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} T(x_1, \dots, x_n) l_n(x_1, \dots, x_n, \vartheta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n =$$
$$= \int_{\mathbb{R}^n} T(x_1, \dots, x_n) \frac{\partial}{\partial \vartheta} l_n(x_1, \dots, x_n, \vartheta) \, \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

vagy

$$\frac{\partial}{\partial\vartheta}\sum_{x_i\in\mathfrak{X}} T(x_1,\dots,x_n)l_n(x_1,\dots,x_n,\vartheta) =$$
$$= \sum_{x_i\in\mathfrak{X}} T(x_1,\dots,x_n)\frac{\partial}{\partial\vartheta}l_n(x_1,\dots,x_n,\vartheta)$$

aszerint, hogy ξ abszolút folytonos vagy diszkrét.

3.31. *Megjegyzés.* Ha $\mathfrak{X}$ véges, akkor Tl_n-re triviálisan teljesül a bederiválási feltétel.

3.32. Lemma. *l_1-re pontosan akkor teljesül a bederiválási feltétel, ha*

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = 0 \quad \textit{vagy} \quad \sum_{x\in\mathfrak{X}}\frac{\partial}{\partial\vartheta}\,\mathrm{P}_\vartheta(\xi = x) = 0$$

aszerint, hogy ξ abszolút folytonos vagy diszkrét.

Bizonyítás. Csak abszolút folytonos esetben bizonyítunk, de diszkrét esetben analóg módon járhatunk el, melyet az Olvasóra bízunk. A bizonyításhoz vegyük észre, hogy $l_1(x,\vartheta) = f_\vartheta(x)$ és $\int_{-\infty}^{\infty} l_1(x,\vartheta)\,\mathrm{d}x = 1$. Most tegyük fel, hogy $\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = 0$. Ebből kapjuk, hogy

$$\frac{\partial}{\partial\vartheta}\int_{-\infty}^{\infty} l_1(x,\vartheta)\,\mathrm{d}x = 0 = \int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = \int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}l_1(x,\vartheta)\,\mathrm{d}x,$$

azaz ekkor l_1-re teljesül a bederiválási feltétel. Megfordítva, ha feltesszük, hogy l_1-re teljesül a bederiválási feltétel, akkor

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}l_1(x,\vartheta)\,\mathrm{d}x = \frac{\partial}{\partial\vartheta}\int_{-\infty}^{\infty} l_1(x,\vartheta)\,\mathrm{d}x = 0.$$

Ezzel teljes a bizonyítás.

3.33. Tétel. *Ha l_1-re teljesül a bederiválási feltétel és I_1 létezik, akkor I_n is létezik és $I_n = nI_1$.*

Bizonyítás. Csak abszolút folytonos esetben bizonyítunk, de diszkrét esetben analóg

módon járhatunk el, melyet az Olvasóra bízunk. Az $l_1(x,\vartheta) = f_\vartheta(x)$, így

$$\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}L_1\big(\xi_1,\vartheta\big)\right) = \int\limits_{-\infty}^{\infty}\left(\frac{\partial}{\partial\vartheta}\ln l_1(x,\vartheta)\right) f_\vartheta(x)\,\mathrm{d}x = \int\limits_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = 0.$$

Ebből

$$I_1(\vartheta) = \mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}L_1\big(\xi_1,\vartheta\big)\right)^2 = \mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}L_1\big(\xi_1,\vartheta\big)\right) = \mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(\xi_1)\right).$$

Másrészt

$$\begin{aligned}
&\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}L_n(\xi_1,\dots,\xi_n,\vartheta)\right) = \mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\sum_{i=1}^{n}\ln f_\vartheta(\xi_i)\right) = \\
&= \sum_{i=1}^{n}\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(\xi_i)\right) = \sum_{i=1}^{n}\int\limits_{-\infty}^{\infty}\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x)\right) f_\vartheta(x)\,\mathrm{d}x = \\
&= \sum_{i=1}^{n}\int\limits_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = 0.
\end{aligned}$$

Ebből

$$\begin{aligned}
I_n(\vartheta) &= \mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}L_n(\xi_1,\dots,\xi_n,\vartheta)\right)^2 = \mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}L_n(\xi_1,\dots,\xi_n,\vartheta)\right) = \\
&= \mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}\sum_{i=1}^{n}\ln f_\vartheta(\xi_i)\right) = \sum_{i=1}^{n}\mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(\xi_i)\right) = \\
&= \sum_{i=1}^{n}\mathrm{D}^2_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(\xi_1)\right) = \sum_{i=1}^{n}I_1(\vartheta) = nI_1(\vartheta).
\end{aligned}$$

3.34. Feladat. Karakterisztikus eloszlás esetén határozza meg a Fisher-féle információmennyiséget.

Megoldás. Legyen tehát ξ egy $0 < p < 1$ paraméterű karakterisztikus eloszlású valószínűségi változó, és a rávonatkozó minta $\xi_1,\dots,\xi_n$. Ekkor $\mathfrak{X} = \{\,0,1\,\}$, $l_1(0,p) = \mathrm{P}_p(\xi_1 = 0) = 1-p$ és $l_1(1,p) = \mathrm{P}_p(\xi_1 = 1) = p$. Így

$$I_1(p) = \mathrm{E}_p\left(\frac{\partial}{\partial p}L_1(\xi_1,p)\right)^2 = \mathrm{E}_p\left(\frac{\partial}{\partial p}\ln l_1(\xi_1,p)\right)^2 =$$

$$= \left(\frac{\partial}{\partial p}\ln \mathrm{P}_p(\xi_1=0)\right)^2 \cdot \mathrm{P}_p(\xi_1=0) + \left(\frac{\partial}{\partial p}\ln \mathrm{P}_p(\xi_1=1)\right)^2 \cdot \mathrm{P}_p(\xi_1=1) =$$
$$= \left(\frac{\partial}{\partial p}\ln(1-p)\right)^2 \cdot (1-p) + \left(\frac{\partial}{\partial p}\ln p\right)^2 \cdot p = \frac{1}{p(1-p)}.$$

Másrészt $\mathfrak{X}$ végessége miatt l_1-re teljesül a bederiválási feltétel, melyből

$$I_n(p) = nI_1(p) = \frac{n}{p(1-p)}.$$

3.35. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$, ahol $\sigma > 0$ rögzített. Határozza meg a Fisher-féle információmennyiséget.

Megoldás. $\int\limits_{-\infty}^{\infty} \frac{\partial}{\partial m} f_m(x)\,\mathrm{d}x = \int\limits_{-\infty}^{\infty} \frac{\partial}{\partial m}\frac{1}{\sigma}\varphi\left(\frac{x-m}{\sigma}\right)\,\mathrm{d}x = \int\limits_{-\infty}^{\infty} \frac{x-m}{\sigma^2} f_m(x)\,\mathrm{d}x = \mathrm{E}_m\left(\frac{\xi-m}{\sigma^2}\right) =$
$= 0$, azaz l_1-re teljesül a bederiválási feltétel. Korábban láttuk, hogy ekkor

$$I_1(m) = \mathrm{D}_m^2\left(\frac{\partial}{\partial m}\ln f_m(\xi)\right) = \mathrm{D}_m^2\left(\frac{1}{f_m(\xi)}\cdot\frac{\partial}{\partial m} f_m(\xi)\right) =$$
$$= \mathrm{D}_m^2\left(\frac{1}{f_m(\xi)}\cdot\frac{\xi-m}{\sigma^2} f_m(\xi)\right) = \mathrm{D}_m^2\left(\frac{\xi-m}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

Ebből kapjuk, hogy $I_n(m) = nI_1(m) = \frac{n}{\sigma^2}$.

3.36. Feladat. Legyen ξ ismeretlen λ paraméterű Poisson-eloszlású. Határozza meg a Fisher-féle információmennyiséget.

Megoldás.

$$I_1(\lambda) = \mathrm{E}_\lambda\left(\frac{\partial}{\partial\lambda}\ln l_1(\xi_1,\lambda)\right)^2 = \sum_{k=0}^{\infty}\left(\frac{\partial}{\partial\lambda}\ln \mathrm{P}_\lambda(\xi_1=k)\right)^2 \mathrm{P}_\lambda(\xi_1=k) =$$
$$= \sum_{k=0}^{\infty}\left(\frac{\partial}{\partial\lambda}\ln\frac{\lambda^k}{k!}e^{-\lambda}\right)^2\frac{\lambda^k}{k!}e^{-\lambda} = \sum_{k=0}^{\infty}\left(\frac{k}{\lambda}-1\right)^2\frac{\lambda^k}{k!}e^{-\lambda} =$$
$$= \sum_{k=0}^{\infty}\left(\frac{k(k-1)}{\lambda^2}+1+\left(\frac{1}{\lambda^2}-\frac{2}{\lambda}\right)k\right)\frac{\lambda^k}{k!}e^{-\lambda} =$$
$$= \left(\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!}+\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}+\left(\frac{1}{\lambda}-2\right)\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}\right)e^{-\lambda} =$$
$$= \left(e^\lambda + e^\lambda + \left(\frac{1}{\lambda}-2\right)e^\lambda\right)e^{-\lambda} = \frac{1}{\lambda}.$$

Másrészt

$$\sum_{k=0}^{\infty} \frac{\partial}{\partial\lambda} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{1}{k!} \left(k\lambda^{k-1} e^{-\lambda} - \lambda^k e^{-\lambda} \right) =$$
$$= \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} - \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) e^{-\lambda} = \left(e^{\lambda} - e^{\lambda} \right) e^{-\lambda} = 0,$$

azaz l_1-re teljesül a bederiválási feltétel. Ebből kapjuk, hogy $I_n(\lambda) = \frac{n}{\lambda}$.

3.37. Feladat. Legyen $\xi \in \mathbf{Exp}(\lambda)$. Határozza meg a Fisher-féle információmennyiséget.

Megoldás.

$$I_1(\lambda) = \mathrm{E}_\lambda \left(\frac{\partial}{\partial\lambda} \ln l_1(\xi_1, \lambda) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial\lambda} \ln l_1(x, \lambda) \right)^2 l_1(x, \lambda)\, \mathrm{d}x =$$
$$= \int_0^{\infty} \left(\frac{\partial}{\partial\lambda} \ln \lambda e^{-\lambda x} \right)^2 \lambda e^{-\lambda x}\, \mathrm{d}x = \int_0^{\infty} \left(\frac{1}{\lambda} - x \right)^2 \lambda e^{-\lambda x}\, \mathrm{d}x =$$
$$= \mathrm{E}_\lambda \left(\frac{1}{\lambda} - \xi \right)^2 = \mathrm{D}^2_\lambda\, \xi = \frac{1}{\lambda^2}.$$

Másrészt

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial\lambda} f_\lambda(x)\, \mathrm{d}x = \int_0^{\infty} \frac{\partial}{\partial\lambda} \lambda e^{-\lambda x}\, \mathrm{d}x = \int_0^{\infty} \left(\frac{1}{\lambda} - x \right) \lambda e^{-\lambda x}\, \mathrm{d}x = \mathrm{E}_\lambda \left(\frac{1}{\lambda} - \xi \right) = 0,$$

azaz l_1-re teljesül a bederiválási feltétel. Ebből kapjuk, hogy $I_n(\lambda) = \frac{n}{\lambda^2}$.

3.38. Feladat. Legyen ξ egyenletes eloszlású a $[0, b]$ intervallumon ($b \in \mathbb{R}_+$). Mutassa meg, hogy ekkor nem teljesül l_1-re a bederiválási feltétel, továbbá az $I_1(b)$ és $I_n(b)$ meghatározásával bizonyítsa be, hogy $I_n \neq nI_1$, ha $n > 1$.

Megoldás. $\int_{-\infty}^{\infty} \frac{\partial}{\partial b} f_b(x)\, \mathrm{d}x = \int_0^b \frac{\partial}{\partial b} \frac{1}{b}\, \mathrm{d}x = \int_0^b \frac{-1}{b^2}\, \mathrm{d}x = -\frac{1}{b} \neq 0$, így l_1-re valóban nem teljesül a bederiválási feltétel.

$$I_1(b) = \mathrm{E}_b \left(\frac{\partial}{\partial b} \ln f_b(\xi_1) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial b} \ln f_b(x) \right)^2 f_b(x)\, \mathrm{d}x =$$
$$= \int_0^b \left(\frac{\partial}{\partial b} \ln \frac{1}{b} \right)^2 \frac{1}{b}\, \mathrm{d}x = \int_0^b \left(\frac{-1}{b} \right)^2 \frac{1}{b}\, \mathrm{d}x = \int_0^b \frac{1}{b^3}\, \mathrm{d}x = \frac{1}{b^2}.$$

$$I_n(b) = \mathrm{E}_b\left(\frac{\partial}{\partial b}\sum_{i=1}^{n}\ln f_b(\xi_i)\right)^2 = \mathrm{E}_b\left(\sum_{i=1}^{n}\frac{\partial}{\partial b}\ln f_b(\xi_i)\right)^2 =$$
$$= \mathrm{E}_b\left(\sum_{i=1}^{n}\frac{\partial}{\partial b}\ln\frac{1}{b}\right)^2 = \mathrm{E}_b\left(\sum_{i=1}^{n}\frac{-1}{b}\right)^2 = \mathrm{E}_b\left(\frac{-n}{b}\right)^2 = \frac{n^2}{b^2}.$$

Tehát ekkor $I_n(b) = n^2 I_1(b)$, azaz $n > 1$ esetén $I_n(b) \neq nI_1(b)$.

3.39. Tétel (Rao–Cramér-egyenlőtlenség)**.** *Legyen $T(\xi_1, \ldots, \xi_n)$ véges szórású torzítatlan becslése $g(\vartheta)$-nak, ahol $g\colon \Theta \to \mathbb{R}$ differenciálható függvény. Tegyük fel, hogy l_1-re és Tl_n-re teljesül a bederiválási feltétel, továbbá, hogy I_1 létezik és pozitív. Ekkor*

$$\mathrm{D}^2_\vartheta\, T(\xi_1, \ldots, \xi_n) \geqslant \frac{\big(g'(\vartheta)\big)^2}{nI_1(\vartheta)}$$

minden $\vartheta \in \Theta$ esetén. A $\frac{(g'(\vartheta))^2}{nI_1(\vartheta)}$ kifejezés az úgynevezett információs határ*.*

Bizonyítás. Csak abszolút folytonos esetben bizonyítunk, de diszkrét esetben analóg módon járhatunk el, melyet az Olvasóra bízunk. Korábban már láttuk, hogy az adott feltételekkel I_n létezik és $I_n = nI_1 > 0$. Legyen

$$\varrho := \frac{g'(\vartheta)}{I_n(\vartheta)}\frac{\partial}{\partial\vartheta}\ln l_n(\xi_1, \ldots, \xi_n, \vartheta).$$

Ekkor

$$\mathrm{E}_\vartheta(\varrho^2) = \left(\frac{g'(\vartheta)}{I_n(\vartheta)}\right)^2 \mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln l_n(\xi_1, \ldots, \xi_n, \vartheta)\right)^2 = \frac{\big(g'(\vartheta)\big)^2}{I_n(\vartheta)},$$

másrészt

$$\mathrm{E}_\vartheta(\varrho) = \frac{g'(\vartheta)}{I_n(\vartheta)}\,\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln l_n(\xi_1, \ldots, \xi_n, \vartheta)\right) =$$
$$= \frac{g'(\vartheta)}{I_n(\vartheta)}\,\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\sum_{i=1}^{n}\ln f_\vartheta(\xi_i)\right) =$$
$$= \frac{g'(\vartheta)}{I_n(\vartheta)}\sum_{i=1}^{n}\mathrm{E}_\vartheta\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(\xi_i)\right) =$$
$$= \frac{g'(\vartheta)}{I_n(\vartheta)}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\vartheta}\ln f_\vartheta(x)\right) f_\vartheta(x)\,\mathrm{d}x =$$
$$= \frac{g'(\vartheta)}{I_n(\vartheta)}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\frac{\partial}{\partial\vartheta}f_\vartheta(x)\,\mathrm{d}x = 0.$$

Ezekből $\mathrm{D}^2_\vartheta(\varrho) = \mathrm{E}_\vartheta(\varrho^2) = \frac{(g'(\vartheta))^2}{I_n(\vartheta)}$, másrészt $\tau := T(\xi_1, \dots, \xi_n)$ jelöléssel

$$\begin{aligned}
\operatorname{cov}_\vartheta(\tau, \varrho) = \mathrm{E}_\vartheta(\tau\varrho) = \frac{g'(\vartheta)}{I_n(\vartheta)}\,\mathrm{E}_\vartheta\left(\tau\frac{\partial}{\partial\vartheta}\ln l_n(\xi_1, \dots, \xi_n, \vartheta)\right) = \\
= \frac{g'(\vartheta)}{I_n(\vartheta)}\int\limits_{\mathbb{R}^n} T(x_1, \dots, x_n)\frac{\partial}{\partial\vartheta}l_n(x_1, \dots, x_n, \vartheta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = \\
= \frac{g'(\vartheta)}{I_n(\vartheta)}\frac{\partial}{\partial\vartheta}\int\limits_{\mathbb{R}^n} T(x_1, \dots, x_n) l_n(x_1, \dots, x_n, \vartheta)\,\mathrm{d}x_1 \cdots \mathrm{d}x_n = \\
= \frac{g'(\vartheta)}{I_n(\vartheta)}\frac{\partial}{\partial\vartheta}\,\mathrm{E}_\vartheta(\tau) = \frac{g'(\vartheta)}{I_n(\vartheta)}\frac{\partial}{\partial\vartheta}g(\vartheta) = \frac{\big(g'(\vartheta)\big)^2}{I_n(\vartheta)}.
\end{aligned}$$

Így $0 \leqslant \mathrm{D}^2_\vartheta(\tau - \varrho) = \mathrm{D}^2_\vartheta(\tau) + \mathrm{D}^2_\vartheta(\varrho) - 2\operatorname{cov}_\vartheta(\tau, \varrho) = \mathrm{D}^2_\vartheta(\tau) - \frac{(g'(\vartheta))^2}{I_n(\vartheta)}$, melyből következik az állítás.

3.40. Lemma (Bederiválhatósági lemma)**.** *Ha $T(\xi_1, \dots, \xi_n)$ véges szórású statisztika, I_1 létezik, pozitív és folytonos, továbbá $\sqrt{l_1(x, \vartheta)}$ a ϑ változóban folytonosan differenciálható minden $x \in \mathfrak{X}$ esetén, akkor l_1-re és Tl_n-re teljesül a bederiválási feltétel.*

A bizonyítást nem közöljük, mert terjedelmes és bonyolult. (Lásd pl. A. A. Borovkov [1, 16. § 1. Lemma, 164. oldal, VI. Tétel bizonyítása, 470. oldal].) A bederiválhatósági lemma I_1-re és l_1-re vonatkozó feltételeit *gyenge regularitási feltételeknek* is nevezzük.

3.41. Feladat. A Rao–Cramér-egyenlőtlenséggel bizonyítsa be, hogy egy $0 < p < 1$ valószínűségű esemény relatív gyakorisága hatásos becslése p-nek.

Megoldás. Legyen ξ egy $0 < p < 1$ paraméterű karakterisztikus eloszlású valószínűségi változó, és a rávonatkozó minta $\xi_1, \dots, \xi_n$. Korábban láttuk, hogy $\overline{\xi}$ véges szórású torzítatlan becslése p-nek és $I_n(p) = \frac{n}{p(1-p)}$. Másrészt $g'(p) = (p)' = 1$ miatt az információs határ $\frac{p(1-p)}{n} = \mathrm{D}^2_p(\overline{\xi})$. Most legyen $T(\xi_1, \dots, \xi_n)$ tetszőleges véges szórású torzítatlan becslése p-nek. Mivel $\mathfrak{X}$ véges, ezért l_1-re és Tl_n-re teljesül a bederiválási feltétel. Így a Rao–Cramér-egyenlőtlenség miatt $\mathrm{D}^2_p T(\xi_1, \dots, \xi_n) \geqslant \mathrm{D}^2_p(\overline{\xi})$. Ebből következik az állítás.

3.42. Feladat. Legyen $\xi \in \mathbf{Norm}(m; \sigma)$, ahol $\sigma > 0$ rögzített. Bizonyítsa be, hogy a mintaátlag hatásos becslése m-nek.

Megoldás. Korábban láttuk, hogy $\overline{\xi}$ véges szórású torzítatlan becslése m-nek és $I_n(m) = \frac{n}{\sigma^2}$. Másrészt $g'(m) = (m)' = 1$ miatt az információs határ $\frac{\sigma^2}{n} = \mathrm{D}^2_m(\overline{\xi})$.

Most legyen $T(\xi_1, \dots, \xi_n)$ tetszőleges véges szórású torzítatlan becslése m-nek. Mivel a bederiválhatósági lemma minden feltétele teljesül, ezért l_1-re és Tl_n-re teljesül a bederiválási feltétel. Így a Rao–Cramér-egyenlőtlenség miatt $\mathrm{D}^2_m T(\xi_1, \dots, \xi_n) \geqslant \geqslant \mathrm{D}^2_m(\overline{\xi})$. Ebből következik az állítás.

3.43. Feladat. Legyen ξ ismeretlen λ paraméterű Poisson-eloszlású. Bizonyítsa be, hogy a mintaátlag hatásos becslése λ-nak.

Megoldás. Tudjuk, hogy $\overline{\xi}$ véges szórású torzítatlan becslése λ-nak és $I_n(\lambda) = \frac{n}{\lambda}$. Másrészt $g'(\lambda) = (\lambda)' = 1$ miatt az információs határ $\frac{\lambda}{n} = \mathrm{D}^2_\lambda(\overline{\xi})$. Most legyen $T(\xi_1, \dots, \xi_n)$ tetszőleges véges szórású torzítatlan becslése λ-nak. Mivel a bederiválhatósági lemma minden feltétele teljesül, ezért l_1-re és Tl_n-re teljesül a bederiválási feltétel. Így a Rao–Cramér-egyenlőtlenség miatt $\mathrm{D}^2_\lambda T(\xi_1, \dots, \xi_n) \geqslant \mathrm{D}^2_\lambda(\overline{\xi})$. Ebből következik az állítás.

3.44. Feladat. Legyen $\xi \in \mathbf{Exp}(\lambda)$. Bizonyítsa be, hogy a mintaátlag hatásos becslése $\frac{1}{\lambda}$-nak.

Megoldás. Korábban láttuk, hogy $\overline{\xi}$ véges szórású torzítatlan becslése $\frac{1}{\lambda}$-nak és $I_n(\lambda) = \frac{n}{\lambda^2}$. Másrészt $g'(\lambda) = (\frac{1}{\lambda})' = \frac{-1}{\lambda^2}$ miatt az információs határ $\frac{1}{n\lambda^2} = \mathrm{D}^2_\lambda(\overline{\xi})$. Most legyen $T(\xi_1, \dots, \xi_n)$ tetszőleges véges szórású torzítatlan becslése $\frac{1}{\lambda}$-nak. Mivel a bederiválhatósági lemma minden feltétele teljesül, ezért l_1-re és Tl_n-re teljesül a bederiválási feltétel. Így a Rao–Cramér-egyenlőtlenség miatt $\mathrm{D}^2_\lambda T(\xi_1, \dots, \xi_n) \geqslant \mathrm{D}^2_\lambda(\overline{\xi})$. Ebből következik az állítás.

3.3. Pontbecslési módszerek

A fejezet hátralévő részében két általános módszert ismertetünk pontbecslések konstruálására.

3.3.1. Momentumok módszere

Ez volt az első általános eljárás pontbecslések készítésére. A módszer *K. Pearson* nevéhez fűződik. Az elve az, hogy r darab ismeretlen paraméter esetén a k-adik momentumot a k-adik tapasztalati momentummal becsüljük ($k = 1, \dots, r$). A következő tétel szerint, bizonyos feltételek esetén az így kapott becslései az ismeretlen paramétereknek erősen konzisztensek.

3.45. Tétel. *Legyen a vizsgált valószínűségi változó ξ és a paramétertér $\Theta \subset \mathbb{R}^r$ nyílt halmaz. Tegyük fel, hogy $\mathrm{E}_\vartheta\, \xi^r$ létezik és véges minden $\vartheta = (\vartheta_1, \dots, \vartheta_r) \in \Theta$*

esetén, $\frac{\partial}{\partial\vartheta_j}\,\mathrm{E}_\vartheta\,\xi^i$ *létezik és folytonos* Θ*-n minden* $i,j\in\{\,1,\dots,r\,\}$ *esetén, továbbá az úgynevezett* Jacobi-determináns

$$\det\left(\frac{\partial}{\partial\vartheta_j}\,\mathrm{E}_\vartheta\,\xi^i\right)\neq 0$$

minden $\vartheta=(\vartheta_1,\dots,\vartheta_r)\in\Theta$ *esetén. Ha az*

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i^k=\mathrm{E}_\vartheta\,\xi^k,\quad k=1,\dots,r$$

egyenletrendszernek 1-hez tartó valószínűséggel létezik $\widehat{\vartheta}_n=(\widehat{\vartheta}_{1n},\dots,\widehat{\vartheta}_{rn})$ *egyértelmű megoldása, amint* $n\to\infty$*, akkor* $\widehat{\vartheta}_{kn}$ *erősen konzisztens becsléssorozata* ϑ_k*-nak* $(k=1,\dots,r)$*.*

Bizonyítás. Legyen

$$G\colon\Theta\to\mathbb{R}^r,\quad G(\vartheta):=(\mathrm{E}_\vartheta\,\xi^1,\dots,\mathrm{E}_\vartheta\,\xi^r).$$

Az adott feltételekkel G folytonos, így Θ nyíltsága miatt $G(\Theta)$ is nyílt. Ebből létezik rögzített $\vartheta\in\Theta$ esetén $G(\vartheta)$-nak olyan $\varepsilon>0$ sugarú környezete, mely részhalmaza $G(\Theta)$-nak. A nagy számok erős törvénye miatt $\frac{1}{n}\sum_{i=1}^{n}\xi_i^k$ erősen konzisztens becslés-sorozata $\mathrm{E}_\vartheta\,\xi^k$-nak $(k=1,\dots,r)$, melyből a konzisztencia is következik. Így bármely $\delta>0$ esetén van olyan $N\in\mathbb{N}$, hogy $n>N$ esetén

$$\mathrm{P}_\vartheta\left(\left|\frac{1}{n}\sum_{i=1}^{n}\xi_i^k-\mathrm{E}_\vartheta\,\xi^k\right|\geqslant\frac{\varepsilon}{\sqrt{r}}\right)<\frac{\delta}{r},\quad k=1,\dots,r.$$

Innen kapjuk, hogy

$$\begin{aligned}&\mathrm{P}_\vartheta\left(\sum_{k=1}^{r}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^k-\mathrm{E}_\vartheta\,\xi^k\right)^2\geqslant\varepsilon^2\right)\leqslant\\&\leqslant\mathrm{P}_\vartheta\left(\bigcup_{k=1}^{r}\left\{\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^k-\mathrm{E}_\vartheta\,\xi^k\right)^2\geqslant\frac{\varepsilon^2}{r}\right\}\right)\leqslant\\&\leqslant\sum_{k=1}^{r}\mathrm{P}_\vartheta\left(\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^k-\mathrm{E}_\vartheta\,\xi^k\right)^2\geqslant\frac{\varepsilon^2}{r}\right)<\delta,\end{aligned}$$

azaz $\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i^1,\dots,\frac{1}{n}\sum_{i=1}^{n}\xi_i^r\right)\in G(\Theta)$ legalább $1-\delta$ valószínűséggel, amennyiben

$n > N$. Ebből következik, hogy

$$\lim_{n\to\infty} \mathrm{P}_\vartheta\Big((\tfrac{1}{n}\sum_{i=1}^n \xi_i^1, \dots, \tfrac{1}{n}\sum_{i=1}^n \xi_i^r) \in G(\Theta)\Big) = 1.$$

Tehát 1-hez tartó valószínűséggel $\widehat{\vartheta}_n = G^{-1}(\frac{1}{n}\sum_{i=1}^n \xi_i^1, \dots, \frac{1}{n}\sum_{i=1}^n \xi_i^r)$, ahol G^{-1} a G inverzét jelenti. Az inverzfüggvény-tétel miatt (lásd W. Rudin [14, 230. oldal]) az adott feltételekkel G^{-1} létezik és folytonos. $\frac{1}{n}\sum_{i=1}^n \xi_i^k$ erősen konzisztens becsléssorozata $\mathrm{E}_\vartheta\, \xi^k$-nak ($k = 1, \dots, r$), melyből a G^{-1} folytonossága miatt 1 valószínűséggel teljesül, hogy

$$\lim_{n\to\infty} G^{-1}(\tfrac{1}{n}\sum_{i=1}^n \xi_i^1, \dots, \tfrac{1}{n}\sum_{i=1}^n \xi_i^r) = G^{-1}(G(\vartheta)) = \vartheta.$$

Mindezekből

$$\mathrm{P}_\vartheta\left(\lim_{n\to\infty} \widehat{\vartheta}_n = \vartheta\right) = 1.$$

(Az utóbbi két határérték koordinátánként értendő.) Ezzel az állítás bizonyított.

3.46. Feladat. Bizonyítsa be, hogy ha $\xi \in \mathbf{Exp}(\lambda)$, akkor $\widehat{\lambda}_n = \frac{n}{\sum_{i=1}^n \xi_i}$ erősen konzisztens becléssorozata λ-nak.

Megoldás. Az előző tétel feltételei teljesülnek, így az

$$\frac{1}{n}\sum_{i=1}^n \xi_i = \mathrm{E}_\lambda\, \xi = \frac{1}{\lambda}$$

megoldása erősen konzisztens becléssorozata λ-nak.

3.47. Feladat. $\xi \in \mathbf{Norm}(m; \sigma)$ esetén számolja ki az m és σ becslését a momentumok módszerével.

Megoldás. A következő egyenletrendszert kapjuk:

$$\frac{1}{n}\sum_{i=1}^n \xi_i = m$$
$$\frac{1}{n}\sum_{i=1}^n \xi_i^2 = m^2 + \sigma^2$$

Ennek a megoldása $\widehat{m}_n = \overline{\xi}$ és $\widehat{\sigma}_n = S_n$. Ezekről már korábban is láttuk, hogy erősen konzisztens becléssorozatok, de az előző tétel is ezt mutatja, hiszen a feltételek teljesülnek.

3.48. Feladat. Legyen ξ egyenletes eloszlású az ismeretlen $[a, b]$ intervallumon. Számolja ki az a és b becslését a momentumok módszerével. Bizonyítsa be, hogy ezek erősen konzisztens becsléssorozatok.

Megoldás. A következő egyenletrendszert kapjuk:

$$\frac{1}{n}\sum_{i=1}^{n}\xi_i = \frac{a+b}{2}$$
$$\frac{1}{n}\sum_{i=1}^{n}\xi_i^2 = \frac{(a-b)^2}{12} + \left(\frac{a+b}{2}\right)^2$$

Ennek a megoldása $\widehat{a}_n = \overline{\xi} - \sqrt{3}S_n$ és $\widehat{b}_n = \overline{\xi} + \sqrt{3}S_n$. Egyszerű számolással kapjuk, hogy a Jacobi-determináns $\frac{b-a}{6}$, így az előző tétel miatt teljesül, hogy ezek a becsléssorozatok erősen konzisztensek.

3.3.2. Maximum likelihood becslés

A maximum likelihood (magyarul: legnagyobb valószínűség) becslés elve az, hogy adott mintarealizációhoz az ismeretlen paramétereknek olyan becslését adjuk meg, amely mellett az adott mintarealizáció a legnagyobb valószínűséggel következik be.

Ennek az elvnek a vizsgálatában feltesszük, hogy a vizsgált ξ valószínűségi változó abszolút folytonos vagy diszkrét, $\Theta \subset \mathbb{R}^r$, a ξ-re vonatkozó minta $\xi_1, \ldots, \xi_n$, továbbá a ξ értékkészlete $\mathfrak{X}$, azaz a mintatér $\mathfrak{X}^n$. Ha ξ abszolút folytonos, akkor f_ϑ jelölje ξ-nek a P_ϑ-ból származó sűrűségfüggvényét, ahol $\vartheta = (\vartheta_1, \ldots, \vartheta_r) \in \Theta$. Először a már korábban definiált likelihood függvényt terjesztjük ki $\Theta \subset \mathbb{R}^r$ esetre.

3.49. Definíció. A $\xi_1, \ldots, \xi_n$ minta *likelihood függvénye*

$$l_n \colon \mathfrak{X}^n \times \Theta \to \mathbb{R},$$
$$l_n(x_1, \ldots, x_n, \vartheta_1, \ldots, \vartheta_r) := \begin{cases} \prod_{i=1}^{n} f_\vartheta(x_i), & \text{ha } \xi \text{ absz. folyt.,} \\ \prod_{i=1}^{n} \mathrm{P}_\vartheta(\xi_i = x_i), & \text{ha } \xi \text{ diszkrét.} \end{cases}$$

3.50. Definíció. A $\widehat{\vartheta}_k = T_k(\xi_1, \ldots, \xi_n)$ statisztika a ϑ_k *maximum likelihood becslése* $(k = 1, \ldots, r)$, ha

$$l_n\Big(\xi_1(\omega), \ldots, \xi_n(\omega), \widehat{\vartheta}_1(\omega), \ldots, \widehat{\vartheta}_r(\omega)\Big) \geqslant l_n\Big(\xi_1(\omega), \ldots, \xi_n(\omega), \vartheta_1, \ldots, \vartheta_r\Big)$$

minden $(\vartheta_1, \ldots, \vartheta_r) \in \Theta$ és $\omega \in \Omega$ esetén.

Tehát a becslés kiszámítása nem más, mint szélsőértékhely keresés. Praktikus okból nem a likelihood függvénynek fogjuk a maximumhelyét keresni, hanem a természetes alapú logaritmusának. Ezzel a szélsőértékhely nem változik, hiszen ln szigorúan monoton növekvő függvény. Az ok az, hogy ekkor nem szorzatot, hanem összeget kell vizsgálni.

3.51. Definíció. A $\xi_1, \dots, \xi_n$ minta *loglikelihood függvénye* $L_n := \ln l_n$.

3.52. Feladat. Legyen ξ egyenletes eloszlású az $[a, b]$ intervallumon. Számolja ki a és b maximum likelihood becslését.

Megoldás. A loglikelihood függvény

$$L_n(\xi_1, \dots, \xi_n, a, b) = \begin{cases} -n \ln(b-a), & \text{ha } \xi_1^* \geqslant a \text{ és } \xi_n^* \leqslant b, \\ 0, & \text{különben.} \end{cases}$$

Ennek maximumhelye $\widehat{a} = \xi_1^*$ és $\widehat{b} = \xi_n^*$, így a maximum likelihood becslése a-nak $\widehat{a} = \xi_1^*$ és b-nek $\widehat{b} = \xi_n^*$.

3.53. Feladat. Legyen ξ Poisson-eloszlású λ paraméterrel. Számolja ki λ maximum likelihood becslését azzal a feltevéssel, hogy a mintarealizációnak van nullától különböző eleme.

Megoldás. $L_n(\xi_1, \dots, \xi_n, \lambda) = \sum_{i=1}^{n} \ln \frac{\lambda^{\xi_i}}{\xi_i!} e^{-\lambda} = \sum_{i=1}^{n} (\xi_i \ln \lambda - \ln \xi_i! - \lambda)$, ami λ változó szerint differenciálható függvény az $\mathbb{R}_+$ halmazon. Mivel

$$\frac{\partial}{\partial \lambda} L_n(\xi_1, \dots, \xi_n, \lambda) = \frac{n\overline{\xi}}{\lambda} - n = 0$$

megoldása $\overline{\xi}$, és $\frac{\partial^2}{\partial \lambda^2} L_n(\xi_1, \dots, \xi_n, \overline{\xi}) = -n/\overline{\xi} < 0$, ezért $\overline{\xi}$ lokális maximumhely. Mivel $\mathbb{R}_+$ összefüggő halmaz, és csak egy lokális szélsőértékhely van, ezért $\overline{\xi}$ globális maximumhely. Tehát a maximum likelihood becslése λ-nak $\widehat{\lambda} = \overline{\xi}$.

3.54. Feladat. Legyen $\xi \in \mathbf{Exp}(\lambda)$. Számolja ki λ maximum likelihood becslését.

Megoldás. $L_n(\xi_1, \dots, \xi_n, \lambda) = \sum_{i=1}^{n} \ln \left(\lambda e^{-\lambda \xi_i}\right) = \sum_{i=1}^{n} (\ln \lambda - \lambda \xi_i) = n \ln \lambda - \lambda n \overline{\xi}$, ami λ változó szerint differenciálható függvény az $\mathbb{R}_+$ halmazon. Mivel

$$\frac{\partial}{\partial \lambda} L_n(\xi_1, \dots, \xi_n, \lambda) = \frac{n}{\lambda} - n\overline{\xi} = 0$$

megoldása $1/\overline{\xi}$, és $\frac{\partial^2}{\partial\lambda^2}L_n(\xi_1,\dots,\xi_n,1/\overline{\xi}) = -n\overline{\xi}^2 < 0$, ezért $1/\overline{\xi}$ lokális maximumhely. Mivel $\mathbb{R}_+$ összefüggő halmaz, és csak egy lokális szélsőértékhely van, ezért $1/\overline{\xi}$ globális maximumhely. Tehát a maximum likelihood becslése λ-nak $\widehat{\lambda} = 1/\overline{\xi}$.

3.55. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$. Számolja ki m és σ maximum likelihood becslését.

Megoldás. A loglikelihood függvény

$$L_n(\xi_1,\dots,\xi_n,m,\sigma) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\xi_i-m)^2}{2\sigma^2}\right)\right) = \\ = \sum_{i=1}^{n}\left(-\ln\sigma - \ln\sqrt{2\pi} - \frac{(\xi_i-m)^2}{2\sigma^2}\right),$$

ami m és σ változók szerint parciálisan differenciálható függvény az $\mathbb{R}\times\mathbb{R}_+$ halmazon. Tekintsük a következő egyenletrendszert:

$$\frac{\partial}{\partial m}L_n(\xi_1,\dots,\xi_n,m,\sigma) = \frac{n}{\sigma^2}(\overline{\xi}-m) = 0$$
$$\frac{\partial}{\partial \sigma}L_n(\xi_1,\dots,\xi_n,m,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(\xi_i-m)^2 = 0$$

Ennek egyetlen megoldása: $\widehat{m} = \overline{\xi}$ és $\widehat{\sigma} = S_n$. Másrészt

$$A := \frac{\partial^2}{\partial m^2}L_n(\xi_1,\dots,\xi_n,\widehat{m},\widehat{\sigma}) = -\frac{n}{S_n^2} < 0$$
$$B := \frac{\partial^2}{\partial \sigma^2}L_n(\xi_1,\dots,\xi_n,\widehat{m},\widehat{\sigma}) = -\frac{2n}{S_n^2}$$
$$C := \frac{\partial^2}{\partial m\partial \sigma}L_n(\xi_1,\dots,\xi_n,\widehat{m},\widehat{\sigma}) = 0$$

továbbá $AB - C^2 = \frac{2n^2}{S_n^4} > 0$, így $(\overline{\xi}, S_n)$ lokális maximumhely. Mivel $\mathbb{R}\times\mathbb{R}_+$ összefüggő halmaz, és csak egy lokális szélsőértékhely van, ezért $(\overline{\xi}, S_n)$ globális maximumhely. Tehát a maximum likelihood becslése m-nek $\widehat{m} = \overline{\xi}$, illetve σ-nak $\widehat{\sigma} = S_n$.

Az utóbbi három példában láttuk, hogy a maximum likelihood becslés meghatározásánál kulcsszerepe lehet a

$$\frac{\partial}{\partial \vartheta_k}L_n(\xi_1,\dots,\xi_n,\vartheta_1,\dots,\vartheta_r) = 0 \quad (k=1,\dots,r)$$

egyenletrendszernek. Ezt az egyenletrendszert *likelihood egyenletrendszernek* nevezzük. Természetesen $r = 1$ esetén egyenletrendszer helyett egyenletet kapunk. Sokszor

a likelihood egyenletrendszer megoldása és a maximum likelihood becslés egybeesik, de ez nem mindig van így. Ilyen példa konstruálása igen bonyolult, most eltekintünk tőle.

A likelihood egyenlet megoldásának a jó tulajdonságát, bizonyos feltételek esetén, a következő tétel fogalmazza meg.

3.56. Tétel (Wald-tétel)**.** *Ha $\Theta \subset \mathbb{R}$, az L_1 differenciálható a valódi ϑ^* paraméter egy $U \subset \Theta$ környezetében, továbbá $\mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta)$ létezik és véges minden $\vartheta \in U$ esetén, akkor a likelihood egyenletnek van olyan $\widehat{\vartheta}$ megoldása, amelyre teljesül, hogy*

$$\mathrm{P}_{\vartheta^*}\left(\lim_{n\to\infty} \widehat{\vartheta} = \vartheta^*\right) = 1,$$

ahol n a minta elemszámát jelenti.

Bizonyítás. Csak abszolút folytonos esetben bizonyítunk, de diszkrét esetben analóg módon járhatunk el, melyet az Olvasóra bízunk. Mivel $-\ln$ konvex függvény, ezért a Jensen-egyenlőtlenség alapján minden $\vartheta \in U$ esetén

$$\mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta) - \mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta^*) = \mathrm{E}_{\vartheta^*} \ln \frac{l_1(\xi, \vartheta)}{l_1(\xi, \vartheta^*)} \leqslant$$

$$\leqslant \ln \mathrm{E}_{\vartheta^*} \frac{l_1(\xi, \vartheta)}{l_1(\xi, \vartheta^*)} = \ln \int_{-\infty}^{\infty} f_\vartheta(x)\,\mathrm{d}x = \ln 1 = 0,$$

azaz az identifikálhatóság miatt minden $\vartheta \in U,\ \vartheta \neq \vartheta^*$ esetén

$$\mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta) < \mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta^*).$$

A Kolmogorov-féle nagy számok erős törvénye és

$$L_n(\xi_1, \ldots, \xi_n, \vartheta) = \sum_{i=1}^{n} \ln f_\vartheta(\xi_i)$$

miatt

$$\mathrm{P}_{\vartheta^*}\left(\lim_{n\to\infty} \frac{1}{n} L_n(\xi_1, \ldots, \xi_n, \vartheta) = \mathrm{E}_{\vartheta^*} L_1(\xi, \vartheta)\right) = 1$$

minden $\vartheta \in U$ esetén. Mindezekből kapjuk, hogy

$$\mathrm{P}_{\vartheta^*}\left(\lim_{n\to\infty} \frac{1}{n} L_n(\xi_1, \ldots, \xi_n, \vartheta) < \lim_{n\to\infty} \frac{1}{n} L_n(\xi_1, \ldots, \xi_n, \vartheta^*)\right) = 1$$

minden $\vartheta \in U,\ \vartheta \neq \vartheta^*$ esetén. Ebből elég nagy n-ekre kapjuk, hogy

$$\mathrm{P}_{\vartheta^*}\big(L_n(\xi_1,\ldots,\xi_n,\vartheta) < L_n(\xi_1,\ldots,\xi_n,\vartheta^*)\big) = 1$$

minden $\vartheta \in U,\ \vartheta \neq \vartheta^*$ esetén. Most legyen $\delta > 0$ olyan, hogy $\vartheta^* \pm \delta \in U$. Ekkor elég nagy n-ekre

$$\mathrm{P}_{\vartheta^*}\big(L_n(\xi_1,\ldots,\xi_n,\vartheta^* \pm \delta) < L_n(\xi_1,\ldots,\xi_n,\vartheta^*)\big) = 1,$$

melyből következik az állítás, hiszen δ tetszőlegesen kicsi lehet.

A likelihood egyenlet egy megoldásának további jó tulajdonságait állítja Cramér tétele, melyet bonyolultsága miatt nem taglalunk (lásd pl. Fazekas I. [2, 90. oldal]).

4. Intervallumbecslések

4.1. Az intervallumbecslés feladata

Legyen ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{ \mathrm{P}_\vartheta : \vartheta \in \Theta \}$ statisztikai mezőn, ahol $\Theta \subset \mathbb{R}^v$ nyílt halmaz. A feladat $(\vartheta_1, \ldots, \vartheta_v) \in \Theta$, $k \in \{1, \ldots, v\}$ jelöléssel ϑ_k valódi értékének becslése.

Amint korábban láttuk a pontbecslés ϑ_k valódi értékét egy számmal becsüli. Mindezt egy statisztika realizációjával tettük meg. Intervallumbecslésnél egy olyan intervallumot adunk meg, amelybe a ϑ_k valódi értéke nagy valószínűséggel beleesik. Ezen intervallum alsó és felső végpontját egy-egy statisztika realizációjával adjuk meg. Magát a becslő intervallumot *konfidenciaintervallumnak* fogjuk nevezni.

4.1. Definíció. Legyen a ξ-re vonatkozó minta $\xi_1, \ldots, \xi_n$, továbbá

$$\tau_1 := T_1(\xi_1, \ldots, \xi_n) \quad \text{és} \quad \tau_2 := T_2(\xi_1, \ldots, \xi_n)$$

statisztikák. Azt mondjuk, hogy $[\tau_1, \tau_2]$ $1-\alpha$ *biztonsági szintű konfidenciaintervallum* a ϑ_k paraméterre, ha

$$\mathrm{P}_\vartheta(\tau_1 \leqslant \vartheta_k \leqslant \tau_2) \geqslant 1 - \alpha$$

minden $\vartheta = (\vartheta_1, \ldots, \vartheta_v) \in \Theta$ esetén, ahol $0 < \alpha < 1$. A $[\tau_1, \tau_2]$ intervallumot *centrált konfidenciaintervallumnak* nevezzük ϑ_k-ra, ha

$$\mathrm{P}_\vartheta(\vartheta_k < \tau_1) = \mathrm{P}_\vartheta(\vartheta_k > \tau_2)$$

minden $\vartheta = (\vartheta_1, \ldots, \vartheta_v) \in \Theta$ esetén. Az

$$\inf_{\vartheta \in \Theta} \mathrm{P}_\vartheta(\tau_1 \leqslant \vartheta_k \leqslant \tau_2)$$

értéket a ϑ_k-ra vonatkozó $[\tau_1, \tau_2]$ konfidenciaintervallum *pontos biztonsági szintjének* nevezzük.

Ha ξ diszkrét, akkor adott α-hoz nem feltétlenül található olyan konfidenciaintervallum, melynek $1 - \alpha$ a pontos biztonsági szintje. Ezért definiáltuk a biztonsági szintet az előző módon.

4.2. Konfidenciaintervallum a normális eloszlás paramétereire

4.2. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1,\dots,\xi_n$ egy ξ-re vonatkozó minta. Tegyük fel, hogy m ismeretlen, de σ ismert. Adjon m-re olyan centrált konfidenciaintervallumot, melynek $1-\alpha$ a pontos biztonsági szintje.

A megoldáshoz szükségünk lesz a következő tételre.

4.3. Tétel. *Ha $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1,\dots,\xi_n$ egy ξ-re vonatkozó minta, akkor*

$$\frac{\overline{\xi}-m}{\sigma}\sqrt{n} \in \mathbf{Norm}(0;1).$$

Bizonyítás. Tudjuk, hogy $\overline{\xi}$ normális eloszlású, $\mathrm{E}\,\overline{\xi} = \mathrm{E}\,\xi = m$ és $\mathrm{D}^2\,\overline{\xi} = \frac{1}{n^2} n\,\mathrm{D}^2\,\xi = \frac{1}{n}\sigma^2$, azaz $\overline{\xi} \in \mathbf{Norm}\big(m;\frac{\sigma}{\sqrt{n}}\big)$. Így, ha F jelöli a $\frac{\overline{\xi}-m}{\sigma}\sqrt{n}$ eloszlásfüggvényét, akkor $x \in \mathbb{R}$ esetén

$$F(x) = \mathrm{P}\left(\overline{\xi} < \frac{\sigma}{\sqrt{n}}x + m\right) = \Phi\left(\frac{\frac{\sigma}{\sqrt{n}}x + m - m}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi(x).$$

Ezzel bizonyított az állítás.

Most térjünk vissza a feladatra.

Megoldás. Legyen $c \in \mathbb{R}_+$. Ekkor az előző tétel szerint

$$\mathrm{P}_m\left(-c \leqslant \frac{\overline{\xi}-m}{\sigma}\sqrt{n} \leqslant c\right) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

Mivel $2\Phi(c)-1 = 1-\alpha$ pontosan akkor teljesül, ha $c = \Phi^{-1}(1-\frac{\alpha}{2})$, ezért ilyen c-re átrendezéssel azt kapjuk, hogy

$$\mathrm{P}_m\left(\overline{\xi} - \frac{\sigma}{\sqrt{n}}c \leqslant m \leqslant \overline{\xi} + \frac{\sigma}{\sqrt{n}}c\right) = 1-\alpha.$$

Könnyű látni, hogy ez centrált konfidenciaintervallum, hiszen

$$\begin{aligned}\mathrm{P}_m\left(m > \overline{\xi} + \frac{\sigma}{\sqrt{n}}c\right) &= \mathrm{P}_m\left(\frac{\overline{\xi}-m}{\sigma}\sqrt{n} < -c\right) = \\ &= \Phi(-c) = 1-\Phi(c) = 1-\left(1-\frac{\alpha}{2}\right) = \frac{\alpha}{2}.\end{aligned}$$

Összefoglalva tehát a megoldás:

$$\tau_1 := \overline{\xi} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

$$\tau_2 := \overline{\xi} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

jelölésekkel $[\tau_1, \tau_2]$ olyan centrált konfidenciaintervallum m-re, melynek $1-\alpha$ a pontos biztonsági szintje.

4.4. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta. Tegyük fel, hogy m ismert és σ ismeretlen. Adjon σ-ra olyan centrált konfidenciaintervallumot, melynek $1-\alpha$ a pontos biztonsági szintje.

A megoldáshoz szükségünk lesz a következő tételre.

4.5. Tétel. *Ha $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta, akkor*

$$\sum_{i=1}^{n} \frac{(\xi_i - m)^2}{\sigma^2} \in \mathbf{Khi}(n).$$

Bizonyítás. Mivel $\frac{\xi_i - m}{\sigma}$ $(i = 1, \ldots, n)$ független standard normális eloszlású valószínűségi változók, ezért a négyzetösszegük n szabadsági fokú khi-négyzet eloszlású valószínűségi változó.

A feladat megoldása előtt bevezetünk egy jelölést, melyet a továbbiakban gyakran fogunk alkalmazni. Legyen η egy tetszőleges valószínűségi változó, és $\mathbf{V}$ az η-val azonos eloszlású valószínűségi változók halmaza. Ekkor $F \sim \mathbf{V}$ jelölje azt, hogy F a $\mathbf{V}$-beli valószínűségi változók közös eloszlásfüggvénye. Például $\Phi \sim \mathbf{Norm}(0;1)$.

Megoldás. Legyen $c_1, c_2 \in \mathbb{R}_+$ és $F \sim \mathbf{Khi}(n)$. Ekkor az előző tétel szerint

$$\mathrm{P}_\sigma\left(\sum_{i=1}^{n} \frac{(\xi_i - m)^2}{\sigma^2} < c_1\right) = F(c_1),$$
$$\mathrm{P}_\sigma\left(\sum_{i=1}^{n} \frac{(\xi_i - m)^2}{\sigma^2} > c_2\right) = 1 - F(c_2).$$

Mivel $F(c_1) = 1 - F(c_2) = \frac{\alpha}{2}$ pontosan akkor teljesül, ha $c_1 = F^{-1}(\frac{\alpha}{2})$ és $c_2 = = F^{-1}(1 - \frac{\alpha}{2})$, ezért ebben az esetben

$$\mathrm{P}_\sigma\left(c_1 \leqslant \sum_{i=1}^{n} \frac{(\xi_i - m)^2}{\sigma^2} \leqslant c_2\right) = 1 - \alpha,$$

azaz átrendezve

$$\mathrm{P}_\sigma\left(\sqrt{\frac{\sum_{i=1}^n(\xi_i-m)^2}{c_2}}\leqslant\sigma\leqslant\sqrt{\frac{\sum_{i=1}^n(\xi_i-m)^2}{c_1}}\right)=1-\alpha.$$

Vegyük észre, hogy $\frac{\alpha}{2}<1-\frac{\alpha}{2}$ miatt $c_1<c_2$. Összefoglalva tehát a megoldás:

$$\begin{aligned}F&\sim\mathbf{Khi}(n)\\ \tau_1&:=\sqrt{\frac{\sum_{i=1}^n(\xi_i-m)^2}{F^{-1}\left(1-\frac{\alpha}{2}\right)}}\\ \tau_2&:=\sqrt{\frac{\sum_{i=1}^n(\xi_i-m)^2}{F^{-1}\left(\frac{\alpha}{2}\right)}}\end{aligned}$$

jelölésekkel $[\tau_1,\tau_2]$ olyan centrált konfidenciaintervallum σ-ra, melynek $1-\alpha$ a pontos biztonsági szintje.

4.6. Feladat. Legyen $\xi\in\mathbf{Norm}(m;\sigma)$ és $\xi_1,\dots,\xi_n$ egy ξ-re vonatkozó minta ($n\geqslant$ $\geqslant 2$). Tegyük fel, hogy m és σ ismeretlenek. Adjon σ-ra centrált konfidenciaintervallumot, melynek $1-\alpha$ a pontos biztonsági szintje.

A megoldáshoz szükségünk lesz a következő tételre.

4.7. Tétel. *Ha $\xi\in\mathbf{Norm}(m;\sigma)$ és $\xi_1,\dots,\xi_n$ egy ξ-re vonatkozó minta ($n\geqslant 2$), akkor $\overline{\xi}$ és S_n^2 függetlenek, továbbá*

$$\frac{S_n^2}{\sigma^2}n\in\mathbf{Khi}(n-1).$$

Bizonyítás. Legyen $X:=(\xi_1-m,\dots,\xi_n-m)^\top$, az U olyan $n\times n$-es ortonormált mátrix (azaz $U^\top U$ egységmátrix), melynek első sorában minden elem $\frac{1}{\sqrt{n}}$, továbbá $Y:=(\eta_1,\dots,\eta_n)^\top:=UX$. Ekkor $\eta_1=\frac{1}{\sqrt{n}}\sum_{i=1}^n(\xi_i-m)$, azaz $\frac{1}{\sqrt{n}}\eta_1=\overline{\xi}-m$, továbbá

$$\sum_{i=1}^n(\xi_i-m)^2=X^\top X=X^\top U^\top UX=Y^\top Y=\sum_{i=1}^n\eta_i^2.$$

Mindezekből a Steiner-formula alapján

$$S_n^2=\frac{1}{n}\sum_{i=1}^n(\xi_i-m)^2-(\overline{\xi}-m)^2=\frac{1}{n}\sum_{i=1}^n\eta_i^2-\frac{1}{n}\eta_1^2=\frac{1}{n}\sum_{i=2}^n\eta_i^2,$$

azaz

$$\frac{S_n^2}{\sigma^2} n = \sum_{i=2}^{n} \left(\frac{\eta_i}{\sigma}\right)^2.$$

Jelölje u_{ij} az U mátrix i-edik sorában és j-edik oszlopában álló elemét. Ekkor

$$\eta_i = \sum_{j=1}^{n} u_{ij}(\xi_j - m),$$

amiből következik, hogy η_i normális eloszlású,

$$\mathrm{E}\,\eta_i = \sum_{j=1}^{n} u_{ij}(\mathrm{E}\,\xi_j - m) = 0$$

és az U ortonormáltsága miatt

$$\mathrm{D}^2\,\eta_i = \sum_{j=1}^{n} u_{ij}^2\,\mathrm{D}^2(\xi_j - m) = \sigma^2 \sum_{j=1}^{n} u_{ij}^2 = \sigma^2.$$

Így $\eta_i \in \mathbf{Norm}(0;\sigma)$. Másrészt $i \neq j$ esetén

$$\mathrm{cov}\,(\eta_i, \eta_j) = \mathrm{E}\,\eta_i\eta_j = \sum_{l=1}^{n}\sum_{t=1}^{n} u_{il}u_{jt}\,\mathrm{cov}(\xi_l, \xi_t) = \sum_{l=1}^{n} u_{il}u_{jl} = 0.$$

Ezekből következik, hogy $\eta_1, \ldots, \eta_n$ függetlenek. Mivel $\overline{\xi}$ csak η_1-től függ, illetve S_n^2 csak $\eta_2, \ldots, \eta_n$-től függ, ezért $\overline{\xi}$ és S_n^2 függetlenek.

Másrészt azt is kaptuk, hogy $\frac{\eta_2}{\sigma}, \ldots, \frac{\eta_n}{\sigma}$ olyan független standard normális eloszlású valószínűségi változók, melyeknek a négyzetösszege $\frac{S_n^2}{\sigma^2} n$. Ebből már következik, hogy $\frac{S_n^2}{\sigma^2} n \in \mathbf{Khi}(n-1)$.

Most rátérünk a feladat megoldására.

Megoldás. Legyen $c_1, c_2 \in \mathbb{R}_+$ és $F \sim \mathbf{Khi}(n-1)$. Ekkor az előző tétel szerint

$$\mathrm{P}_{(m,\sigma)}\left(\frac{S_n^2}{\sigma^2} n < c_1\right) = F(c_1),$$
$$\mathrm{P}_{(m,\sigma)}\left(\frac{S_n^2}{\sigma^2} n > c_2\right) = 1 - F(c_2).$$

Mivel $F(c_1) = 1 - F(c_2) = \frac{\alpha}{2}$ pontosan akkor teljesül, ha $c_1 = F^{-1}(\frac{\alpha}{2})$ és $c_2 =$

$= F^{-1}(1-\frac{\alpha}{2})$, ezért ebben az esetben

$$\mathrm{P}_{(m,\sigma)}\left(c_1 \leqslant \frac{S_n^2}{\sigma^2} n \leqslant c_2\right) = 1-\alpha,$$

azaz átrendezve

$$\mathrm{P}_{(m,\sigma)}\left(S_n\sqrt{\frac{n}{c_2}} \leqslant \sigma \leqslant S_n\sqrt{\frac{n}{c_1}}\right) = 1-\alpha.$$

Vegyük észre, hogy $\frac{\alpha}{2} < 1-\frac{\alpha}{2}$ miatt $c_1 < c_2$. Összefoglalva tehát a megoldás:

$$\begin{aligned} F &\sim \mathbf{Khi}(n-1) \\ \tau_1 &:= S_n\sqrt{\frac{n}{F^{-1}\left(1-\frac{\alpha}{2}\right)}} \\ \tau_2 &:= S_n\sqrt{\frac{n}{F^{-1}\left(\frac{\alpha}{2}\right)}} \end{aligned}$$

jelölésekkel $[\tau_1, \tau_2]$ olyan centrált konfidenciaintervallum σ-ra, melynek $1-\alpha$ a pontos biztonsági szintje.

4.8. *Megjegyzés.* Az előző megoldásban τ_1 és τ_2 független m-től, ezért ez akkor is jó megoldást ad, ha a feladat feltételében m ismert.

4.9. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta ($n \geqslant \geqslant 2$). Tegyük fel, hogy m és σ ismeretlenek. Adjon m-re centrált konfidenciaintervallumot, melynek $1-\alpha$ a pontos biztonsági szintje.

A megoldáshoz szükségünk lesz a következő tételre.

4.10. Tétel. *Ha $\xi \in \mathbf{Norm}(m;\sigma)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta ($n \geqslant 2$), akkor*

$$\frac{\overline{\xi}-m}{S_n^*}\sqrt{n} \in \mathbf{t}(n-1).$$

Bizonyítás. Korábban láttuk, hogy

$$\frac{\overline{\xi}-m}{\sigma}\sqrt{n} \in \mathbf{Norm}(0;1) \quad \text{és} \quad \frac{S_n^2}{\sigma^2} n \in \mathbf{Khi}(n-1),$$

továbbá ezek függetlenek. Így

$$\frac{\sqrt{n-1}\frac{\overline{\xi}-m}{\sigma}\sqrt{n}}{\sqrt{\frac{S_n^2}{\sigma^2}n}} = \frac{\overline{\xi}-m}{S_n}\sqrt{n-1} = \frac{\overline{\xi}-m}{S_n^*}\sqrt{n} \in \mathbf{t}(n-1).$$

Rátérünk a feladat megoldására.

Megoldás. Legyen $c \in \mathbb{R}_+$ és $F \sim \mathbf{t}(n-1)$. Ekkor az előző tétel szerint

$$\mathrm{P}_{(m,\sigma)}\left(-c \leqslant \frac{\overline{\xi}-m}{S_n^*}\sqrt{n} \leqslant c\right) = F(c) - F(-c) = 2F(c) - 1.$$

Mivel $2F(c)-1 = 1-\alpha$ pontosan akkor teljesül, ha $c = F^{-1}(1-\frac{\alpha}{2})$, ezért ilyen c-re átrendezéssel azt kapjuk, hogy

$$\mathrm{P}_{(m,\sigma)}\left(\overline{\xi} - \frac{S_n^*}{\sqrt{n}}c \leqslant m \leqslant \overline{\xi} + \frac{S_n^*}{\sqrt{n}}c\right) = 1-\alpha.$$

Könnyű látni, hogy ez centrált konfidenciaintervallum, hiszen

$$\begin{aligned}&\mathrm{P}_{(m,\sigma)}\left(m > \overline{\xi} + \frac{S_n^*}{\sqrt{n}}c\right) = \mathrm{P}_{(m,\sigma)}\left(\frac{\overline{\xi}-m}{S_n^*}\sqrt{n} < -c\right) = \\ &= F(-c) = 1 - F(c) = 1 - \left(1-\frac{\alpha}{2}\right) = \frac{\alpha}{2}.\end{aligned}$$

Összefoglalva tehát a megoldás:

$$\begin{aligned}&F \sim \mathbf{t}(n-1)\\ &\tau_1 := \overline{\xi} - \frac{S_n^*}{\sqrt{n}}F^{-1}\left(1-\frac{\alpha}{2}\right)\\ &\tau_2 := \overline{\xi} + \frac{S_n^*}{\sqrt{n}}F^{-1}\left(1-\frac{\alpha}{2}\right)\end{aligned}$$

jelölésekkel $[\tau_1, \tau_2]$ olyan centrált konfidenciaintervallum m-re, melynek $1-\alpha$ a pontos biztonsági szintje.

4.11. *Megjegyzés.* Az előző megoldásban τ_1 és τ_2 független σ-tól, ezért ez akkor is jó megoldást ad, ha a feladat feltételében σ ismert.

4.3. Konfidenciaintervallum az exponenciális eloszlás paraméterére

4.12. Feladat. Legyen $\xi \in \mathbf{Exp}(\lambda)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta. Tegyük fel, hogy λ ismeretlen. Adjon λ-ra centrált konfidenciaintervallumot, melynek $1-\alpha$ a pontos biztonsági szintje.

Megoldás. Mivel $x > 0$ esetén

$$\mathrm{P}_\lambda\left(\lambda\xi < x\right) = \mathrm{P}_\lambda\left(\xi < \frac{x}{\lambda}\right) = 1 - e^{-\lambda\frac{x}{\lambda}} = 1 - e^{-x},$$

ezért $\lambda\xi \in \mathbf{Exp}(1)$, következésképpen

$$\lambda\xi_1 + \cdots + \lambda\xi_n = \lambda n\overline{\xi} \in \mathbf{Gamma}(n;1).$$

Így $c_1, c_2 \in \mathbb{R}_+$ és $F \sim \mathbf{Gamma}(n;1)$ esetén

$$\begin{aligned}\mathrm{P}_\lambda\left(\lambda n\overline{\xi} < c_1\right) &= F(c_1),\\ \mathrm{P}_\lambda\left(\lambda n\overline{\xi} > c_2\right) &= 1 - F(c_2).\end{aligned}$$

Mivel $F(c_1) = 1 - F(c_2) = \frac{\alpha}{2}$ pontosan akkor teljesül, ha $c_1 = F^{-1}(\frac{\alpha}{2})$ és $c_2 = F^{-1}(1 - \frac{\alpha}{2})$, ezért ebben az esetben

$$\mathrm{P}_\lambda\left(c_1 \leqslant \lambda n\overline{\xi} \leqslant c_2\right) = 1 - \alpha,$$

azaz átrendezve

$$\mathrm{P}_\lambda\left(\frac{c_1}{n\overline{\xi}} \leqslant \lambda \leqslant \frac{c_2}{n\overline{\xi}}\right) = 1 - \alpha.$$

Vegyük észre, hogy $\frac{\alpha}{2} < 1 - \frac{\alpha}{2}$ miatt $c_1 < c_2$. Összefoglalva tehát a megoldás:

$$\begin{aligned}&F \sim \mathbf{Gamma}(n;1)\\ &\tau_1 := \frac{1}{n\overline{\xi}}F^{-1}\left(\frac{\alpha}{2}\right)\\ &\tau_2 := \frac{1}{n\overline{\xi}}F^{-1}\left(1 - \frac{\alpha}{2}\right)\end{aligned}$$

jelölésekkel $[\tau_1, \tau_2]$ olyan centrált konfidenciaintervallum σ-ra, melynek $1-\alpha$ a pontos biztonsági szintje.

4.4. Konfidenciaintervallum valószínűségre

4.13. Feladat. Legyen $\xi \in \mathbf{Bin}(1;p)$ és $\xi_1, \ldots, \xi_n$ egy ξ-re vonatkozó minta. Tegyük fel, hogy p ismeretlen. Adjon p-re centrált konfidenciaintervallumot, melynek $1 - \alpha$ a biztonsági szintje.

Vegyük észre, hogy ξ egy p valószínűségű esemény indikátorváltozója, így a feladat úgy is megfogalmazható, hogy egy esemény valószínűségére adjon konfidencia-

intervallumot. (Ekkor $\overline{\xi}$ az esemény relatív gyakoriságát jelenti n kísérlet után.)

Megoldás. Bizonyítható, hogy

$$\tau_1 := \frac{1}{n} \max \left\{ z \in \mathbb{N} : \sum_{i=0}^{z} \binom{n}{i} \overline{\xi}^i (1-\overline{\xi})^{n-i} < \frac{\alpha}{2} \right\}$$
$$\tau_2 := \frac{1}{n} \min \left\{ z \in \mathbb{N} : \sum_{i=0}^{z} \binom{n}{i} \overline{\xi}^i (1-\overline{\xi})^{n-i} \geqslant 1 - \frac{\alpha}{2} \right\}$$

jelölésekkel $[\tau_1, \tau_2]$ $1-\alpha$ biztonsági szintű konfidenciaintervallum p-re. Ennek bizonyítása azon múlik, hogy $n\overline{\xi} \in \mathbf{Bin}(n;p)$, de itt nem részletezzük (lásd Kendall, Stuart [7, 103–105. oldal]).

Az előző megoldás kiszámítása nagy n-re komplikált. Ennek kikerülésére ebben az esetben lehetőség van egy másik konfidenciaintervallum szerkesztésére is a Moivre–Laplace-tétel segítségével. Ugyanis $n\overline{\xi} \in \mathbf{Bin}(n;p)$ miatt $c \in \mathbb{R}_+$ esetén

$$\mathrm{P}_p\left(-c \leqslant \frac{n\overline{\xi} - np}{\sqrt{np(1-p)}} \leqslant c \right) \simeq \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

Mivel $2\Phi(c) - 1 = 1 - \alpha$ pontosan akkor teljesül, ha $c = \Phi^{-1}(1 - \frac{\alpha}{2})$, ezért ilyen c-re átrendezéssel azt kapjuk, hogy

$$\mathrm{P}_p\left(\left(1 + \frac{c^2}{n}\right) p^2 - \left(2\overline{\xi} + \frac{c^2}{n}\right) p + \overline{\xi}^2 \leqslant 0 \right) \simeq 1 - \alpha.$$

A p-ben másodfokú

$$\left(1 + \frac{c^2}{n}\right) p^2 - \left(2\overline{\xi} + \frac{c^2}{n}\right) p + \overline{\xi}^2$$

polinom gyökei

$$\frac{\overline{\xi} + \frac{c^2}{2n} \pm \frac{c}{\sqrt{n}} \sqrt{\overline{\xi}(1-\overline{\xi}) + \frac{c^2}{4n}}}{1 + \frac{c^2}{n}},$$

így

$$c := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$
$$\tau_1 := \frac{\overline{\xi} + \frac{c^2}{2n} - \frac{c}{\sqrt{n}} \sqrt{\overline{\xi}(1-\overline{\xi}) + \frac{c^2}{4n}}}{1 + \frac{c^2}{n}}$$
$$\tau_2 := \frac{\overline{\xi} + \frac{c^2}{2n} + \frac{c}{\sqrt{n}} \sqrt{\overline{\xi}(1-\overline{\xi}) + \frac{c^2}{4n}}}{1 + \frac{c^2}{n}}$$

jelölésekkel $[\tau_1, \tau_2]$ $1-\alpha$ biztonsági szintű konfidenciaintervallum p-re. Ha n olyan nagy, hogy $\frac{1}{n}$ elhanyagolhatóan kicsi $\frac{1}{\sqrt{n}}$-hez képest, akkor a megoldás tovább egyszerűsíthető:

$$\tau_1 = \overline{\xi} - \frac{c}{\sqrt{n}}\sqrt{\overline{\xi}(1-\overline{\xi})}$$
$$\tau_2 = \overline{\xi} + \frac{c}{\sqrt{n}}\sqrt{\overline{\xi}(1-\overline{\xi})}.$$

4.5. Általános módszer konfidenciaintervallum készítésére

Legyen ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{ \mathrm{P}_\vartheta : \vartheta \in \Theta \}$ statisztikai mezőn, ahol $\Theta \subset \mathbb{R}$ nyílt halmaz, és a ξ valószínűségi változó F_ϑ eloszlásfüggvénye folytonos minden $\vartheta \in \Theta$ esetén. Mivel $x > 0$ esetén

$$\mathrm{P}_\vartheta\left(-\ln F_\vartheta(\xi) < x\right) = \mathrm{P}_\vartheta\left(\xi > F_\vartheta^{-1}(e^{-x})\right) = 1 - F_\vartheta(F_\vartheta^{-1}(e^{-x})) = 1 - e^{-x},$$

ezért $-\ln F_\vartheta(\xi) \in \mathbf{Exp}(1)$, következésképpen

$$-\sum_{i=1}^{n} \ln F_\vartheta(\xi_i) \in \mathbf{Gamma}(n;1).$$

Így $c_1, c_2 \in \mathbb{R}_+$ és $F \sim \mathbf{Gamma}(n;1)$ esetén

$$\mathrm{P}_\vartheta\left(-\sum_{i=1}^{n} \ln F_\vartheta(\xi_i) < c_1\right) = F(c_1),$$
$$\mathrm{P}_\vartheta\left(-\sum_{i=1}^{n} \ln F_\vartheta(\xi_i) > c_2\right) = 1 - F(c_2).$$

Mivel $F(c_1) = 1 - F(c_2) = \frac{\alpha}{2}$ pontosan akkor teljesül, ha $c_1 = F^{-1}(\frac{\alpha}{2})$ és $c_2 = = F^{-1}(1-\frac{\alpha}{2})$, ezért ebben az esetben

$$\mathrm{P}_\vartheta\left(c_1 \leqslant -\sum_{i=1}^{n} \ln F_\vartheta(\xi_i) \leqslant c_2\right) = 1 - \alpha.$$

Innen a konfidenciaintervallum szerencsés esetben már megadható. Tulajdonképpen ezt alkalmaztuk az exponenciális eloszlás paraméterének intervallumbecslésénél.

4.14. Feladat. Legyen ξ az $[a, b]$ intervallumon egyenletes eloszlású, ahol a ismert, b ismeretlen, és $\xi_1, \ldots, \xi_n$ a ξ-re vonatkozó minta. Adjon b-re centrált konfidenciain-

tervallumot, melynek $1-\alpha$ a biztonsági szintje.

Megoldás. Mivel $F_b(\xi_i) = \frac{\xi_i - a}{b-a}$, így az előzőek miatt a

$$c_1 \leqslant -\sum_{i=1}^{n} \ln \frac{\xi_i - a}{b-a} \leqslant c_2$$

egyenlőtlenséget kell b-re rendezni. Azt kapjuk, hogy

$$a + \left(e^{c_1} \prod_{i=1}^{n} (\xi_i - a) \right)^{\frac{1}{n}} \leqslant b \leqslant a + \left(e^{c_2} \prod_{i=1}^{n} (\xi_i - a) \right)^{\frac{1}{n}},$$

így a feladat megoldása:

$$\begin{aligned}
&F \sim \mathbf{Gamma}(n;1) \\
&c_1 := F^{-1}\left(\frac{\alpha}{2}\right) \\
&c_2 := F^{-1}\left(1-\frac{\alpha}{2}\right) \\
&\tau_1 := a + \left(e^{c_1} \prod_{i=1}^{n} (\xi_i - a) \right)^{\frac{1}{n}} \\
&\tau_2 := a + \left(e^{c_2} \prod_{i=1}^{n} (\xi_i - a) \right)^{\frac{1}{n}}
\end{aligned}$$

jelölésekkel $[\tau_1, \tau_2]$ olyan centrált konfidenciaintervallum b-re, melynek $1-\alpha$ a pontos biztonsági szintje.

5. Hipotézisvizsgálatok

5.1. A hipotézisvizsgálat feladata és jellemzői

Ebben a fejezetben azt vizsgáljuk, hogyan lehet dönteni a mintarealizáció alapján arról, hogy egy a statisztikai mezőre vonatkozó feltételezést, más szóval *hipotézist* elfogadjuk-e igaznak vagy sem. Ez a hipotézis lehet például az, hogy a vizsgált valószínűségi változó normális eloszlású, vagy a valószínűségi változó várható értéke megfelel az előírásnak, vagy két valószínűségi változó független, vagy várható értékeik megegyeznek stb.

5.1.1. Null- illetve ellenhipotézis

Azt a feltételezést, amelyről döntést akarunk hozni, *nullhipotézisnek* nevezzük és H_0-val jelöljük. Legyen $\mathcal{P}_{H_0}$ azon valószínűségek halmaza, melyek a H_0 teljesülése esetén lehetségesek. Feltételezzük, hogy ez nem üres halmaz.

Ha H_0-t elutasítjuk, akkor egy azzal ellentétes állítást fogadunk el, melyet *ellenhipotézisnek* nevezünk, és H_1-gyel jelölünk. Általában H_0 és H_1 közül az egyik mindig bekövetkezik, de ez nem mindig van így (lásd például az úgynevezett egyoldali ellenhipotéziseket). Ennek okát később taglaljuk. Legyen $\mathcal{P}_{H_1}$ azon valószínűségek halmaza, melyek a H_1 teljesülése esetén lehetségesek. Feltételezzük, hogy ez nem üres halmaz.

5.1.2. Statisztikai próba terjedelme és torzítatlansága

Tegyük fel, hogy a $\xi^{(1)}, \dots, \xi^{(k)}$ valószínűségi vektorváltozókra vonatkozik H_0, melyek rendre $d_1, \dots, d_k$ dimenziósak. $\xi^{(i)}$-re vonatkozzon a $\xi_1^{(i)}, \dots, \xi_{n_i}^{(i)}$ minta $(i = 1, \dots, k)$. Legyen

$$C_0 \subset (\mathbb{R}^{d_1})^{n_1} \times \cdots \times (\mathbb{R}^{d_k})^{n_k}.$$

Ha a kísérletben az $\omega \in \Omega$ elemi esemény következett be, és

$$\left(\xi_1^{(1)}(\omega), \dots, \xi_{n_1}^{(1)}(\omega), \dots, \xi_1^{(k)}(\omega), \dots, \xi_{n_k}^{(k)}(\omega)\right) \in C_0,$$

akkor H_0-t elfogadjuk, ellenkező esetben pedig elutasítjuk. Ezt az eljárást *statisztikai próbának* vagy *hipotézisvizsgálatnak* nevezzük. C_0 az úgynevezett *elfogadási tartomány*. C_0 komplementerét C_1-gyel jelöljük, és *kritikus tartománynak* nevezzük.

Döntésünk lehet helyes, vagy helytelen az alábbiak szerint:

	H_0-t elfogadjuk	H_0-t elutasítjuk
H_0 igaz	helyes döntés	*elsőfajú hiba*
H_1 igaz	*másodfajú hiba*	helyes döntés

Legyen $0 < \alpha < \frac{1}{2}$. Az α számot a *próba terjedelmének* nevezzük, ha

$$\mathrm{P}\big((\xi_1^{(1)}, \dots, \xi_{n_1}^{(1)}, \dots, \xi_1^{(k)}, \dots, \xi_{n_k}^{(k)}) \in C_1\big) \leqslant \alpha \quad \forall\, \mathrm{P} \in \mathcal{P}_{H_0}$$

teljesül, azaz az elsőfajú hiba valószínűsége legfeljebb α. Ekkor az $1 - \alpha$ számot a *próba szintjének* nevezzük. Ez azt az értéket jelenti, amelynél nagyobb vagy egyenlő valószínűséggel elfogadjuk H_0-t, ha az igaz. A *próba pontos terjedelme* α, ha

$$\sup_{\mathrm{P} \in \mathcal{P}_{H_0}} \mathrm{P}\big((\xi_1^{(1)}, \dots, \xi_{n_1}^{(1)}, \dots, \xi_1^{(k)}, \dots, \xi_{n_k}^{(k)}) \in C_1\big) = \alpha.$$

Ha a vizsgált valószínűségi (vektor)változók diszkrétek, akkor adott α-hoz nem biztosan található olyan elfogadási tartomány, mellyel a próba pontos terjedelme α. Ezért definiáltuk a próba terjedelmét az előző módon.

Ha egy α terjedelmű próba esetén

$$\mathrm{P}\big((\xi_1^{(1)}, \dots, \xi_{n_1}^{(1)}, \dots, \xi_1^{(k)}, \dots, \xi_{n_k}^{(k)}) \in C_1\big) \geqslant \alpha \quad \forall\, \mathrm{P} \in \mathcal{P}_{H_1}$$

teljesül, akkor a próbát *torzítatlannak* nevezzük. Ez azt jelenti, hogy H_0-t nagyobb valószínűséggel utasítjuk el, ha H_1 igaz, mint amikor H_0 igaz.

5.1.3. Próbastatisztika

Elfogadási tartomány konstruálásához H_0 esetén ismert eloszlású

$$\tau := T(\xi_1^{(1)}, \dots, \xi_{n_1}^{(1)}, \dots, \xi_1^{(k)}, \dots, \xi_{n_k}^{(k)})$$

statisztikára lesz szükségünk, mely lényegesen másképp viselkedik H_0 illetve H_1 teljesülése esetén. Az ilyen statisztikát *próbastatisztikának* nevezzük. Ekkor rögzített α esetén meg tudunk adni egy olyan $I_\tau \subset \mathbb{R}$ intervallumot, melyre

$$\mathrm{P}(\tau \in I_\tau) \geqslant 1 - \alpha \quad \forall\, \mathrm{P} \in \mathcal{P}_{H_0}.$$

Célszerűbb a $\mathrm{P}(\tau \in I_\tau) = 1 - \alpha\ \forall\, \mathrm{P} \in \mathcal{P}_{H_0}$ feltétel, mert ekkor α a pontos terjedelem lesz, de ez nem mindig teljesíthető. Az I_τ végpontjait *kritikus értékeknek* nevezzük.

Ezután legyen

$$C_0 := \left\{ x \in (\mathbb{R}^{d_1})^{n_1} \times \cdots \times (\mathbb{R}^{d_k})^{n_k} : T(x) \in I_\tau \right\}.$$

Mivel a $(\xi_1^{(1)}, \ldots, \xi_{n_1}^{(1)}, \ldots, \xi_1^{(k)}, \ldots, \xi_{n_k}^{(k)}) \in C_0$ esemény pontosan akkor következik be, amikor $\tau \in I_\tau$, ezért ekkor α terjedelmű próbát kapunk.

A gyakorlatban sokkal egyszerűbb a $\tau \in I_\tau$ esemény megadása, mint a C_0 felírása, ezért az előbbit választjuk. Szokás a $\tau \in I_\tau$ eseményt, illetve az I_τ halmazt is elfogadási tartománynak nevezni. Hasonlóan, a $\tau \notin I_\tau$ eseményt, illetve az $\mathbb{R} \setminus I_\tau$ halmazt is szokás kritikus tartománynak nevezni.

5.1.4. A statisztikai próba menete

Amikor a rögzített α próbaterjedelemhez és a választott τ próbastatisztikához megválasztjuk az I_τ intervallumot, akkor ügyelni kell arra, hogy a másodfajú hiba valószínűsége – azaz annak a valószínűsége, hogy H_1 teljesülése esetén H_0-t elfogadjuk – kicsi legyen. Ehhez I_τ megadásánál nem csak H_0-t, hanem H_1-t is figyelembe kell venni. A gyakorlatban a menetrend a következő:

1) H_0 ismeretében kiválasztjuk a τ próbastatisztikát.

2) H_1 és τ ismeretében kiválasztjuk $\mathbb{R} \setminus I_\tau$ jellegét: $(-\infty, a)$, (b, ∞), (c, d) stb. Ez fontos pont, mert ha itt rosszul választunk, akkor a másodfajú hiba valószínűsége túl nagy lesz.

3) A τ próbastatisztika H_0 esetén teljesülő eloszlásának, I_τ jellegének és α-nak az ismeretében meghatározzuk a kritikus értékeket.

4) A próbastatisztika, a mintarealizáció és I_τ ismeretében döntést hozunk. Ha a próbastatisztika realizációja I_τ-ba esik, akkor *H_0-t elfogadjuk H_1 ellenében α terjedelemmel*. Ha a próbastatisztika realizációja nem esik I_τ-ba, akkor *H_0-t elutasítjuk H_1 ellenében α terjedelemmel*, vagyis ilyenkor H_1-gyet fogadjuk el.

5.1.5. A nullhipotézis és az ellenhipotézis megválasztása

A gyakorlatban nem minden esetben érdemes a sejtésünket, vagy az elvárásunkat megválasztani nullhipotézisnek, mert nem találnánk hozzá próbastatisztikát. Ilyenkor ezt ellenhipotézisként kezeljük, és egy olyan ezzel ellentétes állítást fogadunk el nullhipotézisnek, amelyhez már találunk megfelelő próbastatisztikát. Mindez érthetőbbé válik a következő példán:

A tejiparban hasznos lehetne egy olyan eljárás, melynek révén nagyobb arányban születne üszőborjú, mint bikaborjú, hiszen ekkor több fejőstehenet nevelhetnének fel azonos születésszám mellett. Egy kutató javasol egy ilyen eljárást. Hogyan lehetne ellenőrizni az állítását? Jelölje p annak a valószínűségét, hogy az eljárás alkalmazásával üszőborjú születik. Ekkor a kutató állítása az, hogy $p > \frac{1}{2}$. Ezt viszont nem célszerű H_0-nak választani, ugyanis ekkor nem találunk próbastatisztikát. Ehelyett legyen ez az ellenhipotézis, míg $p = \frac{1}{2}$ a nullhipotézis. Ebben az esetben már könnyű próbastatisztikát megadni. Ugyanis ha ξ jelenti az eljárás révén üszőborjú születését az indikátorváltozóját, és a ξ-re vonatkozó minta $\xi_1, \dots, \xi_n$, akkor $n\overline{\xi}$ azt jelenti, hogy n-szer alkalmazva az eljárást hány darab üszőborjú született. Az $n\overline{\xi}$ meg is felel próbastatisztikának, hiszen H_0 esetén n-edrendű $\frac{1}{2}$ paraméterű binomiális eloszlású.

Ebből a példából láthatóan nem feltétlenül kell teljesülnie, hogy H_0 és H_1 közül az egyik mindig bekövetkezik. Nézzünk erre egy másik példát is:

Egy kereskedő egy malomtól nagy tételben lisztet rendel 1 kg-os kiszerelésben. Jelentse ξ a leszállított tételből egy véletlenszerűen kiválasztott zacskó liszt tömegének eltérését az elvárt 1 kg-tól. Ekkor az a nullhipotézis, hogy $\mathrm{E}\,\xi = 0$. Ha ξ jó közelítéssel normális eloszlásúnak tekinthető, akkor a későbbiekben tárgyalt úgynevezett egymintás t-próbánál látni fogjuk, hogy ehhez találhatunk próbastatisztikát. Most az a kérdés, hogy mi legyen az ellenhipotézis. Ha $\mathrm{E}\,\xi \neq 0$ lenne, akkor H_0 elutasítása esetén csak az derülne ki, hogy a zacskók tömege nem felel meg a rendelésnek. Ez azonban nem biztosan jelent rosszat a kereskedőnek. Hiszen, ha valójában $\mathrm{E}\,\xi > 0$ teljesül, akkor a kereskedőtől vásárlók csak ritkán reklamálnának. Ezért célszerűbb $\mathrm{E}\,\xi < 0$ megválasztása H_1-nek. Ekkor ugyanis H_0 elutasítása esetén érdemes megfontolnia a kereskedőnek a leszállított tétel visszautasítását. Vagyis most a kereskedő számára rossz esetet tekintjük ellenhipotézisnek, azt remélvén, hogy a módszer nagy valószínűséggel megvédi őt az előnytelen vételtől. Ehhez persze az kell, hogy a másodfajú hiba valószínűsége kicsi legyen.

5.1.6. A próba erőfüggvénye és konzisztenciája

Ha ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{\, \mathrm{P}_\vartheta : \vartheta \in \Theta \,\}$ statisztikai mezőn, ahol $\Theta \subset \mathbb{R}$, és a ξ-re vonatkozó minta $\xi_1, \dots, \xi_n$, továbbá ha $\vartheta_0 \in \Theta$ rögzített és C_1 kritikus tartomány mellett döntünk a $H_0\colon \vartheta = \vartheta_0$ nullhipotézisről, akkor a

$$\gamma\colon \Theta \to \mathbb{R}, \quad \gamma(\vartheta) := \mathrm{P}_\vartheta\big((\xi_1, \dots, \xi_n) \in C_1\big)$$

függvényt a *próba erőfüggvényének* nevezzük. Ha $H_1\colon \vartheta \in \Theta_1 (\subset \Theta \setminus \{\,\vartheta_0\,\})$ és

$$\lim_{n\to\infty} \mathrm{P}_\vartheta\big((\xi_1,\dots,\xi_n) \in C_1\big) = 1 \quad \forall \vartheta \in \Theta_1,$$

akkor azt mondjuk, hogy a próba *konzisztens*.

Az erőfüggvény a másodfajú hiba vizsgálatában hasznos. Ez az úgynevezett egymintás u-próba kapcsán válik majd világossá. A konzisztencia tulajdonképpen azt jelenti, hogy a másodfajú hiba valószínűsége a mintaelemek számának növelésével 0-hoz tart.

5.2. Paraméteres hipotézisvizsgálatok

Ha a nullhipotézis ismert eloszláscsaládból származó valószínűségi változók eloszlásainak paramétereire vonatkozik, akkor *paraméteres hipotézisvizsgálatról* beszélünk.

5.2.1. Egymintás u-próba

5.1. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$, ahol m ismeretlen és σ ismert, továbbá legyen $\xi_1,\dots,\xi_n$ a ξ-re vonatkozó minta. A

$$H_0\colon m = m_0$$
$$H_1\colon m \neq m_0 \ (\textit{kétoldali ellenhipotézis})$$

hipotézisekre adjon adott α terjedelmű próbát, ahol $m_0 \in \mathbb{R}$ rögzített.

Megoldás. Először próbastatisztikát adunk. Korábban már bizonyítottuk, hogy H_0 teljesülése esetén

$$u := \frac{\overline{\xi} - m_0}{\sigma}\sqrt{n} \in \mathbf{Norm}(0;1).$$

A kritikus tartomány megadásánál vegyük figyelembe, hogy $\overline{\xi}$ az m torzítatlan becslése, így H_1 teljesülése esetén $\overline{\xi} - m_0$ várhatóan kritikus értékben eltávolodik 0-tól. Következésképpen a standard normális eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|u| \leqslant a$ $(a > 0)$ alakú. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(|u| \leqslant a) = 2\Phi(a) - 1,$$

így $\mathrm{P}(|u| \leqslant a) = 1 - \alpha$ esetén $a = \Phi^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát

$$|u| \leqslant \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

elfogadási tartománnyal olyan próbát kapunk, melynek a pontos terjedelme α. Ezt a statisztikai próbát nevezzük *egymintás u-próbának.*

5.2. Feladat. Az előző feladatot oldja meg $H_1\colon m < m_0$ illetve $H_1\colon m > m_0$ úgynevezett *egyoldali ellenhipotézisekre* is.

Megoldás. Itt is az előbbi u próbastatisztikát fogjuk használni. Először legyen az ellenhipotézis $H_1\colon m < m_0$. Ennek teljesülése esetén $\overline{\xi} - m_0$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $u \geqslant b$ $(b < 0)$ jellegű. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(u \geqslant b) = 1 - \Phi(b),$$

így $\mathrm{P}(u \geqslant b) = 1 - \alpha$ esetén $b = \Phi^{-1}(\alpha) < 0$. Tehát

$$u \geqslant \Phi^{-1}(\alpha)$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Ezután legyen az ellenhipotézis $H_1\colon m > m_0$. Ennek teljesülése esetén $\overline{\xi} - m_0$ várhatóan kritikus értékben 0 fölött van. Így az elfogadási tartomány $u \leqslant c$ $(c > 0)$ jellegű. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(u \leqslant c) = \Phi(c),$$

így $\mathrm{P}(u \leqslant c) = 1 - \alpha$ esetén $c = \Phi^{-1}(1 - \alpha) > 0$. Tehát

$$u \leqslant \Phi^{-1}(1 - \alpha)$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

5.3. Feladat. Vizsgálja meg az egymintás u-próbában a másodfajú hiba valószínűségét. Bizonyítsa be, hogy a próba torzítatlan és konzisztens.

Megoldás. Először számoljuk ki az u várható értékét és szórását:

$$\mathrm{E}_m\, u = \frac{\mathrm{E}_m\, \overline{\xi} - m_0}{\sigma}\sqrt{n} = \frac{m - m_0}{\sigma}\sqrt{n}$$
$$\mathrm{D}_m\, u = \frac{\sqrt{n}}{\sigma}\,\mathrm{D}_m\, \overline{\xi} = \frac{\sqrt{n}}{\sigma}\sqrt{\frac{1}{n^2}n\sigma^2} = 1.$$

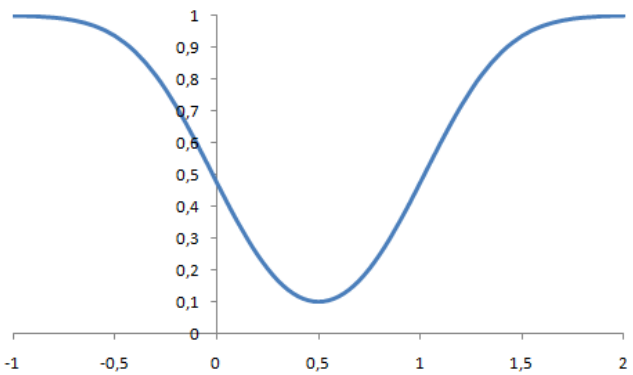
Mivel u az m bármely értéke esetén normális eloszlású, ezért azt kapjuk, hogy

$$u \in \mathbf{Norm}\left(\frac{m - m_0}{\sigma}\sqrt{n}; 1\right).$$

Most tekintsük a kétoldali ellenhipotézis esetét. Ekkor $u_{\alpha/2} := \Phi^{-1}(1-\frac{\alpha}{2})$ jelöléssel az erőfüggvény

$$\begin{aligned}\gamma(m) &= \mathrm{P}_m\big((\xi_1,\dots,\xi_n)\in C_1\big) = \\ &= \mathrm{P}_m(|u|>u_{\alpha/2}) = 1-\mathrm{P}_m(-u_{\alpha/2}\leqslant u\leqslant u_{\alpha/2}) = \\ &= 1-\Phi\left(u_{\alpha/2}-\frac{m-m_0}{\sigma}\sqrt{n}\right)+\Phi\left(-u_{\alpha/2}-\frac{m-m_0}{\sigma}\sqrt{n}\right).\end{aligned}$$

Deriváljuk γ-t, melyből azt kapjuk, hogy γ szigorúan monoton csökken a $(-\infty,m_0]$ intervallumon, illetve szigorúan monoton nő az $[m_0,\infty)$ intervallumon, továbbá minimum helye van m_0-ban, és a minimum értéke α. Az is könnyen látható, hogy $\lim\limits_{m\to\infty}\gamma(m) = \lim\limits_{m\to-\infty}\gamma(m) = 1$. A következő ábrán γ grafikonját láthatjuk $\sigma = 1$, $m_0 = 0{,}5$, $\alpha = 0{,}1$, $n = 10$ paraméterekkel.



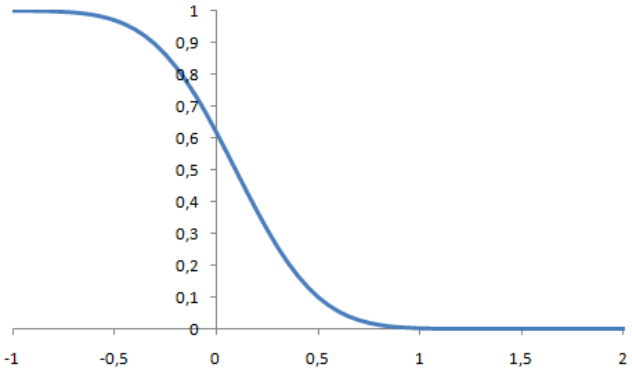
Mindezek alapján tehát, ha $H_1\colon m\neq m_0$ teljesül, akkor $\gamma(m)>\alpha$, melyből következik, hogy a próba torzítatlan. Ha γ-t mint n függvényét tekintjük, akkor könnyen láthatjuk, hogy minden $m\neq m_0$ esetén $\lim\limits_{n\to\infty}\gamma(m)=1$, melyből már következik, hogy a próba konzisztens, azaz a mintaelemek számának növelésével a másodfajú hiba valószínűsége 0-hoz tart.

Érdekes még azt is megvizsgálni, hogy miként változik a másodfajú hiba valószínűsége, ha az első fajú hiba valószínűségét, azaz α-t csökkentjük. Ha α csökken, akkor $u_{\alpha/2}=\Phi^{-1}(1-\frac{\alpha}{2})$ nő, hiszen Φ^{-1} növekvő függvény. Másrészt, ha γ-t mint $u_{\alpha/2}$ függvényét tekintjük, akkor könnyen ellenőrizhető, hogy $\frac{\mathrm{d}\gamma}{\mathrm{d}u_{\alpha/2}}<0$, azaz γ csökkenő. Mindezekből tehát kapjuk, hogy α csökkentésével γ is csökken, azaz a másodfajú hiba valószínűsége nő.

Ezután tekintsük a $H_1\colon m<m_0$ egyoldali ellenhipotézist. Ekkor az erőfüggvény $u_\alpha=\Phi^{-1}(\alpha)$ jelöléssel

$$\gamma(m)=\mathrm{P}_m\big((\xi_1,\dots,\xi_n)\in C_1\big)=\mathrm{P}_m(u<u_\alpha)=\Phi\left(u_\alpha-\frac{m-m_0}{\sigma}\sqrt{n}\right).$$

Φ szigorúan monoton növekvő, ezért γ szigorúan monoton csökkenő. Az is könnyen látható, hogy $\gamma(m_0) = \alpha$, $\lim\limits_{m\to\infty} \gamma(m) = 0$ és $\lim\limits_{m\to-\infty} \gamma(m) = 1$. A következő ábrán γ grafikonját láthatjuk $\sigma = 1,\ m_0 = 0{,}5,\ \alpha = 0{,}1,\ n = 10$ paraméterekkel.



Mindezek alapján, ha $H_1\colon m < m_0$ teljesül, akkor $\gamma(m) > \alpha$, melyből következik, hogy a próba torzítatlan. Ha γ-t mint n függvényét tekintjük, akkor minden $m < m_0$ esetén $\lim\limits_{n\to\infty} \gamma(m) = 1$, melyből már következik, hogy a próba konzisztens.

Ha α csökken, akkor $u_\alpha = \Phi^{-1}(\alpha)$ is csökken, másrészt ekkor Φ növekedése miatt γ csökken. Mindezekből tehát kapjuk, hogy α csökkentésével γ is csökken, azaz a másodfajú hiba valószínűsége nő.

A $H_1\colon m > m_0$ eset tárgyalását az Olvasóra bízzuk.

5.4. *Megjegyzés.* Érdemes még megfontolni a következőket. Tegyük fel, hogy az $0 < < \alpha < \frac{1}{2}$ terjedelmű egymintás u-próbában $H_0\colon m = m_0$-t elutasítjuk a kétoldali ellenhipotézissel szemben, azaz bekövetkezett az $|u| > \Phi^{-1}(1-\frac{\alpha}{2})$ esemény. Ha most még azt is feltesszük, hogy $u > 0$ is bekövetkezett (azaz $\overline{\xi} > m_0$), akkor

$$u > \Phi^{-1}(1-\tfrac{\alpha}{2}) > \Phi^{-1}(1-\alpha) > \Phi^{-1}(\alpha)$$

miatt $H_1\colon m > m_0$ esetén biztosan H_1-gyet, míg $H_1\colon m < m_0$ esetén biztosan H_0-t fogadjuk el.

Viszont, ha a kétoldali ellenhipotézis elfogadása esetén $u < 0$ következett be (azaz $\overline{\xi} < m_0$), akkor

$$u < -\Phi^{-1}(1-\tfrac{\alpha}{2}) = \Phi^{-1}(\tfrac{\alpha}{2}) < \Phi^{-1}(\alpha) < \Phi^{-1}(1-\alpha)$$

miatt $H_1\colon m < m_0$ esetén biztosan H_1-gyet, míg $H_1\colon m > m_0$ esetén biztosan H_0-t fogadjuk el.

Hasonlóan látható be, hogy ha H_0-t elfogadjuk a kétoldali ellenhipotézissel szemben, akkor az egyoldali ellenhipotézisekkel szemben is elfogadjuk.

Így tehát, ha elvégeztük az egymintás u-próbát kétoldali ellenhipotézisre, akkor már fölösleges egyoldalira is megcsinálni, hiszen azok eredménye ebből már megadható a következő táblázat alapján:

$H_0\colon m = m_0$	$H_1\colon m \neq m_0$		
	H_0-t elfogadjuk	H_0-t elutasítjuk	
		$u > 0$	$u < 0$
$H_1\colon m < m_0$	H_0-t elfogadjuk	H_0-t elutasítjuk	H_0-t elfogadjuk
$H_1\colon m > m_0$	H_0-t elfogadjuk	H_0-t elfogadjuk	H_0-t elutasítjuk

A táblázat úgy is értelmezhető, hogy $|u| \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ esetén $m = m_0$, $u < -\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ esetén $m < m_0$, illetve $u > \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ esetén $m > m_0$ mellett döntünk α terjedelemmel.

5.5. *Megjegyzés.* A kritikus értékek kiszámolásánál az eddigiek alapján szükség van a Φ^{-1} ismeretére. Valójában azonban elég csak a Φ használata. Ugyanis $H_1\colon m \neq m_0$ ellenhipotézisre vonatkozó döntés esetén az $|u| \leqslant \Phi^{-1}(1-\frac{\alpha}{2})$ elfogadási tartomány ekvivalens azzal, hogy

$$\alpha \leqslant 2 - 2\Phi(|u|).$$

Hasonlóan, $H_1\colon m < m_0$ illetve $H_1\colon m > m_0$ ellenhipotézisre vonatkozó döntés esetén az elfogadási tartomány ekvivalens azzal, hogy

$$\alpha \leqslant \Phi(u) \quad \text{illetve} \quad \alpha \leqslant 1 - \Phi(u).$$

5.2.2. Kétmintás u-próba

5.6. Feladat. Legyen $\xi \in \mathbf{Norm}(m_1;\sigma_1)$, $\eta \in \mathbf{Norm}(m_2;\sigma_2)$ független valószínűségi változók, ahol m_1, m_2 ismeretlenek és σ_1, σ_2 ismertek. Legyen $\xi_1, \ldots, \xi_{n_1}$ a ξ-re vonatkozó, illetve $\eta_1, \ldots, \eta_{n_2}$ az η-ra vonatkozó minta. A

$$H_0\colon m_1 = m_2$$
$$H_1\colon m_1 \neq m_2 \ (\textit{kétoldali ellenhipotézis})$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1\colon m_1 < m_2 \quad \text{illetve} \quad H_1\colon m_1 > m_2$$

egyoldali ellenhipotézisekre is.

Megoldás. Ha H_0 igaz, akkor könnyen látható, hogy

$$u := \frac{\overline{\xi} - \overline{\eta}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in \mathbf{Norm}(0;1).$$

Először vizsgáljuk a kétoldali ellenhipotézist. Ha ez teljesül, akkor $\overline{\xi} - \overline{\eta}$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a standard normális eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|u| \leqslant a$ $(a > 0)$ alakú. Ebből az egymintás u-próbával megegyező módon bizonyítható, hogy

$$|u| \leqslant \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (\Longleftrightarrow \alpha \leqslant 2 - 2\Phi(|u|))$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Most legyen $H_1 \colon m_1 < m_2$. Ha ez teljesül, akkor $\overline{\xi} - \overline{\eta}$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $u \geqslant b$ $(b < 0)$ jellegű. Ebből az egymintás esettel megegyező módon bizonyítható, hogy $u \geqslant \Phi^{-1}(\alpha)$ $(\Longleftrightarrow \alpha \leqslant \Phi(u))$ elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Hasonlóan, $H_1 \colon m_1 > m_2$ esetén $u \leqslant \Phi^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1 - \Phi(u))$ elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Ezt a statisztikai próbát nevezzük *kétmintás u-próbának*.

Itt is érvényes, hogy ha elvégeztük a kétmintás u-próbát kétoldali ellenhipotézisre, akkor már fölösleges egyoldalira is megcsinálni, mert azok eredménye ebből már megadható a következő táblázat alapján:

$H_0 \colon m_1 = m_2$	$H_1 \colon m_1 \neq m_2$		
	H_0-t elfogadjuk	H_0-t elutasítjuk	
		$u > 0$	$u < 0$
$H_1 \colon m_1 < m_2$	H_0-t elfogadjuk	H_0-t elfogadjuk	H_0-t elutasítjuk
$H_1 \colon m_1 > m_2$	H_0-t elfogadjuk	H_0-t elutasítjuk	H_0-t elfogadjuk

A táblázat szerint $|u| \leqslant \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ esetén $m_1 = m_2$, $u < -\Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ esetén $m_1 < m_2$, illetve $u > \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ esetén $m_1 > m_2$ mellett döntünk α terjedelemmel.

5.7. Feladat. Bizonyítsuk be, hogy a kétmintás u-próba esetén
(1) a próba torzítatlan;
(2) a két minta elemszámainak növelésével a másodfajú hiba valószínűsége 0-hoz tart;

(3) az elsőfajú hiba valószínűségének csökkentésével a másodfajú hiba valószínűsége nő.

Megoldás. Csak kétoldali ellenhipotézisre bizonyítunk, az egyoldaliakat az Olvasóra bízzuk. Könnyen látható, hogy az m_1, m_2 bármely értékei esetén

$$u \in \mathbf{Norm}\left(\frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; 1\right),$$

így $u_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ jelöléssel

$$\begin{aligned}\gamma(m_1, m_2) &:= \mathrm{P}_{(m_1,m_2)}\big((\xi_1, \dots, \xi_{n_1}, \eta_1, \dots, \eta_{n_2}) \in C_1\big) = \\ &= \mathrm{P}_{(m_1,m_2)}(|u| > u_{\alpha/2}) = 1 - \mathrm{P}_{(m_1,m_2)}(-u_{\alpha/2} \leqslant u \leqslant u_{\alpha/2}) = \\ &= 1 - \Phi\left(u_{\alpha/2} - \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) + \Phi\left(-u_{\alpha/2} - \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right).\end{aligned}$$

Tekintsük ezt, mint m_1, m_2 szerinti kétváltozós függvényt. Ekkor a szokásos eljárással kapjuk, hogy pontosan H_0 esetén van minimuma a függvénynek és ott α az értéke. Ebből már adódik, hogy a próba torzítatlan.

Másrészt, ha γ-t, mint n_1, n_2 szerinti kétváltozós függvényt tekintjük, akkor $n_1 \to \infty,\ n_2 \to \infty$ esetén a határértéke 1. Ebből adódik a (2) állítás. Végül a (3) állítást hasonlóan kell belátni, mint az egymintás u-próbánál.

5.2.3. Egymintás t-próba

5.8. Feladat. Legyen $\xi \in \mathbf{Norm}(m; \sigma)$, ahol m és σ ismeretlenek, továbbá legyen $\xi_1, \dots, \xi_n$ a ξ-re vonatkozó minta ($n \geqslant 2$). A

$$H_0 \colon m = m_0$$
$$H_1 \colon m \neq m_0$$

hipotézisekre adjon adott α terjedelmű próbát, ahol $m_0 \in \mathbb{R}$ rögzített. A feladatot oldja meg $H_1 \colon m < m_0$ illetve $H_1 \colon m > m_0$ egyoldali ellenhipotézisekre is.

Megoldás. Korábban bizonyítottuk, hogy H_0 teljesülése esetén

$$t := \frac{\overline{\xi} - m_0}{S_n^*}\sqrt{n} \in \mathbf{t}(n-1).$$

A kétoldali ellenhipotézis teljesülése esetén $\overline{\xi} - m_0$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a t-eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|t| \leqslant a$ $(a > 0)$ alakú. A továbbiakban legyen $F \sim \mathbf{t}(n-1)$. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(|t| \leqslant a) = 2F(a) - 1,$$

így $\mathrm{P}(|t| \leqslant a) = 1 - \alpha$ esetén $a = F^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát

$$|t| \leqslant F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (\Longleftrightarrow \alpha \leqslant 2 - 2F(|t|))$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Most legyen $H_1 \colon m < m_0$. Ennek teljesülésekor $\overline{\xi} - m_0$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $t \geqslant b$ $(b < 0)$ jellegű. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(t \geqslant b) = 1 - F(b),$$

így $\mathrm{P}(t \geqslant b) = 1 - \alpha$ esetén $b = F^{-1}(\alpha) < 0$. Tehát $t \geqslant F^{-1}(\alpha)$ $(\Longleftrightarrow \alpha \leqslant F(t))$ elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Hasonlóan, $H_1 \colon m > m_0$ esetén $t \leqslant F^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1 - F(t))$ elfogadási tartománnyal olyan próbát kapunk, melynek α a pontos terjedelme.

Ezt a statisztikai próbát nevezzük *egymintás t-próbának.*

5.9. *Megjegyzés.* Az egymintás u-próbával vett analógia miatt itt is érvényes, hogy a kétoldali ellenhipotézis esetében meghozott döntés meghatározza az egyoldaliakkal szemben hozott döntéseket az ott található táblázat szerint. Szintén ezen analógia miatt erre a próbára is teljesül, hogy torzítatlan, konzisztens és az elsőfajú hiba valószínűségének csökkentésével a másodfajú hiba valószínűsége nő. Ennek bizonyításában az egymintás u-próbánál leírtakhoz képest csak annyit kell még felhasználni, hogy S_n^* konzisztens becsléssorozata σ-nak.

5.2.4. Kétmintás t-próba, Scheffé-módszer

5.10. Feladat. Legyenek $\xi \in \mathbf{Norm}(m_1; \sigma_1)$, $\eta \in \mathbf{Norm}(m_2; \sigma_2)$ független valószínűségi változók, ahol $m_1, m_2, \sigma_1, \sigma_2$ ismeretlenek és $\sigma_1 = \sigma_2$. Legyen $\xi_1, \ldots, \xi_{n_1}$ a ξ-re vonatkozó, illetve $\eta_1, \ldots, \eta_{n_2}$ az η-ra vonatkozó minta $(n_1 \geqslant 2,\ n_2 \geqslant 2)$. A

$$H_0 \colon m_1 = m_2$$
$$H_1 \colon m_1 \neq m_2$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1\colon m_1 < m_2 \quad \text{illetve} \quad H_1\colon m_1 > m_2$$

egyoldali ellenhipotézisekre is.

A feladat megoldásához szükségünk lesz a következő tételre.

5.11. Tétel. *Legyenek $\xi, \eta \in \mathbf{Norm}(m;\sigma)$ függetlenek, $\xi_1, \dots, \xi_{n_1}$ a ξ-re, illetve $\eta_1, \dots, \eta_{n_2}$ az η-ra vonatkozó minta ($n_1 \geqslant 2,\ n_2 \geqslant 2$). Ekkor*

$$\frac{\overline{\xi} - \overline{\eta}}{\sqrt{n_1 S^2_{\xi,n_1} + n_2 S^2_{\eta,n_2}}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \in \mathbf{t}(n_1 + n_2 - 2).$$

Bizonyítás. Korábban már bizonyítottuk, hogy $\overline{\xi}, \overline{\eta}, S_{\xi,n_1}, S_{\eta,n_2}$ függetlenek, továbbá

$$\begin{aligned} X &:= \frac{\overline{\xi} - \overline{\eta}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \in \mathbf{Norm}(0;1) \\ Y &:= \frac{S^2_{\xi,n_1}}{\sigma^2} n_1 \in \mathbf{Khi}(n_1 - 1) \\ Z &:= \frac{S^2_{\eta,n_2}}{\sigma^2} n_2 \in \mathbf{Khi}(n_2 - 1). \end{aligned}$$

Ezekből kapjuk, hogy $W := Y + Z \in \mathbf{Khi}(n_1 + n_2 - 2)$, továbbá $\frac{X\sqrt{n_1+n_2-2}}{\sqrt{W}} \in \mathbf{t}(n_1 + + n_2 - 2)$. Könnyen látható, hogy

$$\frac{X\sqrt{n_1 + n_2 - 2}}{\sqrt{W}} = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{n_1 S^2_{\xi,n_1} + n_2 S^2_{\eta,n_2}}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

melyből kapjuk a tételt.

Most térjünk vissza a feladat megoldásához.

Megoldás. Az előző tételben bizonyítottuk, hogy H_0 esetén

$$t := \frac{\overline{\xi} - \overline{\eta}}{\sqrt{n_1 S^2_{\xi,n_1} + n_2 S^2_{\eta,n_2}}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \in \mathbf{t}(n_1 + n_2 - 2).$$

Speciálisan $n := n_1 = n_2$ esetén

$$t = \frac{\overline{\xi} - \overline{\eta}}{\sqrt{S^2_{\xi,n} + S^2_{\eta,n}}}\sqrt{n-1} \in \mathbf{t}(2n-2).$$

A kétoldali ellenhipotézis teljesülésekor $\overline{\xi} - \overline{\eta}$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a t-eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|t| \leqslant a$ $(a > 0)$ alakú. A továbbiakban legyen $F \sim \mathbf{t}(n_1 + + n_2 - 2)$. Ha $\mathrm{P} \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(|t| \leqslant a) = 2F(a) - 1,$$

így $\mathrm{P}(|t| \leqslant a) = 1 - \alpha$ esetén $a = F^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát

$$|t| \leqslant F^{-1}\left(1 - \frac{\alpha}{2}\right) \quad (\Longleftrightarrow \alpha \leqslant 2 - 2F(|t|))$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Hasonlóan az egymintás t-próbához kapjuk, hogy $H_1\colon m_1 < m_2$ esetén $t \geqslant$ $\geqslant F^{-1}(\alpha)$ $(\Longleftrightarrow \alpha \leqslant F(t))$ elfogadási tartománnyal, míg $H_1\colon m_1 > m_2$ esetén $t \leqslant F^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1 - F(t))$ elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Ezt a statisztikai próbát nevezzük *kétmintás t-próbának.*

5.12. Feladat. Oldjuk meg az előző feladatot akkor is, ha az ismeretlen szórások viszonyát nem ismerjük.

Szükségünk lesz a következő tételre.

5.13. Tétel. *Legyenek* $\xi \in \mathbf{Norm}(m_1;\sigma_1)$, $\eta \in \mathbf{Norm}(m_2;\sigma_2)$ *független valószínűségi változók és* $\xi_1,\dots,\xi_{n_1}$ *a* ξ*-re vonatkozó, illetve* $\eta_1,\dots,\eta_{n_2}$ *az* η*-ra vonatkozó minta* $(2 \leqslant n_1 \leqslant n_2)$. *Ekkor* $m := m_1 - m_2$ *és* $\sigma := \sqrt{\sigma_1^2 + \frac{n_1}{n_2}\sigma_2^2}$ *jelölésekkel*

$$\xi_i - \sqrt{\frac{n_1}{n_2}}\eta_i + \frac{1}{\sqrt{n_1 n_2}}\sum_{k=1}^{n_1}\eta_k - \overline{\eta} \in \mathbf{Norm}(m;\sigma) \quad (i = 1,\dots,n_1)$$

független valószínűségi változók.

Bizonyítás. Az állítás $n_1 = n_2$ esetén triviális. Legyen $2 \leqslant n_1 < n_2$ és

$$\zeta_i := \xi_i - \sqrt{\frac{n_1}{n_2}}\eta_i + \frac{1}{\sqrt{n_1 n_2}}\sum_{k=1}^{n_1}\eta_k - \overline{\eta} \quad (i = 1,\dots,n_1).$$

Ekkor

$$\mathrm{E}\,\zeta_i = m_1 - \sqrt{\frac{n_1}{n_2}}m_2 + \frac{1}{\sqrt{n_1 n_2}}n_1 m_2 - m_2 = m_1 - m_2 = m,$$

másrészt $K_i := \{1,\dots,n_1\}\setminus\{i\}$ és $K := \{n_1+1,\dots,n_2\}$ jelölésekkel

$$\zeta_i = \xi_i + \left(\tfrac{1}{\sqrt{n_1 n_2}} - \tfrac{1}{n_2}\right)\sum_{k\in K_i}\eta_k - \tfrac{1}{n_2}\sum_{k\in K}\eta_k + \left(\tfrac{1}{\sqrt{n_1 n_2}} - \tfrac{1}{n_2} - \sqrt{\tfrac{n_1}{n_2}}\right)\eta_i,$$

így

$$\mathrm{D}^2\,\zeta_i = \sigma_1^2 + \left((n_1-1)\left(\tfrac{1}{\sqrt{n_1 n_2}} - \tfrac{1}{n_2}\right)^2 + \tfrac{n_2-n_1}{n_2^2} + \left(\tfrac{1}{\sqrt{n_1 n_2}} - \tfrac{1}{n_2} - \sqrt{\tfrac{n_1}{n_2}}\right)^2\right)\sigma_2^2,$$

melyből kapjuk, hogy $\mathrm{D}^2\,\zeta_i = \sigma_1^2 + \frac{n_1}{n_2}\sigma_2^2 = \sigma^2$. Mivel $\zeta_i = \xi_i + \sum_{k=1}^{n_2} a_k^{(i)}\eta_k$ alakú, azaz független normális eloszlású valószínűségi változók lineáris kombinációja, ezért $\zeta_i \in \mathbf{Norm}(m;\sigma)$. Még a függetlenséget kell belátni. Ehhez elég a $\mathrm{cov}(\zeta_i,\zeta_j) = 0$ $(i,j=1,\dots,n_1,\ i\neq j)$ megmutatása.

$$\sum_{k=1}^{n_2} a_k^{(i)}a_k^{(j)} = (n_1-2)\left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2}\right)^2 + \\ + 2\left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2}\right)\left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}}\right) + \frac{n_2-n_1}{n_2^2} = 0,$$

ezért $i,j=1,\dots,n_1,\ i\neq j$ esetén

$$\mathrm{cov}(\zeta_i,\zeta_j) = \mathrm{cov}\left(\sum_{k=1}^{n_2} a_k^{(i)}\eta_k, \sum_{k=1}^{n_2} a_k^{(j)}\eta_k\right) = \\ = \sum_{k=1}^{n_2}\sum_{l=1}^{n_2} a_k^{(i)}a_l^{(j)}\,\mathrm{cov}(\eta_k,\eta_l) = \sum_{k=1}^{n_2} a_k^{(i)}a_k^{(j)}\sigma_2^2 = 0.$$

Ezzel a bizonyítást befejeztük.

Most térjünk rá a feladat megoldására.

Megoldás. Az előző tétel szerint van olyan normális eloszlású $m = m_1 - m_2$ várható értékű ζ valószínűségi változó, hogy

$$\zeta_i := \xi_i - \sqrt{\frac{n_1}{n_2}}\eta_i + \frac{1}{\sqrt{n_1 n_2}}\sum_{k=1}^{n_1}\eta_k - \overline{\eta} \quad (i=1,\dots,n_1)$$

ζ-ra vonatkozó minta. Vegyük észre, hogy $n_1 = n_2$ esetén $\zeta_i = \xi_i - \eta_i$. Végezzük el

erre a mintára az egymintás t-próbát $m_0 = 0$ választással. Ekkor

$$m = 0 \iff m_1 = m_2$$
$$m \neq 0 \iff m_1 \neq m_2$$
$$m < 0 \iff m_1 < m_2$$
$$m > 0 \iff m_1 > m_2$$

miatt ennek a próbának a hipotézisei egybeesnek a feladat hipotéziseivel. Tehát legyen

$$t := \frac{\overline{\zeta}}{S^*_{\zeta,n_1}}\sqrt{n_1}$$

és $F \sim \mathbf{t}(n_1 - 1)$. Ekkor $H_1\colon m_1 \neq m_2$ esetén $|t| \leqslant F^{-1}(1 - \frac{\alpha}{2})$ ($\iff \alpha \leqslant 2 - 2F(|t|)$) elfogadási tartománnyal a próba pontosan α terjedelmű. Másrészt $H_1\colon m_1 < m_2$ esetén $t \geqslant F^{-1}(\alpha)$ ($\iff \alpha \leqslant F(t)$) elfogadási tartománnyal, míg $H_1\colon m_1 > m_2$ esetén $t \leqslant F^{-1}(1 - \alpha)$ ($\iff \alpha \leqslant 1 - F(t)$) elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α.

Ezt az eljárást *Scheffé-módszernek* nevezzük, amely tehát nem egy önálló próba, hanem egy eljárás, melynek révén úgy transzformáljuk a mintát, hogy azon az egymintás t-próba végrehajtható legyen, és ebből dönteni tudjunk a hipotézisekre vonatkozóan.

5.14. *Megjegyzés.* Tegyük fel, hogy a ξ-re és η-ra vonatkozó minták nem függetlenek, hanem úgynevezett párosított minták, azaz valójában a (ξ, η) kétdimenziós vektorváltozóra vonatkozik. A feladat pontosan az, mint a Scheffé-módszernél volt, azaz a várható értékeket kell összehasonlítani. Ha teljesül, hogy $\xi - \eta$ normális eloszlású, akkor könnyen láthatóan, a Scheffé-módszer ($n_1 = n_2$ eset) itt is alkalmazható, azaz a különbség mintára kell végrehajtani az egymintás t-próbát $m_0 = 0$ választással.

5.2.5. F-próba

A kétmintás t-próbát azzal a feltétellel tudjuk alkalmazni, hogy az ismeretlen szórások megegyeznek. Ennek a feltételnek a teljesülését vizsgáljuk ebben az alszakaszban.

5.15. Feladat. Legyenek $\xi \in \mathbf{Norm}(m_1; \sigma_1),\ \eta \in \mathbf{Norm}(m_2; \sigma_2)$ független valószínűségi változók, ahol $m_1, m_2, \sigma_1, \sigma_2$ ismeretlenek. Legyen $\xi_1, \ldots, \xi_{n_1}$ a ξ-re vonatkozó, illetve $\eta_1, \ldots, \eta_{n_2}$ az η-ra vonatkozó minta ($n_1 \geqslant 2,\ n_2 \geqslant 2$). A

$$H_0\colon \sigma_1 = \sigma_2$$

$$H_1\colon \sigma_1 \neq \sigma_2$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1\colon \sigma_1 < \sigma_2 \quad \text{illetve} \quad H_1\colon \sigma_1 > \sigma_2$$

egyoldali ellenhipotézisekre is.

Szükség lesz a következő tételre.

5.16. Tétel. *Legyenek* $\xi \in \mathbf{Norm}(m_1;\sigma)$, $\eta \in \mathbf{Norm}(m_2;\sigma)$ *független valószínűségi változók,* $\xi_1,\dots,\xi_{n_1}$ *a* ξ*-re vonatkozó, illetve* $\eta_1,\dots,\eta_{n_2}$ *az* η*-ra vonatkozó minta* $(n_1 \geqslant 2,\ n_2 \geqslant 2)$*. Ekkor*

$$\frac{S^{*2}_{\xi,n_1}}{S^{*2}_{\eta,n_2}} \in \mathbf{F}(n_1-1;n_2-1).$$

Bizonyítás. Korábban bizonyítottuk, hogy

$$\frac{S^2_{\xi,n_1}}{\sigma^2}n_1 \in \mathbf{Khi}(n_1-1) \quad \text{és} \quad \frac{S^2_{\eta,n_2}}{\sigma^2}n_2 \in \mathbf{Khi}(n_2-1),$$

így ezek függetlensége miatt

$$\frac{(n_2-1)\frac{S^2_{\xi,n_1}}{\sigma^2}n_1}{(n_1-1)\frac{S^2_{\eta,n_2}}{\sigma^2}n_2} = \frac{\frac{n_1}{n_1-1}S^2_{\xi,n_1}}{\frac{n_2}{n_2-1}S^2_{\eta,n_2}} = \frac{S^{*2}_{\xi,n_1}}{S^{*2}_{\eta,n_2}} \in \mathbf{F}(n_1-1;n_2-1).$$

Most térjünk vissza a feladat megoldására.

Megoldás. Az előző tétel szerint, ha $H_0\colon \sigma_1 = \sigma_2$ igaz, akkor

$$\mathsf{F} := \frac{S^{*2}_{\xi,n_1}}{S^{*2}_{\eta,n_2}} \in \mathbf{F}(n_1-1;n_2-1).$$

A $H_1\colon \sigma_1 \neq \sigma_2$ ellenhipotézis teljesülésekor F várhatóan kritikus értékben messze van 1-től, hiszen a korrigált tapasztalati szórás torzítatlan becslése a szórásnak. Ezért az elfogadási tartomány $a \leqslant \mathsf{F} \leqslant b$ alakú, ahol $0 < a < 1 < b$. A továbbiakban legyen $F \sim \mathbf{F}(n_1-1;n_2-1)$. Tegyük fel, hogy $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(\mathsf{F} < a) = F(a) = \frac{\alpha}{2},$$
$$\mathrm{P}(\mathsf{F} > b) = 1 - F(b) = \frac{\alpha}{2},$$

azaz $a = F^{-1}\left(\frac{\alpha}{2}\right) > 0$ és $b = F^{-1}\left(1-\frac{\alpha}{2}\right)$. Mivel $0{,}3 < F(1) < 0{,}7$ (lásd az F-eloszlás leírásánál található lemmát), így $\frac{\alpha}{2} < F(1) < 1-\frac{\alpha}{2}$ biztosan teljesül. Ezért az ezzel ekvivalens $a < 1 < b$ is teljesül. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(a \leqslant \mathsf{F} \leqslant b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leqslant \mathsf{F} \leqslant F^{-1}\left(1-\frac{\alpha}{2}\right)$ $(\Longleftrightarrow \alpha \leqslant 2\min\{\, F(\mathsf{F}), 1 - F(\mathsf{F})\,\})$ elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1\colon \sigma_1 < \sigma_2$ teljesülésekor F várhatóan kritikus értékben kisebb 1-től. Ezért az elfogadási tartomány $\mathsf{F} \geqslant c$ alakú, ahol $0 < c < 1$. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\mathsf{F} \geqslant c) = 1 - F(c),$$

így $\mathrm{P}(\mathsf{F} \geqslant c) = 1-\alpha$ esetén $c = F^{-1}(\alpha) > 0$. Az $\alpha < F(1)$ biztosan teljesül $0 < \alpha \leqslant$ $\leqslant 0{,}3$ esetén, így $c < 1$ is teljesül. Tehát $\mathsf{F} \geqslant F^{-1}(\alpha)$ $(\Longleftrightarrow \alpha \leqslant F(\mathsf{F}))$ elfogadási tartománnyal α terjedelmű próbát kapunk.

Végül $H_1\colon \sigma_1 > \sigma_2$ ellenhipotézis teljesülése esetén F várhatóan kritikus értékben nagyobb 1-től. Ezért az elfogadási tartomány $\mathsf{F} \leqslant d$ alakú, ahol $d > 1$. Ekkor $\mathrm{P} \in$ $\in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\mathsf{F} \leqslant d) = F(d),$$

így $\mathrm{P}(\mathsf{F} \leqslant d) = 1-\alpha$ esetén $d = F^{-1}(1-\alpha)$. Mivel $1-\alpha > F(1)$ biztosan teljesül, ha $0 < \alpha \leqslant 0{,}3$, ezért $d > 1$ is teljesül. Tehát $\mathsf{F} \leqslant F^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1 - F(\mathsf{F}))$ elfogadási tartománnyal α terjedelmű próbát kapunk.

Ezt a statisztikai próbát *F-próbának* nevezzük.

5.17. *Megjegyzés.* Legyen $F_1 \sim \mathbf{F}(n_1-1; n_2-1)$ és $F_2 \sim \mathbf{F}(n_2-1; n_1-1)$. Kétoldali ellenhipotézis esetén láttuk, hogy

$$F_1^{-1}\left(\frac{\alpha}{2}\right) \leqslant \mathsf{F} \leqslant F_1^{-1}\left(1-\frac{\alpha}{2}\right)$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Mivel H_0 teljesülése esetén $\frac{1}{\mathsf{F}} \in \mathbf{F}(n_2-1; n_1-1)$, ezért

$$F_2^{-1}\left(\frac{\alpha}{2}\right) \leqslant \frac{1}{\mathsf{F}} \leqslant F_2^{-1}\left(1-\frac{\alpha}{2}\right)$$

elfogadási tartománnyal szintén α terjedelmű próbát kapunk.

Ha $\mathsf{F} \geqslant 1$, akkor válasszuk az első elfogadási tartományt. De ebben az esetben $\frac{\alpha}{2} < 0{,}3 < F_1(1)$ miatt $F_1^{-1}\left(\frac{\alpha}{2}\right) < 1 \leqslant \mathsf{F}$ biztosan teljesül. Tehát ekkor az elfogadási

tartomány $\mathsf{F} \leqslant F_1^{-1}\left(1-\frac{\alpha}{2}\right)$.

Ha $\mathsf{F} < 1$, azaz $\frac{1}{\mathsf{F}} > 1$, akkor válasszuk a második elfogadási tartományt. De ebben az esetben $\frac{\alpha}{2} < 0{,}3 < F_2(1)$ miatt $F_2^{-1}\left(\frac{\alpha}{2}\right) < 1 < \frac{1}{\mathsf{F}}$ biztosan teljesül. Tehát ekkor az elfogadási tartomány $\frac{1}{\mathsf{F}} \leqslant F_2^{-1}\left(1-\frac{\alpha}{2}\right)$.

Ezzel bizonyítottuk a következőt. Legyen $\mathsf{F}^* := \max\{\,\mathsf{F}, \frac{1}{\mathsf{F}}\,\}$, továbbá $G \sim \mathbf{F}(n_1 - -1; n_2 - 1)$, ha $\mathsf{F}^* = \mathsf{F}$ illetve $G \sim \mathbf{F}(n_2 - 1; n_1 - 1)$, ha $\mathsf{F}^* = \frac{1}{\mathsf{F}}$. Ekkor

$$\mathsf{F}^* \leqslant G^{-1}\left(1-\frac{\alpha}{2}\right)$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ezzel a módszerrel tehát nem két, hanem csak egy kritikus értéket kell számolni.

5.2.6. Khi-négyzet próba normális eloszlás szórására

5.18. Feladat. Legyen $\xi \in \mathbf{Norm}(m;\sigma)$, ahol m és σ ismeretlenek. Legyen $\sigma_0 \in \mathbb{R}_+$ rögzített és $\xi_1, \ldots, \xi_n$ a ξ-re vonatkozó minta ($n \geqslant 2$). A

$$H_0 \colon \sigma = \sigma_0$$
$$H_1 \colon \sigma \neq \sigma_0$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1 \colon \sigma < \sigma_0 \quad \text{illetve} \quad H_1 \colon \sigma > \sigma_0$$

egyoldali ellenhipotézisekre is.

Megoldás. Tudjuk, hogy H_0 esetén

$$\chi^2 := \frac{S_n^2}{\sigma_0^2} n \in \mathbf{Khi}(n-1).$$

Mivel $\frac{S_n^2}{\sigma_0^2} n = \frac{S_n^{*\,2}}{\sigma_0^2}(n-1)$ és S_n^{*2} a szórásnégyzet torzítatlan becslése, így $\sigma \neq \sigma_0$ teljesülése esetén χ^2 várhatóan kritikus mértékben messze van $n-1$-től. Így az elfogadási tartományt válasszuk $a \leqslant \chi^2 \leqslant b$ alakúnak, ahol $0 < a < n-1 < b$. A továbbiakban legyen $F \sim \mathbf{Khi}(n-1)$. Tegyük fel, hogy $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(\chi^2 < a) = F(a) = \frac{\alpha}{2},$$
$$\mathrm{P}(\chi^2 > b) = 1 - F(b) = \frac{\alpha}{2},$$

azaz $a = F^{-1}\left(\frac{\alpha}{2}\right) > 0$ és $b = F^{-1}\left(1-\frac{\alpha}{2}\right)$. Mivel $0{,}5 < F(n-1) < 0{,}7$ (lásd a khi-négyzet eloszlás leírásánál található lemmát), így $\frac{\alpha}{2} < F(n-1) < 1-\frac{\alpha}{2}$ biztosan teljesül. Ezért az ezzel ekvivalens $a < n-1 < b$ is teljesül. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(a \leqslant \chi^2 \leqslant b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leqslant \chi^2 \leqslant F^{-1}\left(1-\frac{\alpha}{2}\right)$ $(\Longleftrightarrow \alpha \leqslant 2\min\{\, F(\chi^2), 1-F(\chi^2)\,\})$ elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1\colon \sigma < \sigma_0$ ellenhipotézis teljesülésekor χ^2 várhatóan kritikus mértékben kisebb $n-1$-től. Azaz az elfogadási tartomány $\chi^2 \geqslant c$ alakú, ahol $0 < c < n-1$. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\chi^2 \geqslant c) = 1 - F(c),$$

így $\mathrm{P}(\chi^2 \geqslant c) = 1-\alpha$ esetén $c = F^{-1}(\alpha) > 0$. Erre teljesül, hogy $c < n-1$, mert $0{,}5 < F(n-1)$. Tehát így $\chi^2 \geqslant F^{-1}(\alpha)$ $(\Longleftrightarrow \alpha \leqslant F(\chi^2))$ elfogadási tartománnyal α terjedelmű próbát kapunk.

Ha $H_1\colon \sigma > \sigma_0$ teljesül, akkor χ^2 várhatóan kritikus mértékben nagyobb $n-1$-től, azaz az elfogadási tartomány $\chi^2 \leqslant d$ alakú, ahol $d > n-1$. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\chi^2 \leqslant d) = F(d),$$

így $\mathrm{P}(\chi^2 \leqslant d) = 1-\alpha$ esetén $d = F^{-1}(1-\alpha)$. Másrészt ekkor $F(n-1) < 0{,}7$ miatt $0 < \alpha \leqslant 0{,}3$ esetén $d > n-1$. Tehát így $\chi^2 \leqslant F^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1-F(\chi^2))$ elfogadási tartománnyal α terjedelmű próbát kapunk. Ez az úgynevezett *khi-négyzet próba.*

5.2.7. Statisztikai próba exponenciális eloszlás paraméterére

5.19. Feladat. Legyen $\xi \in \mathbf{Exp}(\lambda)$, ahol λ ismeretlen. Legyen $\lambda_0 \in \mathbb{R}_+$ rögzített és $\xi_1, \ldots, \xi_n$ a ξ-re vonatkozó minta. A

$$H_0\colon \lambda = \lambda_0$$
$$H_1\colon \lambda \neq \lambda_0$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1\colon \lambda < \lambda_0 \quad \text{illetve} \quad H_1\colon \lambda > \lambda_0$$

egyoldali ellenhipotézisekre is.

Megoldás. A λ intervallumbecslésénél láttuk, hogy H_0 esetén

$$\gamma := \lambda_0 n\overline{\xi} \in \mathbf{Gamma}(n;1).$$

Mivel $\overline{\xi}$ az $\mathrm{E}_\lambda = \frac{1}{\lambda}$ torzítatlan becslése, így a $H_1\colon \lambda \neq \lambda_0$ ($\Longleftrightarrow n \neq \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ várhatóan kritikus mértékben messze van n-től. Azaz az elfogadási tartomány $a \leqslant \gamma \leqslant b$ alakú, ahol $0 < a < n < b$. A továbbiakban legyen $F \sim \mathbf{Gamma}(n;1)$. Tegyük fel, hogy $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(\gamma < a) = F(a) = \frac{\alpha}{2},$$
$$\mathrm{P}(\gamma > b) = 1 - F(b) = \frac{\alpha}{2},$$

azaz $a = F^{-1}\left(\frac{\alpha}{2}\right) > 0$ és $b = F^{-1}\left(1 - \frac{\alpha}{2}\right)$. Mivel $0{,}5 < F(n) < 0{,}7$ (lásd a gamma-eloszlás leírásánál található lemmát), így $\frac{\alpha}{2} < F(n) < 1 - \frac{\alpha}{2}$ biztosan teljesül. Ezért az ezzel ekvivalens $a < n < b$ is teljesül. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$ esetén

$$\mathrm{P}(a \leqslant \gamma \leqslant b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leqslant \gamma \leqslant F^{-1}\left(1 - \frac{\alpha}{2}\right)$ ($\Longleftrightarrow \alpha \leqslant 2\min\{\, F(\gamma), 1 - F(\gamma)\,\}$) elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1\colon \lambda < \lambda_0$ ($\Longleftrightarrow n < \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ (ellentétben a korábbi próbákkal) várhatóan kritikus mértékben nagyobb n-től. Azaz az elfogadási tartomány $\gamma \leqslant c$ alakú, ahol $c > n$. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\gamma \leqslant c) = F(c),$$

így $\mathrm{P}(\gamma \leqslant c) = 1 - \alpha$ esetén $c = F^{-1}(1 - \alpha)$. Másrészt ekkor $F(n) < 0{,}7$ miatt $0 < \alpha \leqslant 0{,}3$ esetén $c > n$. Tehát így $\gamma \leqslant F^{-1}(1 - \alpha)$ ($\Longleftrightarrow \alpha \leqslant 1 - F(\gamma)$) elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1\colon \lambda > \lambda_0$ ($\Longleftrightarrow n > \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ (ellentétben a korábbi próbákkal) várhatóan kritikus mértékben kisebb n-től. Azaz az elfogadási tartomány $\gamma \geqslant d$ alakú, ahol $0 < d < n$. Mivel $\mathrm{P} \in \mathcal{P}_{H_0}$-ra

$$\mathrm{P}(\gamma \geqslant d) = 1 - F(d),$$

így $\mathrm{P}(\gamma \geqslant d) = 1 - \alpha$ esetén $d = F^{-1}(\alpha) > 0$. Erre teljesül, hogy $d < n$, mert $0{,}5 < F(n)$. Tehát így $\gamma \geqslant F^{-1}(\alpha)$ ($\Longleftrightarrow \alpha \leqslant F(\gamma)$) elfogadási tartománnyal α

terjedelmű próbát kapunk.

5.20. *Megjegyzés.* Vigyázzunk arra, hogy itt a egyoldali ellenhipotézisek esetében fordítva vannak az elfogadási tartományok, mint a korábbi próbáknál. Ennek megfelelően változik az is, hogy kétoldali ellenhipotézisnél meghozott döntés ismeretében mik lesznek a egyoldali ellenhipotézisekre a döntések. Ezt a következő táblázatban foglaljuk össze:

$H_0\colon \lambda = \lambda_0$		$H_1\colon \lambda \neq \lambda_0$	
	H_0-t elfogadjuk	H_0-t elutasítjuk	
		$\gamma > F^{-1}(1-\frac{\alpha}{2})$	$\gamma < F^{-1}(\frac{\alpha}{2})$
$H_1\colon \lambda < \lambda_0$	H_0-t elfogadjuk	H_0-t elfogadjuk	H_0-t elutasítjuk
$H_1\colon \lambda > \lambda_0$	H_0-t elfogadjuk	H_0-t elutasítjuk	H_0-t elfogadjuk

A táblázat úgy is értelmezhető, hogy $F^{-1}\left(\frac{\alpha}{2}\right) \leqslant \gamma \leqslant F^{-1}\left(1-\frac{\alpha}{2}\right)$ esetén $\lambda = \lambda_0$, $\gamma < F^{-1}\left(\frac{\alpha}{2}\right)$ esetén $\lambda > \lambda_0$, illetve $\gamma > F^{-1}\left(1-\frac{\alpha}{2}\right)$ esetén $\lambda < \lambda_0$ mellett döntünk α terjedelemmel.

5.2.8. Statisztikai próba valószínűségre

5.21. Feladat. Legyen $\xi \in \mathbf{Bin}(1;p)$, ahol p ismeretlen. Legyen $0 < p_0 < 1$ rögzített és $\xi_1, \ldots, \xi_n$ a ξ-re vonatkozó minta. A

$$H_0\colon p = p_0$$
$$H_1\colon p \neq p_0$$

hipotézisekre adjon adott α terjedelmű próbát. A feladatot oldja meg

$$H_1\colon p < p_0 \quad \text{illetve} \quad H_1\colon p > p_0$$

egyoldali ellenhipotézisekre is.

5.22. *Megjegyzés.* Ha A egy esemény és $\xi = \mathrm{I}_A$, akkor $\xi \in \mathbf{Bin}(1;p)$, ahol p az A valószínűsége. Ezért a feladat úgy is megfogalmazható, hogy adjon az előző hipotézisekre α terjedelmű próbát, ahol p egy esemény valószínűsége.

Megoldás. Ismert, hogy ha H_0 igaz, akkor

$$n\overline{\xi} \in \mathbf{Bin}(n;p_0).$$

Ha ξ egy esemény indikátorváltozója, akkor $n\overline{\xi}$ az esemény bekövetkezéseinek a számát jelenti n kísérlet után. Mivel $\overline{\xi}$ a p torzítatlan becslése, ezért $H_1\colon p \neq p_0$ esetén $\overline{\xi}$ várhatóan kritikus mértékben eltávolodik p_0-tól. Így ekkor az elfogadási tartomány $a \leqslant n\overline{\xi} \leqslant b$ alakú, ahol $a, b \in \mathbb{N}$ és $1 \leqslant a < np_0 < b \leqslant n-1$. Az $1 \leqslant a$ és $b \leqslant n-1$ feltételek azért kellenek, hogy a kritikus tartományban $n\overline{\xi} < a$ illetve $n\overline{\xi} > b$ ne legyenek lehetetlen események. Keressük meg a legkisebb a illetve b pozitív egész számokat, melyekre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

$$\mathrm{P}(n\overline{\xi} \leqslant a) = \sum_{i=0}^{a} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} \geqslant \tfrac{\alpha}{2},$$
$$\mathrm{P}(n\overline{\xi} \leqslant b) = \sum_{i=0}^{b} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} \geqslant 1 - \tfrac{\alpha}{2}.$$

Az így definiált a és b esetén, ha $P \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(a \leqslant n\overline{\xi} \leqslant b) = \sum_{i=a}^{b} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} =$$
$$= \sum_{i=0}^{b} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} - \underbrace{\sum_{i=0}^{a-1} \tbinom{n}{i} p_0^i (1-p_0)^{n-i}}_{<\frac{\alpha}{2}} > 1-\alpha,$$

így ilyenkor $a \leqslant n\overline{\xi} \leqslant b$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $1 \leqslant a < np_0 < b \leqslant n-1$ feltétel mindig teljesíthető α és n alkalmas megválasztásával.

$H_1\colon p < p_0$ esetén $\overline{\xi}$ várhatóan kritikus mértékben p_0 alatt van, azaz az elfogadási tartomány $n\overline{\xi} \geqslant c$ alakú, ahol $c \in \mathbb{N}$ és $1 \leqslant c < np_0$. Legyen c a legkisebb pozitív egész, melyre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

$$\mathrm{P}(n\overline{\xi} \leqslant c) = \sum_{i=0}^{c} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} \geqslant \alpha.$$

Az így definiált c esetén, ha $P \in \mathcal{P}_{H_0}$, akkor

$$\mathrm{P}(n\overline{\xi} \geqslant c) = \sum_{i=c}^{n} \tbinom{n}{i} p_0^i (1-p_0)^{n-i} = 1 - \underbrace{\sum_{i=0}^{c-1} \tbinom{n}{i} p_0^i (1-p_0)^{n-i}}_{<\alpha} > 1-\alpha,$$

azaz $n\overline{\xi} \geqslant c$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $1 \leqslant c < np_0$ feltétel itt is mindig teljesíthető α és n alkalmas megválasztásával.

$H_1\colon p > p_0$ esetén $\overline{\xi}$ várhatóan kritikus mértékben p_0 felett van, azaz az elfogadási tartomány $n\overline{\xi} \leqslant d$ alakú, ahol $d \in \mathbb{N}$ és $np_0 < d \leqslant n-1$. Legyen d a legkisebb pozitív

egész, melyre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

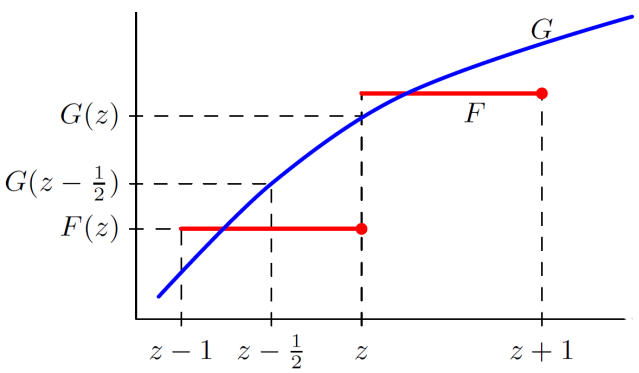
$$\mathrm{P}(n\overline{\xi} \leqslant d) = \sum_{i=0}^{d} \binom{n}{i} p_0^i (1-p_0)^{n-i} \geqslant 1-\alpha.$$

Ekkor tehát $n\overline{\xi} \leqslant d$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $np_0 < d \leqslant n-1$ feltétel itt is mindig teljesíthető α és n alkalmas megválasztásával.

Az első két ellenhipotézisnél azért nem úgy választottuk a kritikus értékeket, hogy az elfogadási tartomány valószínűsége H_0 esetén $1-\alpha$-val egyenlő is lehessen, mert egyrészt ez csak ritkán érhető el az eloszlás diszkrétsége miatt, másrészt ekkor Excellel nehezebben tudnánk számolni.

Ha n elég nagy, akkor az előbbi kritikus értékek kiszámolásához használhatunk egyszerűbb közelítő formulát is. Ehhez szükségünk lesz az úgynevezett folytonossági korrekcióra.

Folytonossági korrekció. Ha a $\min\{\, np, n(1-p)\,\} \geqslant 10$ feltétel teljesül, akkor $F \sim \mathbf{Bin}(n;p)$ és $G \sim \mathbf{Norm}\left(np; \sqrt{np(1-p)}\right)$ jelöléssel $F(z)$ értékét nagyon jól közelíti $G(z)$. Legyen $z \in \mathbb{N}$. Ekkor a következő ábráról látható, hogy az F lépcsőssége és a G folytonossága miatt $G(z-\frac{1}{2})$ még pontosabban megközelíti $F(z)$ értékét.



Tehát $\varrho \in \mathbf{Bin}(n;p)$ és $z \in \mathbb{N}$ esetén

$$\mathrm{P}(\varrho < z) \simeq \Phi\left(\frac{z-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right), \quad \text{illetve}$$

$$\mathrm{P}(\varrho \leqslant z) = \mathrm{P}(\varrho < z+1) \simeq \Phi\left(\frac{z+1-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{z+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

közelítések már nagyon jónak tekinthetők. Például, ha $n = 100$, $p = 0{,}4$, akkor $10 < np = 40 < n(1-p) = 60$ teljesül, ezért használhatjuk a közelítést. Például $\mathrm{P}(\varrho \leqslant 30)$ értéke közelítőleg

$$\Phi\left(\frac{30+\frac{1}{2}-40}{\sqrt{40\cdot 0{,}6}}\right),$$

ami öt tizedesjegyre kerekítve 0,02624. Ha nem használjuk a folytonossági korrekciót, akkor a

$$\Phi\left(\frac{30+1-40}{\sqrt{40\cdot 0{,}6}}\right)$$

értéket kell használni közelítésnek, amely öt tizedesjegyre kerekítve 0,03310. Összehasonlításként $\mathrm{P}(\varrho \leqslant 30)$ igazi értéke 0,02478 öt tizedesjegyre kerekítve. Ebből jól látható, hogy a folytonossági korrekcióval pontosabb közelítést kaptunk.

5.23. Feladat. Az előző megoldásban felírt kritikus értékekre adjunk közelítő képletet $\min\{\, np_0, n(1-p_0)\,\} \geqslant 10$ esetén, a folytonossági korrekciót alkalmazva.

Megoldás. Az előző megoldás jelöléseit fogjuk használni. Az előbbiek miatt

$$\mathrm{P}(n\overline{\xi} \leqslant a) \simeq \Phi\left(\frac{a+\frac{1}{2}-np_0}{\sqrt{np_0(1-p_0)}}\right) = \frac{\alpha}{2},$$

melyből – figyelembe véve, hogy $a \in \mathbb{N}$ és alsó kritikus értéket jelent – kapjuk, hogy

$$h(x) := np_0 - \frac{1}{2} + \sqrt{np_0(1-p_0)}\Phi^{-1}(x)$$

jelöléssel $a \simeq \left[h(\frac{\alpha}{2})\right]$. Hasonlóan kapjuk, hogy $b \simeq \left[h(1-\frac{\alpha}{2})\right] + 1$, $c \simeq [h(\alpha)]$ és $d \simeq [h(1-\alpha)] + 1$.

5.3. Nemparaméteres hipotézisvizsgálatok

A következőkben, ha nem írjuk ki külön az ellenhipotézist, akkor az mindig a nullhipotézis negáltját jelenti. Az itt taglalt illeszkedés-, függetlenség- és homogenitásvizsgálatokat *khi-négyzet próbáknak* is nevezik, mert a próbastatisztika mindegyik esetben hasonló szerkezetű aszimptotikusan khi-négyzet eloszlású.

5.3.1. Tiszta illeszkedésvizsgálat

5.24. Feladat. Legyen $A_1, \dots, A_r$ egy teljes eseményrendszer és $p_1, \dots, p_r \in \mathbb{R}_+$, ahol $p_1 + \dots + p_r = 1$. Készítsünk a

$$H_0\colon\ \mathrm{P}(A_i) = p_i\ (i = 1, \dots, r)$$

nullhipotézisre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti.

Megoldás. Jelölje ϱ_i az A_i esemény gyakoriságát n kísérlet után, és legyen

$$\chi^2 := \sum_{i=1}^{r} \frac{(\varrho_i - np_i)^2}{np_i}.$$

H_0 teljesülése esetén χ^2 várhatóan nem távolodik el kritikus mértékben 0-tól, így az elfogadási tartomány $\chi^2 \leqslant a$ alakú, ahol $a > 0$. Ismert, hogy $\min\{\, \varrho_1, \dots, \varrho_r \,\} \geqslant 10$ teljesülése esetén

$$\mathrm{P}(\chi^2 \leqslant a) \simeq F(a),$$

ahol $\mathrm{P} \in \mathcal{P}_{H_0}$ és $F \sim \mathbf{Khi}(r-1)$. (A bizonyítást lásd pl. Fazekas I. [2, 161–162. oldal].) Így $\mathrm{P}(\chi^2 \leqslant a) = 1-\alpha$ esetén $a \simeq F^{-1}(1-\alpha)$. Tehát $\chi^2 \leqslant F^{-1}(1-\alpha)$ ($\Longleftrightarrow \alpha \leqslant$ $\leqslant 1 - F(\chi^2)$) elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.25. Feladat. Legyen ξ egy ismeretlen eloszlású valószínűségi változó és F_0 egy rögzített eloszlásfüggvény. Készítsünk a

$$H_0\colon\ \mathrm{P}(\xi < x) = F_0(x)\ (x \in \mathbb{R})$$

nullhipotézisre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti.

Megoldás. Ha ξ diszkrét valószínűségi változó $\{\, x_1, x_2, \dots \,\}$ értékkészlettel, ahol $x_1 <$ $< x_2 < \dots$, akkor válasszuk meg a

$$k_0 := 0 < k_1 < k_2 < \dots < k_r$$

egész számokat úgy, hogy az $A_i := \{\, x_{k_{i-1}+1} \leqslant \xi \leqslant x_{k_i} \,\}$ események teljes eseményrendszert alkossanak, továbbá ezek ϱ_i gyakorisága a ξ-re vonatkozó n elemű mintarealizáció alapján legalább 10 legyen minden $i = 1, \dots, r$ esetén.

Ha ξ nem diszkrét valószínűségi változó, akkor válasszuk meg az

$$a_0 := -\infty < a_1 < a_2 < \dots < a_{r-1} < a_r := \infty$$

valós számokat úgy, hogy az $A_i := \{\, a_{i-1} \leqslant \xi < a_i \,\}$ események ϱ_i gyakorisága a ξ-re vonatkozó n elemű mintarealizáció alapján legalább 10 legyen minden $i = 1, \dots, r$ esetén. Ügyeljünk arra, hogy az a_i osztópontok függetlenek legyenek a mintarealizáció elemeitől.

Ezután $p_i := \mathrm{P}(A_i)$ $(\mathrm{P} \in \mathcal{P}_{H_0})$ jelöléssel legyen

$$H_0' \colon \mathrm{P}(A_i) = p_i \ (i = 1, \dots, r).$$

Ha H_0 igaz, akkor H_0' is az. Így az előző feladat megoldásából látható, hogy H_0 teljesülése esetén

$$\chi^2 := \sum_{i=1}^{r} \frac{(\varrho_i - np_i)^2}{np_i}$$

aszimptotikusan $r-1$ szabadsági fokú khi-négyzet eloszlású. Ebből kapjuk, hogy $F \sim \mathbf{Khi}(r-1)$ jelöléssel és $\chi^2 \leqslant F^{-1}(1-\alpha)$ $(\Longleftrightarrow \alpha \leqslant 1 - F(\chi^2))$ elfogadási tartománnyal, H_0-ra közelítőleg α terjedelmű próbát kapunk.

5.3.2. Becsléses illeszkedésvizsgálat

A tiszta illeszkedésvizsgálatban azt vizsgáltuk, hogy egy valószínűségi változónak mi lehet az eloszlása. Azonban legtöbb esetben elég csak azt megmondani, hogy melyik eloszláscsaládba tartozik (egyenletes, normális, Poisson, stb.). Ilyenkor használjuk a becsléses illeszkedésvizsgálatot.

5.26. Feladat. Legyen $v \in \mathbb{N}$, $\Theta \subset \mathbb{R}^v$, $\Theta \neq \emptyset$. Jelöljön F_ϑ eloszlásfüggvényt minden $\vartheta = (\vartheta_1, \dots, \vartheta_v) \in \Theta$ esetén. Legyen az a nullhipotézis, hogy az ismeretlen eloszlású ξ valószínűségi változó az $\{\, F_\vartheta : \vartheta \in \Theta \,\}$ eloszláscsaládba tartozik, azaz

$$H_0 \colon \mathrm{P}(\xi < x) = F_\vartheta(x) \ (x \in \mathbb{R}) \text{ valamely } \vartheta \in \Theta \text{ esetén.}$$

Készítsünk erre a nullhipotézisre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti.

Megoldás. Először konstruáljuk meg az $A_1, \dots, A_r$ teljes eseményrendszert a tiszta illeszkedésvizsgálatban leírtak szerint, és jelölje ϱ_i az A_i esemény gyakoriságát n kísérlet után. Ezután H_0 feltételezésével számoljuk ki ϑ_i maximum likelihood becslését, melyet jelöljön $\widehat{\vartheta}_i$. Legyen $\widehat{\vartheta} := (\widehat{\vartheta}_1, \dots, \widehat{\vartheta}_v)$, $\widehat{p}_i := \mathrm{P}_{\widehat{\vartheta}}(A_i)$, továbbá

$$\chi^2 := \sum_{i=1}^{r} \frac{(\varrho_i - n\widehat{p}_i)^2}{n\widehat{p}_i}.$$

Bizonyítható, hogy ha H_0 igaz, akkor χ^2 eloszlása $r-1-v$ szabadsági fokú khinégyzet eloszláshoz konvergál $n \to \infty$ esetén. (A bizonyítás az úgynevezett likelihood hányados határeloszlásával hozható kapcsolatba, mi nem végezzük el. Lásd például Terdik Gy. [16, 91–93. oldal].) A gyakorlatban ez azt jelenti, hogy $F \sim \mathbf{Khi}(r-1-v)$ jelöléssel

$$\mathrm{P}(\chi^2 < x) \simeq F(x).$$

A közelítés már jónak tekinthető, ha $\min\{\, \varrho_1, \ldots, \varrho_r \,\} \geqslant 10$.

Így hasonlóan a tiszta illeszkedésvizsgálathoz kapjuk, hogy H_0 nullhipotézisre $\chi^2 \leqslant F^{-1}(1-\alpha)$ ($\Longleftrightarrow \alpha \leqslant 1 - F(\chi^2)$) elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.3. Függetlenségvizsgálat

A következő feladatban két teljes eseményrendszer függetlenségét vizsgáljuk.

5.27. Feladat. Legyen $A_1, \ldots, A_r$ és $B_1, \ldots, B_s$ két teljes eseményrendszer. Készítsünk a

$$H_0\colon\ \mathrm{P}(A_i \cap B_j) = \mathrm{P}(A_i)\,\mathrm{P}(B_j)\ (i = 1, \ldots, r,\ j = 1, \ldots, s)$$

nullhipotézisre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti.

Megoldás. Legyen k_i illetve l_j az A_i illetve B_j gyakorisága n kísérlet után. Ekkor $\mathrm{P}(A_i)$ illetve $\mathrm{P}(B_j)$ maximum likelihood becslése $\frac{k_i}{n}$ illetve $\frac{l_j}{n}$. Ez összesen $(r-1)+$ $+\,(s-1)$ darab független becslést jelent a $k_1 + \cdots + k_r = n$ és $l_1 + \cdots + l_s = n$ feltételek miatt. Legyen $\widehat{p}_{ij} := \frac{k_i}{n} \cdot \frac{l_j}{n}$ és

$$\chi^2 := \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(\varrho_{ij} - n\widehat{p}_{ij})^2}{n\widehat{p}_{ij}} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n\varrho_{ij} - k_i l_j)^2}{k_i l_j},$$

ahol ϱ_{ij} az $A_i \cap B_j$ esemény gyakorisága n kísérlet után. A gyakoriságokat a következő úgynevezett *kontingencia táblázatba* szokták összefoglalni.

	B_1	B_2	$\ldots$	B_s	
A_1	ϱ_{11}	ϱ_{12}	$\ldots$	ϱ_{1s}	k_1
A_2	ϱ_{21}	ϱ_{22}	$\ldots$	ϱ_{2s}	k_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	ϱ_{r1}	ϱ_{r2}	$\ldots$	ϱ_{rs}	k_r
	l_1	l_2	$\ldots$	l_s	n

A becsléses illeszkedésvizsgálatnál elmondottak szerint, ha H_0 igaz, akkor χ^2 eloszlása $rs-1-(r-1)-(s-1)=(r-1)(s-1)$ szabadsági fokú khi-négyzet eloszláshoz konvergál $n\to\infty$ esetén. Innen az eddigiekhez hasonlóan, ha $\varrho_{ij}\geqslant 10$ minden i,j esetén és $F\sim\mathbf{Khi}_{((r-1)(s-1))}$, akkor a H_0 nullhipotézisre $\chi^2\leqslant F^{-1}(1-\alpha)$ ($\Longleftrightarrow \alpha\leqslant$ $\leqslant 1-F(\chi^2)$) elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.28. Feladat. Legyen (ξ,η) kétdimenziós valószínűségi vektorváltozó. Az erre vonatkozó $(\xi_1,\eta_1),\dots,(\xi_n,\eta_n)$ minta alapján készítsünk a

$$H_0\colon \xi \text{ és } \eta \text{ független}$$

nullhipotézisre α terjedelmű próbát.

Megoldás. Konstruáljuk meg a $\xi_1,\dots,\xi_n$ illetve az $\eta_1,\dots,\eta_n$ mintákra az $A_1,\dots,A_r$ illetve $B_1,\dots,B_s$ teljes eseményrendszereket a tiszta illeszkedésvizsgálatban leírtak szerint. Ezután legyen

$$H_0'\colon\ \mathrm{P}(A_i\cap B_j)=\mathrm{P}(A_i)\,\mathrm{P}(B_j)\ (i=1,\dots,r,\ j=1,\dots,s).$$

Ha H_0 igaz, akkor H_0' is az. Így az előző feladat megoldásából kapjuk, hogy H_0 teljesülése esetén

$$\chi^2:=\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n\varrho_{ij}-k_il_j)^2}{k_il_j}$$

eloszlása $(r-1)(s-1)$ szabadsági fokú khi-négyzet eloszláshoz konvergál, ha $n\to\infty$. Innen $F\sim\mathbf{Khi}_{((r-1)(s-1))}$ jelöléssel, a H_0 nullhipotézisre $\chi^2\leqslant F^{-1}(1-\alpha)$ ($\Longleftrightarrow \alpha\leqslant$ $\leqslant 1-F(\chi^2)$) elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.4. Homogenitásvizsgálat

5.29. Feladat. Legyenek ξ és η független valószínűségi változók. Az ezekre vonatkozó $\xi_1,\dots,\xi_{n_1}$ illetve $\eta_1,\dots,\eta_{n_2}$ minták alapján készítsünk a

$$H_0\colon \xi \text{ és } \eta \text{ azonos eloszlású}$$

nullhipotézisre α terjedelmű próbát.

Megoldás. Válasszuk meg az

$$c_0:=-\infty<c_1<c_2<\dots<c_{r-1}<c_r:=\infty$$

valós számokat úgy, hogy a $\xi \in C_i := [c_{i-1}, c_i)$ esemény k_i gyakorisága illetve az $\eta \in C_i$ esemény l_i gyakorisága a mintarealizációk alapján legalább 10 legyen minden $i = 1, \dots, r$ esetén.

Most tegyük fel, hogy H_0 teljesül. Ekkor van olyan ζ valószínűségi változó, amelyre vonatkozólag $\xi_1, \dots, \xi_{n_1}, \eta_1, \dots, \eta_{n_2}$ egy $n_1 + n_2$ elemű minta.

Jelentse A_i azt az eseményt, hogy $\zeta \in C_i$. A B_1 illetve B_2 jelentse azt, hogy a mintavétel ξ-re illetve η-ra vonatkozik. De H_0 esetén az, hogy $\zeta \in C_i$ teljesül-e, független attól, hogy a mintavétel valójában ξ-re vagy η-ra történt. Így ekkor

$$H_0' \colon \mathrm{P}(A_i \cap B_j) = \mathrm{P}(A_i)\,\mathrm{P}(B_j) \ (i = 1, \dots, r,\ j = 1, 2)$$

is teljesül. Erre alkalmazhatjuk a függetlenségvizsgálatban leírtakat a következő kontingencia táblázattal:

	B_1	B_2	
A_1	k_1	l_1	$k_1 + l_1$
A_2	k_2	l_2	$k_2 + l_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	k_r	l_r	$k_r + l_r$
	n_1	n_2	$n_1 + n_2$

Ekkor tehát

$$\chi^2 := \tfrac{1}{n_1+n_2} \sum_{i=1}^{r} \left(\tfrac{((n_1+n_2)k_i - (k_i+l_i)n_1)^2}{(k_i+l_i)n_1} + \tfrac{((n_1+n_2)l_i - (k_i+l_i)n_2)^2}{(k_i+l_i)n_2} \right) =$$
$$= n_1 n_2 \sum_{i=1}^{r} \frac{\left(\frac{k_i}{n_1} - \frac{l_i}{n_2}\right)^2}{k_i + l_i}$$

aszimptotikusan $(r-1)(2-1) = r-1$ szabadsági fokú khi-négyzet eloszlású. Tehát $F \sim \mathbf{Khi}(r-1)$ jelöléssel, a H_0 nullhipotézisre $\chi^2 \leqslant F^{-1}(1-\alpha)$ ($\Longleftrightarrow \alpha \leqslant 1 - F(\chi^2)$) elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.5. Kétmintás előjelpróba

5.30. Feladat. Legyen (ξ, η) kétdimenziós valószínűségi vektorváltozó. Az erre vonatkozó $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ minta alapján készítsünk a

$$H_0 \colon \mathrm{P}(\xi > \eta) = \frac{1}{2}$$
$$H_1 \colon \mathrm{P}(\xi > \eta) \neq \frac{1}{2}$$

hipotézisekre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti. A feladatot oldja meg

$$H_1\colon\ \mathrm{P}(\xi > \eta) < \frac{1}{2} \quad \text{illetve} \quad H_1\colon\ \mathrm{P}(\xi > \eta) > \frac{1}{2}$$

egyoldali ellenhipotézisekre is.

Megoldás. Bár a feladatot a nemparaméteres hipotézisvizsgálatokban tárgyaljuk, egyértelmű a kapcsolata a valószínűségre vonatkozó statisztikai próbával, $A := \{\, \xi > \eta \,\}$ és $p_0 = \frac{1}{2}$ választással. Legyen

$$B := \sum_{i=1}^{n} \mathrm{I}_{\xi_i > \eta_i},$$

azaz az A esemény gyakorisága, vagy ha úgy tetszik, azon esetek száma, amikor $\xi_i - \eta_i$ előjele pozitív (innen a próba neve). Ha H_0 teljesül, akkor $B \in \mathbf{Bin}(n; \frac{1}{2})$. Legyenek az $a,\ b,\ c,\ d$ számok a legkisebb olyan pozitív egészek, amelyekre teljesülnek, hogy

$$\begin{aligned}
\sum_{i=0}^{a} \binom{n}{i} \frac{1}{2^n} &\geqslant \frac{\alpha}{2} \\
\sum_{i=0}^{b} \binom{n}{i} \frac{1}{2^n} &\geqslant 1 - \frac{\alpha}{2} \\
\sum_{i=0}^{c} \binom{n}{i} \frac{1}{2^n} &\geqslant \alpha \\
\sum_{i=0}^{d} \binom{n}{i} \frac{1}{2^n} &\geqslant 1 - \alpha.
\end{aligned}$$

Ekkor a valószínűségre vonatkozó statisztikai próbánál leírtak szerint

$$\begin{aligned}
H_1\colon\ \mathrm{P}(\xi > \eta) \neq \frac{1}{2} &\text{ esetén } a \leqslant B \leqslant b, \\
H_1\colon\ \mathrm{P}(\xi > \eta) < \frac{1}{2} &\text{ esetén } B \geqslant c \text{ és} \\
H_1\colon\ \mathrm{P}(\xi > \eta) > \frac{1}{2} &\text{ esetén } B \leqslant d
\end{aligned}$$

elfogadási tartománnyal α terjedelmű próbát kapunk. A kritikus értékek kiszámolásánál itt is alkalmazható $n \geqslant 20$ esetén a folytonossági korrekcióval megadott közelítő számítás. Eszerint

$$h(x) := \frac{1}{2}\left(n - 1 + \sqrt{n}\Phi^{-1}(x)\right)$$

jelöléssel $a \simeq \left[h(\frac{\alpha}{2})\right]$, $b \simeq \left[h(1-\frac{\alpha}{2})\right]+1$, $c \simeq [h(\alpha)]$ és $d \simeq [h(1-\alpha)]+1$.

5.3.6. Kolmogorov – Szmirnov-féle kétmintás próba

5.31. Tétel (Szmirnov-tétel)**.** *Legyenek ξ és η független valószínűségi változók, a rájuk vonatkozó minták $\xi_1, \dots, \xi_n$ és $\eta_1, \dots, \eta_n$, illetve a nekik megfelelő tapasztalati eloszlásfüggvények F_n^* és G_n^*. Ha ξ-nek és η-nak azonos az eloszlásfüggvénye és az folytonos, akkor minden $z \in \mathbb{R}_+$ esetén*

$$\lim_{n\to\infty} \mathrm{P}\Big(\sqrt{\tfrac{n}{2}} \sup_{x\in\mathbb{R}} |F_n^*(x) - G_n^*(x)| < z\Big) = 1 + 2\sum_{i=1}^{\infty} (-1)^i e^{-2i^2z^2}.$$

A bizonyítást lásd például Fazekas I. [2, 194. oldal].

5.32. *Megjegyzés.* A Szmirnov-tétel feltételeivel

$$\mathrm{P}\Big(\sqrt{\tfrac{n}{2}} \sup_{x\in\mathbb{R}} |F_n^*(x) - G_n^*(x)| < z\Big) \simeq 1 + 2\sum_{i=1}^{\infty} (-1)^i e^{-2i^2z^2}$$

közelítés már jónak tekinthető, ha $n > 30$.

A $\sqrt{\frac{n}{2}} \sup_{x\in\mathbb{R}} |F_n^*(x) - G_n^*(x)|$ pontos eloszlása is ismert (lásd pl. Fazekas I. [2, 191. oldal]), melyből $n \leqslant 30$ esetén is tudunk próbát konstruálni. Mi ezzel az esettel nem foglalkozunk.

5.33. Feladat. Legyenek ξ és η folytonos eloszlásfüggvényű független valószínűségi változók. Az ezekre vonatkozó $\xi_1, \dots, \xi_n$ illetve $\eta_1, \dots, \eta_n$ $(n > 30)$ minták alapján készítsünk a

$$H_0 \colon \xi \text{ és } \eta \text{ azonos eloszlású}$$

nullhipotézisre α terjedelmű próbát.

Megoldás. Legyenek a ξ-re illetve η-ra vonatkozó mintákhoz tartozó tapasztalati eloszlásfüggvények F_n^* illetve G_n^*, továbbá legyen

$$D := \sqrt{\tfrac{n}{2}} \sup_{x\in\mathbb{R}} |F_n^*(x) - G_n^*(x)|.$$

Ha H_0 nem teljesül, akkor D várhatóan kritikus mértékben eltávolodik 0-tól. Ezért az elfogadási tartomány legyen $D < z$ alakú, ahol $z \in \mathbb{R}_+$. A Szmirnov-tétel szerint $\mathrm{P} \in \mathcal{P}_{H_0}$ és $n > 30$ esetén

$$\mathrm{P}(D < z) \simeq K(z) \ (z \in \mathbb{R}_+),$$

ahol

$$K(z) = 1 + 2\sum_{i=1}^{\infty}(-1)^i e^{-2i^2z^2}.$$

Így $\mathrm{P}(D < z) = 1 - \alpha$ esetén $z \simeq K^{-1}(1-\alpha)$. Tehát

$$D < K^{-1}(1-\alpha) \ (\Longleftrightarrow K(D) < 1-\alpha)$$

elfogadási tartománnyal körülbelül α terjedelmű próbát kapunk.

5.3.7. Kolmogorov – Szmirnov-féle egymintás próba

A matematikai statisztika alaptétele a tapasztalati eloszlásfüggvény konvergenciájáról szól, de a konvergencia sebességéről nem ad információt. A következő Kolmogorovtól származó tétel ezt a hiányt pótolja, melyet itt bizonyítás nélkül közlünk.

5.34. Tétel (Kolmogorov-tétel)**.** *Legyen a ξ valószínűségi változó F eloszlásfüggvénye folytonos. A ξ-re vonatkozó minta legyen $\xi_1, \ldots, \xi_n$ és a neki megfelelő tapasztalati eloszlásfüggvény F_n^*. Ekkor minden $z \in \mathbb{R}_+$ esetén*

$$\lim_{n\to\infty} \mathrm{P}\Big(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)| < z\Big) = 1 + 2\sum_{i=1}^{\infty}(-1)^i e^{-2i^2z^2}.$$

5.35. *Megjegyzés.* A Kolmogorov-tétel feltételeivel

$$\mathrm{P}\Big(\sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)| < z\Big) \simeq 1 + 2\sum_{i=1}^{\infty}(-1)^i e^{-2i^2z^2}$$

közelítés már jónak tekinthető, ha $n > 30$.

5.36. Feladat. Legyen ξ folytonos eloszlásfüggvényű valószínűségi változó. Az erre vonatkozó $\xi_1, \ldots, \xi_n$ $(n > 30)$ minta alapján készítsünk a

$$H_0\colon \xi \text{ eloszlásfüggvénye } F$$

nullhipotézisre α terjedelmű próbát.

Megoldás. Legyen a tapasztalati eloszlásfüggvény F_n^*, továbbá legyen

$$D := \sqrt{n}\sup_{x\in\mathbb{R}}|F_n^*(x) - F(x)|.$$

Ha H_0 nem teljesül, akkor D várhatóan kritikus mértékben eltávolodik 0-tól. Így a kétmintás esethez hasonlóan kapjuk, hogy

$$D < K^{-1}(1-\alpha) \ (\Longleftrightarrow K(D) < 1-\alpha)$$

elfogadási tartománnyal körülbelül α terjedelmű a próba, ahol

$$K(z) = 1 + 2\sum_{i=1}^{\infty}(-1)^i e^{-2i^2z^2}.$$

6. Regressziószámítás

6.1. Regressziós görbe és regressziós felület

Jelentse η a Duna egy árhullámának tetőző vízállását Budapesten cm-ben, ξ_1 az árhullámot kiváltó csapadék mennyiségét mm-ben és ξ_2 a Duna vízállását Budapestnél az esőzés kezdetekor cm-ben. Joggal gondolhatjuk, hogy ξ_1 és ξ_2 értéke erősen behatárolja az η értékét. Keressünk olyan g függvényt, melyre teljesül, hogy

$$\eta \simeq g(\xi_1, \xi_2).$$

Az eltérés mértéke legyen

$$\mathrm{E}\big(\eta - g(\xi_1, \xi_2)\big)^2,$$

hasonlóan a $\mathrm{D}^2\,\xi = \mathrm{E}(\xi - \mathrm{E}\,\xi)^2$ szórásnégyzethez, ami a ξ és $\mathrm{E}\,\xi$ eltérésének mértéke. Ha sikerülne olyan g függvényt találni, amelyre $\mathrm{E}\big(\eta - g(\xi_1, \xi_2)\big)^2$ a lehető legkisebb, akkor ξ_1 és ξ_2 mérésével közelítőleg meg lehetne jósolni η, azaz az árhullám tetőzésének mértékét.

Általánosítva, ha az $\eta, \xi_1, \ldots, \xi_k$ valószínűségi változók esetén az a feladat, hogy adjuk meg a lehető legjobb

$$\eta \simeq g(\xi_1, \ldots, \xi_k)$$

közelítést adó g függvényt, akkor az úgy értendő, hogy az

$$\mathrm{E}\big(\eta - g(\xi_1, \ldots, \xi_k)\big)^2$$

értékét kell minimalizálni. Ez az úgynevezett *legkisebb négyzetek elve*. Az így kapott g továbbá $\xi_1, \ldots, \xi_k$ ismeretében megbecsülhető lesz η.

6.1. Tétel. *Legyenek $\eta, \xi_1, \ldots, \xi_k$ valószínűségi változók és $\mathrm{E}\,\eta^2 < \infty$. Az összes $g\colon \mathbb{R}^k \to \mathbb{R}$ Borel-mérhető függvényt figyelembe véve $\mathrm{E}\big(\eta - g(\xi_1, \ldots, \xi_k)\big)^2$ akkor a legkisebb, ha*

$$g(\xi_1, \ldots, \xi_k) = \mathrm{E}(\eta \mid \xi_1, \ldots, \xi_k).$$

Bizonyítás. Legyen $\mu := \eta - \mathrm{E}(\eta \mid \xi_1, \ldots, \xi_k)$ és $\nu := \mathrm{E}(\eta \mid \xi_1, \ldots, \xi_k) - g(\xi_1, \ldots, \xi_k)$. Ekkor

$$\begin{aligned}
\mathrm{E}\big(\eta - g(\xi_1, \ldots, \xi_k)\big)^2 &= \mathrm{E}(\mu + \nu)^2 = \mathrm{E}\,\mu^2 + 2\,\mathrm{E}(\mu\nu) + \mathrm{E}\,\nu^2 \geqslant \\
&\geqslant \mathrm{E}\,\mu^2 + 2\,\mathrm{E}(\mu\nu) = \mathrm{E}\,\mu^2 + 2\,\mathrm{E}\big(\mathrm{E}(\mu\nu \mid \xi_1, \ldots, \xi_k)\big) = \\
&= \mathrm{E}\,\mu^2 + 2\,\mathrm{E}\big(\nu\,\mathrm{E}(\mu \mid \xi_1, \ldots, \xi_k)\big),
\end{aligned}$$

másrészt

$$\begin{aligned}\mathrm{E}(\mu \mid \xi_1,\dots,\xi_k) &= \mathrm{E}\big(\eta - \mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) \mid \xi_1,\dots,\xi_k\big) = \\ &= \mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) - \mathrm{E}\big(\mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) \mid \xi_1,\dots,\xi_k\big) = \\ &= \mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) - \mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) = 0.\end{aligned}$$

Így kapjuk, hogy

$$\mathrm{E}\big(\eta - g(\xi_1,\dots,\xi_k)\big)^2 \geqslant \mathrm{E}\,\mu^2 = \mathrm{E}\big(\eta - \mathrm{E}(\eta \mid \xi_1,\dots,\xi_k)\big)^2,$$

melyből adódik az állítás.

6.2. Definíció. Ha $\eta, \xi_1, \dots, \xi_k$ valószínűségi változók, ξ_i értékkészlete R_{ξ_i} $(i = 1,\dots,k)$ és $\mathrm{E}\,\eta^2$ véges, akkor a

$$g\colon R_{\xi_1} \times \cdots \times R_{\xi_k} \to \mathbb{R}, \quad g(x_1,\dots,x_k) := \mathrm{E}(\eta \mid \xi_1 = x_1,\dots,\xi_k = x_k)$$

függvényt az η valószínűségi változó $(\xi_1,\dots,\xi_k)$-ra vonatkozó *regressziós felületének*, illetve ennek meghatározását *regressziószámításnak* nevezzük. Speciálisan $k = 1$ esetén *regressziós görbéről* beszélünk. Ha a regressziós felület lineáris függvénnyel írható le, akkor azt $k = 1$ esetén *(elsőfajú) regressziós egyenesnek*, míg $k = 2$ esetén *(elsőfajú) regressziós síknak* nevezzük.

6.3. *Megjegyzés.* Ismert, hogy $(\eta, \xi_1, \dots, \xi_k) \in \mathbf{Norm}_{k+1}(m; A)$ esetén léteznek olyan $a_1,\dots,a_k \in \mathbb{R}$ konstansok, hogy $\mathrm{E}(\eta \mid \xi_1,\dots,\xi_k) = a_1\xi_1 + \cdots + a_k\xi_k$. Tehát ha $(\eta, \xi_1, \dots, \xi_k)$ valószínűségi vektorváltozó normális eloszlású, akkor a regressziós felület egy lineáris függvénnyel írható le.

6.2. Lineáris regresszió

Ha $(\eta, \xi_1, \dots, \xi_k)$ nem normális eloszlású, akkor a legtöbb esetben a regressziós felület meghatározása igen bonyolult probléma. Ilyen esetekben azzal egyszerűsíthetjük a feladatot, hogy $\mathrm{E}\big(\eta - g(\xi_1,\dots,\xi_k)\big)^2$ minimumát csak a

$$g(x_1,\dots,x_k) = a_0 + a_1x_1 + \cdots + a_kx_k \ (a_0, a_1, \dots, a_k \in \mathbb{R})$$

alakú – azaz lineáris – függvények között keressük. Ezt a típusú regressziószámítást *lineáris regressziónak* nevezzük. A feladat megoldásában szereplő $a_0,\dots,a_k$ konstansokat a *lineáris regresszió együtthatóinak* nevezzük.

A lineáris regresszióval kapott g függvényt $k = 1$ illetve $k = 2$ esetén *másodfajú regressziós egyenesnek* illetve *másodfajú regressziós síknak* nevezzük.

Kérdés, hogy egyáltalán van-e megoldása a lineáris regressziós feladatnak. Erre ad feleletet a következő tétel.

6.4. Tétel. *Legyen* $\xi_0 \equiv 1$, $\mathrm{E}\,\eta^2 \in \mathbb{R}$, $\mathrm{E}(\eta\xi_i) \in \mathbb{R}$, $\mathrm{E}(\xi_i\xi_j) \in \mathbb{R}$ $(i, j = 0, \dots, k)$, *továbbá az*

$$R := \begin{pmatrix} \mathrm{E}(\xi_0\xi_0) & \mathrm{E}(\xi_0\xi_1) & \dots & \mathrm{E}(\xi_0\xi_k) \\ \mathrm{E}(\xi_1\xi_0) & \mathrm{E}(\xi_1\xi_1) & \dots & \mathrm{E}(\xi_1\xi_k) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\xi_k\xi_0) & \mathrm{E}(\xi_k\xi_1) & \dots & \mathrm{E}(\xi_k\xi_k) \end{pmatrix}$$

mátrix pozitív definit, azaz minden bal felső sarokdeterminánsa pozitív. Ekkor a lineáris regressziónak pontosan egy megoldása van, nevezetesen azon $g(x_1, \dots, x_k) = a_0 + a_1x_1 + \dots + a_kx_k$ *függvény, melyre*

$$a_i = \frac{\det R_i}{\det R} \ (i = 0, \dots, k),$$

ahol az R_i *mátrixot úgy kapjuk, hogy az* R *mátrix* i*-edik oszlopát kicseréljük az* $r := \big(\mathrm{E}(\eta\xi_0), \dots, \mathrm{E}(\eta\xi_k)\big)^\top$*-ra.*

Bizonyítás. A feladat azon $a_0, \dots, a_k \in \mathbb{R}$ paraméterek meghatározása, amelyek mellett $\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2$ minimális. Mivel

$$\begin{aligned} &\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = \mathrm{E}(\eta - a_0\xi_0 - \dots - a_k\xi_k)^2 = \\ &= \mathrm{E}\,\eta^2 + \sum_{i=0}^{k} a_i^2\,\mathrm{E}\,\xi_i^2 - 2\sum_{i=0}^{k} a_i\,\mathrm{E}(\eta\xi_i) + 2\sum_{i=0}^{k-1}\sum_{j=i+1}^{k} a_ia_j\,\mathrm{E}(\xi_i\xi_j), \end{aligned}$$

ezért

$$\begin{aligned} &\frac{\partial}{\partial a_l}\,\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = \\ &= 2a_l\,\mathrm{E}\,\xi_l^2 - 2\,\mathrm{E}(\eta\xi_l) + 2\sum_{i\neq l} a_i\,\mathrm{E}(\xi_i\xi_l) = \\ &= 2\sum_{i=0}^{k} a_i\,\mathrm{E}(\xi_i\xi_l) - 2\,\mathrm{E}(\eta\xi_l) \quad (l = 0, \dots, k). \end{aligned}$$

Így azt kapjuk, hogy az

$$\frac{\partial}{\partial a_l}\,\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = 0 \quad (l = 0, \dots, k)$$

egyenletrendszer ekvivalens az

$$R(a_0,\dots,a_k)^\top = r$$

egyenlettel. Mivel R pozitív definit, ezért $\det R > 0$, így a Cramer-szabály alapján ennek pontosan egy megoldása van, nevezetesen az, amely a tételben fel lett írva. Legyen

$$K := \left(\frac{\partial^2}{\partial a_l \partial a_t}\,\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2\right)_{(k+1)\times(k+1)}.$$

Mivel

$$\frac{\partial^2}{\partial a_l \partial a_t}\,\mathrm{E}(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = 2\,\mathrm{E}(\xi_l\xi_t),$$

ezért $K = 2R$. Ebből adódik, hogy K pozitív definit, azaz a kapott megoldás valóban minimumhely. Ezzel bizonyítottuk a tételt.

6.5. *Megjegyzés.* Könnyen látható, hogy $k = 1$ esetén az előző tétel feltételei teljesülnek, ha $\mathrm{E}\,\eta^2 \in \mathbb{R}$, $0 < \mathrm{D}^2\,\xi_1 < \infty$ és $\mathrm{cov}(\eta,\xi_1) \in \mathbb{R}$. Másrészt ekkor $R(a_0,a_1)^\top = r$ ekvivalens a következő egyenletrendszerrel:

$$\begin{aligned} a_0 \qquad &+ a_1\,\mathrm{E}\,\xi_1 = \mathrm{E}\,\eta,\\ a_0\,\mathrm{E}\,\xi_1 &+ a_1\,\mathrm{E}\,\xi_1^2 = \mathrm{E}(\eta\xi_1). \end{aligned}$$

Ennek a megoldása

$$a_0 = \mathrm{E}\,\eta - \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}\,\mathrm{E}\,\xi_1, \quad a_1 = \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}.$$

Így a regressziós egyenes egyenlete

$$g(x) = \mathrm{E}\,\eta - \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}\,\mathrm{E}\,\xi_1 + \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}x,$$

azaz ennek eredményeképpen a továbbiakban az

$$\eta \simeq \mathrm{E}\,\eta - \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}\,\mathrm{E}\,\xi_1 + \frac{\mathrm{cov}(\eta,\xi_1)}{\mathrm{D}^2\,\xi_1}\xi_1$$

lineáris közelítést lehet használni.

6.6. Feladat. Az $\mathrm{E}\big(\eta - g(\xi_1,\dots,\xi_k)\big)^2$ minimumát keresse meg azon g lineáris függ-

vények között, melyek átmennek az origón, azaz a

$$g(x_1,\dots,x_k)=a_1x_1+\dots+a_kx_k\ (a_1,\dots,a_k\in\mathbb{R})$$

alakú függvények között.

Megoldás. Az előző tétel bizonyításához hasonlóan kapjuk a következő állítást. Legyen $\mathrm{E}\,\eta^2\in\mathbb{R}$, $\mathrm{E}(\eta\xi_i)\in\mathbb{R}$, $\mathrm{E}(\xi_i\xi_j)\in\mathbb{R}$ $(i,j=1,\dots,k)$, továbbá az

$$R':=\begin{pmatrix}\mathrm{E}(\xi_1\xi_1) & \mathrm{E}(\xi_1\xi_2) & \dots & \mathrm{E}(\xi_1\xi_k)\\ \mathrm{E}(\xi_2\xi_1) & \mathrm{E}(\xi_2\xi_2) & \dots & \mathrm{E}(\xi_2\xi_k)\\ \vdots & \vdots & \ddots & \vdots\\ \mathrm{E}(\xi_k\xi_1) & \mathrm{E}(\xi_k\xi_2) & \dots & \mathrm{E}(\xi_k\xi_k)\end{pmatrix}$$

mátrix pozitív definit, azaz minden bal felső sarokdeterminánsa pozitív. Ekkor a feladatnak pontosan egy megoldása van, nevezetesen azon $g(x_1,\dots,x_k)=a_1x_1+\dots+a_kx_k$ függvény, melyre

$$a_i=\frac{\det R_i'}{\det R'}\ (i=1,\dots,k),$$

ahol az R_i' mátrixot úgy kapjuk, hogy az R' mátrix i-edik oszlopát kicseréljük az $r':=\big(\mathrm{E}(\eta\xi_1),\dots,\mathrm{E}(\eta\xi_k)\big)^\top$-ra. Speciálisan $k=1$ esetén $a_1=\frac{\mathrm{E}(\eta\xi_1)}{\mathrm{E}(\xi_1^2)}$.

6.7. Feladat. Legyenek $t_0,\dots,t_k\in\mathbb{R}$ rögzített konstansok. Az $\mathrm{E}\big(\eta-g(\xi_1,\dots,\xi_k)\big)^2$ minimumát keresse meg azon lineáris g függvények között, melyekre teljesül, hogy $g(t_1,\dots,t_k)=t_0$. Ez az úgynevezett *fixpontos lineáris regresszió*. A megoldást adó g függvényt $k=1$ illetve $k=2$ esetén *fixpontos regressziós egyenesnek* illetve *fixpontos regressziós síknak* nevezzük.

Megoldás. Könnyen látható, hogy

$$g(x_1,\dots,x_k)=a_0+a_1x_1+\dots+a_kx_k\ (a_0,\dots,a_k\in\mathbb{R})\quad\text{és}\quad g(t_1,\dots,t_k)=t_0$$

pontosan akkor teljesülnek egyszerre, ha

$$g(x_1,\dots,x_k)-t_0=a_1(x_1-t_1)+\dots+a_k(x_k-t_k)\ (a_1,\dots,a_k\in\mathbb{R}).$$

(Vegyük észre, hogy $t_0=\dots=t_k=0$ esetén az előző feladatot kapjuk vissza.) Így az előző feladat megoldásában $\eta,\xi_1,\dots,\xi_k$ helyébe $\eta-t_0,\xi_1-t_1,\dots,\xi_k-t_k$ írva, adódnak a feltételnek eleget tevő $a_1,\dots,a_k$ együtthatók.

6.3. A lineáris regresszió együtthatóinak becslése

Az előzőekben a lineáris regresszió együtthatóit az $\eta, \xi_1, \ldots, \xi_k$ valószínűségi változók és azok kapcsolatának ismeretében határoztuk meg. Ezekről viszont a gyakorlatban csak nagyon ritkán van elegendő információnk. Így ekkor az $(\eta, \xi_1, \ldots, \xi_k)$-ra vonatkozó minta alapján kell ezeket az együtthatókat megbecsülni. Legyen ez a minta

$$(\eta_i, \xi_{i1}, \ldots, \xi_{ik}) \quad i = 1, \ldots, n.$$

Bevezetjük a következő jelöléseket:

$$\begin{aligned} a &:= (a_0, \ldots, a_k)^\top \\ Y &:= (\eta_1, \ldots, \eta_n)^\top \\ X &:= \begin{pmatrix} 1 & \xi_{11} & \ldots & \xi_{1k} \\ 1 & \xi_{21} & \ldots & \xi_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_{n1} & \ldots & \xi_{nk} \end{pmatrix}. \end{aligned}$$

A becslés alapja az, hogy az $\mathrm{E}(\eta - a_0 - a_1\xi_1 - \cdots - a_k\xi_k)^2$ várható értéket az

$$\frac{1}{n}\sum_{i=1}^{n}(\eta_i - a_0 - a_1\xi_{i1} - \cdots - a_k\xi_{ik})^2$$

átlaggal becsüljük. Vegyük észre, hogy ez az átlag $\frac{1}{n}\|Y - Xa\|^2$ alakban is írható, ahol $\|(v_1, \ldots, v_n)^\top\| = \sqrt{v_1^2 + \cdots + v_n^2}$ a $(v_1, \ldots, v_n)^\top$ oszlopvektor hossza. Így a feladat azon a-nak a megtalálása, amely mellett $\|Y - Xa\|$ minimális.

Jelölje L az Xa lineáris leképezés képterét, amely a $\{\, v^\top : v \in \mathbb{R}^n \,\}$ vektortér egy altere. Mivel $\|Y - Xa\|$ az Y és az Xa távolsága, ezért ez akkor lesz minimális, ha Xa az Y merőleges vetülete L-re, azaz $Y - Xa$ merőleges L-re. Ez pontosan azt jelenti, hogy $Y - Xa$ merőleges Xb-re, minden $b_0, \ldots, b_k \in \mathbb{R},\ b = (b_0, \ldots, b_k)^\top$ esetén. Tehát

$$\begin{aligned} (Xb)^\top(Y - Xa) &= 0 \\ b^\top X^\top(Y - Xa) &= 0 \\ b^\top X^\top Y &= b^\top X^\top Xa \\ X^\top Y &= X^\top Xa \end{aligned}$$

Az utolsó lépésben azért hagyható el $b^\top$, mert az egyenlet bármely b-re teljesül. Az a-ra vonatkozó $X^\top Y = X^\top X a$ egyenlet az úgynevezett *normálegyenlet*, melynek $\widehat{a} = (\widehat{a}_0, \dots, \widehat{a}_k)^\top$-val jelölt megoldása szolgáltatja a lineáris regresszió együtthatóinak becslését. Nyilván, ha $X^\top X$ invertálható mátrix, akkor

$$\widehat{a} = (X^\top X)^{-1} X^\top Y.$$

6.8. Példa. Számolja ki $k = 1$ esetén a lineáris regresszió együtthatóinak becslését.

Megoldás. Az (η, ξ_1)-re vonatkozó minta (η_i, ξ_{i1}) $i = 1, \dots, n$,

$$a = (a_0, a_1)^\top$$
$$Y = (\eta_1, \dots, \eta_n)^\top$$
$$X = \begin{pmatrix} 1 & \xi_{11} \\ 1 & \xi_{21} \\ \vdots & \vdots \\ 1 & \xi_{n1} \end{pmatrix}.$$

Némi számolással kapjuk, hogy az $X^\top Y = X^\top X a$ normálegyenlet ekvivalens a következő egyenletrendszerrel:

$$(\xi_{11} + \dots + \xi_{n1}) a_0 + (\xi_{11}^2 + \dots + \xi_{n1}^2) a_1 = \xi_{11}\eta_1 + \dots + \xi_{n1}\eta_n,$$
$$n a_0 + (\xi_{11} + \dots + \xi_{n1}) a_1 = \eta_1 + \dots + \eta_n.$$

Ennek megoldása, és így az a becslése

$$\widehat{a}_0 = \overline{\eta} - \frac{\mathrm{Cov}_n(\eta, \xi_1)}{S^2_{\xi_1, n}} \overline{\xi_1},$$
$$\widehat{a}_1 = \frac{\mathrm{Cov}_n(\eta, \xi_1)}{S^2_{\xi_1, n}}.$$

Ennek alapján a továbbiakban az $\eta \simeq \widehat{a}_0 + \widehat{a}_1 \xi_1$ közelítést fogjuk használni.

6.9. *Megjegyzés.* Összehasonlítva az előbb kapott $\widehat{a}_0$ és $\widehat{a}_1$ becsléseket a korábban kapott elméleti értékekkel, azt láthatjuk, hogy tulajdonképpen a várható értéket mintaátlaggal, a szórásnégyzetet tapasztalati szórásnégyzettel és a kovarianciát a tapasztalati kovarianciával becsültük.

6.10. Feladat. Adjon becslést az $(\eta, \xi_1, \dots, \xi_k)$ valószínűségi vektorváltozóra vonatkozó $(\eta_i, \xi_{i1}, \dots, \xi_{ik})$, $i = 1, \dots, n$ minta alapján a fixpontos lineáris regresszió

együtthatóira.

Megoldás. A feladat tehát rögzített $t_0, \dots, t_k \in \mathbb{R}$ esetén olyan

$$g(x_1, \dots, x_k) = t_0 + a_1(x_1 - t_1) + \dots + a_k(x_k - t_k) \ (a_1, \dots, a_k \in \mathbb{R})$$

függvényt találni, melyre

$$\sum_{i=1}^{n} \big(\eta_i - g(\xi_{i1}, \dots, \xi_{ik})\big)^2$$

minimális. Legyen először $t_0 = \dots = t_k = 0$. Ekkor $g(x_1, \dots, x_k) = a_1 x_1 + \dots + a_k x_k$, így a lineáris regresszió együtthatóinak becsléséhez hasonlóan kapjuk, hogy

$$Y := (\eta_1, \dots, \eta_n)^\top$$
$$X' := \begin{pmatrix} \xi_{11} & \dots & \xi_{1k} \\ \xi_{21} & \dots & \xi_{2k} \\ \vdots & \ddots & \vdots \\ \xi_{n1} & \dots & \xi_{nk} \end{pmatrix}$$

jelölésekkel, ha $X'^\top X'$ invertálható mátrix, akkor

$$(\widehat{a}_1, \dots, \widehat{a}_k)^\top = (X'^\top X')^{-1} X'^\top Y.$$

Speciálisan $k = 1$ esetén

$$\widehat{a}_1 = \frac{\sum_{i=1}^n \xi_{i1}\eta_i}{\sum_{i=1}^n \xi_{i1}^2} = \frac{\mathrm{Cov}_n(\eta, \xi_1) + \overline{\xi_1}\,\overline{\eta}}{S_{\xi_1,n} + \overline{\xi_1}^2},$$

így ekkor az $\eta \simeq \widehat{a}_1 \xi_1$ közelítést fogjuk használni.

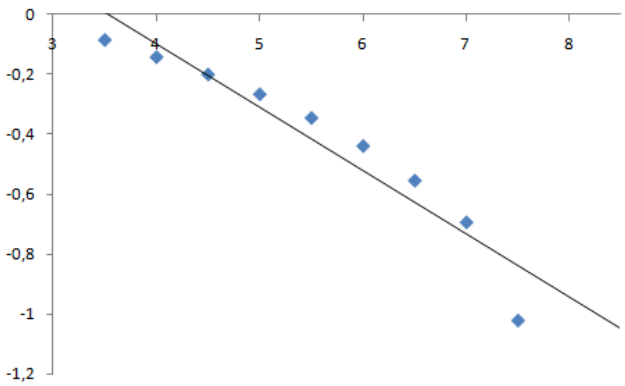
Tetszőleges $t_0, \dots, t_k \in \mathbb{R}$ esetén a fixpontot transzformáljuk az origóra, így az előző megoldásban csak annyit kell változtatni, hogy

$$Y := (\eta_1 - t_0, \dots, \eta_n - t_0)^\top$$
$$X' := \begin{pmatrix} \xi_{11} - t_1 & \dots & \xi_{1k} - t_k \\ \xi_{21} - t_1 & \dots & \xi_{2k} - t_k \\ \vdots & \ddots & \vdots \\ \xi_{n1} - t_1 & \dots & \xi_{nk} - t_k \end{pmatrix}$$

jelöléseket használunk.

6.4. Nemlineáris regresszió

A lineáris regressziós közelítés sokszor nagyon durva becslést adhat. $k = 1$ esetén a mintarealizációt jelentő pontok ábrázolásával jól szemléltethető ez a probléma.



Itt jól látszik, hogy ebben az esetben „hiba" lenne lineáris regressziót alkalmazni. Ilyenkor érdemes megtippelni, hogy milyen típusú függvény közelíti jobban a kapcsolatot a lineárisnál (hatvány, exponenciális, logaritmus, stb.), majd a regressziós függvény keresését le kell szűkíteni erre a csoportra.

Néhány esetben valamilyen transzformációval ez a keresés visszavezethető a lineáris esetre. Most csak ilyen eseteket vizsgálunk, és azt is csak a $k = 1$ (egyváltozós) esetben.

6.4.1. Polinomos regresszió

Ebben az esetben a regressziós függvényt

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_r x^r \quad (a_0, \ldots, a_r \in \mathbb{R}_+)$$

alakban keressük. Ekkor az $a_0, \ldots, a_r$ együtthatókat az $\eta, \xi_1, \xi_1^2, \ldots, \xi_1^r$ között végrehajtott lineáris regresszió adja.

6.4.2. Hatványkitevős regresszió

Ebben az esetben a regressziós függvényt

$$y = ax^b \quad (a \in \mathbb{R}_+,\ b \in \mathbb{R})$$

alakban keressük. Ez azzal ekvivalens, hogy

$$\ln y = \ln a + b \ln x,$$

így ekkor $\ln \eta$ és $\ln \xi_1$ között lineáris regressziót végrehajtva, a kapott a_0, a_1 együtthatókra teljesül, hogy $a_0 = \ln a,\ a_1 = b$, azaz

$$a = e^{a_0}, \quad b = a_1.$$

Ebből a korábbiak alapján

$$a = \exp\left(\mathrm{E}(\ln \eta) - \frac{\mathrm{cov}(\ln \eta, \ln \xi_1)}{\mathrm{D}^2(\ln \xi_1)} \mathrm{E}(\ln \xi_1)\right),$$
$$b = \frac{\mathrm{cov}(\ln \eta, \ln \xi_1)}{\mathrm{D}^2(\ln \xi_1)}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\widehat{a} = \exp\left(\overline{\ln \eta} - \frac{\mathrm{Cov}_n(\ln \eta, \ln \xi_1)}{S^2_{\ln \xi_1, n}} \overline{\ln \xi_1}\right),$$
$$\widehat{b} = \frac{\mathrm{Cov}_n(\ln \eta, \ln \xi_1)}{S^2_{\ln \xi_1, n}}.$$

6.4.3. Exponenciális regresszió

Ebben az esetben a regressziós függvényt

$$y = ab^x \quad (a, b \in \mathbb{R}_+)$$

alakban keressük. Ez azzal ekvivalens, hogy

$$\ln y = \ln a + (\ln b)x,$$

így ekkor $\ln \eta$ és ξ_1 között lineáris regressziót végrehajtva, a kapott a_0, a_1 együtthatókra teljesül, hogy $a_0 = \ln a,\ a_1 = \ln b$, azaz

$$a = e^{a_0}, \quad b = e^{a_1}.$$

Ebből a korábbiak alapján

$$a = \exp\left(\mathrm{E}(\ln\eta) - \frac{\operatorname{cov}(\ln\eta, \xi_1)}{\mathrm{D}^2\,\xi_1}\,\mathrm{E}\,\xi_1\right),$$
$$b = \exp\left(\frac{\operatorname{cov}(\ln\eta, \xi_1)}{\mathrm{D}^2\,\xi_1}\right).$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\widehat{a} = \exp\left(\overline{\ln\eta} - \frac{\operatorname{Cov}_n(\ln\eta, \xi_1)}{S^2_{\xi_1,n}}\overline{\xi_1}\right),$$
$$\widehat{b} = \exp\left(\frac{\operatorname{Cov}_n(\ln\eta, \xi_1)}{S^2_{\xi_1,n}}\right).$$

6.4.4. Logaritmikus regresszió

Ebben az esetben a regressziós függvényt

$$y = a + b\ln x \quad (a, b \in \mathbb{R})$$

alakban keressük. Így ekkor η és $\ln\xi_1$ között lineáris regressziót végrehajtva, a korábbiak alapján

$$a = \mathrm{E}\,\eta - \frac{\operatorname{cov}(\eta, \ln\xi_1)}{\mathrm{D}^2(\ln\xi_1)}\,\mathrm{E}(\ln\xi_1),$$
$$b = \frac{\operatorname{cov}(\eta, \ln\xi_1)}{\mathrm{D}^2(\ln\xi_1)}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\widehat{a} = \overline{\eta} - \frac{\operatorname{Cov}_n(\eta, \ln\xi_1)}{S^2_{\ln\xi_1,n}}\overline{\ln\xi_1},$$
$$\widehat{b} = \frac{\operatorname{Cov}_n(\eta, \ln\xi_1)}{S^2_{\ln\xi_1,n}}.$$

6.4.5. Hiperbolikus regresszió

Ebben az esetben a regressziós függvényt

$$y = \frac{1}{a + bx} \quad (a, b \in \mathbb{R})$$

alakban keressük. Ez azzal ekvivalens, hogy

$$y^{-1} = a + bx,$$

így ekkor η^{-1} és ξ_1 között lineáris regressziót végrehajtva, a korábbiak alapján

$$a = \mathrm{E}(\eta^{-1}) - \frac{\mathrm{cov}(\eta^{-1}, \xi_1)}{\mathrm{D}^2\,\xi_1}\,\mathrm{E}\,\xi_1,$$
$$b = \frac{\mathrm{cov}(\eta^{-1}, \xi_1)}{\mathrm{D}^2\,\xi_1}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\widehat{a} = \overline{\eta^{-1}} - \frac{\mathrm{Cov}_n(\eta^{-1}, \xi_1)}{S^2_{\xi_1,n}}\overline{\xi_1},$$
$$\widehat{b} = \frac{\mathrm{Cov}_n(\eta^{-1}, \xi_1)}{S^2_{\xi_1,n}}.$$

Irodalomjegyzék

[1] Borovkov, A. A.: Matematikai statisztika, Typotex Kiadó, 1999.

[2] Fazekas I. (szerk.): Bevezetés a matematikai statisztikába, Kossuth Egyetemi Kiadó, Debrecen, 2000.

[3] Fazekas I.: Valószínűségszámítás, Kossuth Egyetemi Kiadó, Debrecen, 2000.

[4] Halmos, P. R., Mértékelmélet, Gondolat, Budapest, 1984.

[5] Hunyadi L., Mundruczó Gy., Vita L.: Statisztika, Aula Kiadó, Budapesti Közgazdaságtudományi Egyetem, 1996.

[6] Johnson, N. L., Kotz, S.: Distributions in statistics, Continuous univariate distributions, Houghton Miffin, Boston, 1970.

[7] Kendall, M. G., Stuart, A.: The theory of advanced statistics I–III, Griffin, London, 1961.

[8] Lukács O.: Matematikai statisztika példatár, Műszaki Könyvkiadó, Budapest, 1987.

[9] Meszéna Gy., Ziermann M.: Valószínűségelmélet és matematikai statisztika, Közgazdasági és Jogi Könyvkiadó, Budapest, 1981.

[10] Mogyoródi J., Michaletzky Gy. (szerk.): Matematikai statisztika, Nemzeti Tankönyvkiadó, Budapest, 1995.

[11] Mogyoródi J., Somogyi Á.: Valószínűségszámítás, Tankönyvkiadó, Budapest, 1982.

[12] Prékopa A.: Valószínűségelmélet műszaki alkalmazásokkal, Műszaki Könyvkiadó, Budapest, 1962.

[13] Rényi A.: Valószínűségszámítás, Tankönyvkiadó, Budapest, 1966.

[14] Rudin, W.: A matematikai analízis alapjai, Műszaki Könyvkiadó, Budapest, 1978.

[15] Shiryayev, A. N.: Probability, Springer-Verlag, New York, 1984.

[16] Terdik Gy.: Előadások a matematikai statisztikából, mobiDIÁK könyvtár, Debreceni Egyetem, 2005. http://mobidiak.inf.unideb.hu

[17] Tómács Tibor: Matematikai statisztika gyakorlatok, Eszterházy Károly Főiskola, Eger, 2012.

[18] Vincze I.: Matematikai statisztika, Tankönyvkiadó, Budapest, 1971.