

Tómacs Tibor

Matematikai statisztika



Tómacs Tibor

Matematikai statisztika

Átdolgozott kiadás

Utolsó módosítás:
2025. június 17.

A jegyzet szabadon letölthető az alábbi linkről:

https://tomacstibor.uni-eszterhazy.hu/tananyagok/Matematikai_statisztika.pdf

Eger, 2025

Szerző:
Dr. Tómacs Tibor
egyetemi docens
Eszterházy Károly Katolikus Egyetem

Bíráló:
Dr. Sztrik János
egyetemi tanár
Debreceni Egyetem

Az első kiadás a TÁMOP-4.1.2-08/1/A-2009-0038 támogatásával készült 2012-ben.

Nemzeti Fejlesztési Ügynökség
www.ujszachenyiterv.gov.hu
06 40 638 638



MAGYARORSZÁG MEGÚJUL



A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Tartalomjegyzék

Előszó	6
Jelölések	7
1. Valószínűesszámitás	9
1.1. Valószínűégi mező	9
1.1.1. Véletlen esemény	9
1.1.2. Valószínűség	10
1.2. Valószínűégi változó	11
1.3. Eloszlás- és sűrűségfüggvény	11
1.4. Várható érték, szórásnégyzet	13
1.5. Valószínűégi vektorváltozók	14
1.6. Feltételes várható érték	15
1.7. Független valószínűégi változók	16
1.8. Kovariancia és korrelációs együttható	17
1.9. Nevezetes eloszlások	18
1.9.1. Diszkrét egyenletes eloszlás	18
1.9.2. Karakterisztikus eloszlás	18
1.9.3. Binomiális eloszlás	18
1.9.4. Poisson-eloszlás	19
1.9.5. Hipergeometrikus eloszlás	20
1.9.6. Egyenletes eloszlás	21
1.9.7. Exponenciális eloszlás	23
1.9.8. Gamma-eloszlás	24
1.9.9. Normális eloszlás	26
1.9.10. Többdimenziós normális eloszlás	29
1.9.11. Khi-négyzet eloszlás	30
1.9.12. t-eloszlás	31
1.9.13. Cauchy-eloszlás	32
1.9.14. F-eloszlás	32
1.10. Nagy számok törvényei	33
1.11. Centrális határeloszlási tétel	36
2. A matematikai statisztika alapfogalmai	38
2.1. Minta és mintarealizáció	39
2.2. Tapasztalati eloszlásfüggvény	40

2.3.	Tapasztalati eloszlás, sűrűséghisztogram	44
2.4.	Statisztikák	46
3.	Pontbecslések	50
3.1.	A pontbecslés feladata és jellemzői	50
3.1.1.	Várható érték becslése	53
3.1.2.	Valószínűség becslése	56
3.1.3.	Szórásnégyzet becslése	57
3.2.	Információs határ	58
3.3.	Pontbecslési módszerek	65
3.3.1.	Momentumok módszere	65
3.3.2.	Maximum likelihood becslés	67
4.	Intervallumbecslések	71
4.1.	Az intervallumbecslés feladata	71
4.2.	Konfidenciaintervallum a normális eloszlás paramétereire	72
4.3.	Konfidenciaintervallum az exponenciális eloszlás paraméterére	76
4.4.	Konfidenciaintervallum valószínűségre	77
4.5.	Általános módszer konfidenciaintervallum készítésére	78
5.	Hipotézisvizsgálatok	80
5.1.	A hipotézisvizsgálat feladata és jellemzői	80
5.1.1.	Null- illetve ellenhipotézis	80
5.1.2.	Statisztikai próba, elfogadási és kritikus tartomány	80
5.1.3.	Statisztikai próba terjedelme és torzítatlansága	81
5.1.4.	Próbastatisztika	81
5.1.5.	A statisztikai próba menete	82
5.1.6.	A nullhipotézis és az ellenhipotézis megválasztása	82
5.1.7.	A próba erőfüggvénye és konzisztenciája	83
5.2.	Paraméteres hipotézisvizsgálatok	83
5.2.1.	Egymintás u-próba	83
5.2.2.	Kétmintás u-próba	86
5.2.3.	Egymintás t-próba	88
5.2.4.	Kétmintás t-próba	89
5.2.5.	Scheffé-módszer	91
5.2.6.	Welch-próba	93
5.2.7.	F-próba	93
5.2.8.	Khi-négyzet próba normális eloszlás szórására	96
5.2.9.	Statisztikai próba exponenciális eloszlás paraméterére	97
5.2.10.	Statisztikai próba valószínűségre	99
5.3.	Nemparaméteres hipotézisvizsgálatok	102
5.3.1.	Tiszta illeszkedésvizsgálat	102
5.3.2.	Becsléses illeszkedésvizsgálat	103
5.3.3.	Függetlenségvizsgálat	104
5.3.4.	Homogenitásvizsgálat	106
5.3.5.	Kétmintás előjelpróba	107

5.3.6.	Kolmogorov – Szmirnov-féle kétmintás próba	108
5.3.7.	Kolmogorov – Szmirnov-féle egymintás próba	109
5.4.	Szórásanalízis	109
5.4.1.	Egyszeres osztályozás (I. típusú modell)	111
5.4.2.	Kétszeres osztályozás interakció nélkül (I. típusú modell) . . .	118
5.4.3.	Kétszeres osztályozás interakcióval (I. típusú modell), kiegyen- súlyozott elrendezés esetén	120
6.	Regressziószámítás	124
6.1.	Regressziós görbe és regressziós felület	124
6.2.	Lineáris regresszió	125
6.3.	A lineáris regresszió együtthatóinak becslése	128
6.4.	Nemlineáris regresszió	130
6.4.1.	Polinomos regresszió	131
6.4.2.	Hatványkitevős regresszió	131
6.4.3.	Exponenciális regresszió	131
6.4.4.	Logaritmikus regresszió	132
6.4.5.	Hiperbolikus regresszió	132
	Irodalomjegyzék	134

Előszó

Ez a tananyag az Eszterházy Károly Katolikus Egyetem matematikai statisztika előadásaiból készült. Az összeállításnál nem az volt cél, hogy a matematikai statisztika minden fontos ágát ismertessük, inkább arra törekedtünk, hogy a taglalt témakörök mindegyikére kellő idő jusson az egy féléves kurzus alatt.

A *Valószínűségszámítás* című fejezet nem kerül ismertetésre az előadáson, a célja azoknak a fontos fogalmaknak az összefoglalása, melyekre szükségünk lesz a matematikai statisztika megértéséhez. Ennek átismétlését az Olvasóra bízuk. 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 olyan tételek, amelyeknek a bizonyításán érdemes önállóan is gondolkodni, mielőtt a megoldást elolvasnánk.

Ehhez a tananyaghoz kapcsolódik Tómacs 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
$\lim_{x \rightarrow a+0} f(x)$	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

Valószínűesszámitás

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

$\text{Norm}(m; \sigma)$	az m várható értékű és σ szórású normális eloszlású valószínűségi változók halmaza
$\text{Norm}_d(m; A)$	az m és A paraméterű d -dimenziós normális eloszlású valószínűségi vektorváltozók halmaza
$\text{Gamma}(r; \lambda)$	az r -edrendű λ paraméterű gamma-eloszlású valószínűségi változók halmaza
$\text{Khi}(s)$	az s szabadsági fokú khi-négyzet eloszlású valószínűségi változók halmaza
$\text{T}(s)$	az s szabadsági fokú t-eloszlású valószínűségi változók halmaza
$\text{F}(s_1; s_2)$	az s_1 és s_2 szabadsági fokú F-eloszlású valószínűségi változók halmaza
$F[V]$	Ha ξ valószínűségi változó és V a ξ -vel azonos eloszlású valószínűségi változók halmaza, akkor $F[V]$ a V -beli valószínűségi változók közös eloszlásfüggvényét jelenti. Például $\Phi = F[\text{Norm}(0; 1)]$.

Matematikai statisztika

$(\Omega, \mathcal{F}, \mathcal{P})$	statisztikai mező
F_n^*	tapasztalati eloszlásfüggvény
$\bar{\xi}$	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
$\xi_1^*, \dots, \xi_n^*$	rendezett minta
$\text{Cov}_n(\xi, \eta)$	tapasztalati kovariancia
$\text{Corr}_n(\xi, \eta)$	tapasztalati korrelációs együttható
Θ	paramétertér
P_ϑ	a ϑ paraméterhez tartozó valószínűség
E_ϑ	a ϑ paraméterhez tartozó várható érték
$D_\vartheta, D_\vartheta^2$	a ϑ paraméterhez tartozó szórás illetve szórásnégyzet
f_ϑ, F_ϑ	a ϑ paraméterhez tartozó sűrűség- illetve eloszlásfüggvény
I_n	Fisher-féle információmennyiség
l_n	likelihood függvény
L_n	loglikelihood függvény
$\hat{\vartheta}$	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. fejezet

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 $\bar{A}$ -val. Az $A \cup \bar{A}$ biztosan bekövetkezik, ezért ezt *biztos eseménynek* nevezzük és Ω -val jelöljük. Ebből látható, hogy $\bar{A}$ 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.

1.1. Definíció. Legyen Ω egy nem üres halmaz és $\mathcal{F}$ részhalmaza az Ω hatványhalmazának. Tegyük fel, hogy teljesülnek a következők:

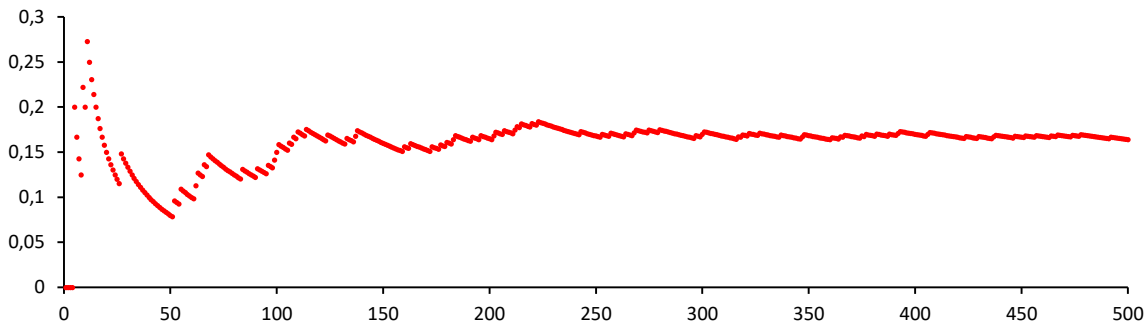
- (1) $\Omega \in \mathcal{F}$;
- (2) Ha $A \in \mathcal{F}$, akkor $\bar{A} \in \mathcal{F}$, ahol $\bar{A} = \Omega \setminus A$;
- (3) Ha $A_i \in \mathcal{F}$ ($i \in \mathbb{N}$), akkor $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Ekkor $\mathcal{F}$ -et 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 A maga után vonja B -t.

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 (lásd az 1.1. ábrát). Bernoulli azt tapasztal



1.1. ábra. Relatív gyakoriság

talta, 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 konstanszt a figyelt esemény *valószínűségének* fogjuk nevezni.

A továbbiakban $P(A)$ jelölje az A esemény bekövetkezésének valószínűségét. Könnyen látható, hogy $P(A) \geq 0$ 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 $P: \mathcal{F} \rightarrow [0, \infty)$ olyan függvény, melyre teljesülnek a következők:

- (1) $P(\Omega) = 1$;
- (2) $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$, ha $A_i \in \mathcal{F}$ páronként diszjunktak.

Ekkor a P függvényt *valószínűségnek*, a $P(A)$ számot az A esemény *valószínűségének*, illetve az $(\Omega, \mathcal{F}, P)$ rendezett hármaszt *valószínűségi mezőnek* nevezzük. Ha egy $A \in \mathcal{F}$ esetén $P(A) = 1$ teljesül, akkor azt mondjuk, hogy A *majdnem biztosan teljesül*.

Ha $(\Omega, \mathcal{F}, P)$ valószínűségi mező, akkor belátható, hogy $P(\emptyset) = 0$, 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 $(\Omega, \mathcal{F}, P)$ 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

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

halmazt, továbbá legyen $P_n : \mathcal{F}_n \rightarrow \mathbb{R}$ olyan valószínűség, melyre minden $A_i \in \mathcal{F}$ ($i = 1, \dots, n$) esetén

$$P_n(A_1 \times \cdots \times A_n) = P(A_1) \cdots P(A_n)$$

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

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

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 $\xi : \Omega \rightarrow \mathbb{R}$ 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 $(\Omega, \mathcal{F}, P)$ valószínűségi mező és $\xi: \Omega \rightarrow \mathbb{R}$ egy valószínűségi változó. Ekkor a ξ *eloszlásfüggvénye*

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

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) $\lim_{x \rightarrow \infty} F(x) = 1$;
- (d) $\lim_{x \rightarrow -\infty} F(x) = 0$.

1.7. Tétel. Ha egy tetszőleges $F: \mathbb{R} \rightarrow \mathbb{R}$ 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 $F: \mathbb{R} \rightarrow \mathbb{R}$ 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) $P(a \leq \xi < b) = F(b) - F(a)$ minden $a, b \in \mathbb{R}$, $a < b$ esetén;
- (2) $\lim_{x \rightarrow a+0} F(x) = F(a) + P(\xi = a)$ minden $a \in \mathbb{R}$ esetén;
- (3) $P(\xi = a) = 0$ 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 $P(\xi = k)$, $k \in R_\xi$ értékekkel. A $k \mapsto P(\xi = k)$, $k \in R_\xi$ 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 $f: \mathbb{R} \rightarrow [0, \infty)$ függvény, melyre

$$F(x) = \int_{-\infty}^x f(t) dt$$

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

1.11. Tétel. Ha a ξ abszolút folytonos valószínűségi változó eloszlásfüggvénye F és sűrűségfüggvénye f , akkor F folytonos (következésképpen $P(\xi = x) = 0$, $\forall x \in \mathbb{R}$) 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

$$(1) P(a < \xi < b) = \int_a^b f(x) dx \text{ minden } a, b \in \mathbb{R}, a < b \text{ esetén};$$

$$(2) \int_{-\infty}^{\infty} f(x) dx = 1.$$

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

Ezen két tétel alapján jogos a következő elnevezés:

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

1.4. Várható érték, szórásnégyzet

A valószínűségi változók fontos paramétere a valószínűség szerinti integrálja.

1.15. Definíció. Legyen $(\Omega, \mathcal{F}, P)$ valószínűségi mező és $\xi: \Omega \rightarrow \mathbb{R}$ egy valószínűségi változó. Ha az $\int \xi dP$ integrál létezik akkor azt $E\xi$ 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 végeességét, 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

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

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) $E\xi$ -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

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

továbbá ekkor

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

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

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

továbbá ekkor

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

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

$$E(a\xi + b\eta) = a E \xi + b E \eta.$$

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

$$g(E \xi) \leq E g(\xi).$$

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

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

feltéve, hogy ezek a várható értékek léteznek.

1.23. Tétel. Ha ξ -nek létezik a szórásnégyzete, akkor

- (1) $D^2 \xi = E \xi^2 - E^2 \xi$;
- (2) $D(a\xi + b) = |a| D \xi$, ahol $a, b \in \mathbb{R}$.

1.5. Valószínűségi vektorváltozók

1.24. Definíció. Ha $\xi_1, \dots, \xi_d$ tetszőleges valószínűségi változók, akkor a $(\xi_1, \dots, \xi_d)$ rendezett elem d -est (d -dimenziós) valószínűségi vektorváltozónak nevezzük.

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

$$F: \mathbb{R}^d \rightarrow \mathbb{R}, \quad F(x_1, \dots, x_d) := P(\xi_1 < x_1, \dots, \xi_d < x_d).$$

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

$$F(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(t_1, \dots, t_d) dt_1 \cdots dt_d$$

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

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

$$E g(\xi_1, \dots, \xi_d) = \int_{\mathbb{R}^d} g(x_1, \dots, x_d) f(x_1, \dots, x_d) dx_1 \cdots dx_d$$

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

1.6. Feltételes várható érték

A feltételes várható értéket az egyszerűség kedvéért csak két speciális esetben definiáljuk. Az általános definíciót lásd pl. [11].

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

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

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

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

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

1.29. Tétel. A feltételes várható értékre teljesülnek a következők:

- (1) $E\eta = E(E(\eta \mid \xi_1, \dots, \xi_k))$;
- (2) $E(a\xi + b\eta \mid \xi_1, \dots, \xi_k) = a E(\xi \mid \xi_1, \dots, \xi_k) + b E(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan, minden $a, b \in \mathbb{R}$ esetén;
- (3) $E(E(\eta \mid \xi_1, \dots, \xi_k) \mid \xi_1, \dots, \xi_k) = E(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan;
- (4) $E(\xi\eta \mid \xi_1, \dots, \xi_k) = \xi E(\eta \mid \xi_1, \dots, \xi_k)$ majdnem biztosan.

1.7. Független valószínűségi változók

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

1.30. Definíció. A $\xi_1, \dots, \xi_n$ valószínűségi változókat *függetleneknek* nevezzük, ha

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

minden $x_1, \dots, x_n \in \mathbb{R}$ esetén teljesül. A $\xi_1, \dots, \xi_n$ 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 $\xi = (\xi_1, \dots, \xi_d)$ egy valószínűségi vektorváltozó és $x = (x_1, \dots, x_d) \in \mathbb{R}^d$. Ekkor a $\xi < x$ esemény alatt azt értjük, hogy a $\xi_k < x_k$ események minden $k = 1, \dots, d$ esetén teljesülnek.

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

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

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

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

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

teljesül minden $x_1 \in R_{\xi_1}, \dots, x_n \in R_{\xi_n}$ esetén.

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

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

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

1.34. 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*

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

1.35. 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

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

1.36. 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

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

1.8. Kovariancia és korrelációs együttható

1.37. Definíció. A ξ és η valószínűségi változók kovarianciája

$$\text{cov}(\xi, \eta) := E((\xi - E\xi)(\eta - E\eta)),$$

feltéve, hogy ezek a várható értékek léteznek.

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

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

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

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

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

1.41. Tétel. Ha a $\xi_1, \dots, \xi_n$ páronként független valószínűségi változóknak léteznek a szórásnégyzeteik, akkor a $\sum_{i=1}^n \xi_i$ valószínűségi változónak is van szórásnégyzete, továbbá

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

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

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

1.43. 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

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

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

1.9. Nevezetes eloszlások

1.9.1. Diszkrét egyenletes eloszlás

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

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

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

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

1.9.2. Karakterisztikus eloszlás

1.47. Definíció. Az A esemény *indikátorváltozójának* az

$$I_A: \Omega \rightarrow \mathbb{R}, \quad I_A(\omega) := \begin{cases} 1, & \text{ha } \omega \in A, \\ 0, & \text{ha } \omega \notin A, \end{cases}$$

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

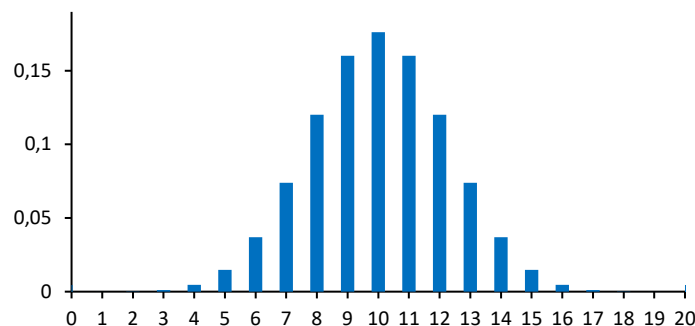
1.48. Tétel. $E I_A = P(A)$ és $D^2 I_A = P(A)(1 - P(A))$.

1.9.3. Binomiális eloszlás

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

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

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



1.2. ábra. $r = 20$ rendű $p = 0,5$ paraméterű binomiális eloszlás vonaldiagramja

Egy tetszőleges A esemény gyakorisága r kísérlet után r -edrendű $P(A)$ 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 $\text{Bin}(1; p)$. Másrészt r darab független p paraméterű karakterisztikus eloszlású valószínűségi változó összege r -edrendű p paraméterű binomiális eloszlású.

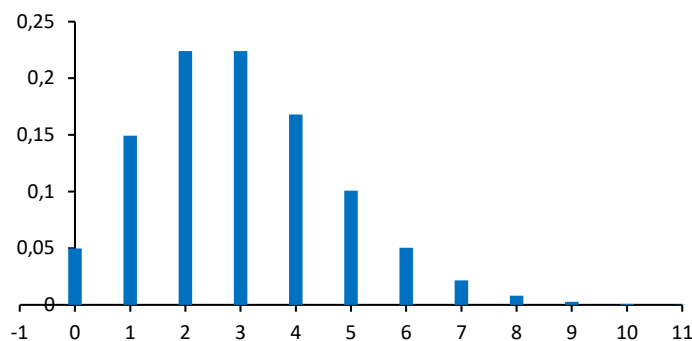
1.50. Tétel. $\xi \in \text{Bin}(r; p)$ esetén $E\xi = rp$ és $D^2\xi = rp(1 - p)$.

1.9.4. Poisson-eloszlás

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

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

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



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

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

1.53. Tétel. Legyenek $\eta_0, \eta_1, \dots$ a $[0, 1]$ intervallumon egyenletes eloszlású független valószínűségi változók és $\lambda > 0$. Ekkor

$$\eta := \min \left\{ r \in \mathbb{N} \cup \{0\} : \eta_0 \eta_1 \cdots \eta_r < e^{-\lambda} \right\}$$

Poisson-eloszlású λ paraméterrel.

Bizonyítás. $P(\eta = 0) = P(\eta_0 < e^{-\lambda}) = e^{-\lambda} = \frac{\lambda^0}{0!} e^{-\lambda}$, illetve az egyenletes eloszlás és a geometriai valószínűségi mező kapcsolata alapján

$$P(\eta = 1) = P(\eta_0 \geq e^{-\lambda}, \eta_0 \eta_1 < e^{-\lambda}) = \int_{e^{-\lambda}}^1 \frac{e^{-\lambda}}{x_0} dx_0 = \frac{\lambda^1}{1!} e^{-\lambda},$$

továbbá ha $k = 2, 3, \dots$, akkor

$$\begin{aligned} P(\eta = k) &= P(\eta_0 \cdots \eta_{k-1} \geq e^{-\lambda}, \eta_0 \cdots \eta_k < e^{-\lambda}) = \\ &= \int_{e^{-\lambda}}^1 \int_{\frac{e^{-\lambda}}{x_0}}^1 \int_{\frac{e^{-\lambda}}{x_0 x_1}}^1 \cdots \int_{\frac{e^{-\lambda}}{x_0 \cdots x_{k-1}}}^1 \frac{e^{-\lambda}}{x_0 \cdots x_{k-1}} dx_{k-1} \cdots dx_0 = \frac{\lambda^k}{k!} e^{-\lambda}. \quad \square \end{aligned}$$

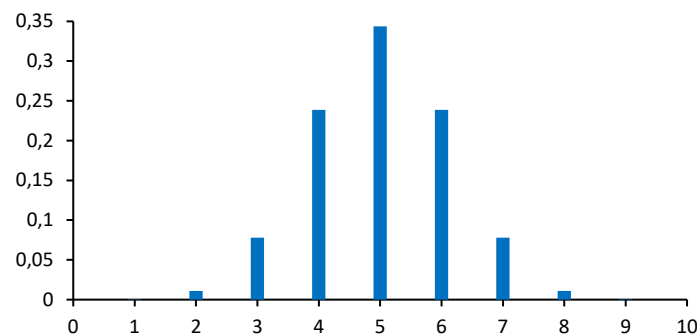
1.9.5. Hipergeometrikus eloszlás

1.54. Definíció. Legyen $\{0, 1, \dots, r\}$ a ξ valószínűségi változó értékkészlete. Ha minden $k \in \{0, 1, \dots, r\}$ esetén

$$P(\xi = k) = \frac{\binom{M}{k} \binom{N-M}{r-k}}{\binom{N}{r}},$$

ahol az M és N egész számokra $0 < M < N$ és $r \leq \min\{M, N - M\}$ áll fenn, akkor ξ -t *hipergeometrikus eloszlású* valószínűségi változónak nevezzük.

1.55. Tétel. Legyen ξ egy hipergeometrikus eloszlású valószínűségi változó N , M és r paraméterekkel. Ekkor $E\xi = \frac{rM}{N}$ és $D^2\xi = \frac{rM}{N} \left(1 - \frac{M}{N}\right) \frac{N-r}{N-1}$.



1.4. ábra. $N = 20$, $M = 10$, $r = 10$ paraméterű hipergeometrikus eloszlás vonaldiagramja

1.9.6. Egyenletes eloszlás

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

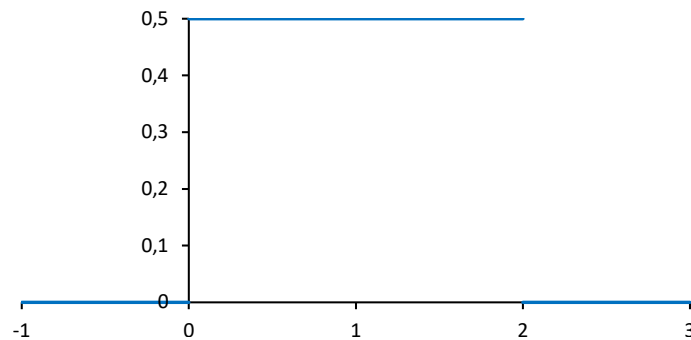
$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = \begin{cases} \frac{1}{b-a}, & \text{ha } a \leq x \leq b, \\ 0 & \text{egyébként,} \end{cases}$$

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

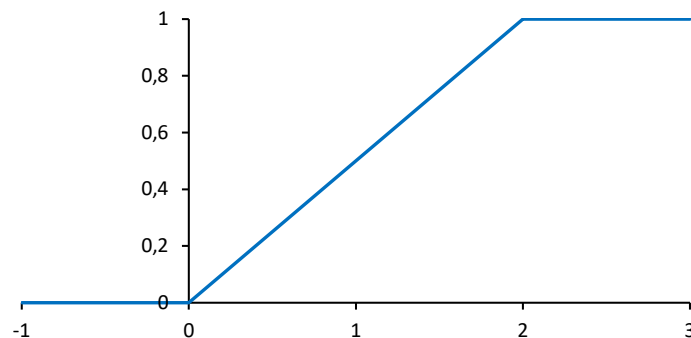
1.57. Tétel. Ha ξ egyenletes eloszlású az $[a, b]$ intervallumon, akkor az eloszlásfüggvénye

$$F: \mathbb{R} \rightarrow \mathbb{R}, \quad F(x) = \begin{cases} 0, & \text{ha } x < a, \\ \frac{x-a}{b-a}, & \text{ha } a \leq x \leq b, \\ 1, & \text{ha } x > b, \end{cases}$$

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



1.5. ábra. $[0, 2]$ intervallumon egyenletes eloszlású valószínűségi változó sűrűségfüggvénye



1.6. ábra. $[0, 2]$ intervallumon egyenletes eloszlású valószínűségi változó eloszlásfüggvénye

A következő állítás szerint bármely eloszlású valószínűségi változó előáll $[0, 1]$ intervallumon egyenletes eloszlású valószínűségi változó valamely transzformáltjaként.

1.58. Tétel. Legyen $F: \mathbb{R} \rightarrow \mathbb{R}$ egy eloszlásfüggvény és

$$G: \mathbb{R} \rightarrow \mathbb{R}, \quad G(x) := \begin{cases} \sup\{y \in \mathbb{R} : F(y) < x\}, & \text{ha } 0 < x < 1, \\ 0, & \text{különben.} \end{cases}$$

Ha ξ egyenletes eloszlású valószínűségi változó a $[0, 1]$ intervallumon, akkor $G(\xi)$ olyan valószínűségi változó, melynek eloszlásfüggvénye F .

Bizonyítás. Legyen

$$A_x := \{y \in \mathbb{R} : F(y) < x\}.$$

Tegyük fel, hogy $A_x = \emptyset$ valamely $0 < x < 1$ esetén. Ekkor $F(y) \geq x \forall y \in \mathbb{R}$, azaz $\lim_{y \rightarrow -\infty} F(y) \geq x > 0$, ami nem lehet, mert $\lim_{y \rightarrow -\infty} F(y) = 0$.

Most tegyük fel, hogy A_x felülről nem korlátos valamely $0 < x < 1$ esetén. Ekkor minden $y \in \mathbb{R}$ -hez létezik $y_0 \in A_x$, hogy $y < y_0$. Ebből $F(y) \leq F(y_0) < x$, hiszen F monoton növekvő. Így $\lim_{y \rightarrow \infty} F(y) \leq x < 1$, ami nem lehet, mert $\lim_{y \rightarrow \infty} F(y) = 1$.

Az eddigieket összegezve tehát A_x nem üres felülről korlátos halmaz minden $0 < x < 1$ esetén, azaz G jól definiált.

Legyen $0 < x_1 < x_2 < 1$. Ekkor $A_{x_1} \subset A_{x_2}$ miatt $G(x_1) = \sup A_{x_1} \leq \sup A_{x_2} = G(x_2)$, azaz G monoton növekvő a $(0, 1)$ intervallumon, míg ezen kívül konstans 0. Ebből kapjuk, hogy G Borel-mérhető függvény, így $G(\xi)$ valószínűségi változó.

Legyen $y \in \mathbb{R}$ és $\omega \in \{G(\xi) < y\} \cap \{0 < \xi < 1\}$. Ekkor $G(\xi(\omega)) = \sup A_{\xi(\omega)} < y$, azaz $y \notin A_{\xi(\omega)}$. Így $F(y) \geq \xi(\omega)$, tehát

$$\{G(\xi) < y\} \cap \{0 < \xi < 1\} \subset \{0 < \xi \leq F(y)\} \quad \forall y \in \mathbb{R}. \quad (1.1)$$

Most legyen $y \in \mathbb{R}$ és $\omega \in \{0 < \xi < F(y)\}$. Tegyük fel, hogy $\omega \notin \{G(\xi) < y\} \cap \{0 < \xi < 1\}$. Mivel $0 < \xi(\omega) < F(y) \leq 1$, ezért ez csak úgy teljesülhet, ha $G(\xi(\omega)) \geq y$. Így $y \leq \sup A_{\xi(\omega)}$, melyből

$$F(y) \leq F(\sup A_{\xi(\omega)}) = \lim_{x \rightarrow \sup A_{\xi(\omega)} - 0} F(x). \quad (1.2)$$

Legyen $x < \sup A_{\xi(\omega)}$ tetszőleges. Ekkor létezik $x_0 \in A_{\xi(\omega)}$, hogy $x < x_0$, azaz $F(x) \leq F(x_0) < \xi(\omega)$. Így $\lim_{x \rightarrow \sup A_{\xi(\omega)} - 0} F(x) \leq \xi(\omega)$. Ebből (1.2) miatt $F(y) \leq \xi(\omega)$, ami nem lehet $\omega \in \{0 < \xi < F(y)\}$ miatt. Így $\omega \in \{G(\xi) < y\} \cap \{0 < \xi < 1\}$, azaz

$$\{0 < \xi < F(y)\} \subset \{G(\xi) < y\} \cap \{0 < \xi < 1\} \quad \forall y \in \mathbb{R}. \quad (1.3)$$

Ebből (1.1) és (1.3) miatt minden $y \in \mathbb{R}$ esetén

$$F(y) = P(0 < \xi < F(y)) \leq P(G(\xi) < y \text{ és } 0 < \xi < 1) \leq P(0 < \xi \leq F(y)) = F(y),$$

azaz

$$P(G(\xi) < y) = P(G(\xi) < y \text{ és } 0 < \xi < 1) = F(y),$$

hiszen $P(0 < \xi < 1) = 1$. Tehát $G(\xi)$ eloszlásfüggvénye F . $\square$

1.59. *Megjegyzés.* Az 1.58. tételben, ha F invertálható eloszlásfüggvény, azaz szigorúan monoton növekvő, akkor $0 < x < 1$ esetén

$$G(x) = \sup\{y \in \mathbb{R} : F(y) < x\} = \sup\{y \in \mathbb{R} : y < F^{-1}(x)\} = F^{-1}(x).$$

Ezért a G függvény $(0, 1)$ -re vett leszűkítettjét az F általánosított inverzének is nevezik.

1.60. *Megjegyzés.* Az 1.58. tételben, ha F egy olyan valószínűségi változó eloszlásfüggvénye, amely az $x_1 < x_2 < \dots < x_r$ értékeket veheti fel rendre $p_1, p_2, \dots, p_r$ valószínűségekkel, akkor

$$G(x) = \begin{cases} x_1, & \text{ha } 0 < x \leq p_1, \\ x_2, & \text{ha } p_1 < x \leq p_1 + p_2, \\ x_3, & \text{ha } p_1 + p_2 < x \leq p_1 + p_2 + p_3, \\ \vdots & \\ x_{r-1}, & \text{ha } p_1 + \dots + p_{r-2} < x \leq p_1 + \dots + p_{r-1}, \\ x_r, & \text{ha } p_1 + \dots + p_{r-1} < x < 1. \end{cases}$$

Könnyen látható, hogy a $G(x)$ felírásában a $<$ és $\leq$ relációs jelek tetszőlegesen felcserélhetőek, hiszen ez nem változtat a $G(\xi)$ eloszlásán.

Hasonló állítás fogalmazható meg akkor is, ha megszámlálhatóan végtelen sok értéket felvevő valószínűségi változót akarunk transzformálni egyenletes eloszlásból.

1.9.7. Exponenciális eloszlás

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

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

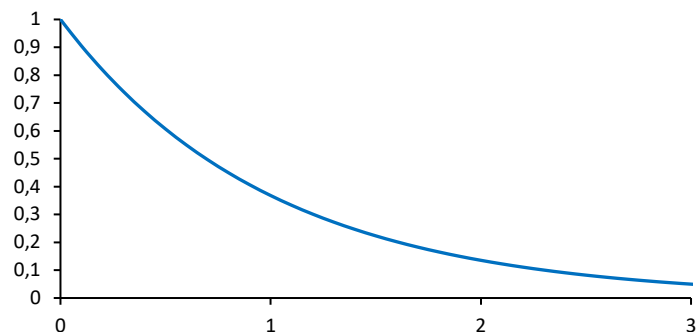
akkor ξ -t λ paraméterű *exponenciális eloszlású* valószínűségi változónak nevezzük. Az ilyen valószínűségi változók halmazát $\text{Exp}(\lambda)$ módon jelöljük.

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

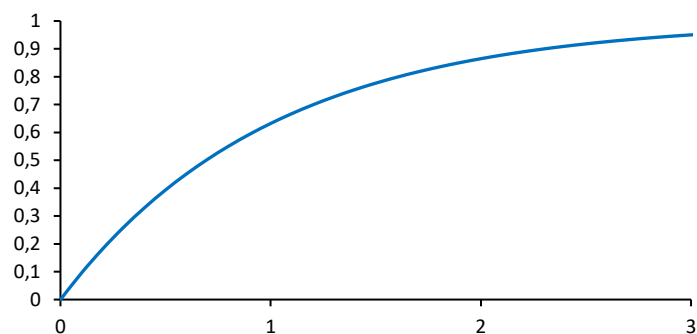
$$F: \mathbb{R} \rightarrow \mathbb{R}, \quad F(x) = \begin{cases} 0, & \text{ha } x \leq 0, \\ 1 - e^{-\lambda x}, & \text{ha } x > 0. \end{cases}$$

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

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



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



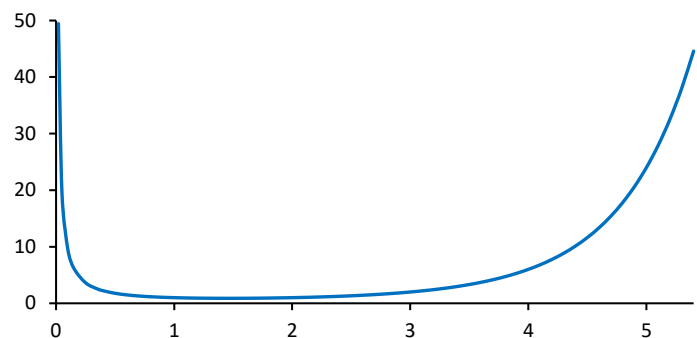
1.8. ábra. $\lambda = 1$ paraméterű exponenciális eloszlású valószínűségi változó eloszlásfüggvénye

1.9.8. Gamma-eloszlás

1.65. Definíció. A következő függvényt *gamma-függvénynek* nevezzük:

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

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



1.9. ábra. A gamma-függvény grafikonja

1.67. Definíció. A következő függvényt *nem teljes gamma-függvénynek* nevezzük:

$$\Gamma^*: \mathbb{R}_+ \times \mathbb{R}_+ \rightarrow \mathbb{R}, \quad \Gamma^*(x, y) := \int_0^y u^{x-1} e^{-u} du.$$

1.68. *Megjegyzés.* $\Gamma(x) = \lim_{y \rightarrow \infty} \Gamma^*(x, y)$.

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

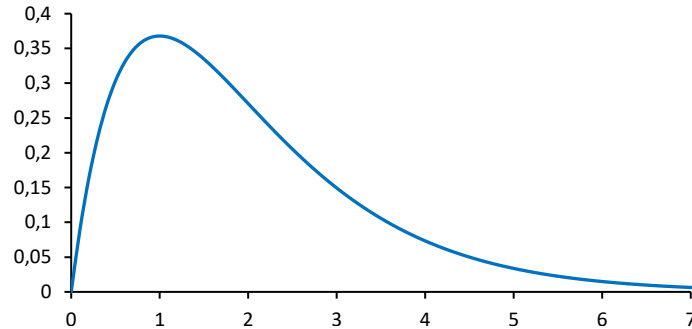
$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) := \begin{cases} 0, & \text{ha } x \leq 0, \\ \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{ha } x > 0. \end{cases}$$

Ekkor ξ -t r -edrendű λ paraméterű gamma-eloszlásúnak nevezzük. Az ilyen valószínűségi változók halmazát $\text{Gamma}(r; \lambda)$ módon jelöljük. Az r helyett szokás α jelölést is használni, amit *alakparaméternek* neveznek, továbbá $\frac{1}{\lambda}$ helyett β jelölést használni, amit *skálaparaméternek* is neveznek.

1.70. *Megjegyzés.* A definíció következménye, hogy $\text{Exp}(\lambda) = \text{Gamma}(1; \lambda)$.

1.71. Tétel. Ha $\xi \in \text{Gamma}(r; \lambda)$, akkor $E\xi = \frac{r}{\lambda}$, $D^2\xi = \frac{r}{\lambda^2}$ és az eloszlásfüggvénye

$$F: \mathbb{R} \rightarrow \mathbb{R}, \quad F(x) = \begin{cases} 0, & \text{ha } x \leq 0, \\ \frac{\Gamma^*(r, \lambda x)}{\Gamma(r)}, & \text{ha } x > 0. \end{cases}$$

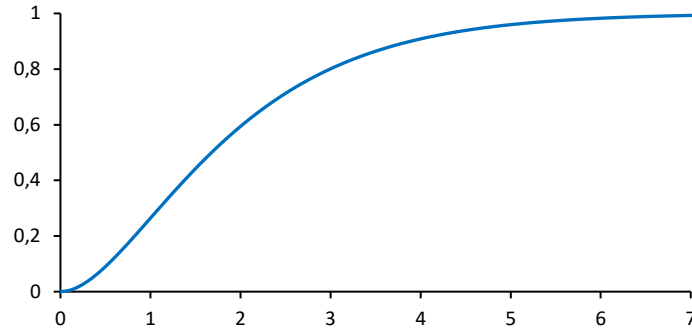


1.10. ábra. $r = 2$ rendű $\lambda = 1$ paraméterű gamma-eloszlású valószínűségi változó sűrűségfüggvénye

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

$$\xi_1 + \dots + \xi_r \in \text{Gamma}(r; \lambda).$$

1.73. Lemma. Ha $r \geq 1$ és $\xi \in \text{Gamma}(r; 1)$ eloszlásfüggvénye F_r , akkor $0,5 < F_r(r) < 0,7$.



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

1.9.9. Normális eloszlás

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

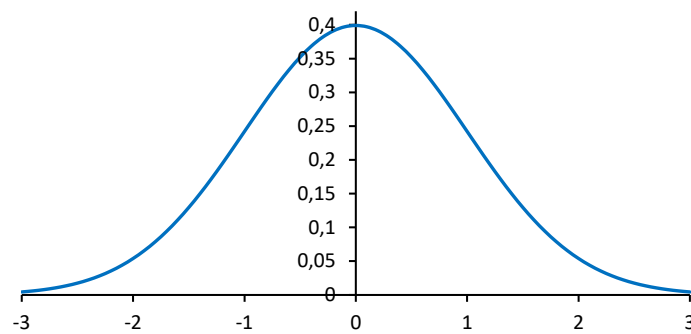
$$\varphi: \mathbb{R} \rightarrow \mathbb{R}, \quad \varphi(x) := \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

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

$$\Phi: \mathbb{R} \rightarrow \mathbb{R}, \quad \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt.$$

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

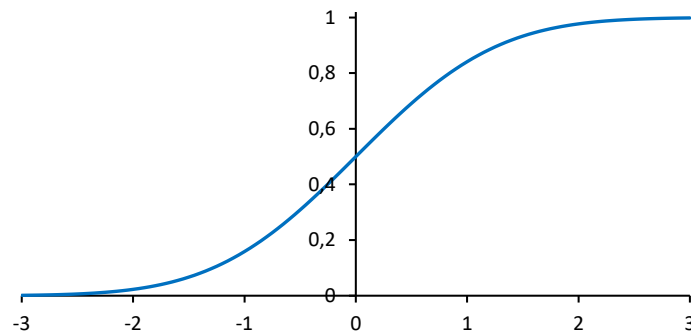
$$\Phi(x) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{\infty} \frac{(-1)^k}{2^k(2k+1)k!} x^{2k+1}.$$



1.12. ábra. Standard normális eloszlású valószínűségi változó sűrűségfüggvénye

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

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



1.13. ábra. Standard normális eloszlású valószínűségi változó eloszlásfüggvénye

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

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

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

1.75. Tétel. Ha ξ standard normális eloszlású valószínűségi változó, akkor $E\xi = 0$ és $D\xi = 1$.

1.76. 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 $\text{Norm}(m; \sigma)$ módon jelöljük.

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

1.77. Tétel. $\xi \in \text{Norm}(m; \sigma)$ esetén $E\xi = m$, $D\xi = \sigma$, továbbá ξ eloszlásfüggvénye

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

illetve sűrűségfüggvénye

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

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

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

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

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

feltéve, hogy ezek a kifejezések léteznek.

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

1.82. Megjegyzés. Ha $\xi \in \text{Bin}(n; p)$, akkor $\frac{\xi - np}{\sqrt{np(1-p)}}$ közelítőleg standard normális eloszlású (lásd Moivre–Laplace-tétel). A közelítés akkor tekinthető megfelelően pontosnak, ha $\min\{np, n(1-p)\} \geq 10$.

1.83. Tétel (Box–Muller-transzformáció). Ha ξ és η a $[0, 1]$ intervallumon egyenletes eloszlású független valószínűségi változók, akkor $\sqrt{-2 \ln \xi} \cos(2\pi\eta)$ standard normális eloszlású.

Bizonyítás. Ha $x > 0$, akkor $P(\sqrt{-2 \ln \xi} < x) = P(\xi > e^{-\frac{x^2}{2}}) = 1 - e^{-\frac{x^2}{2}}$, illetve ha $x \leq 0$, akkor $P(\sqrt{-2 \ln \xi} < x) = 0$. Így $\sqrt{-2 \ln \xi}$ sűrűségfüggvénye

$$f(x) = \begin{cases} 0, & \text{ha } x \leq 0, \\ \left(1 - e^{-\frac{x^2}{2}}\right)' = xe^{-\frac{x^2}{2}}, & \text{ha } x > 0. \end{cases}$$

Ha $-1 < x \leq 1$, akkor

$$\begin{aligned} P(\cos(2\pi\eta) < x) &= \\ &= P\left(\frac{1}{2\pi} \arccos x < \eta < 1 - \frac{1}{2\pi} \arccos x\right) = 1 - \frac{1}{\pi} \arccos x. \end{aligned}$$

$P(\cos(2\pi\eta) < x) = 1$, ha $x > 1$, illetve $P(\cos(2\pi\eta) < x) = 0$, ha $x \leq -1$. Így $\cos(2\pi\eta)$ sűrűségfüggvénye

$$g(x) = \begin{cases} \left(1 - \frac{1}{\pi} \arccos x\right)' = \frac{1}{\pi\sqrt{1-x^2}}, & \text{ha } -1 < x \leq 1, \\ 0 & \text{különben.} \end{cases}$$

Ismert, hogy f és g sűrűségfüggvényű független valószínűségi változók szorzatának sűrűségfüggvénye $h(z) = \int_{-\infty}^{\infty} f(y)g\left(\frac{z}{y}\right) \frac{1}{|y|} dy$. (Lásd például Rényi A. [13, 189. oldal].)

Így $\sqrt{-2 \ln \xi} \cos(2\pi\eta)$ sűrűségfüggvénye

$$\begin{aligned} h(z) &= \int_{-\infty}^{\infty} f(y)g\left(\frac{z}{y}\right) \frac{1}{|y|} dy = \int_0^{\infty} e^{-\frac{y^2}{2}} g\left(\frac{z}{y}\right) dy = \\ &= \int_{|z|}^{\infty} e^{-\frac{y^2}{2}} \frac{1}{\pi\sqrt{1-\left(\frac{z}{y}\right)^2}} dy = \frac{1}{\pi} \int_{|z|}^{\infty} ye^{-\frac{y^2}{2}} \frac{1}{\sqrt{y^2 - z^2}} dy = \\ &= \frac{1}{\pi} e^{-\frac{z^2}{2}} \int_{|z|}^{\infty} ye^{-\frac{y^2 - z^2}{2}} \frac{1}{\sqrt{y^2 - z^2}} dy = \frac{1}{\pi} e^{-\frac{z^2}{2}} \int_0^{\infty} e^{-\frac{x^2}{2}} dx = \\ &= \frac{1}{2\pi} e^{-\frac{z^2}{2}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \end{aligned}$$

Az integrálásban $x = \sqrt{y^2 - z^2}$ helyettesítést alkalmaztunk. □

1.9.10. Többdimenziós normális eloszlás

1.84. 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.

1.85. 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 $\text{Norm}_d(m; A)$ módon jelöljük.

1.86. Tétel. Ha $\xi = (\xi_1, \dots, \xi_d) \in \text{Norm}_d(m; A)$, akkor $m = (E \xi_1, \dots, E \xi_d)$, továbbá

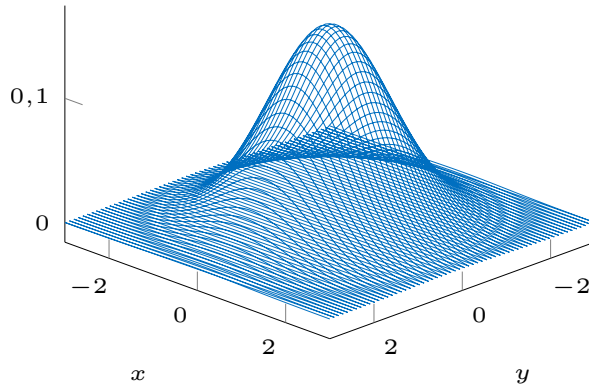
$$D := A^\top A = (\text{cov}(\xi_i, \xi_j))_{d \times d},$$

jelöléssel, ha $\det D \neq 0$, akkor ξ sűrűségfüggvénye

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

Speciálisan, a kétdimenziós standard normális eloszlású valószínűségi vektorváltozó sűrűségfüggvénye

$$f(x, y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)}.$$



1.14. ábra. Kétdimenziós standard normális eloszlású valószínűségi vektorváltozó sűrűségfüggvénye

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

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

1.9.11. Khi-négyzet eloszlás

1.89. 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 eloszlásúnak¹ nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $\text{Khi}(s)$ módon jelöljük.

1.90. Tétel (Khi-négyzet addíciós tétel). *Ha $\xi_1 \in \text{Khi}(s_1)$ és $\xi_2 \in \text{Khi}(s_2)$ függetlenek, akkor*

$$\xi_1 + \xi_2 \in \text{Khi}(s_1 + s_2).$$

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

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = \begin{cases} 0, & \text{ha } x \leq 0, \\ \frac{2^{-\frac{s}{2}} x^{\frac{s}{2}-1}}{\Gamma(\frac{s}{2})} e^{-\frac{x}{2}}, & \text{ha } x > 0. \end{cases}$$

1.92. Következmény. *Ha $\xi \in \text{Khi}(s)$, akkor $E\xi = s$ és $D^2\xi = 2s$.*

1.93. 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 $p_i := P(A_i) > 0$ 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 $n \rightarrow \infty$ esetén.

A bizonyítás a karakterisztikus függvények elméletén és lineáris algebrán alapul (lásd pl. [2, 161–162. oldal]). A közelítés már jónak tekinthető, ha $\min\{\varrho_1, \dots, \varrho_r\} \geq 10$.

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

1.95. Tétel (Fisher–Cochran-tétel). *Legyenek $\xi_1, \xi_2, \dots, \xi_n$ független standard normális eloszlású valószínűségi változók, továbbá a belőlük képzett $Q_1, Q_2, \dots, Q_k$ rendre $s_1, s_2, \dots, s_k$ szabadsági fokú kvadratikus formák olyanok, hogy*

$$Q_1 + Q_2 + \dots + Q_k = \xi_1^2 + \xi_2^2 + \dots + \xi_n^2.$$

Ekkor szükséges és elégséges feltétele annak, hogy $Q_1, Q_2, \dots, Q_k$ rendre $s_1, s_2, \dots, s_k$ szabadsági fokú khi-négyzet eloszlású független valószínűségi változók legyenek az, hogy $n = s_1 + s_2 + \dots + s_k$ teljesüljön.

¹ Szokták χ^2 -eloszlásnak is írni.

1.96. *Megjegyzés.* Q a $\xi_1, \xi_2, \dots, \xi_n$ független standard normális eloszlású valószínűségi változókból álló *kvadratikussá alakba*, ha előáll

$$Q = \eta_1^2 + \eta_2^2 + \dots + \eta_m^2$$

alakban, ahol minden η_i a $\xi_1, \xi_2, \dots, \xi_n$ valószínűségi változók egy lineáris kombinációja nem csupa 0 együtthatókkal. Tekintsük azon B mátrixokat, melyekre teljesül, hogy

$$B(\eta_1, \eta_2, \dots, \eta_m)^\top = (0, 0, \dots, 0)^\top.$$

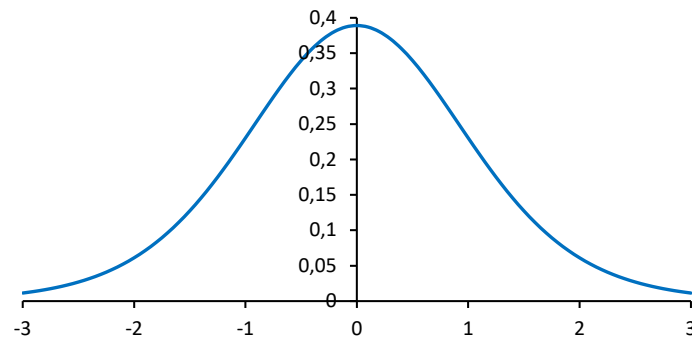
Ezen B mátrixok rangjai közül a legnagyobbat jelölje r . Ekkor a Q szabadsági foka $m - r$. Belátható, hogy a szabadsági fok értéke egyértelműen meghatározott. Ha B csak a nullmátrix lehet, akkor $r = 0$, így ebben az esetben a Q szabadsági foka m . Ha $m = 1$, akkor B csak a nullmátrix lehet, így ekkor Q szabadsági foka 1.

1.9.12. t-eloszlás

1.97. Definíció. Ha $\xi \in \text{Norm}(0; 1)$ és $\eta \in \text{Khi}(s)$ függetlenek, akkor a $\xi \sqrt{\frac{s}{\eta}}$ valószínűségi változót s szabadsági fokú *t-eloszlásúnak*² nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $T(s)$ módon jelöljük.

1.98. Tétel. Ha $\xi \in T(s)$, akkor a *sűrűségfüggvénye*

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

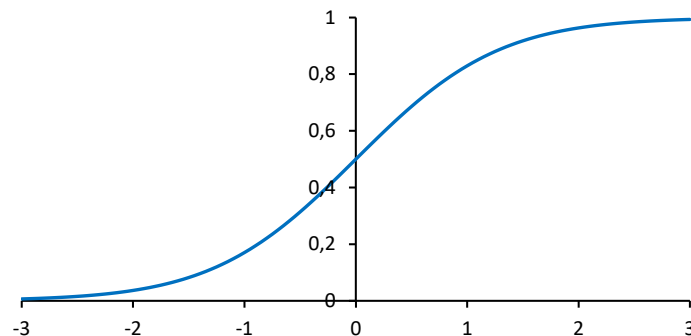


1.15. ábra. $s = 10$ szabadsági fokú t-eloszlású valószínűségi változó sűrűségfüggvénye

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

1.100. Tétel. Ha $\xi \in T(s)$, akkor $s \geq 2$ esetén $E\xi = 0$, illetve $s \geq 3$ esetén $D^2\xi = \frac{s}{s-2}$. Ezekről eltérő esetekben nem létezik ξ várható értéke illetve szórása.

² A t-eloszlás WILLIAM SEALY GOSSET (1876–1937) nevéhez köthető, aki STUDENT álnéven publikált. Ezért a t-eloszlás *Student-eloszlás* néven is ismert.



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

1.101. Tétel. Ha $\xi_s \in T(s)$ minden $s \in \mathbb{N}$ esetén, akkor $\lim_{s \rightarrow \infty} P(\xi_s < x) = \Phi(x)$ minden $x \in \mathbb{R}$ -re, azaz a t -eloszlás konvergál a standard normális eloszláshoz, ha a szabadsági fok tart ∞ -be.

1.102. *Megjegyzés.* Gyakorlatilag $s \geq 50$ esetén a $\xi_s \in T(s)$ eloszlásfüggvénye és Φ között elhanyagolhatóan kicsi a különbség.

1.9.13. Cauchy-eloszlás

1.103. Definíció. Legyenek ξ és η független standard normális eloszlású valószínűségi változók, $\mu \in \mathbb{R}$ és $\sigma > 0$. Ekkor a $\mu + \sigma \frac{\xi}{\eta}$ eloszlását μ helyparaméterű és σ skálaparaméterű Cauchy-eloszlásnak nevezzük. Ha $\mu = 0$ és $\sigma = 1$, akkor standard Cauchy-eloszlásról beszélünk.

1.104. Tétel. A standard Cauchy-eloszlás az 1 szabadsági fokú t -eloszlással egyezik meg.

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

1.106. Tétel. A μ és σ paraméterű Cauchy-eloszlású valószínűségi változó sűrűség- és eloszlásfüggvénye

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = \frac{\sigma}{\pi \sigma^2 + \pi(x - \mu)^2}$$

és

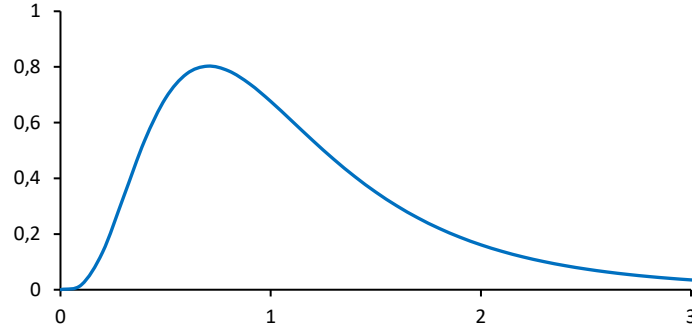
$$F: \mathbb{R} \rightarrow \mathbb{R}, \quad F(x) = \frac{1}{2} + \frac{1}{\pi} \arctg \left(\frac{x - \mu}{\sigma} \right).$$

1.9.14. F-eloszlás

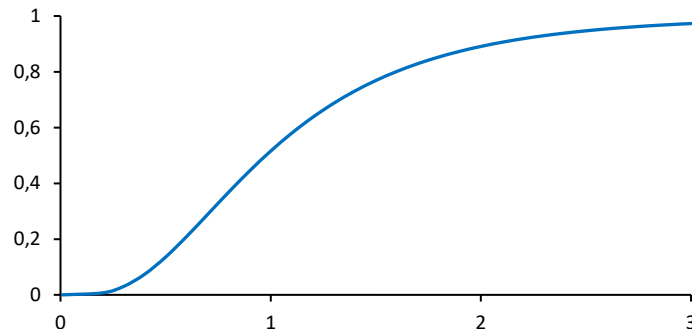
1.107. Definíció. Ha $\xi_1 \in \text{Khi}(s_1)$ és $\xi_2 \in \text{Khi}(s_2)$ függetlenek, akkor az $\frac{s_2 \xi_1}{s_1 \xi_2}$ valószínűségi változót s_1 és s_2 szabadsági fokú F -eloszlásúnak nevezzük. Az ilyen eloszlású valószínűségi változók halmazát $F(s_1; s_2)$ módon jelöljük.

1.108. Tétel. Ha $\xi \in F(s_1; s_2)$, akkor a sűrűségfüggvénye

$$f: \mathbb{R} \rightarrow \mathbb{R}, \quad f(x) = \begin{cases} 0, & \text{ha } x \leq 0, \\ \frac{\Gamma(\frac{s_1+s_2}{2})}{\Gamma(\frac{s_1}{2})\Gamma(\frac{s_2}{2})} \sqrt{\frac{s_1^{s_1} s_2^{s_2} x^{s_1-2}}{(s_1 x + s_2)^{s_1+s_2}}}, & \text{ha } x > 0. \end{cases}$$



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



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

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

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

1.111. Tétel. Ha $\xi \in T(s)$, akkor $\xi^2 \in F(1; s)$.

1.112. Lemma. Legyen $\xi \in F(s_1; s_2)$ 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 $0,3 < F_{s_1, 1}(1) \leq F_{s_1, s_2}(1) \leq F_{1, s_2}(1) < 0,7$.

1.10. Nagy számok törvényei

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

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

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

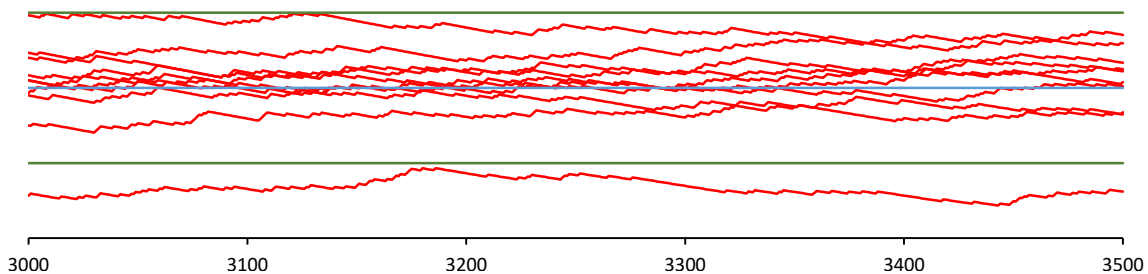
1.114. Tétel (Bernoulli-féle nagy számok törvénye). *Legyen $\frac{\varrho_n}{n}$ az A esemény relatív gyakorisága n kísérlet után. Ekkor*

$$P\left(\left|\frac{\varrho_n}{n} - P(A)\right| \geq \varepsilon\right) \leq \frac{P(A)P(\bar{A})}{n\varepsilon^2}$$

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

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

Az 1.19. ábrán a hatos dobás relatív gyakoriságát láthatjuk szabályos kockával 10 dobássorozat után 3000-től 3500 dobásig. A kék vonal jelzi a hatos dobás valószínűsége-



1.19. ábra

gét, míg a zöld vonalak annak $\varepsilon = 0,01$ sugarú környezetét. Az ábrán láthatjuk, hogy a 10 dobássorozatból 9 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:



A videóban használt program letölthető innen:

<https://tomacstibor.uni-eszterhazy.hu/tananyagok/valdem/valdem.zip>

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 $P(A)$ paraméterű karakterisztikus eloszlású páronként független valószínűségi változók, melyeknek a számtani közepe az A relatív gyakorisága, másrészt ekkor $E \xi_1 = P(A)$ és $D^2 \xi_1 = P(A)P(\bar{A})$. Így tehát bármely $\varepsilon \in \mathbb{R}_+$ esetén

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

Más eloszlású valószínűségi változók számtani közepe is hasonló tulajdonságot mutat.

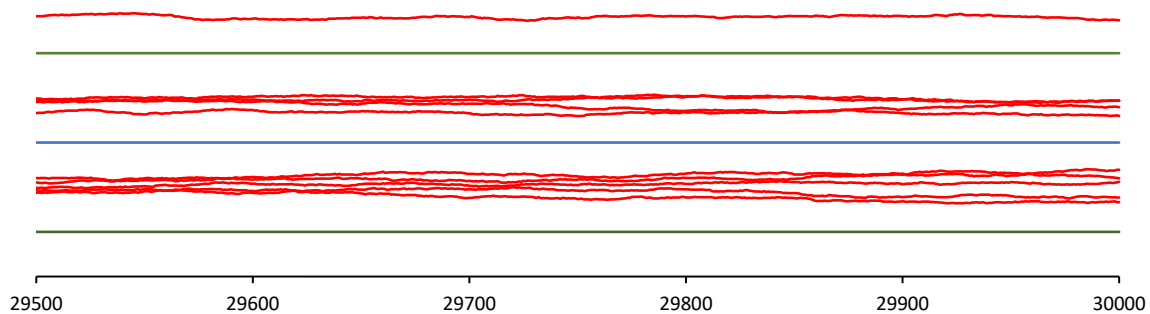
1.115. Tétel (Nagy számok gyenge törvénye). *Legyenek $\xi_1, \xi_2, \dots, \xi_n$ véges várható értékű és szórású, azonos eloszlású, páronként független valószínűségi változók. Ekkor*

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

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.

Az 1.20. ábrán n darab standard normális eloszlású páronként független valószínűségi változó számtani közepét láthatjuk n függvényében $n = 29\,500$ -tól $n = 30\,000$ -ig 10 kísérletsorozat után. A kék vonal jelzi a várható értéket (ez most 0), míg a zöld



1.20. ábra

vonalak annak $\varepsilon = 0,01$ sugarú környezetét. Az ábrán láthatjuk, hogy a 10 kísérletsorozatból 9 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.



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.



1.116. Tétel (Nagy számok Kolmogorov-féle erős törvénye). *Legyenek $\xi_1, \xi_2, \dots$ független azonos eloszlású véges várható értékű valószínűségi változók. Ekkor*

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

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.

1.117. Tétel (Centrális határeloszlási tétel). *Legyenek $\xi_1, \xi_2, \dots$ független, azonos eloszlású, pozitív véges szórású valószínűségi változók. Ekkor $S_n := \xi_1 + \xi_2 + \dots + \xi_n$ standardizáltjának határeloszlása standard normális, azaz*

$$\lim_{n \rightarrow \infty} P \left(\frac{S_n - E S_n}{D S_n} < x \right) = \Phi(x)$$

minden $x \in \mathbb{R}$ esetén.

Speciálisan, ha $\xi_1, \xi_2, \dots$ függetlenek és p paraméterű karakterisztikus eloszlásúak, akkor S_n 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

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

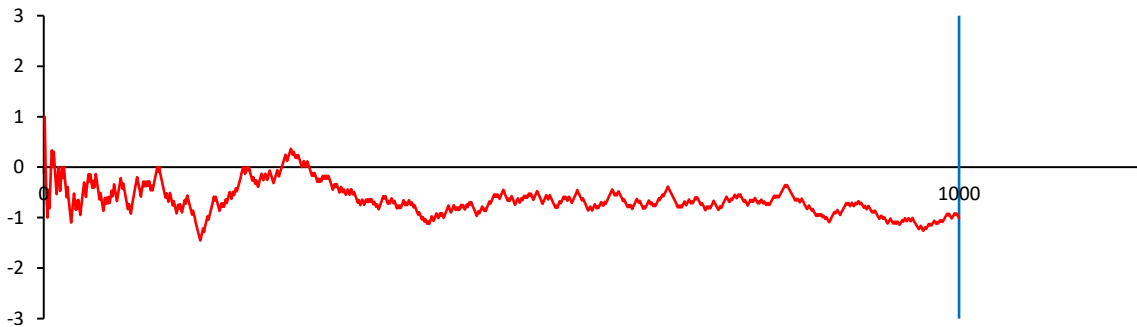
Ez az ún. *Moiivre – Laplace-tétel*. Ez ekvivalens azzal, hogy $x \in \mathbb{R}$ és $\Delta x > 0$ esetén

$$\lim_{n \rightarrow \infty} P \left(x \leq \frac{S_n - np}{\sqrt{np(1-p)}} < x + \Delta x \right) = \frac{1}{\sqrt{2\pi}} \int_x^{x+\Delta x} e^{-\frac{t^2}{2}} dt.$$

Így nagy n és kicsiny Δx esetén

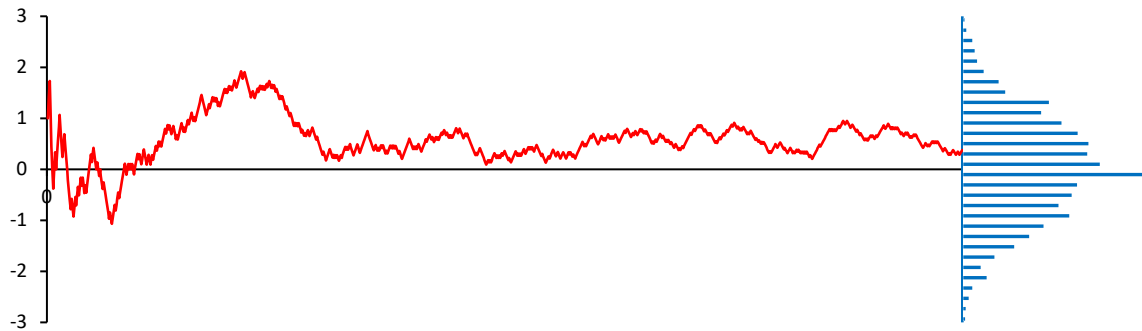
$$\frac{1}{\Delta x} P \left(x \leq \frac{S_n - np}{\sqrt{np(1-p)}} < x + \Delta x \right) \simeq \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

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 $m = 1, 2, \dots, n$.



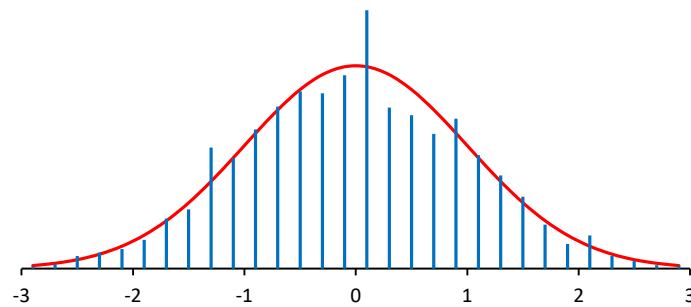
1.21. ábra

Az 1.21. ábra ezt mutatja $p = 0,5$ és $n = 1000$ esetén. A kísérletsorozatot megismétljük N -szer. A kék vonalon ábrázoljuk a becsapódások számát vonaldiagrammal. Az 1.22. ábrán ez látható $N = 3000$ esetén.



1.22. ábra

Végül a vonaldiagramot normáljuk N -nel és Δx -szel, mely már összehasonlítható a standard normális eloszlás sűrűségfüggvényével (lásd az 1.23. ábrát).



1.23. ábra

A következő videóban az előző kísérletsorozatot folyamatában vizsgáljuk.



2. fejezet

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 $P: \mathcal{F} \rightarrow \mathbb{R}$ függvények halmaza, melyekre $(\Omega, \mathcal{F}, P)$ valószínűségi mező. Ekkor az $(\Omega, \mathcal{F}, P)$ 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 $P(A \cap B) = P(A)P(B)$.

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 $\xi: \Omega \rightarrow \mathbb{R}$, $\xi(k) = k$. 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 $P(\xi = k) = \frac{1}{6}$.

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

$$P(\xi = i, \eta = j) = P(\xi = i) P(\eta = j)$$

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 $E I_A = P(A)$ és $E I_B = P(B)$. Így tehát azt kell eldöntenünk, hogy a két esemény indikátorváltozói közül 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 $(\Omega_0, \mathcal{F}_0, P_0)$ 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 $(\Omega_n, \mathcal{F}_n, P_n)$ azon független kísérletek valószínűségi mezője, amely az $(\Omega_0, \mathcal{F}_0, P_0)$ kísérlet n -szeri független elvégzését modellezi. Tegyük fel, hogy ξ d -dimenziós. Legyen

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

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

$$\begin{aligned} P_n(\xi_i < x) &= P_n(\{(\omega_1, \dots, \omega_n) \in \Omega_n : \xi(\omega_i) < x\}) = \\ &= P_n(\Omega_0 \times \dots \times \Omega_0 \times \{\xi < x\} \times \Omega_0 \times \dots \times \Omega_0) = \\ &= P_0(\Omega_0) \cdots P_0(\Omega_0) P_0(\xi < x) P_0(\Omega_0) \cdots P_0(\Omega_0) = P_0(\xi < x), \end{aligned}$$

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

$$\begin{aligned} P_n(\xi_1 < x_1, \dots, \xi_n < x_n) &= \\ &= P_n(\{(\omega_1, \dots, \omega_n) \in \Omega_n : \xi(\omega_1) < x_1, \dots, \xi(\omega_n) < x_n\}) = \\ &= P_n((\xi < x_1) \times \dots \times (\xi < x_n)) = \prod_{i=1}^n P_0(\xi < x_i) = \prod_{i=1}^n P_n(\xi_i < x_i), \end{aligned}$$

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 $(\Omega, \mathcal{F}, P)$ 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 válaszukban 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 -et meghatározni a mintarealizáció alapján nem tudjuk, de becsülni igen. Egy rögzített $x \in \mathbb{R}$ esetén $F(x) = P(\xi < x)$. 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 $\xi_1, \dots, \xi_n$ minta alapján könnyen megadható indikátorváltozókkal: $\frac{1}{n} \sum_{i=1}^n I_{\xi_i < x}$. Itt $\sum_{i=1}^n I_{\xi_i < x}$ azon mintaelemek számát jelenti, melyek kisebbek x -nél. A későbbiekben látni fogjuk, hogy ez a becslés valóban megfelelő lesz számunkra.

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

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

függvényt a ξ -re vonatkozó n elemű mintához tartozó *tapasztalati eloszlásfüggvénynek* 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 $\xi_1(\omega), \dots, \xi_n(\omega)$, akkor az

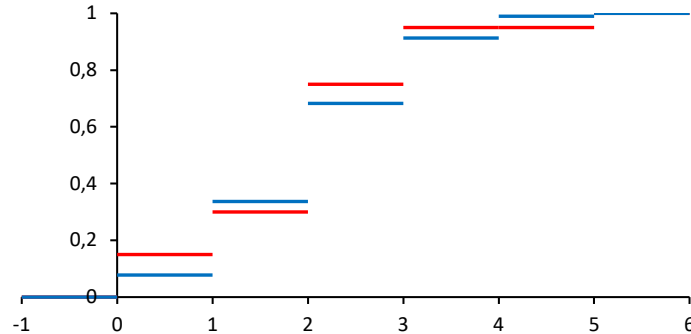
$$x \mapsto (F_n^*(x))(\omega) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{\xi_i(\omega) \leq x} \quad (x \in \mathbb{R})$$

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

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

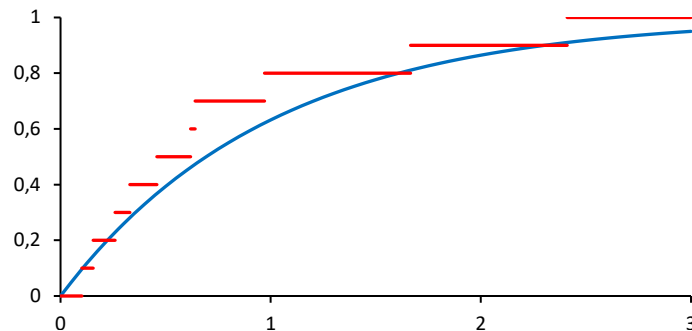
A 2.1. ábrán egy $\text{Bin}(5; 0,4)$ -beli valószínűségi változóra vonatkozó 20 elemű mintához tartozó tapasztalati eloszlásfüggvényt láthatunk. A kék grafikon a valódi



2.1. ábra

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 2.2. ábrán egy $\text{Exp}(1)$ -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.



2.2. ábra

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*

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

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 $k \in \{1, \dots, m-1\}$, akkor az F balról való folytonossága miatt az $\{x \in \mathbb{R} : F(x) \leq \frac{k}{m}\}$ halmaznak létezik maximuma. Ezt a maximumot jelöljük x_k -val. Legyen továbbá $x_0 := -\infty$ és $x_m := \infty$. Ekkor

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

Így

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

Jelentse A_k azt az eseményt, hogy $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{\xi_i < x_k} = P(\xi < x_k)$, illetve B_k azt, hogy

$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{\xi_i \leq x_k} = P(\xi \leq x_k)$. A nagy számok erős törvénye miatt $P(A_k) = P(B_k) = 1$ ($k = 0, \dots, m$). Ebből

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

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

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

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

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

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

$$\begin{aligned}
F(x) - F_n^*(x) &= P(\xi < x) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \leq \\
&\leq P(\xi < x_t) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \leq \\
&\leq \frac{1}{m} + P(\xi \leq x_{t-1}) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \leq \\
&\leq \frac{1}{m} + P(\xi \leq x_{t-1}) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i \leq x_{t-1}} < \frac{1}{m} + \frac{\varepsilon}{2} < \varepsilon.
\end{aligned}$$

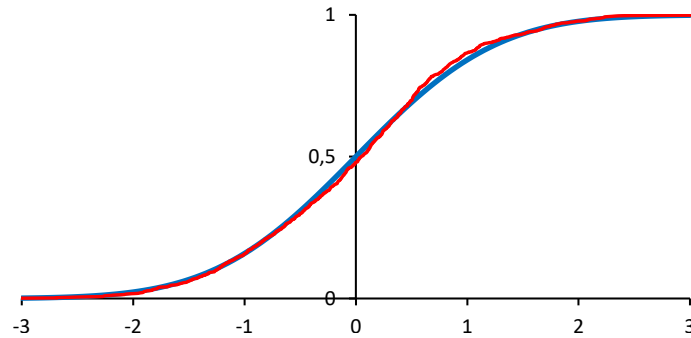
Hasonlóan teljesül minden $n > N$ egész esetén az A -n, hogy

$$\begin{aligned}
F(x) - F_n^*(x) &= P(\xi < x) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \geq \\
&\geq P(\xi \leq x_{t-1}) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \geq \\
&\geq -\frac{1}{m} + P(\xi < x_t) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x} \geq \\
&\geq -\frac{1}{m} + P(\xi < x_t) - \frac{1}{n} \sum_{i=1}^n I_{\xi_i < x_t} > -\frac{1}{m} - \frac{\varepsilon}{2} > -\varepsilon.
\end{aligned}$$

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

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

A 2.3. ábrán egy standard normális eloszlású valószínűségi változóra vonatkozó 1000 elemű mintának a tapasztalati eloszlásfüggvényét látjuk. A kék grafikon a



2.3. ábra

valódi eloszlásfüggvényt jelenti, míg a piros a tapasztalatit. Látható, hogy 1000-es mintaelemszám esetén már gyakorlatilag megegyezik a tapasztalati és a valódi

eloszlásfüggvény. Itt 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:



A videóban használt program letölthető innen:

<https://tomacstibor.uni-eszterhazy.hu/tananyagok/valdem/valdem.zip>

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 értékészlete $\{x_1, \dots, x_k\}$ és a ξ -re vonatkozó minta $\xi_1, \dots, \xi_n$, akkor a tapasztalati eloszlás az

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

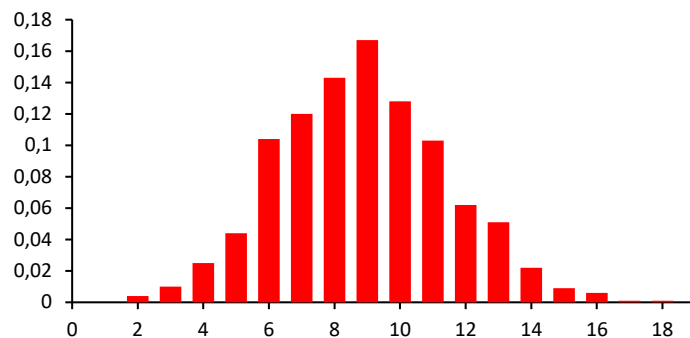
hozzárendelés. (Tehát nr_t a mintában az x_t -vel egyenlő elemek számát jelenti.)

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

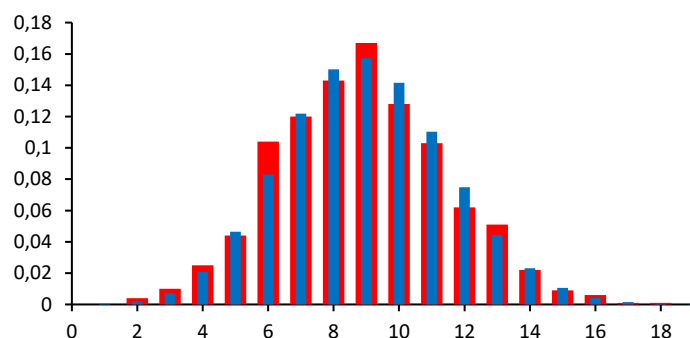
$$x_t \mapsto r_t(\omega) = \frac{1}{n} \sum_{i=1}^n \mathbf{I}_{\xi_i=x_t}(\omega) \quad (t = 1, \dots, k)$$

hozzárendelést (ahol $\mathbf{I}_{\xi_i=x_t}(\omega)$ értéke aszerint 1 vagy 0, hogy $\xi_i(\omega) = x_t$ teljesül vagy sem) 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 $(x_t, r_t(\omega))$ ponttal minden t -re. A 2.4. ábrán egy $\text{Bin}(30; 0,3)$ -beli valószínűségi változóra vonatkozó 1000 elemű mintarealizációból számolt tapasztalati eloszlást láthatunk vonaldiagrammal ábrázolva. Ugyanezen az ábrán késsel felrajzoljuk a valódi eloszlást is, mely jól mutatja a hasonlóságot (lásd a 2.4. ábrát).

Abszolút folytonos ξ valószínűségi változó esetén az ún. sűrűséghisztogram vizsgálata is célravezető lehet a tapasztalati eloszlásfüggvény mellett. Legyen $r \in \mathbb{N}$, $x_0, x_1, \dots, x_r \in \mathbb{R}$ és $x_0 < x_1 < \dots < x_r$. Tegyük fel, hogy a ξ -re vonatkozó $\xi_1(\omega), \dots, \xi_n(\omega)$ 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



2.4. ábra



2.5. ábra

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

$$\int_{x_{j-1}}^{x_j} f(x) dx = P(x_{j-1} \leq \xi < x_j) \simeq \frac{1}{n} \sum_{i=1}^n I_{x_{j-1} \leq \xi_i < x_j}(\omega) = y_j(x_j - x_{j-1}),$$

melyből

$$y_j = \frac{\sum_{i=1}^n I_{x_{j-1} \leq \xi_i < x_j}(\omega)}{n(x_j - x_{j-1})} \quad (j = 1, \dots, r).$$

Itt $I_{x_{j-1} \leq \xi_i < x_j}(\omega)$ értéke aszerint 1 vagy 0, hogy $x_{j-1} \leq \xi_i(\omega) < x_j$ teljesül vagy sem. A kapott oszlopdiagramot *sűrűséghisztogramnak* nevezzük, amely tehát a valódi f sűrűségfüggvényt a j -edik részintervallumon az y_j konstanssal közelíti.

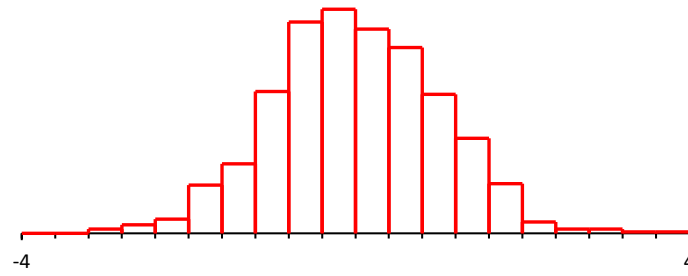
A sűrűséghisztogram megadása a mintarealizáció alapján nem egyértelmű, függ az osztópontok választásától. Az osztópontok felvételéhez csak annyi általános irányelv mondható, hogy függetlennek kell lennie a minta értékeitől.

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.

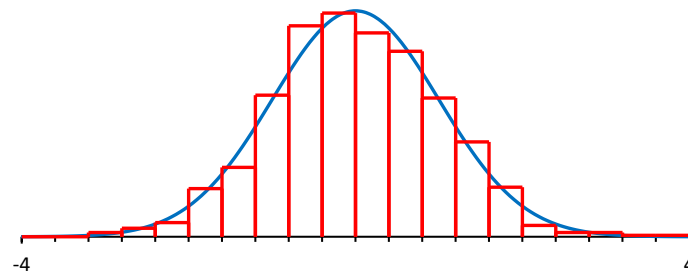
A 2.6. ábrán standard normális eloszlású 1000 elemű mintára vonatkozó sűrűséghisztogramot láthatunk $r = 20$, $x_0 = -4$, $x_{20} = 4$ választással, továbbá a

részintervallumok egyenlő hosszúságúak. Ugyanezen az ábrán kékkel felrajzoljuk



2.6. ábra

a standard normális eloszlás sűrűségfüggvényét a $[-4, 4]$ intervallumon, mely jól mutatja a hasonlóságot (lásd a 2.7. ábrát).



2.7. ábra

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 $\frac{1}{n}(\xi_1 + \dots + \xi_n)$. 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.

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

$$T: \mathbb{R}^n \rightarrow \mathbb{R}$$

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 $T(\xi_1(\omega), \dots, \xi_n(\omega))$ 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.

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

<i>mintaátlag</i>	$\bar{\xi} := \frac{1}{n} \sum_{i=1}^n \xi_i$
<i>tapasztalati szórásnégyzet</i>	$S_n^2 := \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$
<i>tapasztalati szórás</i>	$S_n := \sqrt{\frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^2}$
<i>korrigált tapasztalati szórásnégyzet</i>	$S_n^{*2} := \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2$
<i>korrigált tapasztalati szórás</i>	$S_n^* := \sqrt{\frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})^2}$
<i>k-adik tapasztalati momentum ($k \in \mathbb{N}$)</i>	$\frac{1}{n} \sum_{i=1}^n \xi_i^k$
<i>k-adik tapasztalati centrált momentum ($k \in \mathbb{N}$)</i>	$\frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{\xi})^k$
<i>tapasztalati ferdeség</i>	$\frac{1}{n S_n^3} \sum_{i=1}^n (\xi_i - \bar{\xi})^3$
<i>tapasztalati lapultság</i>	$\frac{1}{n S_n^4} \sum_{i=1}^n (\xi_i - \bar{\xi})^4 - 3$

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.

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

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

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

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

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

$$x_{r_1} \leq x_{r_2} \leq \dots \leq x_{r_n}.$$

Legyen

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

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

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 $\frac{1}{2}(\xi_{\frac{n}{2}}^* + \xi_{\frac{n}{2}+1}^*)$, ha n páros.

Legyen $0 \leq t \leq 1$. 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.

2.10. 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 $\xi_1(\omega), \dots, \xi_n(\omega)$, akkor a $\bar{\xi}(\omega) = \frac{1}{n} \sum_{i=1}^n \xi_i(\omega)$ 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.

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

$$T: (\mathbb{R}^d)^n \rightarrow \mathbb{R}$$

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 $T(\xi_1(\omega), \dots, \xi_n(\omega))$ számot az előbbi *statisztika egy realizációjának* nevezzük.

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

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

illetve *tapasztalati korrelációs együtthatója*

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

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

2.14. Definíció. Legyenek $\xi_1, \xi_2, \dots, \xi_k$ valószínűségi változók, melyekre vonatkozó minták rendre a következők:

$$\begin{aligned} &\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}, \\ &\xi_{21}, \xi_{22}, \dots, \xi_{2n_2}, \\ &\vdots \\ &\xi_{k1}, \xi_{k2}, \dots, \xi_{kn_k}. \end{aligned}$$

Ha

$$T: \mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_k} \rightarrow \mathbb{R}$$

olyan függvény, melyre $T(\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}, \dots, \xi_{k1}, \xi_{k2}, \dots, \xi_{kn_k})$ valószínűségi változó (azaz T Borel-mérhető), akkor ezt a valószínűségi változót az előbbi k darab minta egy *statisztikájának* nevezzük.

Ilyen statisztikákra példát, majd a hipotézisvizsgálatoknál fogunk látni.

3. fejezet

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

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

mezőre, ahol $\Theta = \{ (a, b) \in \mathbb{R}^2 : a < b \}$ és P_ϑ olyan valószínűség az $(\Omega, \mathcal{F})$ téren, melyre $P_\vartheta(\xi < x) = \frac{x-a}{b-a}$ 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 $\xi_1(\omega), \dots, \xi_n(\omega)$ 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 *paraméterter*. Feltesszük, hogy $\Theta \neq \emptyset$. Jelöljön F_ϑ eloszlásfüggvényt minden $\vartheta = (\vartheta_1, \dots, \vartheta_v) \in \Theta$ 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

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

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

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

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

$$P_\vartheta(\xi < x) = F_\vartheta(x)$$

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 $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ statisztikai mezőn.

Legyen $g: \Theta \rightarrow \mathbb{R}$ 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

$$E_\vartheta T(\xi_1, \dots, \xi_n) = g(\vartheta)$$

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

3.2. Feladat. Bizonyítsuk be, hogy $F_n^*(x)$ torzítatlan becslése $F(x)$ -nek bármely $x \in \mathbb{R}$ 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 $E_p F_n^*(x) = \frac{1}{n} E_p(nF_n^*(x)) = \frac{1}{n} np = p = F(x)$. $\square$

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 $h: \mathbb{R}^r \rightarrow \mathbb{R}$ olyan függvény, melyre $h(\tau_1, \dots, \tau_r)$ valószínűségi változó. Bizonyítsuk 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 $E_p \bar{\xi} = p$, azaz $\bar{\xi}$ torzítatlan becslése p -nek. Másrészt $h: \mathbb{R} \rightarrow \mathbb{R}$, $h(x) := x^2$ jelöléssel

$$\begin{aligned} E_p h(\bar{\xi}) &= E_p \bar{\xi}^2 = D_p^2 \bar{\xi} + E_p^2 \bar{\xi} = \frac{1}{n^2} n D_p^2 \xi + E_p^2 \xi = \\ &= \frac{1}{n} p(1-p) + p^2 \neq p^2 = h(p), \end{aligned}$$

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

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

$$\lim_{n \rightarrow \infty} E_{\vartheta} T_n(\xi_1, \dots, \xi_n) = g(\vartheta).$$

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

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

$$D_{\vartheta} T_1(\xi_1, \dots, \xi_n) \leq D_{\vartheta} T_2(\xi_1, \dots, \xi_n).$$

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

$$P_{\vartheta}(T_1(\xi_1, \dots, \xi_n) = T_2(\xi_1, \dots, \xi_n)) = 1.$$

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

$$E_{\vartheta} \tau = \frac{1}{2}(E_{\vartheta} \tau_1 + E_{\vartheta} \tau_2) = \frac{1}{2}(g(\vartheta) + g(\vartheta)) = g(\vartheta),$$

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

$$\begin{aligned} D_{\vartheta}^2 \tau_1 &\leq D_{\vartheta}^2 \tau = D_{\vartheta}^2 \frac{\tau_1 + \tau_2}{2} = \\ &= \frac{1}{4}(D_{\vartheta}^2 \tau_1 + D_{\vartheta}^2 \tau_2 + 2 \operatorname{cov}_{\vartheta}(\tau_1, \tau_2)) = \frac{1}{4}(2 D_{\vartheta}^2 \tau_1 + 2 \operatorname{cov}_{\vartheta}(\tau_1, \tau_2)). \end{aligned}$$

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

$$P_{\vartheta}(\tau_1 - \tau_2 = E_{\vartheta}(\tau_1 - \tau_2)) = 1.$$

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

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

$$\lim_{n \rightarrow \infty} P_{\vartheta}(|T_n(\xi_1, \dots, \xi_n) - g(\vartheta)| \geq \varepsilon) = 0.$$

3.10. Tétel. *Létezik nem konzisztens torzítatlan becsléssorozat.*

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

$$P_m(|\xi_n - m| \geq \varepsilon) = 1 - P(|\xi_n - m| < \varepsilon) = 2 - 2\Phi(\varepsilon),$$

azaz $\lim_{n \rightarrow \infty} P_m(|\xi_n - m| \geq \varepsilon) \neq 0$. Így ξ_n nem konzisztens becsléssorozat. $\square$

A torzítatlan becsléssorozatok konzisztenciájához tudunk adni elégséges feltételt.

3.11. Tétel. *Ha $T_n(\xi_1, \dots, \xi_n)$ torzítatlan becslése $g(\vartheta)$ -nak minden $n \in \mathbb{N}$ esetén, és $\lim_{n \rightarrow \infty} D_\vartheta^2 T_n(\xi_1, \dots, \xi_n) = 0$ minden $\vartheta \in \Theta$ esetén, akkor $T_n(\xi_1, \dots, \xi_n)$ a $g(\vartheta)$ konzisztens becsléssorozata.*

Bizonyítás. Legyen $\tau_n := T_n(\xi_1, \dots, \xi_n)$, $\varepsilon > 0$ és $\vartheta \in \Theta$. Ekkor τ_n torzítatlansága, a Csebisev-egyenlőtlenség és $\lim_{n \rightarrow \infty} D_\vartheta^2 \tau_n = 0$ miatt

$$\lim_{n \rightarrow \infty} P_\vartheta(|\tau_n - g(\vartheta)| \geq \varepsilon) = \lim_{n \rightarrow \infty} P_\vartheta(|\tau_n - E_\vartheta \tau_n| \geq \varepsilon) \leq \lim_{n \rightarrow \infty} \frac{D_\vartheta^2 \tau_n}{\varepsilon^2} = 0.$$

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

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

$$P_\vartheta \left(\lim_{n \rightarrow \infty} T_n(\xi_1, \dots, \xi_n) = g(\vartheta) \right) = 1.$$

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 $E_\vartheta \xi \in \mathbb{R}$ minden $\vartheta \in \Theta$ esetén.

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

Bizonyítás. $E_\vartheta \sum_{i=1}^n c_i \xi_i = \sum_{i=1}^n c_i E_\vartheta \xi_i = \sum_{i=1}^n c_i E_\vartheta \xi = E_\vartheta \xi \sum_{i=1}^n c_i = E_\vartheta \xi$. $\square$

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

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

3.16. Feladat. Bizonyítsuk 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

$$\lim_{n \rightarrow \infty} D_{\vartheta}^2 \bar{\xi} = \lim_{n \rightarrow \infty} \frac{1}{n} D_{\vartheta}^2 \xi = 0,$$

melyből következik az állítás. $\square$

3.17. Feladat. Bizonyítsuk be, hogy a mintaátlag erősen konzisztens becsléssorozata a várható értéknek.

Bizonyítás. A bizonyításhoz elég észrevenni, hogy az állítás a Kolmogorov-féle nagy számok erős törvényével ekvivalens. $\square$

3.18. Feladat. Bizonyítsuk be, hogy $\bar{\xi}$ 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. $D_{\vartheta}^2 \left(\sum_{i=1}^n c_i \xi_i \right) = \sum_{i=1}^n c_i^2 D_{\vartheta}^2 \xi = D_{\vartheta}^2 \xi \sum_{i=1}^n c_i^2 \geq D_{\vartheta}^2 \xi \frac{1}{n} (c_1 + \dots + c_n)^2 = \frac{1}{n} D_{\vartheta}^2 \xi = D_{\vartheta}^2 \bar{\xi}$. Itt felhasználtuk a számtani és a négyzetes közép közötti relációt, mely szerint tetszőleges $a_1, \dots, a_n \in \mathbb{R}$ esetén $\frac{a_1 + \dots + a_n}{n} \leq \sqrt{\frac{a_1^2 + \dots + a_n^2}{n}}$. (Ez a Cauchy-egyenlőtlenségből következik.) $\square$

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 tétel erre ad általánosságban nemleges választ.

3.19. Tétel. Ha ξ egyenletes eloszlású a $[0, b]$ intervallumon ($b \in \mathbb{R}_+$), akkor a terjedelmeközép 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 terjedelmeközép 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

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

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

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

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

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

Ebből belátható, hogy ξ_k^* sűrűségfüggvénye x helyen

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

Így $k \in \{1, \dots, n\}$ esetén

$$E_b \xi_k^* = \int_0^b x \frac{n}{b} \binom{n-1}{k-1} \left(\frac{x}{b}\right)^{k-1} \left(1 - \frac{x}{b}\right)^{n-k} dx = \dots = \frac{kb}{n+1}.$$

Ebből $E_b \frac{\xi_1^* + \xi_n^*}{2} = \frac{1}{2} \left(\frac{b}{n+1} + \frac{nb}{n+1}\right) = \frac{b}{2} = E_b \xi$. Tehát a terjedelmeközép a várható érték torzítatlan becslése. Most rátérünk a szórás meghatározására. A korábbiak alapján

$$E_b \xi_k^{*2} = \int_0^b x^2 \frac{n}{b} \binom{n-1}{k-1} \left(\frac{x}{b}\right)^{k-1} \left(1 - \frac{x}{b}\right)^{n-k} dx = \dots = \frac{k(k+1)b^2}{(n+1)(n+2)}$$

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

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

Ebből bizonyítható, hogy

$$E_b(\xi_k^* \xi_l^*) = \frac{k(l+1)b^2}{(n+1)(n+2)}, \quad 1 \leq k < l \leq n.$$

Így a szórásnégyzet:

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

Mivel $D_b^2 \bar{\xi} = \frac{1}{n} D_b^2 \xi = \frac{b^2}{12n}$, ezért az állítás ekvivalens az

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

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 $\frac{\xi_1^* + \xi_n^*}{2} = \bar{\xi}$. Ezzel bizonyított az állítás. $\square$

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. Bizonyítsuk be, hogy egy esemény relatív gyakorisága torzítatlan becslése az esemény valószínűségének.

Bizonyítás. Legyen ξ a vizsgált esemény indikátorváltozója. Ekkor az esemény relatív gyakorisága $\bar{\xi}$ -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. $\square$

3.21. Feladat. Bizonyítsuk 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. $\square$

3.22. Tétel. 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 $\bar{\xi}$, továbbá az eddigiek alapján $\bar{\xi}$ 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 $S(\xi_1(\omega), \dots, \xi_n(\omega))$ 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

$$0 = E_p(S(\xi_1, \dots, \xi_n) - \bar{\xi}) = \sum_{k=0}^n \left(S_k - \frac{k}{n} \right) \binom{n}{k} p^k (1-p)^{n-k},$$

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

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

Így azt kell belátni, hogy $D_p^2 S(\xi_1, \dots, \xi_n) \leq D_p^2 T(\xi_1, \dots, \xi_n)$, amely azzal ekvivalens a torzítatlanság miatt, hogy $E_p S^2(\xi_1, \dots, \xi_n) \leq E_p T^2(\xi_1, \dots, \xi_n)$. Legyen

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

Ekkor az előzőekhez hasonlóan látható, hogy

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

Másrészt

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

így elég azt belátni, hogy

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

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

3.1.3. Szórásnégyzet becslése

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

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

Bizonyítás. A Steiner-formula és $E_\vartheta \xi^2 = D_\vartheta^2 \xi + E_\vartheta^2 \xi$ miatt

$$E_\vartheta S_n^2 = E_\vartheta \left(\frac{1}{n} \sum_{i=1}^n \xi_i^2 - \bar{\xi}^2 \right) = \frac{1}{n} \sum_{i=1}^n E_\vartheta \xi_i^2 - E_\vartheta \bar{\xi}^2 =$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{i=1}^n E_{\vartheta} \xi^2 - D_{\vartheta}^2 \bar{\xi} - E_{\vartheta}^2 \bar{\xi} = E_{\vartheta} \xi^2 - D_{\vartheta}^2 \bar{\xi} - E_{\vartheta}^2 \xi = \\
&= D_{\vartheta}^2 \xi - D_{\vartheta}^2 \bar{\xi} = D_{\vartheta}^2 \xi - \frac{1}{n^2} \sum_{i=1}^n D_{\vartheta}^2 \xi_i = D_{\vartheta}^2 \xi - \frac{1}{n^2} \sum_{i=1}^n D_{\vartheta}^2 \xi = \\
&= D_{\vartheta}^2 \xi - \frac{1}{n} D_{\vartheta}^2 \xi = \frac{n-1}{n} D_{\vartheta}^2 \xi \neq D_{\vartheta}^2 \xi. \quad \square
\end{aligned}$$

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

Bizonyítás. Láttuk, hogy $E_{\vartheta} S_n^2 = \frac{n-1}{n} D_{\vartheta}^2 \xi$, így $\lim_{n \rightarrow \infty} E_{\vartheta} S_n^2 = D_{\vartheta}^2 \xi$. $\square$

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

Bizonyítás. A Kolmogorov-féle nagy számok törvénye miatt

$$P_{\vartheta} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_i^2 = E_{\vartheta} \xi^2 \right) = 1 \quad \text{és} \quad P_{\vartheta} \left(\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \xi_i = E_{\vartheta} \xi \right) = 1.$$

Így a Steiner-formulából kapjuk az állítást. $\square$

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

Bizonyítás. Láttuk, hogy $E_{\vartheta} S_n^2 = \frac{n-1}{n} D_{\vartheta}^2 \xi$, így $E_{\vartheta} S_n^{*2} = E_{\vartheta} \frac{n}{n-1} S_n^2 = D_{\vartheta}^2 \xi$. $\square$

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

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

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 $\bar{\xi}$ hatásos becslése p -nek. Mivel $D_p^2 \bar{\xi} = \frac{1}{n} D_p^2 \xi = \frac{p(1-p)}{n}$, ezért azt kapjuk, hogy a p összes véges szórással torzítatlan becslésének szórása nagyobb vagy egyenlő, mint $\frac{p(1-p)}{n}$.

Általánosságban, ha $g(\vartheta)$ összes véges szórással $T(\xi_1, \dots, \xi_n)$ 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 értékészlete legyen $\mathfrak{X}$, azaz a mintatér $\mathfrak{X}^n$.

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

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

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*

$$I_n: \Theta \rightarrow \mathbb{R}, \quad I_n(\vartheta) := E_{\vartheta} \left(\frac{\partial}{\partial \vartheta} L_n(\xi_1, \dots, \xi_n, \vartheta) \right)^2,$$

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

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

$$\begin{aligned} \frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} T(x_1, \dots, x_n) l_n(x_1, \dots, x_n, \vartheta) dx_1 \cdots dx_n = \\ = \int_{\mathbb{R}^n} T(x_1, \dots, x_n) \frac{\partial}{\partial \vartheta} l_n(x_1, \dots, x_n, \vartheta) dx_1 \cdots dx_n \end{aligned}$$

vagy

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

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

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \vartheta} f_{\vartheta}(x) dx = 0 \quad \text{vagy} \quad \sum_{x \in \mathfrak{X}} \frac{\partial}{\partial \vartheta} P_{\vartheta}(\xi = x) = 0$$

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 $\int_{-\infty}^{\infty} l_1(x, \vartheta) dx = 1$. Most tegyük fel, hogy $\int_{-\infty}^{\infty} \frac{\partial}{\partial \vartheta} f_{\vartheta}(x) dx = 0$. Ebből kapjuk, hogy

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

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

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

Ezzel teljes a bizonyítás. $\square$

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

$$E_\vartheta \left(\frac{\partial}{\partial \vartheta} L_1(\xi_1, \vartheta) \right) = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \vartheta} \ln l_1(x, \vartheta) \right) f_\vartheta(x) dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \vartheta} f_\vartheta(x) dx = 0.$$

Ebből

$$I_1(\vartheta) = E_\vartheta \left(\frac{\partial}{\partial \vartheta} L_1(\xi_1, \vartheta) \right)^2 = D_\vartheta^2 \left(\frac{\partial}{\partial \vartheta} L_1(\xi_1, \vartheta) \right) = D_\vartheta^2 \left(\frac{\partial}{\partial \vartheta} \ln f_\vartheta(\xi_1) \right).$$

Másrészt

$$\begin{aligned} E_\vartheta \left(\frac{\partial}{\partial \vartheta} L_n(\xi_1, \dots, \xi_n, \vartheta) \right) &= E_\vartheta \left(\frac{\partial}{\partial \vartheta} \sum_{i=1}^n \ln f_\vartheta(\xi_i) \right) = \\ &= \sum_{i=1}^n E_\vartheta \left(\frac{\partial}{\partial \vartheta} \ln f_\vartheta(\xi_i) \right) = \sum_{i=1}^n \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \vartheta} \ln f_\vartheta(x) \right) f_\vartheta(x) dx = \\ &= \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{\partial}{\partial \vartheta} f_\vartheta(x) dx = 0. \end{aligned}$$

Ebből

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

3.34. Feladat. Karakterisztikus eloszlás esetén határozzuk 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 $\mathfrak{X} = \{0, 1\}$, $l_1(0, p) = P_p(\xi_1 = 0) = 1 - p$ és $l_1(1, p) = P_p(\xi_1 = 1) = p$. Így

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

$$\begin{aligned}
&= \left(\frac{\partial}{\partial p} \ln P_p(\xi_1 = 0) \right)^2 \cdot P_p(\xi_1 = 0) + \left(\frac{\partial}{\partial p} \ln P_p(\xi_1 = 1) \right)^2 \cdot P_p(\xi_1 = 1) = \\
&= \left(\frac{\partial}{\partial p} \ln(1-p) \right)^2 \cdot (1-p) + \left(\frac{\partial}{\partial p} \ln p \right)^2 \cdot p = \frac{1}{p(1-p)}.
\end{aligned}$$

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

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

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

Megoldás. Az l_1 -re teljesül a bederiválási feltétel, hiszen

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{\partial}{\partial m} f_m(x) dx &= \int_{-\infty}^{\infty} \frac{\partial}{\partial m} \frac{1}{\sigma} \varphi\left(\frac{x-m}{\sigma}\right) dx = \\
&= \int_{-\infty}^{\infty} \frac{x-m}{\sigma^2} f_m(x) dx = E_m\left(\frac{\xi-m}{\sigma^2}\right) = 0.
\end{aligned}$$

Korábban láttuk, hogy ekkor

$$\begin{aligned}
I_1(m) &= D_m^2 \left(\frac{\partial}{\partial m} \ln f_m(\xi) \right) = D_m^2 \left(\frac{1}{f_m(\xi)} \cdot \frac{\partial}{\partial m} f_m(\xi) \right) = \\
&= D_m^2 \left(\frac{1}{f_m(\xi)} \cdot \frac{\xi-m}{\sigma^2} f_m(\xi) \right) = D_m^2 \left(\frac{\xi-m}{\sigma^2} \right) = \frac{1}{\sigma^2}.
\end{aligned}$$

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

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

Megoldás.

$$\begin{aligned}
I_1(\lambda) &= E_\lambda \left(\frac{\partial}{\partial \lambda} \ln l_1(\xi_1, \lambda) \right)^2 = \sum_{k=0}^{\infty} \left(\frac{\partial}{\partial \lambda} \ln P_\lambda(\xi_1 = k) \right)^2 P_\lambda(\xi_1 = k) = \\
&= \sum_{k=0}^{\infty} \left(\frac{\partial}{\partial \lambda} \ln \frac{\lambda^k}{k!} e^{-\lambda} \right)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \left(\frac{k}{\lambda} - 1 \right)^2 \frac{\lambda^k}{k!} e^{-\lambda} = \\
&= \sum_{k=0}^{\infty} \left(\frac{k(k-1)}{\lambda^2} + 1 + \left(\frac{1}{\lambda^2} - \frac{2}{\lambda} \right) k \right) \frac{\lambda^k}{k!} e^{-\lambda} = \\
&= \left(\sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} + \left(\frac{1}{\lambda} - 2 \right) \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) e^{-\lambda} = \\
&= \left(e^\lambda + e^\lambda + \left(\frac{1}{\lambda} - 2 \right) e^\lambda \right) e^{-\lambda} = \frac{1}{\lambda}.
\end{aligned}$$

Másrészt

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

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

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

Megoldás.

$$\begin{aligned} I_1(\lambda) &= \mathbb{E}_{\lambda} \left(\frac{\partial}{\partial \lambda} \ln l_1(\xi_1, \lambda) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \lambda} \ln l_1(x, \lambda) \right)^2 l_1(x, \lambda) dx = \\ &= \int_0^{\infty} \left(\frac{\partial}{\partial \lambda} \ln \lambda e^{-\lambda x} \right)^2 \lambda e^{-\lambda x} dx = \int_0^{\infty} \left(\frac{1}{\lambda} - x \right)^2 \lambda e^{-\lambda x} dx = \\ &= \mathbb{E}_{\lambda} \left(\frac{1}{\lambda} - \xi \right)^2 = D_{\lambda}^2 \xi = \frac{1}{\lambda^2}. \end{aligned}$$

Másrészt

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

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 ($b \in \mathbb{R}_+$). Mutassuk meg, hogy ekkor nem teljesül l_1 -re a bederiválási feltétel, továbbá az $I_1(b)$ és $I_n(b)$ meghatározásával bizonyítsuk be, hogy $I_n \neq nI_1$, ha $n > 1$.

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

$$\begin{aligned} I_1(b) &= \mathbb{E}_b \left(\frac{\partial}{\partial b} \ln f_b(\xi_1) \right)^2 = \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial b} \ln f_b(x) \right)^2 f_b(x) dx = \\ &= \int_0^b \left(\frac{\partial}{\partial b} \ln \frac{1}{b} \right)^2 \frac{1}{b} dx = \int_0^b \left(\frac{-1}{b} \right)^2 \frac{1}{b} dx = \int_0^b \frac{1}{b^3} dx = \frac{1}{b^2}. \end{aligned}$$

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

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 $T(\xi_1, \dots, \xi_n)$ véges szórású torzítatlan becslése $g(\vartheta)$ -nak, ahol $g: \Theta \rightarrow \mathbb{R}$ differenciálható függvény. Tegyük fel, hogy l_1 -re és Tl_n -re teljesül a bederiválási feltétel, továbbá, hogy I_1 létezik és pozitív. Ekkor*

$$D_{\vartheta}^2 T(\xi_1, \dots, \xi_n) \geq \frac{(g'(\vartheta))^2}{nI_1(\vartheta)}$$

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

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

Ekkor

$$E_{\vartheta}(\varrho^2) = \left(\frac{g'(\vartheta)}{I_n(\vartheta)} \right)^2 E_{\vartheta} \left(\frac{\partial}{\partial \vartheta} \ln l_n(\xi_1, \dots, \xi_n, \vartheta) \right)^2 = \frac{(g'(\vartheta))^2}{I_n(\vartheta)},$$

másrészt

$$\begin{aligned} E_{\vartheta}(\varrho) &= \frac{g'(\vartheta)}{I_n(\vartheta)} E_{\vartheta} \left(\frac{\partial}{\partial \vartheta} \ln l_n(\xi_1, \dots, \xi_n, \vartheta) \right) = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} E_{\vartheta} \left(\frac{\partial}{\partial \vartheta} \sum_{i=1}^n \ln f_{\vartheta}(\xi_i) \right) = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \sum_{i=1}^n E_{\vartheta} \left(\frac{\partial}{\partial \vartheta} \ln f_{\vartheta}(\xi_i) \right) = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \sum_{i=1}^n \int_{-\infty}^{\infty} \left(\frac{\partial}{\partial \vartheta} \ln f_{\vartheta}(x) \right) f_{\vartheta}(x) dx = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \sum_{i=1}^n \int_{-\infty}^{\infty} \frac{\partial}{\partial \vartheta} f_{\vartheta}(x) dx = 0. \end{aligned}$$

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

$$\begin{aligned} \text{cov}_{\vartheta}(\tau, \varrho) &= E_{\vartheta}(\tau \varrho) = \frac{g'(\vartheta)}{I_n(\vartheta)} E_{\vartheta} \left(\tau \frac{\partial}{\partial \vartheta} \ln l_n(\xi_1, \dots, \xi_n, \vartheta) \right) = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \int_{\mathbb{R}^n} T(x_1, \dots, x_n) \frac{\partial}{\partial \vartheta} l_n(x_1, \dots, x_n, \vartheta) dx_1 \cdots dx_n = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \frac{\partial}{\partial \vartheta} \int_{\mathbb{R}^n} T(x_1, \dots, x_n) l_n(x_1, \dots, x_n, \vartheta) dx_1 \cdots dx_n = \\ &= \frac{g'(\vartheta)}{I_n(\vartheta)} \frac{\partial}{\partial \vartheta} E_{\vartheta}(\tau) = \frac{g'(\vartheta)}{I_n(\vartheta)} \frac{\partial}{\partial \vartheta} g(\vartheta) = \frac{(g'(\vartheta))^2}{I_n(\vartheta)}. \end{aligned}$$

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

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. [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 bizonyítsuk 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 $\bar{\xi}$ véges szórású torzítatlan becslése p -nek és $I_n(p) = \frac{n}{p(1-p)}$. Másrészt $g'(p) = (p)' = 1$ miatt az információs határ $\frac{p(1-p)}{n} = D_p^2(\bar{\xi})$. 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 $D_p^2 T(\xi_1, \dots, \xi_n) \geq D_p^2(\bar{\xi})$. Ebből következik az állítás.

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

Megoldás. Korábban láttuk, hogy $\bar{\xi}$ véges szórású torzítatlan becslése m -nek és $I_n(m) = \frac{n}{\sigma^2}$. Másrészt $g'(m) = (m)' = 1$ miatt az információs határ $\frac{\sigma^2}{n} = D_m^2(\bar{\xi})$. 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 $D_m^2 T(\xi_1, \dots, \xi_n) \geq D_m^2(\bar{\xi})$. Ebből következik az állítás.

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

Megoldás. Tudjuk, hogy $\bar{\xi}$ véges szórású torzítatlan becslése λ -nak és $I_n(\lambda) = \frac{n}{\lambda}$. Másrészt $g'(\lambda) = (\lambda)' = 1$ miatt az információs határ $\frac{\lambda}{n} = D_{\lambda}^2(\bar{\xi})$. 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 $D_{\lambda}^2 T(\xi_1, \dots, \xi_n) \geq D_{\lambda}^2(\bar{\xi})$. Ebből következik az állítás.

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

Megoldás. Korábban láttuk, hogy $\bar{\xi}$ 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 $\frac{1}{n\lambda^2} = D_{\lambda}^2(\bar{\xi})$. 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 $D_{\lambda}^2 T(\xi_1, \dots, \xi_n) \geq D_{\lambda}^2(\bar{\xi})$. 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 paraméterter $\Theta \subset \mathbb{R}^r$ nyílt halmaz. Tegyük fel, hogy $E_{\vartheta} \xi^r$ létezik és véges minden $\vartheta = (\vartheta_1, \dots, \vartheta_r) \in \Theta$ esetén, $\frac{\partial}{\partial \vartheta_j} E_{\vartheta} \xi^i$ létezik és folytonos Θ -n minden $i, j \in \{1, \dots, r\}$ esetén, továbbá az úgynevezett Jacobi-determináns*

$$\det \left(\frac{\partial}{\partial \vartheta_j} E_{\vartheta} \xi^i \right) \neq 0$$

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

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

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

Bizonyítás. Legyen

$$G: \Theta \rightarrow \mathbb{R}^r, \quad G(\vartheta) := (E_{\vartheta} \xi^1, \dots, E_{\vartheta} \xi^r).$$

Az inverzfüggvény-tétel miatt (lásd [14, 230. oldal]) az adott feltételekkel G^{-1} (G inverze) létezik és 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 becsléssorozata $E_{\vartheta} \xi^k$ -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

$$P_{\vartheta} \left(\left| \frac{1}{n} \sum_{i=1}^n \xi_i^k - E_{\vartheta} \xi^k \right| \geq \frac{\varepsilon}{\sqrt{r}} \right) < \frac{\delta}{r}, \quad k = 1, \dots, r.$$

Innen kapjuk, hogy

$$P_{\vartheta} \left(\sum_{k=1}^r \left(\frac{1}{n} \sum_{i=1}^n \xi_i^k - E_{\vartheta} \xi^k \right)^2 \geq \varepsilon^2 \right) \leq$$

$$\begin{aligned} &\leq P_{\vartheta} \left(\bigcup_{k=1}^r \left\{ \left(\frac{1}{n} \sum_{i=1}^n \xi_i^k - E_{\vartheta} \xi^k \right)^2 \geq \frac{\varepsilon^2}{r} \right\} \right) \leq \\ &\leq \sum_{k=1}^r P_{\vartheta} \left(\left(\frac{1}{n} \sum_{i=1}^n \xi_i^k - E_{\vartheta} \xi^k \right)^2 \geq \frac{\varepsilon^2}{r} \right) < \delta, \end{aligned}$$

azaz $m_k := \frac{1}{n} \sum_{i=1}^n \xi_i^k$ jelöléssel $(m_1, \dots, m_k) \in G(\Theta)$ legalább $1 - \delta$ valószínűséggel, amennyiben $n > N$. Ebből következik, hogy

$$\lim_{n \rightarrow \infty} P_{\vartheta}((m_1, \dots, m_k) \in G(\Theta)) = 1.$$

Tehát 1-hez tartó valószínűséggel $\hat{\vartheta}_n = G^{-1}(m_1, \dots, m_k)$. $\frac{1}{n} \sum_{i=1}^n \xi_i^k$ erősen konzisztens becsléssorozat $E_{\vartheta} \xi^k$ -nak ($k = 1, \dots, r$), melyből a G^{-1} folytonossága miatt 1 valószínűséggel teljesül, hogy

$$\lim_{n \rightarrow \infty} G^{-1}(m_1, \dots, m_k) = G^{-1}(G(\vartheta)) = \vartheta.$$

Mindezekből

$$P_{\vartheta} \left(\lim_{n \rightarrow \infty} \hat{\vartheta}_n = \vartheta \right) = 1.$$

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

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

Megoldás. Az előző tétel feltételei teljesülnek, így az

$$\frac{1}{n} \sum_{i=1}^n \xi_i = E_{\lambda} \xi = \frac{1}{\lambda}$$

megoldása erősen konzisztens becsléssorozata λ -nak.

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

Megoldás. A következő egyenletrendszert kapjuk:

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

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

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

Megoldás. A következő egyenletrendszert kapjuk:

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

Ennek a megoldása $\hat{a}_n = \bar{\xi} - \sqrt{3}S_n$ és $\hat{b}_n = \bar{\xi} + \sqrt{3}S_n$. 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 $\xi_1, \dots, \xi_n$, továbbá a ξ értékészlete $\mathfrak{X}$, azaz a mintatér $\mathfrak{X}^n$. Ha ξ abszolút folytonos, akkor f_ϑ jelölje ξ -nek a P_ϑ -ból származó sűrűségfüggvényét, ahol $\vartheta = (\vartheta_1, \dots, \vartheta_r) \in \Theta$. Először a már korábban definiált likelihood függvényt terjesztjük ki $\Theta \subset \mathbb{R}^r$ esetre.

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

$$l_n: \mathfrak{X}^n \times \Theta \rightarrow \mathbb{R},$$

$$l_n(x_1, \dots, x_n, \vartheta_1, \dots, \vartheta_r) := \begin{cases} \prod_{i=1}^n f_\vartheta(x_i), & \text{ha } \xi \text{ absz. folyt.}, \\ \prod_{i=1}^n P_\vartheta(\xi_i = x_i), & \text{ha } \xi \text{ diszkrét.} \end{cases}$$

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

$$l_n(\xi_1(\omega), \dots, \xi_n(\omega), \hat{\vartheta}_1(\omega), \dots, \hat{\vartheta}_r(\omega)) \geq l_n(\xi_1(\omega), \dots, \xi_n(\omega), \vartheta_1, \dots, \vartheta_r)$$

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

Tehát a becslés kiszámítása nem más, mint szélsőérték hely keresés. Praktikus okból nem a likelihood függvénynek fogjuk a maximum helyét keresni, hanem a természetes alapú logaritmusának. Ezzel a szélsőérték hely nem változik, hiszen $\ln$ szigorúan monoton növekvő függvény. Az ok az, hogy ekkor nem szorzatot, hanem összeget kell vizsgálni, ami a szélsőérték keresésének deriválásos módszerénél sokkal előnyösebb alak.

3.51. Definíció. A $\xi_1, \dots, \xi_n$ minta *loglikelihood függvénye* $L_n := \ln l_n$, ahol L_n értelmezési tartománya azon pontokból áll, ahol l_n pozitív.

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

Megoldás. A loglikelihood függvény

$$L_n(\xi_1, \dots, \xi_n, a, b) = -n \ln(b - a), \text{ ha } \xi_1^* \geq a \text{ és } \xi_n^* \leq b,$$

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

3.53. Feladat. Legyen ξ Poisson-eloszlású λ paraméterrel. Számoljuk 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

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

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

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

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

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

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

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

Megoldás. A loglikelihood függvény

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

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

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

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

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

továbbá $AB - C^2 = \frac{2n^2}{S_n^4} > 0$, így $(\bar{\xi}, S_n)$ lokális maximumhely. Mivel $\mathbb{R} \times \mathbb{R}_+$ összefüggő halmaz, és csak egy lokális szélsőérték hely van, ezért $(\bar{\xi}, S_n)$ globális maximumhely. Tehát a maximum likelihood becslése m -nek $\widehat{m} = \bar{\xi}$, 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 $E_{\vartheta^*} L_1(\xi, \vartheta)$ létezik és véges minden $\vartheta \in U$ esetén, akkor a likelihood egyenletnek van olyan $\widehat{\vartheta}$ megoldása, amelyre teljesül, hogy*

$$P_{\vartheta^*} \left(\lim_{n \rightarrow \infty} \widehat{\vartheta} = \vartheta^* \right) = 1,$$

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

$$E_{\vartheta^*} L_1(\xi, \vartheta) - E_{\vartheta^*} L_1(\xi, \vartheta^*) = E_{\vartheta^*} \ln \frac{l_1(\xi, \vartheta)}{l_1(\xi, \vartheta^*)} \leq$$

$$\leq \ln E_{\vartheta^*} \frac{l_1(\xi, \vartheta)}{l_1(\xi, \vartheta^*)} = \ln \int_{-\infty}^{\infty} f_{\vartheta}(x) dx = \ln 1 = 0,$$

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

$$E_{\vartheta^*} L_1(\xi, \vartheta) < E_{\vartheta^*} L_1(\xi, \vartheta^*).$$

A Kolmogorov-féle nagy számok erős törvénye és

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

miatt

$$P_{\vartheta^*} \left(\lim_{n \rightarrow \infty} \frac{1}{n} L_n(\xi_1, \dots, \xi_n, \vartheta) = E_{\vartheta^*} L_1(\xi, \vartheta) \right) = 1$$

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

$$P_{\vartheta^*} \left(\lim_{n \rightarrow \infty} \frac{1}{n} L_n(\xi_1, \dots, \xi_n, \vartheta) < \lim_{n \rightarrow \infty} \frac{1}{n} L_n(\xi_1, \dots, \xi_n, \vartheta^*) \right) = 1$$

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

$$P_{\vartheta^*} \left(L_n(\xi_1, \dots, \xi_n, \vartheta) < L_n(\xi_1, \dots, \xi_n, \vartheta^*) \right) = 1$$

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

$$P_{\vartheta^*} \left(L_n(\xi_1, \dots, \xi_n, \vartheta^* \pm \delta) < L_n(\xi_1, \dots, \xi_n, \vartheta^*) \right) = 1,$$

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. [2, 90. oldal]).

4. fejezet

Intervallumbecslések

4.1. Az intervallumbecslés feladata

Legyen ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ statisztikai mezőn, ahol $\Theta \subset \mathbb{R}^v$ nyílt halmaz. A feladat $(\vartheta_1, \dots, \vartheta_v) \in \Theta$, $k \in \{1, \dots, v\}$ 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 $\xi_1, \dots, \xi_n$, továbbá

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

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

$$P_\vartheta(\tau_1 \leq \vartheta_k \leq \tau_2) \geq 1 - \alpha$$

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

$$P_\vartheta(\vartheta_k < \tau_1) = P_\vartheta(\vartheta_k > \tau_2)$$

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

$$\inf_{\vartheta \in \Theta} P_\vartheta(\tau_1 \leq \vartheta_k \leq \tau_2)$$

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

A megoldáshoz szükségünk lesz a következő tételre.

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

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

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

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

Ezzel bizonyított az állítás. □

Most térjünk vissza a feladatra.

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

$$P_m\left(-c \leq \frac{\bar{\xi} - m}{\sigma} \sqrt{n} \leq c\right) = \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

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

$$P_m\left(\bar{\xi} - \frac{\sigma}{\sqrt{n}}c \leq m \leq \bar{\xi} + \frac{\sigma}{\sqrt{n}}c\right) = 1 - \alpha.$$

Könnyű látni, hogy ez centrált konfidenciaintervallum, hiszen

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

Összefoglalva tehát a megoldás:

$$\begin{aligned} \tau_1 &:= \bar{\xi} - \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \\ \tau_2 &:= \bar{\xi} + \frac{\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \end{aligned}$$

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

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

A megoldáshoz szükségünk lesz a következő tételre.

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

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

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

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

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

$$\begin{aligned} P_\sigma \left(\sum_{i=1}^n \frac{(\xi_i - m)^2}{\sigma^2} < c_1 \right) &= F(c_1), \\ P_\sigma \left(\sum_{i=1}^n \frac{(\xi_i - m)^2}{\sigma^2} > c_2 \right) &= 1 - F(c_2). \end{aligned}$$

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

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

azaz átrendezve

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

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

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

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

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

A megoldáshoz szükségünk lesz a következő tételre.

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

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

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 $\frac{1}{\sqrt{n}}\eta_1 = \bar{\xi} - m$, továbbá

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

Mindezekből a Steiner-formula alapján

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

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,

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

és az U ortonormáltsága miatt

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

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

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

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

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

Most rátérünk a feladat megoldására.

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

$$\begin{aligned} P_{(m,\sigma)} \left(\frac{S_n^2}{\sigma^2} n < c_1 \right) &= F(c_1), \\ P_{(m,\sigma)} \left(\frac{S_n^2}{\sigma^2} n > c_2 \right) &= 1 - F(c_2). \end{aligned}$$

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

$$P_{(m,\sigma)} \left(c_1 \leq \frac{S_n^2}{\sigma^2} n \leq c_2 \right) = 1 - \alpha,$$

azaz átrendezve

$$P_{(m,\sigma)} \left(S_n \sqrt{\frac{n}{c_2}} \leq \sigma \leq S_n \sqrt{\frac{n}{c_1}} \right) = 1 - \alpha.$$

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

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

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

A megoldáshoz szükségünk lesz a következő tételre.

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

$$\frac{\bar{\xi} - m}{S_n^*} \sqrt{n} \in T(n-1).$$

Bizonyítás. Korábban láttuk, hogy

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

továbbá ezek függetlenek. Így

$$\frac{\sqrt{n-1} \frac{\bar{\xi} - m}{\sigma} \sqrt{n}}{\sqrt{\frac{S_n^2}{\sigma^2} n}} = \frac{\bar{\xi} - m}{S_n} \sqrt{n-1} = \frac{\bar{\xi} - m}{S_n^*} \sqrt{n} \in T(n-1). \quad \square$$

Rátérünk a feladat megoldására.

Megoldás. Legyen $c \in \mathbb{R}_+$ és $F = F[T(n-1)]$. Ekkor az előző tétel szerint

$$P_{(m,\sigma)} \left(-c \leq \frac{\bar{\xi} - m}{S_n^*} \sqrt{n} \leq c \right) = F(c) - F(-c) = 2F(c) - 1.$$

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

$$P_{(m,\sigma)} \left(\bar{\xi} - \frac{S_n^*}{\sqrt{n}} c \leq m \leq \bar{\xi} + \frac{S_n^*}{\sqrt{n}} c \right) = 1 - \alpha.$$

Könnyű látni, hogy ez centrált konfidenciaintervallum, hiszen

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

Összefoglalva tehát a megoldás:

$$\begin{aligned} F &:= F[T(n-1)] \\ \tau_1 &:= \bar{\xi} - \frac{S_n^*}{\sqrt{n}} F^{-1} \left(1 - \frac{\alpha}{2} \right) \\ \tau_2 &:= \bar{\xi} + \frac{S_n^*}{\sqrt{n}} F^{-1} \left(1 - \frac{\alpha}{2} \right) \end{aligned}$$

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

Megoldás. Mivel $x > 0$ esetén

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

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

$$\lambda\xi_1 + \dots + \lambda\xi_n = \lambda n\bar{\xi} \in \text{Gamma}(n; 1).$$

Így $c_1, c_2 \in \mathbb{R}_+$ és $F = F[\text{Gamma}(n; 1)]$ esetén

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

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

$$P_\lambda \left(c_1 \leq \lambda n \bar{\xi} \leq c_2 \right) = 1 - \alpha,$$

azaz átrendezve

$$P_\lambda \left(\frac{c_1}{n \bar{\xi}} \leq \lambda \leq \frac{c_2}{n \bar{\xi}} \right) = 1 - \alpha.$$

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

$$\begin{aligned} F &:= F[\text{Gamma}(n; 1)] \\ \tau_1 &:= \frac{1}{n \bar{\xi}} F^{-1} \left(\frac{\alpha}{2} \right) \\ \tau_2 &:= \frac{1}{n \bar{\xi}} F^{-1} \left(1 - \frac{\alpha}{2} \right) \end{aligned}$$

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

4.4. Konfidenciaintervallum valószínűségre

4.13. Feladat. Legyen $\xi \in \text{Bin}(1; p)$ és $\xi_1, \dots, \xi_n$ egy ξ -re vonatkozó minta. Tegyük fel, hogy p ismeretlen. Adjunk 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 konfidenciaintervallumot. (Ekkor $\bar{\xi}$ az esemény relatív gyakoriságát jelenti n kísérlet után.)

Megoldás. Bizonyítható, hogy

$$\begin{aligned} \tau_1 &:= \frac{1}{n} \max \left\{ z \in \mathbb{N} : \sum_{i=0}^z \binom{n}{i} \bar{\xi}^i (1 - \bar{\xi})^{n-i} < \frac{\alpha}{2} \right\} \\ \tau_2 &:= \frac{1}{n} \min \left\{ z \in \mathbb{N} : \sum_{i=0}^z \binom{n}{i} \bar{\xi}^i (1 - \bar{\xi})^{n-i} \geq 1 - \frac{\alpha}{2} \right\} \end{aligned}$$

jelölésekkel $[\tau_1, \tau_2]$ $1 - \alpha$ biztonsági szintű konfidenciaintervallum p -re. Ennek bizonyítása azon múlik, hogy $n \bar{\xi} \in \text{Bin}(n; p)$, de itt nem részletezzük (lásd [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 $n \bar{\xi} \in \text{Bin}(n; p)$ miatt $c \in \mathbb{R}_+$ esetén

$$P_p \left(-c \leq \frac{n \bar{\xi} - np}{\sqrt{np(1-p)}} \leq c \right) \simeq \Phi(c) - \Phi(-c) = 2\Phi(c) - 1.$$

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

$$P_p \left(\left(1 + \frac{c^2}{n} \right) p^2 - \left(2\bar{\xi} + \frac{c^2}{n} \right) p + \bar{\xi}^2 \leq 0 \right) \simeq 1 - \alpha.$$

A p -ben másodfokú

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

polinom gyökei

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

így

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

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:

$$\begin{aligned} \tau_1 &= \bar{\xi} - \frac{c}{\sqrt{n}} \sqrt{\bar{\xi}(1 - \bar{\xi})} \\ \tau_2 &= \bar{\xi} + \frac{c}{\sqrt{n}} \sqrt{\bar{\xi}(1 - \bar{\xi})}. \end{aligned}$$

4.5. Általános módszer konfidenciaintervallum készítésére

Legyen ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{ P_{\vartheta} : \vartheta \in \Theta \}$ 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

$$P_{\vartheta} (-\ln F_{\vartheta}(\xi) < x) = P_{\vartheta} (\xi > F_{\vartheta}^{-1}(e^{-x})) = 1 - F_{\vartheta}(F_{\vartheta}^{-1}(e^{-x})) = 1 - e^{-x},$$

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

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

Így $c_1, c_2 \in \mathbb{R}_+$ és $F = F[\text{Gamma}(n; 1)]$ esetén

$$\begin{aligned} P_\vartheta \left(-\sum_{i=1}^n \ln F_\vartheta(\xi_i) < c_1 \right) &= F(c_1), \\ P_\vartheta \left(-\sum_{i=1}^n \ln F_\vartheta(\xi_i) > c_2 \right) &= 1 - F(c_2). \end{aligned}$$

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

$$P_\vartheta \left(c_1 \leq -\sum_{i=1}^n \ln F_\vartheta(\xi_i) \leq c_2 \right) = 1 - \alpha.$$

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 $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta. Adjunk b -re centrált konfidenciaintervallumot, melynek $1 - \alpha$ a biztonsági szintje.

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

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

egyenlőtlenséget kell b -re rendezni. Azt kapjuk, hogy

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

így a feladat megoldása:

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

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

5. fejezet

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, elfogadási és kritikus tartomány

Tegyük fel, hogy a $\xi_1, \xi_2, \dots, \xi_k$ valószínűségi változókkal kapcsolatos H_0 . A ξ_i -re vonatkozzon a $\xi_{i1}, \xi_{i2}, \dots, \xi_{in_i}$ minta ($i = 1, \dots, k$). Legyen

$$C_0 \subset \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \dots \times \mathbb{R}^{n_k}.$$

Vezessük be a

$$\xi_{\text{minta}} := (\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}, \xi_{21}, \xi_{22}, \dots, \xi_{2n_2}, \dots, \xi_{k1}, \xi_{k2}, \dots, \xi_{kn_k})$$

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

$$\xi_{\text{minta}}(\omega) \in C_0,$$

azaz a mintarealizáció benne van C_0 -ban, 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.

5.1.3. Statisztikai próba terjedelme és torzítatlansága

Döntésünk lehet helyes vagy helytelen az alábbiak szerint:

	H_0 -t elfogadjuk	H_0 -t elutasítjuk
H_0 igaz	helyes döntés	<i>elsőfajú hiba</i>
H_1 igaz	<i>másodfajú hiba</i>	helyes döntés

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

$$P(\xi_{\text{minta}} \in C_1) \leq \alpha \quad \forall P \in \mathcal{P}_{H_0}$$

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

$$\sup_{P \in \mathcal{P}_{H_0}} P(\xi_{\text{minta}} \in C_1) = \alpha.$$

Ha a vizsgált valószínűségi változók diszkrét, 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

$$P(\xi_{\text{minta}} \in C_1) \geq \alpha \quad \forall P \in \mathcal{P}_{H_1}$$

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.

5.1.4. Próbastatisztika

Elfogadási tartomány konstruálásához H_0 esetén ismert eloszlású

$$\tau := T(\xi_{\text{minta}})$$

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

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

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

$$C_0 := \{x \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \times \cdots \times \mathbb{R}^{n_k} : T(x) \in I_\tau\}.$$

Mivel a $\xi_{\text{minta}} \in C_0$ 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.

5.1.5. 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.

5.1.6. 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 születésének az indikátorváltozóját, és a ξ -re vonatkozó minta $\xi_1, \dots, \xi_n$, akkor $n\bar{\xi}$ azt jelenti, hogy n -szer alkalmazva az eljárást hány darab üszőborjú született. Az $n\bar{\xi}$ 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. Jelentsé ξ a leszállított tételből egy véletlenszerűen kiválasztott zacskó liszt tömegének eltérését az elvárt 1 kg-tól. Ekkor az a nullhipotézis, hogy $E\xi = 0$. Ha ξ jó közelítéssel normális eloszlásúnak tekinthető, akkor a későbbiekben tárgyalt úgynevezett egy-mintás t-próbánál látni fogjuk, hogy ehhez találhatunk próbastatisztikát. Most az a kérdés, hogy mi legyen az ellenhipotézis. Ha $E\xi \neq 0$ 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 $E\xi > 0$ teljesül, akkor a kereskedőtől vásárlók csak ritkán reklamálnának. Ezért célszerűbb $E\xi < 0$ 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.

5.1.7. A próba erőfüggvénye és konzisztenciája

Ha ξ a vizsgált valószínűségi változó az $(\Omega, \mathcal{F}, \mathcal{P})$, $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$ 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 $H_0: \vartheta = \vartheta_0$ nullhipotézisről, akkor a

$$\gamma: \Theta \rightarrow \mathbb{R}, \quad \gamma(\vartheta) := P_\vartheta((\xi_1, \dots, \xi_n) \in C_1)$$

függvényt a *próba erőfüggvényének* nevezzük. Ha $H_1: \vartheta \in \Theta_1 (\subset \Theta \setminus \{\vartheta_0\})$ és

$$\lim_{n \rightarrow \infty} P_\vartheta((\xi_1, \dots, \xi_n) \in C_1) = 1 \quad \forall \vartheta \in \Theta_1,$$

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 eloszlástípusból származó valószínűségi változók eloszlásainak paramétereire vonatkozik, akkor *paraméteres hipotézisvizsgálatról* beszélünk.

5.2.1. Egymintás u-próba

5.1. Feladat. Legyen $\xi \in \text{Norm}(m; \sigma)$, ahol m ismeretlen és σ ismert, továbbá legyen $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta. A

$$H_0: m = m_0$$

$$H_1: m \neq m_0 \text{ (kétoldali ellenhipotézis)}$$

hipotézisekre adjunk adott α terjedelmű próbát, ahol $m_0 \in \mathbb{R}$ rögzített. A feladatot oldjuk meg $H_1: m < m_0$ és $H_1: m > m_0$ úgynevezett *egyoldali ellenhipotézisekre* is.

Megoldás. Először próbastatisztikát adunk. Korábban már bizonyítottuk, hogy H_0 teljesülése esetén

$$u := \frac{\bar{\xi} - m_0}{\sigma} \sqrt{n} \in \text{Norm}(0; 1).$$

A kritikus tartomány megadásánál vegyük figyelembe, hogy $\bar{\xi}$ az m torzítatlan becslése, így H_1 teljesülése esetén $\bar{\xi} - m_0$ várhatóan kritikus értékben eltávolodik 0-tól. Következésképpen a standard normális eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|u| \leq a$ ($a > 0$) alakú. Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(|u| \leq a) = 2\Phi(a) - 1,$$

így $P(|u| \leq a) = 1 - \alpha$ esetén $a = \Phi^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát $|u| \leq \Phi^{-1}(1 - \frac{\alpha}{2})$, azaz

$$2 - 2\Phi(|u|) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek a pontos terjedelme α . Ezt a statisztikai próbát nevezzük *egymintás u-próbának*.

Most legyen az ellenhipotézis $H_1: m < m_0$. Ennek teljesülése esetén $\bar{\xi} - m_0$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $u \geq b$ ($b < 0$) jellegű. Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(u \geq b) = 1 - \Phi(b),$$

így $P(u \geq b) = 1 - \alpha$ esetén $b = \Phi^{-1}(\alpha) < 0$. Tehát $u \geq \Phi^{-1}(\alpha)$, azaz

$$\Phi(u) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \leq m_0$, azaz $u \leq 0$, akkor $\Phi(u) = 1 - \Phi(-u) = 1 - \Phi(|u|)$ miatt az elfogadási tartomány

$$1 - \Phi(|u|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \geq m_0$, azaz $u \geq 0$, továbbá $0 < \alpha < 0,5$, akkor $\Phi(u) \geq \Phi(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m < m_0$ ellenében.

Ezután legyen az ellenhipotézis $H_1: m > m_0$. Ennek teljesülése esetén $\bar{\xi} - m_0$ várhatóan kritikus értékben 0 fölött van. Így az elfogadási tartomány $u \leq c$ ($c > 0$) jellegű. Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(u \leq c) = \Phi(c),$$

így $P(u \leq c) = 1 - \alpha$ esetén $c = \Phi^{-1}(1 - \alpha) > 0$. Tehát $u \leq \Phi^{-1}(1 - \alpha)$, azaz

$$1 - \Phi(u) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \geq m_0$, azaz $u \geq 0$, akkor $1 - \Phi(u) = 1 - \Phi(|u|)$ miatt az elfogadási tartomány

$$1 - \Phi(|u|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \leq m_0$, azaz $u \leq 0$, továbbá $0 < \alpha < 0,5$, akkor $1 - \Phi(u) = \Phi(-u) \geq \Phi(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m > m_0$ ellenében.

5.2. Tétel. *Az egymintás u -próba torzítatlan és konzisztens, továbbá az elsőfajú hiba valószínűségének csökkentésével a másodfajú hiba valószínűsége nő.*

Bizonyítás. Először számoljuk ki az u várható értékét és szórását:

$$\begin{aligned} E_m u &= \frac{E_m \bar{\xi} - m_0}{\sigma} \sqrt{n} = \frac{m - m_0}{\sigma} \sqrt{n} \\ D_m u &= \frac{\sqrt{n}}{\sigma} D_m \bar{\xi} = \frac{\sqrt{n}}{\sigma} \sqrt{\frac{1}{n^2} n \sigma^2} = 1. \end{aligned}$$

Mivel u az m bármely értéke esetén normális eloszlású, ezért azt kapjuk, hogy

$$u \in \text{Norm} \left(\frac{m - m_0}{\sigma} \sqrt{n}; 1 \right).$$

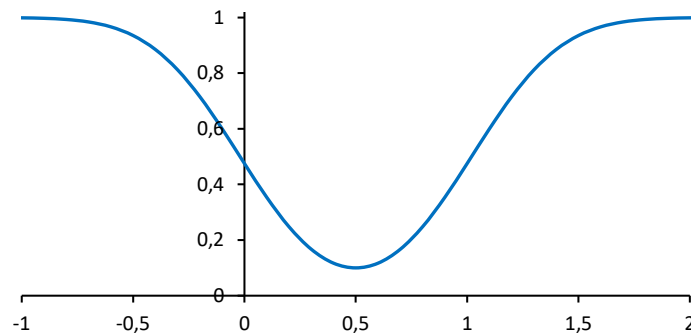
Most tekintsük a kétoldali ellenhipotézis esetét. Ekkor $u_{\alpha/2} := \Phi^{-1}(1 - \frac{\alpha}{2})$ jelöléssel az erőfüggvény

$$\begin{aligned} \gamma(m) &= P_m((\xi_1, \dots, \xi_n) \in C_1) = \\ &= P_m(|u| > u_{\alpha/2}) = 1 - P_m(-u_{\alpha/2} \leq u \leq u_{\alpha/2}) = \\ &= 1 - \Phi \left(u_{\alpha/2} - \frac{m - m_0}{\sigma} \sqrt{n} \right) + \Phi \left(-u_{\alpha/2} - \frac{m - m_0}{\sigma} \sqrt{n} \right). \end{aligned}$$

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 $\lim_{m \rightarrow \infty} \gamma(m) = \lim_{m \rightarrow -\infty} \gamma(m) = 1$. Az 5.1. ábrán γ grafikonját láthatjuk $\sigma = 1$, $m_0 = 0,5$, $\alpha = 0,1$, $n = 10$ paraméterekkel.

Mindezek alapján tehát, ha $H_1: m \neq m_0$ 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 $\lim_{n \rightarrow \infty} \gamma(m) = 1$, melyből már következik, hogy a próba konzisztens, azaz a mintaelemek számának növelésével a másodfajú hiba valószínűsége 0-hoz tart.

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 $\frac{d\gamma}{du_{\alpha/2}} < 0$, azaz γ csökkenő. Mindezekből tehát kapjuk, hogy α csökkentésével γ is csökken, azaz a másodfajú hiba valószínűsége nő.

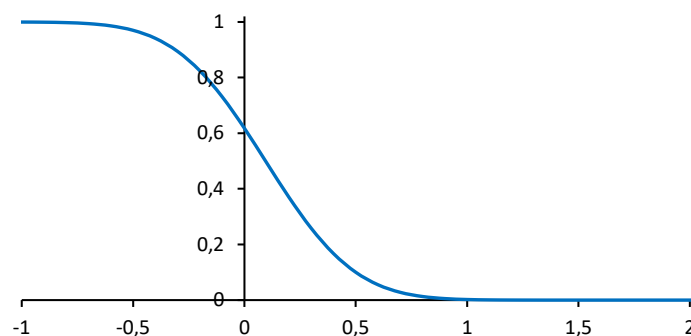


5.1. ábra

Ezután tekintsük a $H_1: m < m_0$ egyoldali ellenhipotézist. Ekkor az erőfüggvény $u_\alpha = \Phi^{-1}(\alpha)$ jelöléssel

$$\gamma(m) = P_m((\xi_1, \dots, \xi_n) \in C_1) = P_m(u < u_\alpha) = \Phi\left(u_\alpha - \frac{m - m_0}{\sigma} \sqrt{n}\right).$$

Φ szigorúan monoton növekvő, ezért γ szigorúan monoton csökkenő. Az is könnyen látható, hogy $\gamma(m_0) = \alpha$, $\lim_{m \rightarrow \infty} \gamma(m) = 0$ és $\lim_{m \rightarrow -\infty} \gamma(m) = 1$. Az 5.2. ábrán γ grafikonját láthatjuk $\sigma = 1$, $m_0 = 0,5$, $\alpha = 0,1$, $n = 10$ paraméterekkel.



5.2. ábra

Mindezek alapján, ha $H_1: m < m_0$ teljesül, akkor $\gamma(m) > \alpha$, melyből következik, hogy a próba torzítatlan. Ha γ -t mint n függvényét tekintjük, akkor minden $m < m_0$ esetén $\lim_{n \rightarrow \infty} \gamma(m) = 1$, 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 $H_1: m > m_0$ eset tárgyalását az Olvasóra bízuk. □

5.2.2. Kétmintás u-próba

5.3. Feladat. Legyen $\xi \in \text{Norm}(m_1; \sigma_1)$, $\eta \in \text{Norm}(m_2; \sigma_2)$ független valószínűségi változók, ahol m_1, m_2 ismeretlenek és σ_1, σ_2 ismertek. Legyen $\xi_1, \dots, \xi_{n_1}$ a ξ -re vonatkozó, illetve $\eta_1, \dots, \eta_{n_2}$ az η -ra vonatkozó minta. A

$$H_0: m_1 = m_2$$

$$H_1: m_1 \neq m_2 \text{ (kétoldali ellenhipotézis)}$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: m_1 < m_2$ illetve $H_1: m_1 > m_2$ *egyoldali ellenhipotézisekre* is.

Megoldás. Ha H_0 igaz, akkor könnyen látható, hogy

$$u := \frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in \text{Norm}(0; 1).$$

Először vizsgáljuk a kétoldali ellenhipotézist. Ha ez teljesül, akkor $\bar{\xi} - \bar{\eta}$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a standard normális eloszlás szimmetriája miatt az elfogadási tartomány $|u| \leq a$ ($a > 0$) alakú. Ebből az egymintás u -próbával megegyező módon bizonyítható, hogy az elfogadási tartomány $|u| \leq \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$, azaz

$$2 - 2\Phi(|u|) \geq \alpha.$$

Ezzel olyan próbát kapunk, melynek pontos terjedelme α .

Most legyen $H_1: m_1 < m_2$. Ha ez teljesül, akkor $\bar{\xi} - \bar{\eta}$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $u \geq b$ ($b < 0$) jellegű. Ebből az egymintás esettel megegyező módon bizonyítható, hogy $u \geq \Phi^{-1}(\alpha)$, azaz

$$\Phi(u) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \leq \bar{\eta}$, azaz $u \leq 0$, akkor $\Phi(u) = 1 - \Phi(-u) = 1 - \Phi(|u|)$ miatt az elfogadási tartomány

$$1 - \Phi(|u|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \geq \bar{\eta}$, azaz $u \geq 0$, továbbá $0 < \alpha < 0,5$, akkor $\Phi(u) \geq \Phi(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m_1 < m_2$ ellenében.

Hasonlóan, $H_1: m_1 > m_2$ esetén $u \leq \Phi^{-1}(1 - \alpha)$ azaz

$$1 - \Phi(u) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \geq \bar{\eta}$, azaz $u \geq 0$, akkor $1 - \Phi(u) = 1 - \Phi(|u|)$ miatt az elfogadási tartomány

$$1 - \Phi(|u|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \leq \bar{\eta}$, azaz $u \leq 0$, továbbá $0 < \alpha < 0,5$, akkor $1 - \Phi(u) = \Phi(-u) \geq \Phi(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m_1 > m_2$ ellenében.

Ezt a statisztikai próbát nevezzük *kétmintás u -próbának*.

5.4. Tétel. *A kétmintás u -próba torzítatlan és konzisztens, továbbá az elsőfajú hiba valószínűségének csökkentésével a másodfajú hiba valószínűsége nő.*

Megoldás. Csak kétoldali ellenhipotézisre bizonyítunk, az egyoldaliakat az Olvasóra bízunk. Könnyen látható, hogy az m_1, m_2 bármely értékei esetén

$$u \in \text{Norm} \left(\frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}; 1 \right),$$

így $u_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$ jelöléssel

$$\begin{aligned} \gamma(m_1, m_2) &:= P_{(m_1, m_2)}((\xi_1, \dots, \xi_{n_1}, \eta_1, \dots, \eta_{n_2}) \in C_1) = \\ &= P_{(m_1, m_2)}(|u| > u_{\alpha/2}) = 1 - P_{(m_1, m_2)}(-u_{\alpha/2} \leq u \leq u_{\alpha/2}) = \\ &= 1 - \Phi \left(u_{\alpha/2} - \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right) + \Phi \left(-u_{\alpha/2} - \frac{m_1 - m_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right). \end{aligned}$$

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.

Ha γ -t, mint n_1, n_2 szerinti kétváltozós függvényt tekintjük, akkor $n_1 \rightarrow \infty, n_2 \rightarrow \infty$ esetén a határértéke 1, azaz a próba konzisztens.

Végül az utolsó állítást hasonlóan kell belátni, mint az egymintás u-próbánál.

5.2.3. Egymintás t-próba

5.5. Feladat. Legyen $\xi \in \text{Norm}(m; \sigma)$, ahol m és σ ismeretlenek, továbbá legyen $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta ($n \geq 2$). A

$$H_0: m = m_0$$

$$H_1: m \neq m_0$$

hipotézisekre adjunk adott α terjedelmű próbát, ahol $m_0 \in \mathbb{R}$ rögzített. A feladatot oldjuk meg $H_1: m < m_0$ illetve $H_1: m > m_0$ egyoldali ellenhipotézisekre is.

Megoldás. Korábban bizonyítottuk, hogy H_0 teljesülése esetén

$$t := \frac{\bar{\xi} - m_0}{S_n^*} \sqrt{n} \in T(n-1).$$

A kétoldali ellenhipotézis teljesülése esetén $\bar{\xi} - m_0$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a t-eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|t| \leq a$ ($a > 0$) alakú. A továbbiakban legyen

$$F = F[T(n-1)].$$

Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(|t| \leq a) = 2F(a) - 1,$$

így $P(|t| \leq a) = 1 - \alpha$ esetén $a = F^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát $|t| \leq F^{-1}(1 - \frac{\alpha}{2})$, azaz

$$2 - 2F(|t|) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α .

Most legyen $H_1: m < m_0$. Ennek teljesülésekor $\bar{\xi} - m_0$ várhatóan kritikus értékben 0 alatt van. Így az elfogadási tartomány $t \geq b$ ($b < 0$) jellegű. Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(t \geq b) = 1 - F(b),$$

így $P(t \geq b) = 1 - \alpha$ esetén $b = F^{-1}(\alpha) < 0$. Tehát $t \geq F^{-1}(\alpha)$, azaz

$$F(t) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \leq m_0$, azaz $t \leq 0$, akkor $F(t) = 1 - F(-t) = 1 - F(|t|)$ miatt az elfogadási tartomány

$$1 - F(|t|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \geq m_0$, azaz $t \geq 0$, továbbá $0 < \alpha < 0,5$, akkor $F(t) \geq F(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m < m_0$ ellenében.

Hasonlóan, $H_1: m > m_0$ esetén $t \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(t) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek α a pontos terjedelme. Ha $\bar{\xi} \geq m_0$, azaz $t \geq 0$, akkor $1 - F(t) = 1 - F(|t|)$ miatt az elfogadási tartomány

$$1 - F(|t|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \leq m_0$, azaz $t \leq 0$, továbbá $0 < \alpha < 0,5$, akkor $1 - F(t) = F(-t) \geq F(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m > m_0$ ellenében.

Ezt a statisztikai próbát nevezzük *egymintás t-próbának*.

5.6. Megjegyzés. 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

5.7. Feladat. Legyenek $\xi \in \text{Norm}(m_1; \sigma_1)$, $\eta \in \text{Norm}(m_2; \sigma_2)$ független valószínűségi változók, ahol $m_1, m_2, \sigma_1, \sigma_2$ ismeretlenek és $\sigma_1 = \sigma_2$. Legyen $\xi_1, \dots, \xi_{n_1}$ a ξ -re vonatkozó, illetve $\eta_1, \dots, \eta_{n_2}$ az η -ra vonatkozó minta ($n_1 \geq 2$, $n_2 \geq 2$). A

$$H_0: m_1 = m_2$$

$$H_1: m_1 \neq m_2$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: m_1 < m_2$ illetve $H_1: m_1 > m_2$ egyoldali ellenhipotézisekre is.

A feladat megoldásához szükségünk lesz a következő tételre.

5.8. Tétel. Ha $\xi, \eta \in \text{Norm}(m; \sigma)$ függetlenek, $\xi_1, \dots, \xi_{n_1}$ a ξ -re, illetve $\eta_1, \dots, \eta_{n_2}$ az η -ra vonatkozó minta ($n_1 \geq 2, n_2 \geq 2$), akkor

$$\frac{\bar{\xi} - \bar{\eta}}{\sqrt{n_1 S_{\xi, n_1}^2 + n_2 S_{\eta, n_2}^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \in T(n_1 + n_2 - 2).$$

Bizonyítás. Korábban már bizonyítottuk, hogy $\bar{\xi}, \bar{\eta}, S_{\xi, n_1}, S_{\eta, n_2}$ függetlenek, továbbá

$$\begin{aligned} X &:= \frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}} \in \text{Norm}(0; 1) \\ Y &:= \frac{S_{\xi, n_1}^2}{\sigma^2} n_1 \in \text{Khi}(n_1 - 1) \\ Z &:= \frac{S_{\eta, n_2}^2}{\sigma^2} n_2 \in \text{Khi}(n_2 - 1). \end{aligned}$$

Ezekből kapjuk, hogy $W := Y + Z \in \text{Khi}(n_1 + n_2 - 2)$, továbbá $\frac{X\sqrt{n_1+n_2-2}}{\sqrt{W}} \in T(n_1 + n_2 - 2)$. Könnyen látható, hogy

$$\frac{X\sqrt{n_1+n_2-2}}{\sqrt{W}} = \frac{\bar{\xi} - \bar{\eta}}{\sqrt{n_1 S_{\xi, n_1}^2 + n_2 S_{\eta, n_2}^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}},$$

melyből kapjuk a tételt. □

Most térjünk vissza a feladat megoldásához.

Megoldás. Az előző tételben bizonyítottuk, hogy H_0 esetén

$$t := \frac{\bar{\xi} - \bar{\eta}}{\sqrt{n_1 S_{\xi, n_1}^2 + n_2 S_{\eta, n_2}^2}} \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \in T(n_1 + n_2 - 2).$$

Speciálisan $n := n_1 = n_2$ esetén

$$t = \frac{\bar{\xi} - \bar{\eta}}{\sqrt{S_{\xi, n}^2 + S_{\eta, n}^2}} \sqrt{n-1} \in T(2n-2).$$

A kétoldali ellenhipotézis teljesülésekor $\bar{\xi} - \bar{\eta}$ várhatóan kritikus értékben messze van 0-tól. Következésképpen a t-eloszlás szimmetriája miatt célszerűnek tűnik, ha az elfogadási tartomány $|t| \leq a$ ($a > 0$) alakú. A továbbiakban legyen

$$F = F[T(n_1 + n_2 - 2)].$$

Ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(|t| \leq a) = 2F(a) - 1,$$

így $P(|t| \leq a) = 1 - \alpha$ esetén $a = F^{-1}(1 - \frac{\alpha}{2}) > 0$. Tehát $|t| \leq F^{-1}(1 - \frac{\alpha}{2})$, azaz

$$2 - 2F(|t|) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α .

Hasonlóan az egymintás t-próbához kapjuk, hogy $H_1: m_1 < m_2$ esetén $t \geq \geq F^{-1}(\alpha)$, azaz

$$F(t) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \leq \bar{\eta}$, azaz $t \leq 0$, akkor $F(t) = 1 - F(-t) = 1 - F(|t|)$ miatt az elfogadási tartomány

$$1 - F(|t|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \geq \bar{\eta}$, azaz $t \geq 0$, továbbá $0 < \alpha < 0,5$, akkor $F(t) \geq \geq F(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m_1 < m_2$ ellenében.

Szintén az egymintás t-próbához hasonlóan $H_1: m_1 > m_2$ esetén $t \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(t) \geq \alpha$$

elfogadási tartománnyal olyan próbát kapunk, melynek pontos terjedelme α . Ha $\bar{\xi} \geq \bar{\eta}$, azaz $t \geq 0$, akkor $1 - F(t) = 1 - F(|t|)$ miatt az elfogadási tartomány

$$1 - F(|t|) \geq \alpha$$

alakban is írható. Ha $\bar{\xi} \leq \bar{\eta}$, azaz $t \leq 0$, továbbá $0 < \alpha < 0,5$, akkor $1 - F(t) = = F(-t) \geq F(0) = 0,5 > \alpha$, így ebben az esetben mindig H_0 mellett döntünk $H_1: m_1 > m_2$ ellenében.

Ezt a statisztikai próbát nevezzük *kétmintás t-próbának*.

5.2.5. Scheffé-módszer

5.9. Feladat. Oldjuk meg az 5.7. feladatot akkor is, ha az ismeretlen szórások viszonyát nem ismerjük.

Szükségünk lesz a következő tételre.

5.10. Tétel. Legyenek $\xi \in \text{Norm}(m_1; \sigma_1)$, $\eta \in \text{Norm}(m_2; \sigma_2)$ 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 ($2 \leq n_1 \leq n_2$). Ekkor $m := m_1 - m_2$ és $\sigma := \sqrt{\sigma_1^2 + \frac{n_1}{n_2} \sigma_2^2}$ jelölésekkel

$$\xi_i - \sqrt{\frac{n_1}{n_2}} \eta_i + \frac{1}{\sqrt{n_1 n_2}} \sum_{k=1}^{n_1} \eta_k - \bar{\eta} \in \text{Norm}(m; \sigma) \quad (i = 1, \dots, n_1)$$

független valószínűségi változók.

Bizonyítás. Az állítás $n_1 = n_2$ esetén triviális. Legyen $2 \leq n_1 < n_2$ és

$$\zeta_i := \xi_i - \sqrt{\frac{n_1}{n_2}} \eta_i + \frac{1}{\sqrt{n_1 n_2}} \sum_{k=1}^{n_1} \eta_k - \bar{\eta} \quad (i = 1, \dots, n_1).$$

Ekkor

$$E \zeta_i = m_1 - \sqrt{\frac{n_1}{n_2}} m_2 + \frac{1}{\sqrt{n_1 n_2}} n_1 m_2 - m_2 = m_1 - m_2 = m,$$

másrészt $K_i := \{1, \dots, n_1\} \setminus \{i\}$ és $K := \{n_1 + 1, \dots, n_2\}$ jelölésekkel

$$\zeta_i = \xi_i + \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} \right) \sum_{k \in K_i} \eta_k - \frac{1}{n_2} \sum_{k \in K} \eta_k + \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}} \right) \eta_i,$$

így

$$D^2 \zeta_i = \sigma_1^2 + \left((n_1 - 1) \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} \right)^2 + \frac{n_2 - n_1}{n_2^2} + \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}} \right)^2 \right) \sigma_2^2,$$

melyből kapjuk, hogy $D^2 \zeta_i = \sigma_1^2 + \frac{n_1}{n_2} \sigma_2^2 = \sigma^2$. 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 $\zeta_i \in \text{Norm}(m; \sigma)$. Még a függetlenséget kell belátni. Ehhez elég a $\text{cov}(\zeta_i, \zeta_j) = 0$ ($i, j = 1, \dots, n_1, i \neq j$) megmutatása.

$$\begin{aligned} \sum_{k=1}^{n_2} a_k^{(i)} a_k^{(j)} &= (n_1 - 2) \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} \right)^2 + \\ &+ 2 \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} \right) \left(\frac{1}{\sqrt{n_1 n_2}} - \frac{1}{n_2} - \sqrt{\frac{n_1}{n_2}} \right) + \frac{n_2 - n_1}{n_2^2} = 0, \end{aligned}$$

ezért $i, j = 1, \dots, n_1, i \neq j$ esetén

$$\begin{aligned} \text{cov}(\zeta_i, \zeta_j) &= \text{cov} \left(\sum_{k=1}^{n_2} a_k^{(i)} \eta_k, \sum_{k=1}^{n_2} a_k^{(j)} \eta_k \right) = \\ &= \sum_{k=1}^{n_2} \sum_{l=1}^{n_2} a_k^{(i)} a_l^{(j)} \text{cov}(\eta_k, \eta_l) = \sum_{k=1}^{n_2} a_k^{(i)} a_k^{(j)} \sigma_2^2 = 0. \end{aligned}$$

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

$$\zeta_i := \xi_i - \sqrt{\frac{n_1}{n_2}} \eta_i + \frac{1}{\sqrt{n_1 n_2}} \sum_{k=1}^{n_1} \eta_k - \bar{\eta} \quad (i = 1, \dots, n_1)$$

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

$$\begin{aligned} m = 0 &\iff m_1 = m_2 \\ m \neq 0 &\iff m_1 \neq m_2 \\ m < 0 &\iff m_1 < m_2 \\ m > 0 &\iff m_1 > m_2 \end{aligned}$$

miatt ennek a próbának a hipotézisei egybeesnek a feladat hipotéziseivel.

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.

5.11. *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.

5.2.6. Welch-próba

Az 5.9. feladat egy széles körben elterjedt közelítő megoldását B. L. WELCH 1947-ben adta meg (lásd [19]), amely szerint a

$$t := \frac{\bar{\xi} - \bar{\eta}}{\sqrt{\frac{S_{\xi, n_1}^{*2}}{n_1} + \frac{S_{\eta, n_2}^{*2}}{n_2}}}$$

statisztika H_0 teljesülése esetén megközelítőleg s szabadsági fokú t-eloszlású, ahol

$$a := \frac{S_{\xi, n_1}^{*2}}{n_1}, \quad b := \frac{S_{\eta, n_2}^{*2}}{n_2}, \quad c := \frac{(a + b)^2}{\frac{a^2}{n_1 - 1} + \frac{b^2}{n_2 - 1}}$$

jelölésekkel az s szabadsági fok a c értékének kerekítése a legközelebbi egészre. Ezután a kétmintás t-próbánál leírtak szerint dönthetünk a hipotéziseinkről.

Ezt a próbát akkor szokták alkalmazni, ha az F-próba a szórások különbözőségét mutatja ki, ugyanis a szórások egyezése esetén a kétmintás t-próba megbízhatóbb.

Legyen $F_k := F[T(k)]$, ahol $k \in \mathbb{N}$. Az előbb az $F(|t|) \simeq F_s(|t|)$ közelítést alkalmaztuk, de ezt lehet finomítani *polinomiális interpolációval*. Ehhez tekintsük azt a g polinomot, melyre $g(k) = F_k(|t|)$ minden $k \in \mathbb{N}$ esetén. Ezután az $F(|t|) \simeq g(c)$ közelítéssel számolhatunk tovább.

5.2.7. 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.

5.12. Feladat. Legyenek $\xi \in \text{Norm}(m_1; \sigma_1)$, $\eta \in \text{Norm}(m_2; \sigma_2)$ független valószínűségi változók, ahol $m_1, m_2, \sigma_1, \sigma_2$ ismeretlenek. Legyen $\xi_1, \dots, \xi_{n_1}$ a ξ -re vonatkozó, illetve $\eta_1, \dots, \eta_{n_2}$ az η -ra vonatkozó minta ($n_1 \geq 2$, $n_2 \geq 2$). A

$$H_0: \sigma_1 = \sigma_2$$

$$H_1: \sigma_1 \neq \sigma_2$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: \sigma_1 < \sigma_2$ illetve $H_1: \sigma_1 > \sigma_2$ egyoldali ellenhipotézisekre is.

Szükség lesz a következő tételre.

5.13. Tétel. Legyenek $\xi \in \text{Norm}(m_1; \sigma)$, $\eta \in \text{Norm}(m_2; \sigma)$ 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 ($n_1 \geq 2$, $n_2 \geq 2$). Ekkor

$$\frac{S_{\xi, n_1}^{*2}}{S_{\eta, n_2}^{*2}} \in F(n_1 - 1; n_2 - 1).$$

Bizonyítás. Korábban bizonyítottuk, hogy

$$\frac{S_{\xi, n_1}^2}{\sigma^2} n_1 \in \text{Khi}(n_1 - 1) \quad \text{és} \quad \frac{S_{\eta, n_2}^2}{\sigma^2} n_2 \in \text{Khi}(n_2 - 1),$$

így ezek függetlensége miatt

$$\frac{(n_2 - 1) \frac{S_{\xi, n_1}^2}{\sigma^2} n_1}{(n_1 - 1) \frac{S_{\eta, n_2}^2}{\sigma^2} n_2} = \frac{\frac{n_1}{n_1 - 1} S_{\xi, n_1}^2}{\frac{n_2}{n_2 - 1} S_{\eta, n_2}^2} = \frac{S_{\xi, n_1}^{*2}}{S_{\eta, n_2}^{*2}} \in F(n_1 - 1; n_2 - 1). \quad \square$$

Most térjünk vissza a feladat megoldására.

Megoldás. Az előző tétel szerint, ha $H_0: \sigma_1 = \sigma_2$ igaz, akkor

$$F := \frac{S_{\xi, n_1}^{*2}}{S_{\eta, n_2}^{*2}} \in F(n_1 - 1; n_2 - 1).$$

A $H_1: \sigma_1 \neq \sigma_2$ ellenhipotézis teljesülésekor F várhatóan kritikus értékben messze van 1-től, hiszen a korrigált tapasztalati szórás torzítatlan becslése a szórásnak. Ezért az elfogadási tartomány $a \leq F \leq b$ alakú, ahol $0 < a < 1 < b$. A továbbiakban legyen

$$F = F[F(n_1 - 1; n_2 - 1)].$$

Tegyük fel, hogy $P \in \mathcal{P}_{H_0}$ esetén

$$\begin{aligned} P(F < a) &= F(a) = \frac{\alpha}{2}, \\ P(F > b) &= 1 - F(b) = \frac{\alpha}{2}, \end{aligned}$$

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 1.112. lemmát), így $\frac{\alpha}{2} < F(1) < 1 - \frac{\alpha}{2}$ biztosan teljesül. Ezért az ezzel ekvivalens $a < 1 < b$ is teljesül. Mivel $P \in \mathcal{P}_{H_0}$ esetén

$$P(a \leq F \leq b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leq F \leq F^{-1}\left(1 - \frac{\alpha}{2}\right)$, azaz

$$2 \min\{F(F), 1 - F(F)\} \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1: \sigma_1 < \sigma_2$ teljesülésekor F várhatóan kritikus értékben kisebb 1-től. Ezért az elfogadási tartomány $F \geq c$ alakú, ahol $0 < c < 1$. Mivel $P \in \mathcal{P}_{H_0}$ -ra

$$P(F \geq c) = 1 - F(c),$$

így $P(F \geq c) = 1 - \alpha$ esetén $c = F^{-1}(\alpha) > 0$. Az $\alpha < F(1)$ biztosan teljesül $0 < \alpha \leq 0,3$ esetén, így $c < 1$ is teljesül. Tehát $F \geq F^{-1}(\alpha)$, azaz

$$F(F) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $S_{\xi, n_1}^* \leq S_{\eta, n_2}^*$ és $0 < \alpha \leq 0,3$ esetén ekvivalens azzal, hogy

$$\min\{F(F), 1 - F(F)\} \geq \alpha.$$

Valóban, ha $\min\{F(F), 1 - F(F)\} \geq \alpha$, akkor $F(F) \geq \min\{F(F), 1 - F(F)\}$ miatt $F(F) \geq \alpha$ is teljesül. Megfordítva, ha $F(F) \geq \alpha$, akkor $F \leq 1$ miatt

$$\min\{F(F), 1 - F(F)\} = \begin{cases} F(F) \geq \alpha, & \text{ha } F(F) \leq \frac{1}{2}, \\ 1 - F(F) \geq 1 - F(1) > 0,3 \geq \alpha, & \text{ha } F(F) > \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $S_{\xi, n_1}^* \geq S_{\eta, n_2}^*$ és $0 < \alpha \leq 0,3$ esetén $F(F) \geq F(1) > 0,3 \geq \alpha$. Tehát ekkor $H_1: \sigma_1 < \sigma_2$ ellenhipotézissel szemben mindig H_0 mellett döntünk.

A $H_1: \sigma_1 > \sigma_2$ ellenhipotézis teljesülése esetén F várhatóan kritikus értékben nagyobb 1-től. Ezért az elfogadási tartomány $F \leq d$ alakú, ahol $d > 1$. Ekkor $P \in \mathcal{P}_{H_0}$ -ra

$$P(F \leq d) = F(d),$$

így $P(F \leq d) = 1 - \alpha$ esetén $d = F^{-1}(1 - \alpha)$. Mivel $1 - \alpha > F(1)$ biztosan teljesül, ha $0 < \alpha \leq 0,3$, ezért $d > 1$ is teljesül. Tehát $F \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(F) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $S_{\xi, n_1}^* \geq S_{\eta, n_2}^*$ és $0 < \alpha \leq 0,3$ esetén ekvivalens azzal, hogy

$$\min\{F(F), 1 - F(F)\} \geq \alpha.$$

Valóban, ha $\min\{F(F), 1 - F(F)\} \geq \alpha$, akkor $1 - F(F) \geq \min\{F(F), 1 - F(F)\}$ miatt $1 - F(F) \geq \alpha$ is teljesül. Megfordítva, ha $1 - F(F) \geq \alpha$, akkor $F \geq 1$ miatt

$$\min\{F(F), 1 - F(F)\} = \begin{cases} 1 - F(F) \geq \alpha, & \text{ha } F(F) \geq \frac{1}{2}, \\ F(F) \geq F(1) > 0,3 \geq \alpha, & \text{ha } F(F) < \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $S_{\xi, n_1}^* \leq S_{\eta, n_2}^*$ és $0 < \alpha \leq 0,3$ esetén $1 - F(F) \geq 1 - F(1) > 0,3 \geq \alpha$. Tehát ekkor $H_1: \sigma_1 > \sigma_2$ ellenhipotézissel szemben mindig H_0 mellett döntünk. Ezt a statisztikai próbát *F-próbának* nevezzük.

5.2.8. Khi-négyzet próba normális eloszlás szórására

5.14. Feladat. Legyen $\xi \in \text{Norm}(m; \sigma)$, ahol m és σ ismeretlenek. Legyen $\sigma_0 \in \mathbb{R}_+$ rögzített és $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta ($n \geq 2$). A

$$H_0: \sigma = \sigma_0$$

$$H_1: \sigma \neq \sigma_0$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: \sigma < \sigma_0$ illetve $H_1: \sigma > \sigma_0$ egyoldali ellenhipotézisekre is.

Megoldás. Tudjuk, hogy H_0 esetén

$$\chi^2 := \frac{S_n^2}{\sigma_0^2} n \in \text{Khi}(n-1).$$

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 $a \leq \chi^2 \leq b$ alakúnak, ahol $0 < a < n-1 < b$. A továbbiakban legyen

$$F = F[\text{Khi}(n-1)].$$

Tegyük fel, hogy $P \in \mathcal{P}_{H_0}$ esetén

$$P(\chi^2 < a) = F(a) = \frac{\alpha}{2},$$

$$P(\chi^2 > b) = 1 - F(b) = \frac{\alpha}{2},$$

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 az 1.94. 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 $P \in \mathcal{P}_{H_0}$ esetén

$$P(a \leq \chi^2 \leq b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leq \chi^2 \leq F^{-1}\left(1 - \frac{\alpha}{2}\right)$, azaz

$$2 \min\{F(\chi^2), 1 - F(\chi^2)\} \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1: \sigma < \sigma_0$ ellenhipotézis teljesülésekor χ^2 várhatóan kritikus mértékben kisebb $n-1$ -től. Azaz az elfogadási tartomány $\chi^2 \geq c$ alakú, ahol $0 < c < n-1$. Mivel $P \in \mathcal{P}_{H_0}$ -ra

$$P(\chi^2 \geq c) = 1 - F(c),$$

így $P(\chi^2 \geq c) = 1 - \alpha$ esetén $c = F^{-1}(\alpha) > 0$. Erre teljesül, hogy $c < n-1$, mert $0,5 < F(n-1)$. Tehát így $\chi^2 \geq F^{-1}(\alpha)$, azaz

$$F(\chi^2) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $S_n^* \leq \sigma_0$ és $0 < \alpha \leq 0,3$ esetén ekvivalens azzal, hogy

$$\min\{F(\chi^2), 1 - F(\chi^2)\} \geq \alpha.$$

Valóban, ha $\min\{F(\chi^2), 1 - F(\chi^2)\} \geq \alpha$, akkor $F(\chi^2) \geq \min\{F(\chi^2), 1 - F(\chi^2)\}$ miatt $F(\chi^2) \geq \alpha$ is teljesül. Megfordítva, ha $F(\chi^2) \geq \alpha$, akkor $\chi^2 \leq n - 1$ miatt

$$\min\{F(\chi^2), 1 - F(\chi^2)\} = \begin{cases} F(\chi^2) \geq \alpha, & \text{ha } F(\chi^2) \leq \frac{1}{2}, \\ 1 - F(\chi^2) \geq 1 - F(n - 1) > 0,3 \geq \alpha, & \text{ha } F(\chi^2) > \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $S_n^* \geq \sigma_0$ és $0 < \alpha \leq 0,5$ esetén $F(\chi^2) \geq F(n - 1) > 0,5 \geq \alpha$. Tehát ekkor $H_1: \sigma < \sigma_0$ ellenhipotézissel szemben mindig H_0 mellett döntünk.

Ha $H_1: \sigma > \sigma_0$ teljesül, akkor χ^2 várhatóan kritikus mértékben nagyobb $n - 1$ -től, azaz az elfogadási tartomány $\chi^2 \leq d$ alakú, ahol $d > n - 1$. Mivel $P \in \mathcal{P}_{H_0}$ -ra

$$P(\chi^2 \leq d) = F(d),$$

így $P(\chi^2 \leq d) = 1 - \alpha$ esetén $d = F^{-1}(1 - \alpha)$. Másrészt ekkor $F(n - 1) < 0,7$ miatt $0 < \alpha \leq 0,3$ esetén $d > n - 1$. Tehát így $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $S_n^* \geq \sigma_0$ és $0 < \alpha \leq 0,5$ esetén ekvivalens azzal, hogy

$$\min\{F(\chi^2), 1 - F(\chi^2)\} \geq \alpha.$$

Valóban, ha $\min\{F(\chi^2), 1 - F(\chi^2)\} \geq \alpha$, akkor $1 - F(\chi^2) \geq \min\{F(\chi^2), 1 - F(\chi^2)\}$ miatt $1 - F(\chi^2) \geq \alpha$ is teljesül. Megfordítva, ha $1 - F(\chi^2) \geq \alpha$, akkor $\chi^2 \geq n - 1$ miatt

$$\min\{F(\chi^2), 1 - F(\chi^2)\} = \begin{cases} 1 - F(\chi^2) \geq \alpha, & \text{ha } F(\chi^2) \geq \frac{1}{2}, \\ F(\chi^2) \geq F(n - 1) > 0,5 \geq \alpha, & \text{ha } F(\chi^2) < \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $S_n^* \leq \sigma_0$ és $0 < \alpha \leq 0,3$ esetén $1 - F(\chi^2) \geq 1 - F(n - 1) > 0,3 \geq \alpha$. Tehát ekkor $H_1: \sigma > \sigma_0$ ellenhipotézissel szemben mindig H_0 mellett döntünk. Ez az úgynevezett *khi-négyzet próba*.

5.2.9. Statisztikai próba exponenciális eloszlás paraméterére

5.15. Feladat. Legyen $\xi \in \text{Exp}(\lambda)$, ahol λ ismeretlen. Legyen $\lambda_0 \in \mathbb{R}_+$ rögzített és $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta. A

$$H_0: \lambda = \lambda_0$$

$$H_1: \lambda \neq \lambda_0$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: \lambda < \lambda_0$ illetve $H_1: \lambda > \lambda_0$ egyoldali ellenhipotézisekre is.

Megoldás. A λ intervallumbecslésénél láttuk, hogy H_0 esetén

$$\gamma := \lambda_0 n \bar{\xi} \in \text{Gamma}(n; 1).$$

Mivel $\bar{\xi}$ az $E_\lambda = \frac{1}{\lambda}$ torzítatlan becslése, így a $H_1: \lambda \neq \lambda_0$ ($\iff n \neq \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ várhatóan kritikus mértékben messze van n -től. Azaz az elfogadási tartomány $a \leq \gamma \leq b$ alakú, ahol $0 < a < n < b$. A továbbiakban legyen

$$F = F[\text{Gamma}(n; 1)].$$

Tegyük fel, hogy $P \in \mathcal{P}_{H_0}$ esetén

$$\begin{aligned} P(\gamma < a) &= F(a) = \frac{\alpha}{2}, \\ P(\gamma > b) &= 1 - F(b) = \frac{\alpha}{2}, \end{aligned}$$

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 az 1.73. lemmát), így $\frac{\alpha}{2} < F(n) < 1 - \frac{\alpha}{2}$ biztosan teljesül. Ezért az ezzel ekvivalens $a < n < b$ is teljesül. Mivel $P \in \mathcal{P}_{H_0}$ esetén

$$P(a \leq \gamma \leq b) = F(b) - F(a) = 1 - \frac{\alpha}{2} - \frac{\alpha}{2} = 1 - \alpha,$$

így $F^{-1}\left(\frac{\alpha}{2}\right) \leq \gamma \leq F^{-1}\left(1 - \frac{\alpha}{2}\right)$, azaz

$$2 \min\{F(\gamma), 1 - F(\gamma)\} \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk.

A $H_1: \lambda < \lambda_0$ ($\iff n < \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ (ellentétben a korábbi próbákkal) várhatóan kritikus mértékben nagyobb n -től. Azaz az elfogadási tartomány $\gamma \leq c$ alakú, ahol $c > n$. Mivel $P \in \mathcal{P}_{H_0}$ -ra

$$P(\gamma \leq c) = F(c),$$

így $P(\gamma \leq c) = 1 - \alpha$ esetén $c = F^{-1}(1 - \alpha)$. Másrészt ekkor $F(n) < 0,7$ miatt $0 < \alpha \leq 0,3$ esetén $c > n$. Tehát így $\gamma \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\gamma) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $1/\bar{\xi} \leq \lambda_0$ és $0 < \alpha \leq 0,5$ esetén ekvivalens azzal, hogy

$$\min\{F(\gamma), 1 - F(\gamma)\} \geq \alpha.$$

Valóban, ha $\min\{F(\gamma), 1 - F(\gamma)\} \geq \alpha$, akkor $1 - F(\gamma) \geq \min\{F(\gamma), 1 - F(\gamma)\}$ miatt $1 - F(\gamma) \geq \alpha$ is teljesül. Megfordítva, ha $1 - F(\gamma) \geq \alpha$, akkor $\gamma \geq n$ miatt

$$\min\{F(\gamma), 1 - F(\gamma)\} = \begin{cases} 1 - F(\gamma) \geq \alpha, & \text{ha } F(\gamma) \leq \frac{1}{2}, \\ F(\gamma) \geq F(n) > 0,5 \geq \alpha, & \text{ha } F(\gamma) > \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $1/\bar{\xi} \geq \lambda_0$ és $0 < \alpha \leq 0,3$ esetén $1 - F(\gamma) \geq 1 - F(n) > 0,3 \geq \alpha$. Tehát ekkor $H_1: \lambda < \lambda_0$ ellenhipotézissel szemben mindig H_0 mellett döntünk.

A $H_1: \lambda > \lambda_0$ ($\iff n > \lambda_0 \frac{1}{\lambda} n$) ellenhipotézis teljesülésekor γ (ellentétben a korábbi próbákkal) várhatóan kritikus mértékben kisebb n -től. Azaz az elfogadási tartomány $\gamma \geq d$ alakú, ahol $0 < d < n$. Mivel $P \in \mathcal{P}_{H_0}$ -ra

$$P(\gamma \geq d) = 1 - F(d),$$

így $P(\gamma \geq d) = 1 - \alpha$ esetén $d = F^{-1}(\alpha) > 0$. Erre teljesül, hogy $d < n$, mert $0,5 < F(n)$. Tehát így $\gamma \geq F^{-1}(\alpha)$, azaz

$$F(\gamma) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Ez $1/\bar{\xi} \geq \lambda_0$ és $0 < \alpha \leq 0,3$ esetén ekvivalens azzal, hogy

$$\min\{F(\gamma), 1 - F(\gamma)\} \geq \alpha.$$

Valóban, ha $\min\{F(\gamma), 1 - F(\gamma)\} \geq \alpha$, akkor $F(\gamma) \geq \min\{F(\gamma), 1 - F(\gamma)\}$ miatt $F(\gamma) \geq \alpha$ is teljesül. Megfordítva, ha $F(\gamma) \geq \alpha$, akkor $\gamma \leq n$ miatt

$$\min\{F(\gamma), 1 - F(\gamma)\} = \begin{cases} F(\gamma) \geq \alpha, & \text{ha } F(\gamma) \leq \frac{1}{2}, \\ 1 - F(\gamma) \geq 1 - F(n) > 0,3 \geq \alpha, & \text{ha } F(\gamma) > \frac{1}{2}. \end{cases}$$

Vegyük még észre, hogy $1/\bar{\xi} \leq \lambda_0$ és $0 < \alpha \leq 0,5$ esetén $F(\gamma) \geq F(n) > 0,5 \geq \alpha$. Tehát ekkor $H_1: \lambda > \lambda_0$ ellenhipotézissel szemben mindig H_0 mellett döntünk.

5.2.10. Statisztikai próba valószínűségre

5.16. Feladat. Legyen $\xi \in \text{Bin}(1; p)$, ahol p ismeretlen. Legyen $0 < p_0 < 1$ rögzített és $\xi_1, \dots, \xi_n$ a ξ -re vonatkozó minta. A

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

hipotézisekre adjunk adott α terjedelmű próbát. A feladatot oldjuk meg $H_1: p < p_0$ illetve $H_1: p > p_0$ egyoldali ellenhipotézisekre is.

5.17. Megjegyzés. Ha A egy esemény és $\xi = I_A$, akkor $\xi \in \text{Bin}(1; p)$, 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

$$n\bar{\xi} \in \text{Bin}(n; p_0).$$

Ha ξ egy esemény indikátorváltozója, akkor $n\bar{\xi}$ az esemény bekövetkezéseinek a számát jelenti n kísérlet után. Mivel $\bar{\xi}$ a p torzítatlan becslése, ezért $H_1: p \neq p_0$

esetén $\bar{\xi}$ várhatóan kritikus mértékben eltávolodik p_0 -tól. Így ekkor az elfogadási tartomány $a \leq n\bar{\xi} \leq b$ alakú, ahol $a, b \in \mathbb{N}$ és $1 \leq a < np_0 < b \leq n-1$. Az $1 \leq a$ és $b \leq n-1$ feltételek azért kellenek, hogy a kritikus tartományban $n\bar{\xi} < a$ illetve $n\bar{\xi} > b$ ne legyenek lehetetlen események. Keressük meg a legkisebb a illetve b pozitív egész számokat, melyekre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

$$\begin{aligned} P(n\bar{\xi} \leq a) &= \sum_{i=0}^a \binom{n}{i} p_0^i (1-p_0)^{n-i} \geq \frac{\alpha}{2}, \\ P(n\bar{\xi} \leq b) &= \sum_{i=0}^b \binom{n}{i} p_0^i (1-p_0)^{n-i} \geq 1 - \frac{\alpha}{2}. \end{aligned}$$

Az így definiált a és b esetén, ha $P \in \mathcal{P}_{H_0}$, akkor

$$\begin{aligned} P(a \leq n\bar{\xi} \leq b) &= \sum_{i=a}^b \binom{n}{i} p_0^i (1-p_0)^{n-i} = \\ &= \sum_{i=0}^b \binom{n}{i} p_0^i (1-p_0)^{n-i} - \underbrace{\sum_{i=0}^{a-1} \binom{n}{i} p_0^i (1-p_0)^{n-i}}_{< \frac{\alpha}{2}} > 1 - \alpha, \end{aligned}$$

így ilyenkor $a \leq n\bar{\xi} \leq b$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $1 \leq a < np_0 < b \leq n-1$ feltétel mindig teljesíthető α és n alkalmas megválasztásával.

$H_1: p < p_0$ esetén $\bar{\xi}$ várhatóan kritikus mértékben p_0 alatt van, azaz az elfogadási tartomány $n\bar{\xi} \geq c$ alakú, ahol $c \in \mathbb{N}$ és $1 \leq c < np_0$. Legyen c a legkisebb pozitív egész, melyre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

$$P(n\bar{\xi} \leq c) = \sum_{i=0}^c \binom{n}{i} p_0^i (1-p_0)^{n-i} \geq \alpha.$$

Az így definiált c esetén, ha $P \in \mathcal{P}_{H_0}$, akkor

$$P(n\bar{\xi} \geq c) = \sum_{i=c}^n \binom{n}{i} p_0^i (1-p_0)^{n-i} = 1 - \underbrace{\sum_{i=0}^{c-1} \binom{n}{i} p_0^i (1-p_0)^{n-i}}_{< \alpha} > 1 - \alpha,$$

azaz $n\bar{\xi} \geq c$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $1 \leq c < np_0$ feltétel itt is mindig teljesíthető α és n alkalmas megválasztásával.

$H_1: p > p_0$ esetén $\bar{\xi}$ várhatóan kritikus mértékben p_0 felett van, azaz az elfogadási tartomány $n\bar{\xi} \leq d$ alakú, ahol $d \in \mathbb{N}$ és $np_0 < d \leq n-1$. Legyen d a legkisebb pozitív egész, melyre $P \in \mathcal{P}_{H_0}$ esetén teljesül, hogy

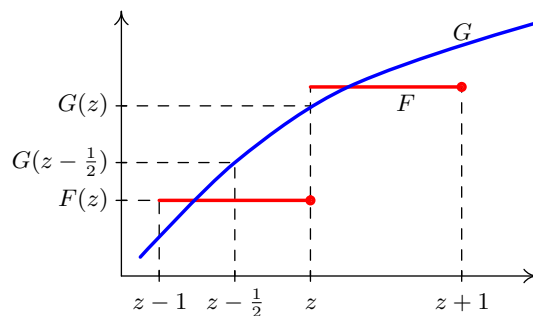
$$P(n\bar{\xi} \leq d) = \sum_{i=0}^d \binom{n}{i} p_0^i (1-p_0)^{n-i} \geq 1 - \alpha.$$

Ekkor tehát $n\bar{\xi} \leq d$ elfogadási tartománnyal α terjedelmű próbát kapunk. Az $np_0 < d \leq n-1$ 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 diszkrétisége 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 $\min\{np, n(1-p)\} \geq 10$ feltétel teljesül, akkor $F = F[\text{Bin}(n; p)]$ és $G = F[\text{Norm}(np; \sqrt{np(1-p)})]$ jelöléssel $F(z)$ értékét nagyon jól közelíti $G(z)$. Legyen $z \in \mathbb{N}$. Ekkor az 5.3. ábráról látható, hogy az F lépcsőssége és a G folytonossága miatt $G(z - \frac{1}{2})$ még pontosabban megközelíti $F(z)$ értékét.



5.3. ábra. Folytonossági korrekció

Tehát $\varrho \in \text{Bin}(n; p)$ és $z \in \mathbb{N}$ esetén

$$P(\varrho < z) \simeq \Phi\left(\frac{z - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right), \quad \text{illetve}$$

$$P(\varrho \leq z) = P(\varrho < z+1) \simeq \Phi\left(\frac{z+1 - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) = \Phi\left(\frac{z + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

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 $P(\varrho \leq 30)$ 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 $P(\varrho \leq 30)$ igazi értéke 0,02478 öt tizedesjegyre kerekítve. Ebből jól látható, hogy a folytonossági korrekcióval pontosabb közelítést kaptunk.

5.18. Feladat. Az előző megoldásban felírt kritikus értékekre adjunk közelítő képletet $\min\{np_0, n(1-p_0)\} \geq 10$ esetén, a folytonossági korrekciót alkalmazva.

Megoldás. Az előző megoldás jelöléseit fogjuk használni. Az előbbieket miatt

$$P(n\bar{\xi} \leq a) \simeq \Phi\left(\frac{a + \frac{1}{2} - np_0}{\sqrt{np_0(1-p_0)}}\right) = \frac{\alpha}{2},$$

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 $a \simeq \left\lceil h\left(\frac{\alpha}{2}\right) \right\rceil$. Hasonlóan kapjuk, hogy $b \simeq \left\lceil h\left(1 - \frac{\alpha}{2}\right) \right\rceil + 1$, $c \simeq \lceil h(\alpha) \rceil$ és $d \simeq \lceil h(1 - \alpha) \rceil + 1$.

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

5.19. 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

$$H_0: P(A_i) = p_i \quad (i = 1, \dots, r)$$

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 $\chi^2 \leq a$ alakú, ahol $a > 0$. Ismert, hogy $\min\{\varrho_1, \dots, \varrho_r\} \geq 10$ teljesülése esetén

$$P(\chi^2 \leq a) \simeq F(a),$$

ahol $P \in \mathcal{P}_{H_0}$ és

$$F = F[\text{Khi}(r-1)].$$

(A bizonyítást lásd pl. [2, 161–162. oldal].) Így $P(\chi^2 \leq a) = 1 - \alpha$ esetén $a \simeq F^{-1}(1 - \alpha)$. Tehát $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.20. Feladat. Legyen ξ egy ismeretlen eloszlású valószínűségi változó és F_0 egy rögzített eloszlásfüggvény. Készítsünk a

$$H_0: P(\xi < x) = F_0(x) \quad (x \in \mathbb{R})$$

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 $A_i := \{x_{k_{i-1}+1} \leq \xi \leq x_{k_i}\}$ 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 $A_i := \{a_{i-1} \leq \xi < a_i\}$ 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 $p_i := P(A_i)$ ($P \in \mathcal{P}_{H_0}$) jelöléssel legyen

$$H'_0: P(A_i) = p_i \quad (i = 1, \dots, r).$$

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

$$F = F[\text{Khi}(r - 1)]$$

jelöléssel $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal 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 eloszlástípusba tartozik (egyenletes, normális, Poisson, stb.). Ilyenkor használjuk a becsléses illeszkedésvizsgálatot.

5.21. 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 eloszlásfüggvénye az $\{F_\vartheta : \vartheta \in \Theta\}$ halmazba tartozik, azaz

$$H_0: P(\xi < x) = F_\vartheta(x) \quad (x \in \mathbb{R}) \text{ valamely } \vartheta \in \Theta \text{ esetén.}$$

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 $\hat{\vartheta}_i$. Legyen $\hat{\vartheta} := (\hat{\vartheta}_1, \dots, \hat{\vartheta}_v)$, $\hat{p}_i := P_{\hat{\vartheta}}(A_i)$, továbbá

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

Bizonyítható, hogy ha H_0 igaz, akkor χ^2 eloszlása $r - 1 - v$ szabadsági fokú khi-négyszet eloszláshoz konvergál $n \rightarrow \infty$ esetén. (A bizonyítás az úgynevezett likelihood hányados határeloszlásával hozható kapcsolatba, mi nem végezzük el. Lásd pl. [16, 91–93. oldal].) A gyakorlatban ez azt jelenti, hogy

$$F = F[\text{Khi}(r - 1 - v)]$$

jelöléssel

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

A közelítés már jónak tekinthető, ha $\min\{\varrho_1, \dots, \varrho_r\} \geq 10$. Így hasonlóan a tiszta illeszkedésvizsgálathoz, $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.3. Függelenségvizsgálat

A következő feladatban két teljes eseményrendszer függetlenségét vizsgáljuk.

5.22. Feladat. Legyen $A_1, \dots, A_r$ és $B_1, \dots, B_s$ két teljes eseményrendszer. Készítsünk a

$$H_0: P(A_i \cap B_j) = P(A_i)P(B_j) \quad (i = 1, \dots, r, \quad j = 1, \dots, s)$$

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 $P(A_i)$ illetve $P(B_j)$ maximum likelihood becslése $\frac{k_i}{n}$ illetve $\frac{l_j}{n}$. Ez összesen $(r - 1) + (s - 1)$ darab független becslést jelent a $k_1 + \dots + k_r = n$ és $l_1 + \dots + l_s = n$ feltételek miatt. Legyen $\hat{p}_{ij} := \frac{k_i}{n} \cdot \frac{l_j}{n}$ és

$$\chi^2 := \sum_{i=1}^r \sum_{j=1}^s \frac{(\varrho_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s \frac{(n\varrho_{ij} - k_i l_j)^2}{k_i l_j},$$

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	$\dots$	B_s	
A_1	ϱ_{11}	ϱ_{12}	$\dots$	ϱ_{1s}	k_1
A_2	ϱ_{21}	ϱ_{22}	$\dots$	ϱ_{2s}	k_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	ϱ_{r1}	ϱ_{r2}	$\dots$	ϱ_{rs}	k_r
	l_1	l_2	$\dots$	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 $n \rightarrow \infty$ esetén. Innen az eddigiekhez hasonlóan, ha $\varrho_{ij} \geq 10$ minden i, j esetén és

$$F = F[\text{Khi}((r - 1)(s - 1))],$$

akkor a H_0 nullhipotézisre $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.23. Feladat. Legyen (ξ, η) kétdimenziós valószínűségi vektorváltozó. Az erre vonatkozó $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ minta alapján készítsünk a

$$H_0: \xi \text{ és } \eta \text{ független}$$

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

$$H'_0: P(A_i \cap B_j) = P(A_i)P(B_j) \quad (i = 1, \dots, r, \quad j = 1, \dots, s).$$

Ha H_0 igaz, akkor H'_0 is az. Így az előző feladat megoldásából kapjuk, hogy H_0 teljesülése esetén

$$\chi^2 := \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^s \frac{(n\varrho_{ij} - k_i l_j)^2}{k_i l_j}$$

eloszlása $(r - 1)(s - 1)$ szabadsági fokú khi-négyzet eloszláshoz konvergál, ha $n \rightarrow \infty$. Innen

$$F = F[\text{Khi}((r - 1)(s - 1))]$$

jelöléssel, a H_0 nullhipotézisre $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.4. Homogenitásvizsgálat

5.24. 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

$$H_0: \xi \text{ és } \eta \text{ azonos eloszlású}$$

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

$$H'_0: P(A_i \cap B_j) = P(A_i)P(B_j) \quad (i = 1, \dots, r, \quad j = 1, 2)$$

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

$$\begin{aligned} \chi^2 &:= \frac{1}{n_1 + n_2} \sum_{i=1}^r \left(\frac{((n_1 + n_2)k_i - (k_i + l_i)n_1)^2}{(k_i + l_i)n_1} + \frac{((n_1 + n_2)l_i - (k_i + l_i)n_2)^2}{(k_i + l_i)n_2} \right) = \\ &= n_1 n_2 \sum_{i=1}^r \frac{\left(\frac{k_i}{n_1} - \frac{l_i}{n_2} \right)^2}{k_i + l_i} \end{aligned}$$

aszimptotikusan $(r - 1)(2 - 1) = r - 1$ szabadsági fokú khi-négyzet eloszlású. Tehát

$$F = F[\text{Khi}(r - 1)]$$

jelöléssel, a H_0 nullhipotézisre $\chi^2 \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(\chi^2) \geq \alpha$$

elfogadási tartománnyal közelítőleg α terjedelmű próbát kapunk.

5.3.5. Kétmintás előjelpróba

5.25. Feladat. Legyen (ξ, η) kétdimenziós valószínűségi vektorváltozó. Az erre vonatkozó $(\xi_1, \eta_1), \dots, (\xi_n, \eta_n)$ minta alapján készítsünk a

$$H_0: P(\xi > \eta) = \frac{1}{2}$$

$$H_1: P(\xi > \eta) \neq \frac{1}{2}$$

hipotézisekre α terjedelmű próbát, ahol P a valódi valószínűséget jelenti. A feladatot oldjuk meg $H_1: P(\xi > \eta) < \frac{1}{2}$ illetve $H_1: P(\xi > \eta) > \frac{1}{2}$ egyoldali ellenhipotézisekre is.

Megoldás. Bár a feladatot a nemparaméteres hipotézisvizsgálatokban tárgyaljuk, de $A := \{ \xi > \eta \}$ és $p_0 = \frac{1}{2}$ választással egyértelmű a kapcsolata a valószínűségre vonatkozó statisztikai próbával. Legyen

$$B := \sum_{i=1}^n I_{\xi_i > \eta_i},$$

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 $B \in \text{Bin}(n; \frac{1}{2})$. Legyenek az a, b, c, d számok a legkisebb olyan pozitív egészek, amelyekre teljesülnek, hogy

$$\sum_{i=0}^a \binom{n}{i} \frac{1}{2^n} \geq \frac{\alpha}{2}$$

$$\sum_{i=0}^b \binom{n}{i} \frac{1}{2^n} \geq 1 - \frac{\alpha}{2}$$

$$\sum_{i=0}^c \binom{n}{i} \frac{1}{2^n} \geq \alpha$$

$$\sum_{i=0}^d \binom{n}{i} \frac{1}{2^n} \geq 1 - \alpha.$$

Ekkor a valószínűségre vonatkozó statisztikai próbánál leírtak szerint

$$H_1: P(\xi > \eta) \neq \frac{1}{2} \text{ esetén } a \leq B \leq b,$$

$$H_1: P(\xi > \eta) < \frac{1}{2} \text{ esetén } B \geq c \text{ és}$$

$$H_1: P(\xi > \eta) > \frac{1}{2} \text{ esetén } B \leq d$$

elfogadási tartománnyal α terjedelmű próbát kapunk. A kritikus értékek kiszámolásánál itt is alkalmazható $n \geq 20$ esetén a folytonossági korrekcióval megadott közelítő számítás. Eszerint

$$h(x) := \frac{1}{2} \left(n - 1 + \sqrt{n} \Phi^{-1}(x) \right)$$

jelöléssel

$$a \simeq \left\lceil h\left(\frac{\alpha}{2}\right) \right\rceil, \quad b \simeq \left\lceil h\left(1 - \frac{\alpha}{2}\right) \right\rceil + 1, \quad c \simeq [h(\alpha)] \quad \text{és} \quad d \simeq [h(1 - \alpha)] + 1.$$

5.3.6. Kolmogorov – Szmirnov-féle kétmintás próba

5.26. 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*

$$\lim_{n \rightarrow \infty} P\left(\sqrt{\frac{n}{2}} \sup_{x \in \mathbb{R}} |F_n^*(x) - G_n^*(x)| < z\right) = 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}.$$

A bizonyítást lásd pl. [2, 194. oldal].

5.27. *Megjegyzés.* A Szmirnov-tétel feltételeivel

$$P\left(\sqrt{\frac{n}{2}} \sup_{x \in \mathbb{R}} |F_n^*(x) - G_n^*(x)| < z\right) \simeq 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}$$

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. [2, 191. oldal]), melyből $n \leq 30$ esetén is tudunk próbát konstruálni. Mi ezzel az esettel nem foglalkozunk.

5.28. 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

$$H_0: \xi \text{ és } \eta \text{ azonos eloszlású}$$

nullhipotézisre α terjedelmű próbát.

Megoldás. Legyenek a ξ -re illetve η -ra vonatkozó mintákhoz tartozó tapasztalati eloszlásfüggvények F_n^* illetve G_n^* , továbbá legyen

$$D := \sqrt{\frac{n}{2}} \sup_{x \in \mathbb{R}} |F_n^*(x) - G_n^*(x)|.$$

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 $P \in \mathcal{P}_{H_0}$ és $n > 30$ esetén

$$P(D < z) \simeq K(z) \quad (z \in \mathbb{R}_+),$$

ahol

$$K(z) = 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}.$$

Így $P(D < z) = 1 - \alpha$ esetén $z \simeq K^{-1}(1 - \alpha)$. Tehát $D < K^{-1}(1 - \alpha)$, azaz

$$1 - K(D) > \alpha$$

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.

5.29. Tétel (Kolmogorov-tétel). *Legyen a ξ valószínűségi változó F eloszlásfüggvénye folytonos. A ξ -re vonatkozó minta legyen $\xi_1, \dots, \xi_n$ és a neki megfelelő tapasztalati eloszlásfüggvény F_n^* . Ekkor minden $z \in \mathbb{R}_+$ esetén*

$$\lim_{n \rightarrow \infty} P\left(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| < z\right) = 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}.$$

5.30. *Megjegyzés.* A Kolmogorov-tétel feltételeivel

$$P\left(\sqrt{n} \sup_{x \in \mathbb{R}} |F_n^*(x) - F(x)| < z\right) \simeq 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}$$

közelítés már jónak tekinthető, ha $n > 30$.

5.31. Feladat. Legyen ξ folytonos eloszlásfüggvényű valószínűségi változó. Az erre vonatkozó $\xi_1, \dots, \xi_n$ ($n > 30$) minta alapján készítsünk a

$$H_0: \xi \text{ eloszlásfüggvénye } F$$

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 $D < K^{-1}(1 - \alpha)$, azaz

$$1 - K(D) > \alpha$$

elfogadási tartománnyal körülbelül α terjedelmű a próba, ahol

$$K(z) = 1 + 2 \sum_{i=1}^{\infty} (-1)^i e^{-2i^2 z^2}.$$

5.4. Szórásanalízis

5.32. Példa. Tegyük fel, hogy egy gazdaságban a búza termés hozamára vagyunk kíváncsiak (tonna/hektár). Jelölje ezt a ξ valószínűségi változó. Azt vizsgáljuk, hogy egyetlen *tényező*, például a búza fajtája milyen hatással van a termés hozamra. Tegyük fel, hogy a vizsgált gazdaság $r = 3$ különböző búzafajtát termeszt. A későbbiekben ezt úgy fogjuk mondani, hogy a vizsgált tényezőnek 3 különböző *szintje* van. Az 1. fajtát $n_1 = 4$, a 2. fajtát $n_2 = 3$, végül a 3. fajtát $n_3 = 5$ különböző parcellán termesztik. A

ξ_{ij} jelentse az i . fajta j . parcellán kapott termés hozamát tonna/hektár-ban mérve. A kapott mintarealizációk legyenek például a következők:

$\xi_{11}(\omega) = 5,24$	$\xi_{12}(\omega) = 4,17$	$\xi_{13}(\omega) = 4,35$	$\xi_{14}(\omega) = 4,77$
$\xi_{21}(\omega) = 5,09$	$\xi_{22}(\omega) = 6,05$	$\xi_{23}(\omega) = 5,89$	
$\xi_{31}(\omega) = 4,18$	$\xi_{32}(\omega) = 4,10$	$\xi_{33}(\omega) = 4,17$	$\xi_{34}(\omega) = 3,98$ $\xi_{35}(\omega) = 3,60$

Legyen a nullhipotézis az, hogy a különböző búzafajták (azaz a tényező szintjei) nincsenek hatással a termés hozamra. Ezt a vizsgálatot *egyszeres osztályozású szórásanalízisnek* nevezzük.

5.33. Példa. Az előző példát tovább gondolva, tegyük fel, hogy nem csak a búza fajtáját, hanem a parcella talajtípusát is vizsgálni szeretnénk a termés hozamot illetően, vagyis nem egy, hanem két tényező hatását figyeljük. Tegyük fel, hogy $r_1 = 3$ fajta búzát $r_2 = 4$ típusú talajba vetették. Azaz az 1. tényezőnek 3, a 2. tényezőnek pedig 4 szintje van. A ξ_{ij} jelentse az i . búzafajta j . talajtípuson vett termés hozamát. A kapott mintarealizációk legyenek például a következők:

$\xi_{11}(\omega) = 7,51$	$\xi_{12}(\omega) = 6,34$	$\xi_{13}(\omega) = 5,07$	$\xi_{14}(\omega) = 6,17$
$\xi_{21}(\omega) = 5,43$	$\xi_{22}(\omega) = 4,81$	$\xi_{23}(\omega) = 3,42$	$\xi_{24}(\omega) = 4,00$
$\xi_{31}(\omega) = 5,76$	$\xi_{32}(\omega) = 4,71$	$\xi_{33}(\omega) = 4,45$	$\xi_{34}(\omega) = 4,33$

Itt kétféle nullhipotézist vizsgálhatunk: a különböző búzafajták nincsenek hatással a termés hozamra, illetve, hogy a különböző talajtípusok nincsenek hatással a termés hozamra. Ezt a vizsgálatot *kétszeres osztályozású szórásanalízisnek* nevezzük.

Az előző mintarealizációval azt nem tudjuk vizsgálni, hogy a két tényező milyen hatással van egymásra, azaz, hogy egy konkrét búzafajta különbözőképpen terem-e a különböző talajtípusokon, vagy hogy egy konkrét talajtípuson különbözőképpen teremnek-e a különböző búzafajták, mert minden búzafajta–talajtípus kombinációból csak egy mintaelemünk van. Ezért az ilyen vizsgálatot *interakció nélküli kétszeres osztályozású szórásanalízisnek* is nevezik.

5.34. Példa. Folytatva az előző példát, ha a két tényező közötti kapcsolatot is vizsgálni szeretnénk, akkor minden búzafajta–talajtípus kombinációból több mérést kell végeznünk. Ha minden kombinációra ugyanannyi megfigyelésünk van, akkor *kiegyensúlyozott elrendezésről* beszélünk. Legyen az előző példában minden kombinációra $s = 3$ mérésünk. Jelentse ξ_{ijk} az i . búzafajta j . talajtípuson vett termés hozamára vonatkozó, azaz az (i, j) cellában végzett k . mérési eredményt. A kapott mintarealizációk

legyenek például a következők:

$\xi_{111}(\omega) = 7,51$	$\xi_{121}(\omega) = 6,34$	$\xi_{131}(\omega) = 5,07$	$\xi_{141}(\omega) = 6,17$
$\xi_{112}(\omega) = 7,03$	$\xi_{122}(\omega) = 5,81$	$\xi_{132}(\omega) = 4,19$	$\xi_{142}(\omega) = 5,90$
$\xi_{113}(\omega) = 6,91$	$\xi_{123}(\omega) = 6,61$	$\xi_{133}(\omega) = 5,27$	$\xi_{143}(\omega) = 6,28$
$\xi_{211}(\omega) = 5,43$	$\xi_{221}(\omega) = 4,81$	$\xi_{231}(\omega) = 3,42$	$\xi_{241}(\omega) = 4,00$
$\xi_{212}(\omega) = 4,95$	$\xi_{222}(\omega) = 3,82$	$\xi_{232}(\omega) = 3,19$	$\xi_{242}(\omega) = 3,80$
$\xi_{213}(\omega) = 5,48$	$\xi_{223}(\omega) = 4,18$	$\xi_{233}(\omega) = 2,02$	$\xi_{243}(\omega) = 3,94$
$\xi_{311}(\omega) = 5,76$	$\xi_{321}(\omega) = 4,71$	$\xi_{331}(\omega) = 4,45$	$\xi_{341}(\omega) = 4,33$
$\xi_{312}(\omega) = 5,90$	$\xi_{322}(\omega) = 5,24$	$\xi_{332}(\omega) = 4,65$	$\xi_{342}(\omega) = 5,41$
$\xi_{313}(\omega) = 6,01$	$\xi_{323}(\omega) = 4,07$	$\xi_{333}(\omega) = 4,59$	$\xi_{343}(\omega) = 5,70$

Itt háromféle nullhipotézist vizsgálhatunk: a különböző búzafajták nincsenek hatással a terméshozamra, a különböző talajtípusok nincsenek hatással a terméshozamra, illetve, hogy a búzafajta és a talajtípus között nincs kapcsolat a terméshozamot illetően. Ez a vizsgálat az ún. *kétszeres osztályozású szórásanalízis interakcióval*.

5.4.1. Egyszeres osztályozás (I. típusú modell)

Vizsgáljuk a ξ valószínűségi változót egyetlen tényező r darab különböző szintjén. Az i . szinthez tartozó valószínűségi változót jelölje ξ_i . Feltesszük, hogy $\xi_i \in \text{Norm}(m_i; \sigma)$ ($i = 1, 2, \dots, r$) függetlenek, ahol az $m_1, m_2, \dots, m_r, \sigma$ paraméterek ismeretlenek. A ξ_i -re ($i = 1, 2, \dots, r$) vonatkozó minta legyen

$$\xi_{i1}, \xi_{i2}, \dots, \xi_{in_i}.$$

Feltételezzük, hogy ξ_{ij} ($i = 1, 2, \dots, r$; $j = 1, 2, \dots, n_i$) előáll

$$\xi_{ij} = \beta_i + \varepsilon_{ij}$$

alakban, ahol az ε_{ij} *hibaváltozók* normális eloszlású, 0 várható értékű, független valószínűségi változók. Az *I. típusú modellben* $\beta_i = m_i$, azaz konstansok, míg a *II. típusú modellben* β_i normális eloszlású valószínűségi változók m_i várható értékekkel. Mi csak az I. típusú modellel foglalkozunk. Legyen

$$\begin{aligned} n &:= n_1 + n_2 + \dots + n_r, \\ m &:= \frac{1}{n}(n_1 m_1 + n_2 m_2 + \dots + n_r m_r), \\ a_i &:= m_i - m \quad (i = 1, 2, \dots, r). \end{aligned}$$

Vegyük észre, hogy ekkor

$$n_1 a_1 + n_2 a_2 + \dots + n_r a_r = 0,$$

ami speciálisan $n_1 = n_2 = \dots = n_r$ esetén azt jelenti, hogy $a_1 + a_2 + \dots + a_r = 0$.

Az m az ún. *teljes átlag*, az a_i az egyetlen tényező i . szintjének hatása a mért eredményre, míg ε_{ij} a véletlen hibák mértéke. Ezekkel tehát a modellünk

$$\xi_{ij} = m + a_i + \varepsilon_{ij} \quad (i = 1, 2, \dots, r; \quad j = 1, 2, \dots, n_i)$$

alakú. A nullhipotézis az lesz, hogy az egyetlen tényező különböző szintjei nincsenek hatással a mért értékekre, azaz

$$H_0: a_1 = a_2 = \dots = a_r = 0.$$

Nyilván ez az $m_1 = m_2 = \dots = m_r$ állítással ekvivalens, vagyis, hogy a várható értékek megegyeznek.

A következőkben ezen nullhipotézishez szeretnénk $1 - \alpha$ szintű próbát találni. Ennek érdekében vezessük be a következő jelöléseket:

$$\begin{aligned} \bar{\xi}_{..} &:= \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} \xi_{ij} \quad \text{az } m \text{ becslése,} \\ \bar{\xi}_{i.} &:= \frac{1}{n_i} \sum_{j=1}^{n_i} \xi_{ij} \quad \text{az } m_i \text{ becslése.} \end{aligned}$$

Ekkor

$$\begin{aligned} \bar{\xi}_{i.} - \bar{\xi}_{..} &\quad \text{az } a_i = m_i - m \text{ becslése,} \\ \xi_{ij} - \bar{\xi}_{i.} &\quad \text{az } \varepsilon_{ij} = \xi_{ij} - m_i \text{ becslése.} \end{aligned}$$

5.35. Tétel. *Bevezetve a*

$$\begin{aligned} Q &:= \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{..})^2, \\ Q_1 &:= \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{\xi}_{i.} - \bar{\xi}_{..})^2 = \sum_{i=1}^r n_i (\bar{\xi}_{i.} - \bar{\xi}_{..})^2, \\ Q_2 &:= \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{i.})^2 \end{aligned}$$

jelöléseket, teljesül, hogy

$$Q = Q_1 + Q_2.$$

5.36. Megjegyzés. Vegyük észre, hogy Q a teljes eltérések négyzetösszege, Q_1 a szintek közötti eltérések négyzetösszege és Q_2 a véletlen hibák (szinteken belüli eltérések) négyzetösszege.

Bizonyítás. Mivel

$$\xi_{ij} - \bar{\xi}_{..} = (\xi_{ij} - \bar{\xi}_{i.}) + (\bar{\xi}_{i.} - \bar{\xi}_{..}),$$

ezért mindkét oldalt négyzetre emelve és összegezve kapjuk, hogy

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{..})^2 =$$

$$= \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{\xi}_{i.} - \bar{\xi}_{..})^2 + \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{i.})^2 + 2 \sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{\xi}_{i.} - \bar{\xi}_{..}) (\xi_{ij} - \bar{\xi}_{i.}).$$

Másrészt

$$\sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{i.}) = \sum_{j=1}^{n_i} \xi_{ij} - n_i \bar{\xi}_{i.} = \sum_{j=1}^{n_i} \xi_{ij} - \sum_{j=1}^{n_i} \xi_{ij} = 0,$$

így

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{\xi}_{i.} - \bar{\xi}_{..}) (\xi_{ij} - \bar{\xi}_{i.}) = \sum_{i=1}^r (\bar{\xi}_{i.} - \bar{\xi}_{..}) \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{i.}) = 0.$$

Ezekből adódik az állítás. □

5.37. Tétel. *Az előző tétel jelöléseivel*

$$\begin{aligned} \mathbb{E} Q_1 &= (r-1)\sigma^2 + \sum_{i=1}^r n_i a_i^2, \\ \mathbb{E} Q_2 &= (n-r)\sigma^2. \end{aligned}$$

Bizonyítás. Először számoljuk ki $\text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..})$ értékét. Figyelembe véve, hogy a kovariancia bilineáris forma és

$$\text{cov}(\xi_{ij}, \xi_{kl}) = \begin{cases} D^2 \xi_{ij} = \sigma^2, & \text{ha } k = i \text{ és } l = j, \\ 0, & \text{különben,} \end{cases}$$

kapjuk, hogy

$$\begin{aligned} \text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..}) &= \text{cov}\left(\sum_{j=1}^{n_i} \frac{1}{n_i} \xi_{ij}, \sum_{k=1}^r \sum_{l=1}^{n_k} \frac{1}{n} \xi_{kl}\right) = \sum_{j=1}^{n_i} \sum_{k=1}^r \sum_{l=1}^{n_k} \frac{1}{nn_i} \text{cov}(\xi_{ij}, \xi_{kl}) = \\ &= \frac{1}{nn_i} \sum_{j=1}^{n_i} \left(\sum_{k=1}^r \sum_{l=1}^{n_k} \text{cov}(\xi_{ij}, \xi_{kl})\right) = \frac{1}{nn_i} \sum_{j=1}^{n_i} \sigma^2 = \frac{1}{nn_i} n_i \sigma^2 = \frac{1}{n} \sigma^2. \end{aligned}$$

Emiatt

$$\sum_{i=1}^r n_i \text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..}) = \sum_{i=1}^r \sum_{j=1}^{n_i} \text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..}) = \sum_{i=1}^r \sum_{j=1}^{n_i} \frac{1}{n} \sigma^2 = \frac{1}{n} n \sigma^2 = \sigma^2.$$

Ezt felhasználva adódik, hogy

$$\begin{aligned} \mathbb{E} Q_1 &= \sum_{i=1}^r n_i \mathbb{E} (\bar{\xi}_{i.} - \bar{\xi}_{..})^2 = \sum_{i=1}^r n_i \left(D^2 (\bar{\xi}_{i.} - \bar{\xi}_{..}) + \mathbb{E}^2 (\bar{\xi}_{i.} - \bar{\xi}_{..}) \right) = \\ &= \sum_{i=1}^r n_i \left(D^2 \bar{\xi}_{i.} + D^2 \bar{\xi}_{..} - 2 \text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..}) + (\mathbb{E} \bar{\xi}_{i.} - \mathbb{E} \bar{\xi}_{..})^2 \right) = \\ &= \sum_{i=1}^r n_i \left(\frac{1}{n_i} \sigma^2 + \frac{1}{n} \sigma^2 - 2 \text{cov}(\bar{\xi}_{i.}, \bar{\xi}_{..}) + (m_i - m)^2 \right) = \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=1}^r \sigma^2 + \frac{1}{n} \sigma^2 \sum_{i=1}^r n_i - 2 \sum_{i=1}^r n_i \operatorname{cov}(\bar{\xi}_i, \bar{\xi}_{..}) + \sum_{i=1}^r n_i a_i^2 = \\
&= r\sigma^2 + \sigma^2 - 2\sigma^2 + \sum_{i=1}^r n_i a_i^2 = (r-1)\sigma^2 + \sum_{i=1}^r n_i a_i^2.
\end{aligned}$$

A Q_2 várható értéke hasonló módon számolható ki. □

5.38. Tétel. Ha H_0 igaz, akkor

$$F := \frac{n-r}{r-1} \cdot \frac{Q_1}{Q_2} \in F(r-1; n-r).$$

Bizonyítás. Legyen

$$\zeta_{ij} := \frac{\xi_{ij} - m}{\sigma}.$$

Mivel H_0 igaz, azaz $m = m_1 = m_2 = \dots = m_r$, ezért a ζ_{ij} valószínűségi változók standard normális eloszlásúak és függetlenek. Vezessük be a

$$\bar{\zeta}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} \zeta_{ij} \quad \text{és} \quad \bar{\zeta}_{..} := \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} \zeta_{ij} = \frac{1}{n} \sum_{i=1}^r n_i \bar{\zeta}_i.$$

jelöléseket. Ekkor

$$\begin{aligned}
\bar{\zeta}_i - \bar{\zeta}_{..} &= \frac{1}{\sigma} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} (\xi_{ij} - m) - \frac{1}{n} \sum_{k=1}^r \sum_{j=1}^{n_k} (\xi_{kj} - m) \right) = \\
&= \frac{1}{\sigma} \left(\frac{1}{n_i} \sum_{j=1}^{n_i} \xi_{ij} - m - \frac{1}{n} \sum_{k=1}^r \sum_{j=1}^{n_k} \xi_{kj} + m \right) = \\
&= \frac{1}{\sigma} (\bar{\xi}_i - \bar{\xi}_{..}),
\end{aligned}$$

melyből kapjuk, hogy

$$\frac{Q_1}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^r n_i (\bar{\xi}_i - \bar{\xi}_{..})^2 = \sum_{i=1}^r n_i (\bar{\zeta}_i - \bar{\zeta}_{..})^2 = \sum_{i=1}^r \eta_i^2,$$

ahol az

$$\eta_i := \sqrt{n_i} (\bar{\zeta}_i - \bar{\zeta}_{..})$$

valószínűségi változók a ζ_{kl} valószínűségi változók lineáris kombinációi. Másrészt

$$\sum_{i=1}^r \sqrt{n_i} \eta_i = \sum_{i=1}^r n_i (\bar{\zeta}_i - \bar{\zeta}_{..}) = \sum_{i=1}^r \sum_{j=1}^{n_i} \zeta_{ij} - n \bar{\zeta}_{..} = 0,$$

így $B_0 := (\sqrt{n_1}, \dots, \sqrt{n_r})$ választással $B_0(\eta_1, \dots, \eta_r)^\top = 0$. Most legyen B tetszőleges olyan mátrix, melyre $B(\eta_1, \dots, \eta_r)^\top = (0, \dots, 0)^\top$. Jelölje $b_1, \dots, b_r$ a B valamelyik sorában található elemeket. Ekkor

$$\sum_{i=1}^r b_i \eta_i = \sum_{i=1}^r b_i \sqrt{n_i} (\bar{\zeta}_i - \bar{\zeta}_{..}) =$$

$$\begin{aligned}
&= \sum_{i=1}^r \bar{\zeta}_i b_i \sqrt{n_i} - \bar{\zeta}_{..} \sum_{i=1}^r b_i \sqrt{n_i} = \\
&= \sum_{i=1}^r \bar{\zeta}_i b_i \sqrt{n_i} - \frac{1}{n} \sum_{j=1}^r n_j \bar{\zeta}_j \sum_{i=1}^r b_i \sqrt{n_i} = \\
&= \sum_{i=1}^r \bar{\zeta}_i b_i \sqrt{n_i} - \frac{1}{n} \sum_{i=1}^r n_i \bar{\zeta}_i \sum_{j=1}^r b_j \sqrt{n_j} = \\
&= \sum_{i=1}^r \left(\bar{\zeta}_i b_i \sqrt{n_i} - \frac{1}{n} n_i \bar{\zeta}_i \sum_{j=1}^r b_j \sqrt{n_j} \right) = \\
&= \sum_{i=1}^r \left(b_i \sqrt{n_i} - \frac{n_i}{n} \sum_{j=1}^r b_j \sqrt{n_j} \right) \bar{\zeta}_i = 0
\end{aligned}$$

Mivel $\bar{\zeta}_i$ bármilyen értéket felvehet, ezért ez csak úgy teljesülhet, ha minden együtthatója a szummában 0, azaz $c := \frac{1}{n} \sum_{j=1}^r b_j \sqrt{n_j}$ jelöléssel $b_i = c \sqrt{n_i}$. Vagyis a B tetszőleges sora előáll $(\sqrt{n_1}, \dots, \sqrt{n_r})$ konstansszorosaként. Így $B = B_0$ a legnagyobb rangú mátrix, melyre $B(\eta_1, \dots, \eta_r)^\top = (0, \dots, 0)^\top$ teljesül. Ennek rangja 1, így az 1.96. megjegyzésből kapjuk, hogy Q_1/σ^2 a ζ_{kl} valószínűségi változókból képzett $r-1$ szabadsági fokú kvadratikus forma.

Most legyen

$$\eta_{ij} := \zeta_{ij} - \bar{\zeta}_i,$$

mely a ζ_{kl} valószínűségi változók lineáris kombinációja. Másrészt

$$\eta_{ij} = \frac{1}{\sigma} \left(\xi_{ij} - m - \frac{1}{n_i} \sum_{k=1}^{n_i} (\xi_{ik} - m) \right) = \frac{1}{\sigma} \left(\xi_{ij} - m - \frac{1}{n_i} \sum_{k=1}^{n_i} \xi_{ik} + m \right) = \frac{1}{\sigma} (\xi_{ij} - \bar{\xi}_i),$$

melyből

$$\frac{Q_2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_i)^2 = \sum_{i=1}^r \sum_{j=1}^{n_i} \eta_{ij}^2,$$

továbbá

$$\sum_{j=1}^{n_i} \eta_{ij} = \frac{1}{\sigma} \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_i) = \frac{1}{\sigma} \left(\sum_{j=1}^{n_i} \xi_{ij} - n_i \bar{\xi}_i \right) = 0 \quad (i = 1, 2, \dots, r).$$

Ez azzal ekvivalens, hogy $B_0 V^\top = (0, \dots, 0)^\top$, ahol

$$V := (\eta_{11}, \dots, \eta_{1n_1}, \eta_{21}, \dots, \eta_{2n_2}, \dots, \eta_{r1}, \dots, \eta_{rn_r})$$

és B_0 azon $r \times n$ -es mátrix, melynek k -adik sorában először $n_1 + \dots + n_{k-1}$ darab 0 van, majd n_k darab 1, utána pedig végig 0. Ennek minden sora független, így a rangja r . Megmutatjuk, hogy ez a legnagyobb rangú B mátrix, melyre $BV^\top = (0, \dots, 0)^\top$ teljesül. Ehhez most legyen B tetszőleges ilyen tulajdonságú mátrix, melynek valamelyik sorában található elemeit jelölje $b_{11}, \dots, b_{rn_r}$. Ekkor

$$\sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \eta_{ij} = \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \left(\zeta_{ij} - \frac{1}{n_i} \sum_{k=1}^{n_i} \zeta_{ik} \right) =$$

$$\begin{aligned}
&= \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \zeta_{ij} - \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \left(\frac{1}{n_i} \sum_{k=1}^{n_i} \zeta_{ik} \right) = \\
&= \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \zeta_{ij} - \sum_{i=1}^r \frac{1}{n_i} \sum_{k=1}^{n_i} \zeta_{ik} \left(\sum_{j=1}^{n_i} b_{ij} \right) = \\
&= \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \zeta_{ij} - \sum_{i=1}^r \left(\frac{1}{n_i} \sum_{j=1}^{n_i} b_{ij} \right) \sum_{k=1}^{n_i} \zeta_{ik} = \\
&= \sum_{i=1}^r \sum_{j=1}^{n_i} b_{ij} \zeta_{ij} - \sum_{i=1}^r \sum_{k=1}^{n_i} c_i \zeta_{ik} = \\
&= \sum_{i=1}^r \sum_{j=1}^{n_i} (b_{ij} - c_i) \zeta_{ij} = 0,
\end{aligned}$$

ahol $c_i := \frac{1}{n_i} \sum_{k=1}^{n_i} b_{ik}$. Mivel ζ_{ij} bármilyen értéket felvehet, ezért ez csak úgy lehetséges, ha a szummában minden együtthatója 0, azaz $b_{ij} = c_i$. Ez azt jelenti, hogy rögzített i -re minden b_{ij} ugyanazt az értéket veszi fel. Így B minden sora előáll B_0 sorainak lineáris kombinációjaként. Tehát beláttuk, hogy $B = B_0$ választással megkapjuk a legnagyobb rangú mátrixot, melyre $BV^\top = (0, \dots, 0)^\top$. Így az 1.96. megjegyzés alapján Q_2/σ^2 a ζ_{kl} valószínűségi változókból képzett $n - r$ szabadsági fokú kvadratikus forma.

Ezután legyenek

$$\eta := \sqrt{n} \bar{\zeta}_{..} \quad \text{és} \quad Q_0 := \sigma^2 \eta^2.$$

Ekkor $Q_0/\sigma^2 = \eta^2$, így az 1.96. megjegyzés alapján Q_0/σ^2 a ζ_{kl} valószínűségi változókból képzett 1 szabadsági fokú kvadratikus forma.

Mivel

$$\bar{\zeta}_{..} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} \zeta_{ij} = \frac{1}{\sigma} \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - m) = \frac{1}{\sigma} \left(\frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} \xi_{ij} - m \right) = \frac{1}{\sigma} (\bar{\xi}_{..} - m),$$

ezért

$$Q_0 = \sigma^2 n \bar{\zeta}_{..}^2 = n (\bar{\xi}_{..} - m)^2.$$

Ezután vizsgáljuk a Q_0/σ^2 , Q_1/σ^2 és Q_2/σ^2 kvadratikus formák összegét. Mivel

$$\xi_{ij} - m = (\xi_{ij} - \bar{\xi}_{..}) + (\bar{\xi}_{..} - m),$$

ezért

$$(\xi_{ij} - m)^2 = (\xi_{ij} - \bar{\xi}_{..})^2 + (\bar{\xi}_{..} - m)^2 + 2(\xi_{ij} - \bar{\xi}_{..})(\bar{\xi}_{..} - m),$$

azaz

$$\begin{aligned}
\sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - m)^2 &= \underbrace{\sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{..})^2}_Q + \underbrace{\sum_{i=1}^r \sum_{j=1}^{n_i} (\bar{\xi}_{..} - m)^2}_{Q_0} + \\
&\quad + 2(\bar{\xi}_{..} - m) \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{..}).
\end{aligned}$$

De

$$\sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - \bar{\xi}_{..}) = \sum_{i=1}^r \sum_{j=1}^{n_i} \xi_{ij} - n\bar{\xi}_{..} = n\bar{\xi}_{..} - n\bar{\xi}_{..} = 0,$$

így

$$\sum_{i=1}^r \sum_{j=1}^{n_i} \zeta_{ij}^2 = \frac{1}{\sigma^2} \sum_{i=1}^r \sum_{j=1}^{n_i} (\xi_{ij} - m)^2 = \frac{1}{\sigma^2} (Q + Q_0) = \frac{Q_0}{\sigma^2} + \frac{Q_1}{\sigma^2} + \frac{Q_2}{\sigma^2}.$$

Mivel $n = 1 + (r - 1) + (n - r)$, ezért a Fisher–Cochran-tétel alapján (1.95. tétel)

$$\frac{Q_1}{\sigma^2} \in \text{Khi}(r - 1) \quad \text{és} \quad \frac{Q_2}{\sigma^2} \in \text{Khi}(n - r)$$

független valószínűségi változók. Ebből pedig az F-eloszlás definíciója alapján

$$\frac{n - r}{r - 1} \cdot \frac{Q_1/\sigma^2}{Q_2/\sigma^2} = F \in F(r - 1; n - r)$$

adódik. □

Ezek alapján, ha H_0 igaz, akkor F értéke várhatóan közel van 1-hez. Ugyanakkor, ha H_0 nem igaz, akkor F várhatóan kritikus mértékben 1 fölé nő, ugyanis ekkor $\frac{Q_1}{r-1}$ várható értéke megnő, míg $\frac{Q_2}{n-r}$ várható értéke változatlan marad. Ezért az elfogadási tartomány $F \leq c$ alakú, ahol $c > 1$. Ekkor $P \in \mathcal{P}_{H_0}$ -ra $P(F \leq c) = F(c)$, ahol

$$F = F[F(r - 1; n - r)].$$

Így $P(F \leq c) = 1 - \alpha$ esetén $c = F^{-1}(1 - \alpha)$. Mivel $1 - \alpha > F(1)$ biztosan teljesül, ha $0 < \alpha \leq 0,3$, ezért $c > 1$ is teljesül. Tehát $F \leq F^{-1}(1 - \alpha)$, azaz

$$1 - F(F) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk. Az elemzéshez a következő *szórásanalízis táblázat*ot szokták elkészíteni:

szóródás forrása	szabadsági fok	négyzetösszeg	korr. tap. szórásnégyzet	próbat statisztika
szintek közötti eltérések	$f_1 = r - 1$	Q_1	$S_1^{*2} = \frac{Q_1}{f_1}$	$F = \frac{S_1^{*2}}{S_2^{*2}}$
véletlen hibák	$f_2 = n - r$	Q_2	$S_2^{*2} = \frac{Q_2}{f_2}$	
teljes	$f_1 + f_2$	$Q_1 + Q_2$		

Az 5.32. példa esetén a következőt kapjuk:

szóródás forrása	szabadsági fok	négyzetösszeg	korr. tap. szórásnégyzet	próbat statisztika
szintek közötti eltérések	2	5,2334	2,6167	$F = 16,3286$
véletlen hibák	9	1,4423	0,1603	
teljes	11	6,6757		

Ebből $F = F[F(2; 9)]$ jelöléssel $1 - F(F) = 0,0010 < 0,05$. Így $1 - \alpha = 0,95$ szinten elutasítjuk azt a hipotézist, hogy a búza fajtája nincs hatással a terméshozamra.

5.4.2. Kétszeres osztályozás interakció nélkül (I. típusú modell)

Vizsgáljuk két tényező hatását egy ξ valószínűségi változóra. Legyen az 1. tényezőnek r_1 , míg a 2. tényezőnek r_2 különböző szintje. Jelölje ξ_{ij} az 1. tényező i . szintjéhez és a 2. tényező j . szintjéhez tartozó valószínűségi változót. Feltesszük, hogy $\xi_{ij} \in \text{Norm}(m_{ij}; \sigma)$ ($i = 1, 2, \dots, r_1$; $j = 1, 2, \dots, r_2$) függetlenek, ahol minden paraméter ismeretlen. Az I. típusú modellben a várható értékeket

$$m_{ij} = m + a_i + b_j$$

alakban írjuk fel, ahol

$$\sum_{i=1}^{r_1} a_i = 0 \quad \text{és} \quad \sum_{j=1}^{r_2} b_j = 0.$$

Az m a teljes átlag, az a_i az 1. tényező i . szintjének hatása a mért eredményre, b_j a 2. tényező j . szintjének hatása a mért eredményre, és

$$\varepsilon_{ij} := \xi_{ij} - m_{ij}$$

a véletlen hibák mértéke. Tehát ebben a modellben

$$\xi_{ij} = m + a_i + b_j + \varepsilon_{ij} \quad (i = 1, 2, \dots, r_1; j = 1, 2, \dots, r_2)$$

alakú. Két nullhipotézist fogunk vizsgálni. Az első

$$H_0^{(1)}: a_1 = a_2 = \dots = a_{r_1} = 0,$$

azaz az 1. tényező különböző szintjei nincsenek hatással ξ -re. A második

$$H_0^{(2)}: b_1 = b_2 = \dots = b_{r_2} = 0,$$

azaz a 2. tényező különböző szintjei nincsenek hatással ξ -re.

A nullhipotézisekre vonatkozó próbákban vezessük be a következő jelöléseket:

$$\begin{aligned} \bar{\xi}_{..} &:= \frac{1}{r_1 r_2} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \xi_{ij}, \\ \bar{\xi}_{i.} &:= \frac{1}{r_2} \sum_{j=1}^{r_2} \xi_{ij} \quad (i = 1, 2, \dots, r_1), \\ \bar{\xi}_{.j} &:= \frac{1}{r_1} \sum_{i=1}^{r_1} \xi_{ij} \quad (j = 1, 2, \dots, r_2), \\ Q_1 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\bar{\xi}_{i.} - \bar{\xi}_{..})^2 = r_2 \sum_{i=1}^{r_1} (\bar{\xi}_{i.} - \bar{\xi}_{..})^2, \\ Q_2 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\bar{\xi}_{.j} - \bar{\xi}_{..})^2 = r_1 \sum_{j=1}^{r_2} (\bar{\xi}_{.j} - \bar{\xi}_{..})^2, \\ Q_3 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\xi_{ij} - \bar{\xi}_{i.} - \bar{\xi}_{.j} + \bar{\xi}_{..})^2. \end{aligned}$$

A Q_1 az 1. tényező szintjei közötti eltérések négyzetösszege, Q_2 a 2. tényező szintjei közötti eltérések négyzetösszege és Q_3 a hibatag. Ekkor teljesülnek az alábbiak:

$$Q := Q_1 + Q_2 + Q_3 = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\xi_{ij} - \bar{\xi}_{..})^2 \text{ a teljes eltérések négyzetösszege,}$$

$$F_1 := (r_2 - 1) \cdot \frac{Q_1}{Q_3} \in F(r_1 - 1; (r_1 - 1)(r_2 - 1)), \text{ ha } H_0^{(1)} \text{ igaz,}$$

$$F_2 := (r_1 - 1) \cdot \frac{Q_2}{Q_3} \in F(r_2 - 1; (r_1 - 1)(r_2 - 1)), \text{ ha } H_0^{(2)} \text{ igaz.}$$

Belátható továbbá, hogy ha $H_0^{(1)}$ igaz, akkor F_1 értéke közel van 1-hez, ellenkező esetben F_1 kritikus mértékben 1 fölé nő. Hasonló állítás teljesül F_2 -re is.

Ezek alapján

$$F = F[F(r_1 - 1; (r_1 - 1)(r_2 - 1))]$$

jelöléssel

$$1 - F(F_1) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk $H_0^{(1)}$ esetén, illetve

$$F = F[F(r_2 - 1; (r_1 - 1)(r_2 - 1))]$$

jelöléssel

$$1 - F(F_2) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk $H_0^{(2)}$ esetén.

Az elemzéshez a következő *szórásanalízis táblázat*ot szokták elkészíteni:

szóródás forrása	szabadsági fok	négyzetösszeg	korr. tap. szórásnégyzet	próbat statisztika
1. tényező szintjei közötti eltérések	$f_1 = r_1 - 1$	Q_1	$S_1^{*2} = \frac{Q_1}{f_1}$	$F_1 = \frac{S_1^{*2}}{S_3^{*2}}$
2. tényező szintjei közötti eltérések	$f_2 = r_2 - 1$	Q_2	$S_2^{*2} = \frac{Q_2}{f_2}$	$F_2 = \frac{S_2^{*2}}{S_3^{*2}}$
véletlen hibák	$f_3 = f_1 f_2$	Q_3	$S_3^{*2} = \frac{Q_3}{f_3}$	
teljes	$f_1 + f_2 + f_3$	$Q_1 + Q_2 + Q_3$		

Az 5.33. példa esetén a következőt kapjuk:

szóródás forrása	szabadsági fok	négyzetösszeg	kor. tap. szórásnégyzet	próbat statisztika
1. tényező szintjei közötti eltérések	2	7,6532	3,8266	$F_1 = 35,9278$
2. tényező szintjei közötti eltérések	3	5,9744	1,9915	$F_2 = 18,6978$
véletlen hibák	6	0,6391	0,1065	
teljes	11	14,2667		

Ebből $F = F[F(2; 6)]$ jelöléssel $1 - F(F_1) = 0,0005 < 0,05$. Így $1 - \alpha = 0,95$ szinten elutasítjuk azt a hipotézist, hogy a búza fajtája nincs hatással a terméshozamra. Másrészt $F = F[F(3; 6)]$ jelöléssel $1 - F(F_2) = 0,0019 < 0,05$. Így $1 - \alpha = 0,95$ szinten elutasítjuk azt a hipotézist, hogy a talaj típusa nincs hatással a terméshozamra.

5.4.3. Kétszeres osztályozás interakcióval (I. típusú modell), kiegyensúlyozott elrendezés esetén

Vizsgáljuk két tényező hatását egy ξ valószínűségi változóra. Legyen az 1. tényezőnek r_1 , míg a 2. tényezőnek r_2 különböző szintje. Jelölje ξ_{ij} az 1. tényező i . szintjéhez és a 2. tényező j . szintjéhez tartozó valószínűségi változót. Feltesszük, hogy $\xi_{ij} \in \text{Norm}(m_{ij}; \sigma)$ ($i = 1, 2, \dots, r_1$; $j = 1, 2, \dots, r_2$) függetlenek, ahol minden paraméter ismeretlen. Az I. típusú modellben a várható értékeket

$$m_{ij} = m + a_i + b_j + c_{ij}$$

alakban írjuk fel, ahol

$$\sum_{i=1}^{r_1} a_i = 0, \quad \sum_{j=1}^{r_2} b_j = 0, \\ \sum_{i=1}^{r_1} c_{ij} = 0 \quad (j = 1, 2, \dots, r_2) \quad \text{és} \quad \sum_{j=1}^{r_2} c_{ij} = 0 \quad (i = 1, 2, \dots, r_1).$$

Az m a teljes átlag, az a_i az 1. tényező i . szintjének hatása a mért eredményre, b_j a 2. tényező j . szintjének hatása a mért eredményre, c_{ij} a két tényező együttes hatása a mért eredményre.

Minden ξ_{ij} -hez készítsünk egy s elemű mintát:

$$\xi_{ij1}, \xi_{ij2}, \dots, \xi_{ijs}.$$

Feltesszük, hogy

$$\varepsilon_{ijk} := \xi_{ijk} - m_{ij}$$

a véletlen hibákból adódó valószínűségi változó. Tehát ebben a modellben

$$\xi_{ijk} = m + a_i + b_j + c_{ij} + \varepsilon_{ijk} \quad (i = 1, 2, \dots, r_1; j = 1, 2, \dots, r_2; k = 1, 2, \dots, s)$$

alakú. Három nullhipotézist fogunk vizsgálni. Az első

$$H_0^{(1)}: a_1 = a_2 = \dots = a_{r_1} = 0,$$

azaz az 1. tényező különböző szintjei nincsenek hatással ξ -re. A második

$$H_0^{(2)}: b_1 = b_2 = \dots = b_{r_2} = 0,$$

azaz a 2. tényező különböző szintjei nincsenek hatással ξ -re. A harmadik

$$H_0^{(3)}: c_{ij} = 0 \text{ minden } i = 1, 2, \dots, r_1 \text{ és } j = 1, 2, \dots, r_2 \text{ esetén,}$$

azaz a két tényező együttes hatása nem befolyásolja a ξ értékét.

A nullhipotézisekre vonatkozó próbákban vezessük be a következő jelöléseket:

$$\begin{aligned}\bar{\xi}_{...} &:= \frac{1}{r_1 r_2 s} \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s \xi_{ijk}, \\ \bar{\xi}_{i..} &:= \frac{1}{r_2 s} \sum_{j=1}^{r_2} \sum_{k=1}^s \xi_{ijk} \quad (i = 1, 2, \dots, r_1), \\ \bar{\xi}_{.j.} &:= \frac{1}{r_1 s} \sum_{i=1}^{r_1} \sum_{k=1}^s \xi_{ijk} \quad (j = 1, 2, \dots, r_2), \\ \bar{\xi}_{ij.} &:= \frac{1}{s} \sum_{k=1}^s \xi_{ijk} \quad (i = 1, 2, \dots, r_1; j = 1, 2, \dots, r_2), \\ Q_1 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s (\bar{\xi}_{i..} - \bar{\xi}_{...})^2 = r_2 s \sum_{i=1}^{r_1} (\bar{\xi}_{i..} - \bar{\xi}_{...})^2, \\ Q_2 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s (\bar{\xi}_{.j.} - \bar{\xi}_{...})^2 = r_1 s \sum_{j=1}^{r_2} (\bar{\xi}_{.j.} - \bar{\xi}_{...})^2, \\ Q_3 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s (\bar{\xi}_{ij.} - \bar{\xi}_{i..} - \bar{\xi}_{.j.} + \bar{\xi}_{...})^2 = s \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} (\bar{\xi}_{ij.} - \bar{\xi}_{i..} - \bar{\xi}_{.j.} + \bar{\xi}_{...})^2, \\ Q_4 &:= \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s (\xi_{ijk} - \bar{\xi}_{ij.})^2.\end{aligned}$$

A Q_1 az 1. tényező szintjei közötti eltérések négyzetösszege, Q_2 a 2. tényező szintjei közötti eltérések négyzetösszege, Q_3 a két tényező együttes hatásaiból adódó eltérések négyzetösszege és Q_4 a hibatag. Ekkor teljesülnek az alábbiak:

$$Q := Q_1 + Q_2 + Q_3 + Q_4 = \sum_{i=1}^{r_1} \sum_{j=1}^{r_2} \sum_{k=1}^s (\xi_{ijk} - \bar{\xi}_{...})^2 \text{ a teljes eltérések négyzetösszege,}$$

$$F_1 := \frac{r_1 r_2 (s-1)}{r_1 - 1} \cdot \frac{Q_1}{Q_4} \in F(r_1 - 1; r_1 r_2 (s-1)), \text{ ha } H_0^{(1)} \text{ igaz,}$$

$$F_2 := \frac{r_1 r_2 (s-1)}{r_2 - 1} \cdot \frac{Q_2}{Q_4} \in F(r_2 - 1; r_1 r_2 (s-1)), \text{ ha } H_0^{(2)} \text{ igaz,}$$

$$F_3 := \frac{r_1 r_2 (s-1)}{(r_1 - 1)(r_2 - 1)} \cdot \frac{Q_3}{Q_4} \in F((r_1 - 1)(r_2 - 1); r_1 r_2 (s-1)), \text{ ha } H_0^{(3)} \text{ igaz.}$$

Belátható továbbá, hogy ha $H_0^{(1)}$ igaz, akkor F_1 értéke közel van 1-hez, ellenkező esetben F_1 kritikus mértékben 1 fölé nő. Hasonló állítás teljesül F_2 -re és F_3 -ra is.

Ezek alapján

$$F = F\left[F\left(r_1 - 1; r_1 r_2 (s - 1)\right)\right]$$

jelöléssel

$$1 - F(F_1) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk $H_0^{(1)}$ esetén,

$$F = F\left[F\left(r_2 - 1; r_1 r_2 (s - 1)\right)\right]$$

jelöléssel

$$1 - F(F_2) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk $H_0^{(2)}$ esetén, végül

$$F = F\left[F\left((r_1 - 1)(r_2 - 1); r_1 r_2 (s - 1)\right)\right]$$

jelöléssel

$$1 - F(F_3) \geq \alpha$$

elfogadási tartománnyal α terjedelmű próbát kapunk $H_0^{(3)}$ esetén.

Az elemzéshez a következő *szórásanalízis táblázat*ot szokták elkészíteni:

szóródás forrása	szabadsági fok	négyzetösszeg	korr. tap. szórásnégyzet	próbat statisztika
1. tényező szintjei közötti eltérések	$f_1 = r_1 - 1$	Q_1	$S_1^{*2} = \frac{Q_1}{f_1}$	$F_1 = \frac{S_1^{*2}}{S_4^{*2}}$
2. tényező szintjei közötti eltérések	$f_2 = r_2 - 1$	Q_2	$S_2^{*2} = \frac{Q_2}{f_2}$	$F_2 = \frac{S_2^{*2}}{S_4^{*2}}$
két tényező közötti interakció	$f_3 = f_1 f_2$	Q_3	$S_3^{*2} = \frac{Q_3}{f_3}$	$F_3 = \frac{S_3^{*2}}{S_4^{*2}}$
véletlen hibák	$f_4 = r_1 r_2 (s - 1)$	Q_4	$S_4^{*2} = \frac{Q_4}{f_4}$	
teljes	$f_1 + f_2 + f_3 + f_4$	$Q_1 + Q_2 + Q_3 + Q_4$		

Az 5.34. példa esetén a következőt kapjuk:

szóródás forrása	szabadsági fok	négyzetösszeg	kor. tap. szórásnégyzet	próbatatisztika
1. tényező szintjei közötti eltérések	2	24,1034	12,0517	$F_1 = 59,3583$
2. tényező szintjei közötti eltérések	3	18,2751	6,0917	$F_2 = 30,0035$
két tényező közötti interakció	6	2,0206	0,3368	$F_3 = 1,6587$
véletlen hibák	24	4,8728	0,2030	
teljes	35	49,2720		

Ebből $F = F[F(2; 24)]$ jelöléssel $1 - F(F_1) = 5 \cdot 10^{-10} < \alpha$, így elutasítjuk azt a hipotézist, hogy a búza fajtája nincs hatással a termés hozamra. Másrészt $F = F[F(3; 24)]$ jelöléssel $1 - F(F_2) = 3 \cdot 10^{-8} < \alpha$, így elutasítjuk azt a hipotézist, hogy a talaj típusa nincs hatással a termés hozamra. Végül $F = F[F(6; 24)]$ jelöléssel $1 - F(F_3) = 0,1746 \geq \alpha$, így elfogadjuk azt a hipotézist, hogy a búza fajtájának és a talaj típusának nincs együttes hatása a termés hozamra.

6. fejezet

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

$$E\left(\eta - g(\xi_1, \xi_2)\right)^2,$$

hasonlóan a $D^2 \xi = E(\xi - E \xi)^2$ szórásnégyzethez, ami a ξ és $E \xi$ eltérésének mértéke. Ha sikerülne olyan g függvényt találni, amelyre $E\left(\eta - g(\xi_1, \xi_2)\right)^2$ a lehető legkisebb, akkor ξ_1 és ξ_2 mérésével közelítőleg meg lehetne jósolni η , azaz az árhullám tetőzésének mértékét.

Általánosítva, ha az $\eta, \xi_1, \dots, \xi_k$ valószínűségi változók esetén az a feladat, hogy adjuk meg a lehető legjobb

$$\eta \simeq g(\xi_1, \dots, \xi_k)$$

közelítést adó g függvényt, akkor az úgy értendő, hogy az

$$E\left(\eta - g(\xi_1, \dots, \xi_k)\right)^2$$

értékét kell minimalizálni. Ez az úgynevezett *legkisebb négyzetek elve*. Az így kapott g továbbá $\xi_1, \dots, \xi_k$ ismeretében megbecsülhető lesz η .

6.1. Tétel. *Legyenek $\eta, \xi_1, \dots, \xi_k$ valószínűségi változók és $E \eta^2 < \infty$. Az összes $g: \mathbb{R}^k \rightarrow \mathbb{R}$ Borel-mérhető függvényt figyelembe véve $E\left(\eta - g(\xi_1, \dots, \xi_k)\right)^2$ akkor a legkisebb, ha*

$$g(\xi_1, \dots, \xi_k) = E(\eta \mid \xi_1, \dots, \xi_k).$$

Bizonyítás. Legyen $\mu := \eta - E(\eta \mid \xi_1, \dots, \xi_k)$ és $\nu := E(\eta \mid \xi_1, \dots, \xi_k) - g(\xi_1, \dots, \xi_k)$. Ekkor

$$E\left(\eta - g(\xi_1, \dots, \xi_k)\right)^2 = E(\mu + \nu)^2 = E\mu^2 + 2E(\mu\nu) + E\nu^2 \geq$$

$$\begin{aligned} &\geq E \mu^2 + 2 E(\mu \nu) = E \mu^2 + 2 E(E(\mu \nu \mid \xi_1, \dots, \xi_k)) = \\ &= E \mu^2 + 2 E(\nu E(\mu \mid \xi_1, \dots, \xi_k)), \end{aligned}$$

másrészt

$$\begin{aligned} E(\mu \mid \xi_1, \dots, \xi_k) &= E(\eta - E(\eta \mid \xi_1, \dots, \xi_k) \mid \xi_1, \dots, \xi_k) = \\ &= E(\eta \mid \xi_1, \dots, \xi_k) - E(E(\eta \mid \xi_1, \dots, \xi_k) \mid \xi_1, \dots, \xi_k) = \\ &= E(\eta \mid \xi_1, \dots, \xi_k) - E(\eta \mid \xi_1, \dots, \xi_k) = 0. \end{aligned}$$

Így kapjuk, hogy

$$E(\eta - g(\xi_1, \dots, \xi_k))^2 \geq E \mu^2 = E(\eta - E(\eta \mid \xi_1, \dots, \xi_k))^2,$$

melyből adódik az állítás. $\square$

6.2. Definíció. Ha $\eta, \xi_1, \dots, \xi_k$ valószínűségi változók, ξ_i értékészlete R_{ξ_i} ($i = 1, \dots, k$) és $E \eta^2$ véges, akkor a

$$g: R_{\xi_1} \times \dots \times R_{\xi_k} \rightarrow \mathbb{R}, \quad g(x_1, \dots, x_k) := E(\eta \mid \xi_1 = x_1, \dots, \xi_k = x_k)$$

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 függvényvel írható le, akkor azt $k = 1$ esetén (*elsőfajú*) *regressziós egyenesnek*, míg $k = 2$ esetén (*elsőfajú*) *regressziós síknak* nevezzük.

6.3. Megjegyzés. Ismert, hogy $(\eta, \xi_1, \dots, \xi_k) \in \text{Norm}_{k+1}(m; A)$ esetén léteznek olyan $a_1, \dots, a_k \in \mathbb{R}$ konstansok, hogy $E(\eta \mid \xi_1, \dots, \xi_k) = a_1 \xi_1 + \dots + a_k \xi_k$. Tehát ha $(\eta, \xi_1, \dots, \xi_k)$ valószínűségi vektorváltozó normális eloszlású, akkor a regressziós felület egy lineáris függvényvel 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 $E(\eta - g(\xi_1, \dots, \xi_k))^2$ minimumát csak a

$$g(x_1, \dots, x_k) = a_0 + a_1 x_1 + \dots + a_k x_k \quad (a_0, a_1, \dots, a_k \in \mathbb{R})$$

alakú – azaz lineáris – függvények között keressük. Ezt a típusú regressziószámítást *lineáris regresszió*nak 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 $\xi_0 \equiv 1$, $E\eta^2 \in \mathbb{R}$, $E(\eta\xi_i) \in \mathbb{R}$, $E(\xi_i\xi_j) \in \mathbb{R}$ ($i, j = 0, \dots, k$), továbbá az

$$R := \begin{pmatrix} E(\xi_0\xi_0) & E(\xi_0\xi_1) & \dots & E(\xi_0\xi_k) \\ E(\xi_1\xi_0) & E(\xi_1\xi_1) & \dots & E(\xi_1\xi_k) \\ \vdots & \vdots & \ddots & \vdots \\ E(\xi_k\xi_0) & E(\xi_k\xi_1) & \dots & E(\xi_k\xi_k) \end{pmatrix}$$

mátrix pozitív definit, azaz minden bal felső sarokdeterminánsa pozitív. Ekkor a lineáris regresszióknak pontosan egy megoldása van, nevezetesen azon $g(x_1, \dots, x_k) = a_0 + a_1x_1 + \dots + a_kx_k$ függvény, melyre

$$a_i = \frac{\det R_i}{\det R} \quad (i = 0, \dots, k),$$

ahol az R_i mátrixot úgy kapjuk, hogy az R mátrix i -edik oszlopát kicseréljük az $r := (E(\eta\xi_0), \dots, E(\eta\xi_k))^\top$ -ra.

Bizonyítás. A feladat azon $a_0, \dots, a_k \in \mathbb{R}$ paraméterek meghatározása, amelyek mellett $E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2$ minimális. Mivel

$$\begin{aligned} E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 &= E(\eta - a_0\xi_0 - \dots - a_k\xi_k)^2 = \\ &= E\eta^2 + \sum_{i=0}^k a_i^2 E\xi_i^2 - 2 \sum_{i=0}^k a_i E(\eta\xi_i) + 2 \sum_{i=0}^{k-1} \sum_{j=i+1}^k a_i a_j E(\xi_i\xi_j), \end{aligned}$$

ezért

$$\begin{aligned} \frac{\partial}{\partial a_l} E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 &= \\ &= 2a_l E\xi_l^2 - 2E(\eta\xi_l) + 2 \sum_{i \neq l} a_i E(\xi_i\xi_l) = \\ &= 2 \sum_{i=0}^k a_i E(\xi_i\xi_l) - 2E(\eta\xi_l) \quad (l = 0, \dots, k). \end{aligned}$$

Így azt kapjuk, hogy az

$$\frac{\partial}{\partial a_l} E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = 0 \quad (l = 0, \dots, k)$$

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

$$K := \left(\frac{\partial^2}{\partial a_l \partial a_t} E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 \right)_{(k+1) \times (k+1)}.$$

Mivel

$$\frac{\partial^2}{\partial a_l \partial a_t} E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2 = 2E(\xi_l\xi_t),$$

ezért $K = 2R$. Ebből adódik, hogy K pozitív definit, azaz a kapott megoldás valóban minimumhely. Ezzel bizonyítottuk a tételt. $\square$

6.5. *Megjegyzés.* Könnyen látható, hogy $k = 1$ esetén az előző tétel feltételei teljesülnek, ha $E\eta^2 \in \mathbb{R}$, $0 < D^2 \xi_1 < \infty$ és $\text{cov}(\eta, \xi_1) \in \mathbb{R}$. Másrészt ekkor $R(a_0, a_1)^\top = r$ ekvivalens a következő egyenletrendszerrel:

$$\begin{aligned} a_0 + a_1 E \xi_1 &= E \eta, \\ a_0 E \xi_1 + a_1 E \xi_1^2 &= E(\eta \xi_1). \end{aligned}$$

Ennek a megoldása

$$a_0 = E \eta - \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1} E \xi_1, \quad a_1 = \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1}.$$

Így a regressziós egyenes egyenlete

$$g(x) = E \eta - \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1} E \xi_1 + \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1} x,$$

azaz ennek eredményeképpen a továbbiakban az

$$\eta \simeq E \eta - \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1} E \xi_1 + \frac{\text{cov}(\eta, \xi_1)}{D^2 \xi_1} \xi_1$$

lineáris közelítést lehet használni.

6.6. Feladat. Az $E(\eta - g(\xi_1, \dots, \xi_k))^2$ minimumát keressük meg azon g lineáris függvények között, melyek átmennek az origón, azaz a

$$g(x_1, \dots, x_k) = a_1 x_1 + \dots + a_k x_k \quad (a_1, \dots, a_k \in \mathbb{R})$$

alakú függvények között.

Megoldás. Az előző tétel bizonyításához hasonlóan kapjuk a következő állítást. Legyen $E\eta^2 \in \mathbb{R}$, $E(\eta \xi_i) \in \mathbb{R}$, $E(\xi_i \xi_j) \in \mathbb{R}$ ($i, j = 1, \dots, k$), továbbá az

$$R' := \begin{pmatrix} E(\xi_1 \xi_1) & E(\xi_1 \xi_2) & \dots & E(\xi_1 \xi_k) \\ E(\xi_2 \xi_1) & E(\xi_2 \xi_2) & \dots & E(\xi_2 \xi_k) \\ \vdots & \vdots & \ddots & \vdots \\ E(\xi_k \xi_1) & E(\xi_k \xi_2) & \dots & E(\xi_k \xi_k) \end{pmatrix}$$

mátrix pozitív definit, azaz minden bal felső sarokdeterminánsa pozitív. Ekkor a feladatnak pontosan egy megoldása van, nevezetesen azon $g(x_1, \dots, x_k) = a_1 x_1 + \dots + a_k x_k$ függvény, melyre

$$a_i = \frac{\det R'_i}{\det R'} \quad (i = 1, \dots, k),$$

ahol az R'_i mátrixot úgy kapjuk, hogy az R' mátrix i -edik oszlopát kicseréljük az $r' := (E(\eta \xi_1), \dots, E(\eta \xi_k))^\top$ -ra. Speciálisan $k = 1$ esetén $a_1 = \frac{E(\eta \xi_1)}{E(\xi_1^2)}$.

6.7. Feladat. Legyenek $t_0, \dots, t_k \in \mathbb{R}$ rögzített konstansok. Az $E(\eta - g(\xi_1, \dots, \xi_k))^2$ minimumát keressük 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

$$g(x_1, \dots, x_k) = a_0 + a_1x_1 + \dots + a_kx_k \quad (a_0, \dots, a_k \in \mathbb{R}) \quad \text{és} \quad g(t_1, \dots, t_k) = t_0$$

pontosan akkor teljesülnek egyszerre, ha

$$g(x_1, \dots, x_k) - t_0 = a_1(x_1 - t_1) + \dots + a_k(x_k - t_k) \quad (a_1, \dots, a_k \in \mathbb{R}).$$

(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 $\eta, \xi_1, \dots, \xi_k$ valószínűségi változók és azok kapcsolatának ismeretében határoztuk meg. Ezekről viszont a gyakorlatban csak nagyon ritkán van elegendő információnk. Így ekkor az $(\eta, \xi_1, \dots, \xi_k)$ -ra vonatkozó minta alapján kell ezeket az együtthatókat megbecsülni. Legyen ez a minta

$$(\eta_i, \xi_{i1}, \dots, \xi_{ik}) \quad i = 1, \dots, n.$$

Bevezetjük a következő jelöléseket:

$$\begin{aligned} a &:= (a_0, \dots, a_k)^\top \\ Y &:= (\eta_1, \dots, \eta_n)^\top \\ X &:= \begin{pmatrix} 1 & \xi_{11} & \dots & \xi_{1k} \\ 1 & \xi_{21} & \dots & \xi_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \xi_{n1} & \dots & \xi_{nk} \end{pmatrix}. \end{aligned}$$

A becslés alapja az, hogy az $E(\eta - a_0 - a_1\xi_1 - \dots - a_k\xi_k)^2$ várható értéket az

$$\frac{1}{n} \sum_{i=1}^n (\eta_i - a_0 - a_1\xi_{i1} - \dots - a_k\xi_{ik})^2$$

átlaggal becsüljük. Vegyük észre, hogy ez az átlag $\frac{1}{n}\|Y - Xa\|^2$ alakban is írható, ahol $\|(v_1, \dots, v_n)^\top\| = \sqrt{v_1^2 + \dots + v_n^2}$ a $(v_1, \dots, v_n)^\top$ 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 altére. 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 $b_0, \dots, b_k \in \mathbb{R}$, $b = (b_0, \dots, b_k)^\top$ 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 $X^\top Y = X^\top Xa$ egyenlet az úgynevezett *normálegyenlet*, melynek $\hat{a} = (\hat{a}_0, \dots, \hat{a}_k)^\top$ -val jelölt megoldása szolgáltatja a lineáris regresszió együtthatóinak becslését. Nyilván, ha $X^\top X$ invertálható mátrix, akkor

$$\hat{a} = (X^\top X)^{-1} X^\top Y.$$

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}) $i = 1, \dots, n$,

$$\begin{aligned}a &= (a_0, a_1)^\top \\ Y &= (\eta_1, \dots, \eta_n)^\top \\ X &= \begin{pmatrix} 1 & \xi_{11} \\ 1 & \xi_{21} \\ \vdots & \vdots \\ 1 & \xi_{n1} \end{pmatrix}.\end{aligned}$$

Némi számolással kapjuk, hogy az $X^\top Y = X^\top Xa$ normálegyenlet ekvivalens a következő egyenletrendszerrel:

$$\begin{aligned}(\xi_{11} + \dots + \xi_{n1})a_0 + (\xi_{11}^2 + \dots + \xi_{n1}^2)a_1 &= \xi_{11}\eta_1 + \dots + \xi_{n1}\eta_n, \\ na_0 + (\xi_{11} + \dots + \xi_{n1})a_1 &= \eta_1 + \dots + \eta_n.\end{aligned}$$

Ennek megoldása, és így az a becslése

$$\begin{aligned}\hat{a}_0 &= \bar{\eta} - \frac{\text{Cov}_n(\eta, \xi_1)}{S_{\xi_1, n}^2} \bar{\xi}_1, \\ \hat{a}_1 &= \frac{\text{Cov}_n(\eta, \xi_1)}{S_{\xi_1, n}^2}.\end{aligned}$$

Ennek alapján a továbbiakban az $\eta \simeq \hat{a}_0 + \hat{a}_1 \xi_1$ közelítést fogjuk használni.

6.9. Megjegyzés. Összehasonlítva az előbb kapott $\hat{a}_0$ és $\hat{a}_1$ 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.

Megjegyezzük még, hogy ebben az esetben meg szokták adni az η és ξ_1 közötti tapasztalati korrelációs együttható négyzetét (az ún. *determinációs együtthatót*). Ennek értéke minél közelebb van 1-hez, annál pontosabbnak tekinthető a lineáris közelítés.

6.10. Feladat. Adjunk 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

$$g(x_1, \dots, x_k) = t_0 + a_1(x_1 - t_1) + \dots + a_k(x_k - t_k) \quad (a_1, \dots, a_k \in \mathbb{R})$$

függvényt találni, melyre

$$\sum_{i=1}^n (\eta_i - g(\xi_{i1}, \dots, \xi_{ik}))^2$$

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

$$(\hat{a}_1, \dots, \hat{a}_k)^\top = (X'^\top X')^{-1} X'^\top Y.$$

Speciálisan $k = 1$ esetén

$$\hat{a}_1 = \frac{\sum_{i=1}^n \xi_{i1} \eta_i}{\sum_{i=1}^n \xi_{i1}^2} = \frac{\text{Cov}_n(\eta, \xi_1) + \bar{\xi}_1 \bar{\eta}}{S_{\xi_1, n} + \bar{\xi}_1^2},$$

így ekkor az $\eta \simeq \hat{a}_1 \xi_1$ 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 sok esetben nem célszerű, mert valamilyen elvi vagy tapasztalati tény a lineáris kapcsolatnak ellentmond. Ilyenkor meg kell tippelni, 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 esetekkel foglalkozunk $k = 1$ esetén.

6.4.1. Polinomos regresszió

Ebben az esetben a regressziós függvényt

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_rx^r \quad (a_0, \dots, a_r \in \mathbb{R}_+)$$

alakban keressük. Ekkor az $a_0, \dots, a_r$ együtthatókat az $\eta, \xi_1, \xi_1^2, \dots, \xi_1^r$ 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 végrehajtván, 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

$$a = \exp \left(E(\ln \eta) - \frac{\text{cov}(\ln \eta, \ln \xi_1)}{D^2(\ln \xi_1)} E(\ln \xi_1) \right),$$
$$b = \frac{\text{cov}(\ln \eta, \ln \xi_1)}{D^2(\ln \xi_1)}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\hat{a} = \exp \left(\overline{\ln \eta} - \frac{\text{Cov}_n(\ln \eta, \ln \xi_1)}{S_{\ln \xi_1, n}^2} \overline{\ln \xi_1} \right),$$
$$\hat{b} = \frac{\text{Cov}_n(\ln \eta, \ln \xi_1)}{S_{\ln \xi_1, n}^2}.$$

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

$$a = \exp \left(E(\ln \eta) - \frac{\text{cov}(\ln \eta, \xi_1)}{D^2 \xi_1} E \xi_1 \right),$$

$$b = \exp \left(\frac{\text{cov}(\ln \eta, \xi_1)}{D^2 \xi_1} \right).$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\hat{a} = \exp \left(\overline{\ln \eta} - \frac{\text{Cov}_n(\ln \eta, \xi_1)}{S_{\xi_1, n}^2} \overline{\xi_1} \right),$$

$$\hat{b} = \exp \left(\frac{\text{Cov}_n(\ln \eta, \xi_1)}{S_{\xi_1, n}^2} \right).$$

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

$$a = E \eta - \frac{\text{cov}(\eta, \ln \xi_1)}{D^2(\ln \xi_1)} E(\ln \xi_1),$$

$$b = \frac{\text{cov}(\eta, \ln \xi_1)}{D^2(\ln \xi_1)}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\hat{a} = \bar{\eta} - \frac{\text{Cov}_n(\eta, \ln \xi_1)}{S_{\ln \xi_1, n}^2} \overline{\ln \xi_1},$$

$$\hat{b} = \frac{\text{Cov}_n(\eta, \ln \xi_1)}{S_{\ln \xi_1, n}^2}.$$

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

$$a = E(\eta^{-1}) - \frac{\text{cov}(\eta^{-1}, \xi_1)}{D^2 \xi_1} E \xi_1,$$

$$b = \frac{\text{cov}(\eta^{-1}, \xi_1)}{D^2 \xi_1}.$$

Ezen paraméterek becslése, szintén a korábbiak alapján

$$\hat{a} = \overline{\eta^{-1}} - \frac{\text{Cov}_n(\eta^{-1}, \xi_1)}{S_{\xi_1, n}^2} \overline{\xi_1},$$

$$\hat{b} = \frac{\text{Cov}_n(\eta^{-1}, \xi_1)}{S_{\xi_1, n}^2}.$$

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.: *Valószínűesszámitás*, 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 Mifflin, 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.: *Valószínűesszámitás*, Tankönyvkiadó, Budapest, 1982.
- [12] PRÉKOPA A.: *Valószínűségelmélet műszaki alkalmazásokkal*, Műszaki Könyvkiadó, Budapest, 1962.
- [13] RÉNYI A.: *Valószínűesszámitás*, 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. <https://gyires.inf.unideb.hu/mobiDiak/Terdik-Gyorgy/Eloadasok-a-matematikai-statisztikabol/Stat2JegyTGy.pdf>
- [17] TÓMÁCS T.: *Matematikai statisztika gyakorlatok* https://tomacstibor.uni-eszterhazy.hu/tananyagok/Matematikai_statisztika_gyakorlatok.pdf
- [18] VINCZE I.: *Matematikai statisztika*, Tankönyvkiadó, Budapest, 1971.
- [19] WELCH, B. L.: *The generalization of 'Student's' problem when several different population variances are involved* *Biometrika*, Volume 34, Issue 1–2, January 1947, Pages 28–35, <https://doi.org/10.1093/biomet/34.1-2.28>.