Statisztika (jegyzet)

Csiszár Villő

2009. május 6.

1. Statisztikai mező

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

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

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

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

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

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

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

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

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

- egészségügyben: gyógyszerek hatásosságának bizonyítása;
- ipar: selejtarány ellenőrzése: le kell-e a gépsort cserélni?
- irodalomtudomány: 2 szövegről el kell dönteni, hogy ugyanaz írta-e őket;
- szociológia: pártok népszerűsége: van-e szignifikáns különbség?
- szociológia: pártpreferencia és iskolázottság között van-e összefüggés?

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

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

ahol P_ϑ valószínűségi mértékek.

A Θ halmazt *paramétertérnek* nevezzük. Legtöbbször Θ véges dimenziós euklideszi tér részhalmaza, ekkor azt mondjuk, hogy paraméteres a feladat (pl.: 1.-5. paraméteres feladatok). Θ lehet ennél jóval "nagyobb", pl.: ha $\mathcal{P}$ az összes lehetséges valószínűségi mérték, ekkor nemparaméteres a feladat.

1.2. Definíció. Egy $X = (X_1, \ldots, X_n) : \Omega \to \mathcal{X} \subseteq \mathbb{R}^n$ valószínűségi változót *(n elemű) mintának* nevezünk, ahol $\mathcal{X}$ a mintatér, n pedig a minta nagysága vagy elemszáma. Az X_i koordináták a minta elemei.

Mi (majdnem) mindig olyan mintákkal fogunk foglalkozni, amikor ugyanazt a véletlen jelenséget, egymástól függetlenül, n-szer figyeljük meg.

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

A minta eloszlásfüggvényeinek családja $\{F_{n;\vartheta} | \vartheta \in \Theta\}$, ahol $F_{n;\vartheta}(x_1, ..., x_n) = P_\vartheta(X_1 < x_1, ..., X_n < x_n)$. Ha X *n elemű, független, azonos eloszlású minta*, akkor

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

ahol $F_{1;\vartheta}$ az X_i koordináták közös eloszlásfüggvénye. Emlékezzünk rá, hogy az eloszlásfüggvény helyett diszkrét esetben a

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

valószínűségeket, abszolút folytonos esetben pedig az

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

sűrűségfüggvényt is használhatjuk. Független elemű minta esetén ezek is szorzatra bomlanak.

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

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

■

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

1.2. Példa. Néhány gyakran használt statisztika (X mindenhol n elemű minta):

1) $T(X) = \overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$ az X minta *mintaátlaga.*

2) $T(X) = S_X^2 = \dfrac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$ az X minta *tapasztalati szórásnégyzete.*

3) $T(X) = (X_1^{(n)}, X_2^{(n)}, ..., X_n^{(n)})$ az X minta *rendezett mintája*, ahol $X_1^{(n)} \leq X_2^{(n)} \leq \quad \leq X_n^{(n)}$
(pl. $T(2, 4, 1, 3) = (1, 2, 3, 4)$).

4) $T(X) = X_n^{(n)} - X_1^{(n)}$ az X minta *mintaterjedelme.*

5) $T(X) = \begin{cases} X_{\frac{n+1}{2}}^{(n)} & \text{ha } n \text{ páratlan} \\ \dfrac{X_{\frac{n}{2}}^{(n)} + X_{\frac{n}{2}+1}^{(n)}}{2} & \text{ha } n \text{ páros} \end{cases}$ az X minta *tapasztalati mediánja.* ■

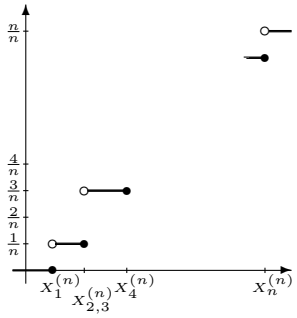
1.5. Definíció. Az $(X_1, ..., X_n)$ minta *tapasztalati eloszlása* az a véletlen diszkrét eloszlás, melynek lehetséges értéki az X_i értékek, az értékekhez tartozó valószínűségek pedig a megfigyelt relatív gyakoriságok. Azaz jelölje $x_1 < x_2 < \cdots < x_m$ a megfigyelt (különböző) értékeket ($m \leq n$), ekkor az x_j-hez tartozó valószínűség:

$$\frac{|\{i|X_i = x_j\}|}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i = x_j),$$

ahol $I(X_i = x_j) = 1$, ha $X_i = x_j$, és $I(X_i = x_j) = 0$, ha $X_i \neq x_j$.
Az X minta *tapasztalati eloszlásfüggvénye* a tapasztalati eloszláshoz tartozó eloszlásfüggvény:

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

Hogy néz ez ki?



$$\hat{F}_n(x) = \begin{cases} 0 & \text{ha } x \leq X_1^{(n)}, \\ \dfrac{k}{n} & \text{ha } X_k^{(n)} < x \leq X_{k+1}^{(n)}, \\ 1 & \text{ha } X_n^{(n)} < x. \end{cases}$$

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

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

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

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

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

1.1. Feladat. Az 2-5 példákra adjuk meg a mintateret, a paraméterteret, és a minta eloszlásainak családját!
Mo: az X minta mindegyik példában független, azonos eloszlású.
2) X_i az i. fonalszál élettartama.
$\mathcal{X} = \mathbb{R}_+^{20}$, $X_i \sim Exp(\vartheta)$, $\Theta = (0, \infty)$,

$$f_{20;\vartheta}(x_1, ..., x_{20}) = \prod_{i=1}^{20} f_{1;\vartheta}(x_i) = \prod_{i=1}^{20} \vartheta e^{-\vartheta x_i} = \vartheta^{20} e^{-\vartheta \sum x_i}.$$

3. X_i az i. dobás pontszáma.
$\mathcal{X} = \{0, 1\}^{10}$, $X_i \sim Ind(\vartheta)$, $\Theta = [0, 1]$,

$$p_{10;\vartheta}(x_1, \ldots, x_{10}) = \prod_{i=1}^{10} p_{1;\vartheta}(x_i) = \prod_{i=1}^{10} \vartheta^{x_i}(1-\vartheta)^{1-x_i} = \vartheta^{\sum x_i}(1-\vartheta)^{10-\sum x_i}.$$

4. X_i az i. nap fogyasztása.
$\mathcal{X} = \mathbb{R}_+^5$, $X_i \sim N(\vartheta_1, \vartheta_2)$, $\Theta = \mathbb{R} \times \mathbb{R}_+$,

$$f_{5;\vartheta}(x_1, ..., x_5) = \prod_{i=1}^{5} f_{1;\vartheta}(x_i) = \prod_{i=1}^{5} \frac{1}{\sqrt{2\pi\vartheta_2^2}} e^{-\frac{(x_i-\vartheta_1)^2}{2\vartheta_2^2}}.$$

5. X_i az i. napon a várakozási idő.
$\mathcal{X} = \mathbb{R}_+^5$, $X_i \sim E(0, \vartheta)$, $\Theta = (0, \infty)$,

$$f_{5;\vartheta}(x_1, ..., x_5) = \prod_{i=1}^{5} f_{1;\vartheta}(x_i) = \prod_{i=1}^{5} \frac{1}{\vartheta} I(0 \le x_i \le \vartheta) = \frac{1}{\vartheta^5} I(x_1^{(1)} \ge 0 \text{ és } x_5^{(5)} \le \vartheta).$$

■

1.2. Feladat. Végezzük el 8-szor a következő kísérletet: addig dobunk egy érmével, amíg fejet nem kapunk. Jelölje X_i, hogy az i. kísérletben hányszor kellett dobni. Az X minta egy konkrét realizációjára adjuk meg a tapasztalati eloszlásfüggvényt, és számítsuk ki az 1.2 Példában szereplő statisztikákat!

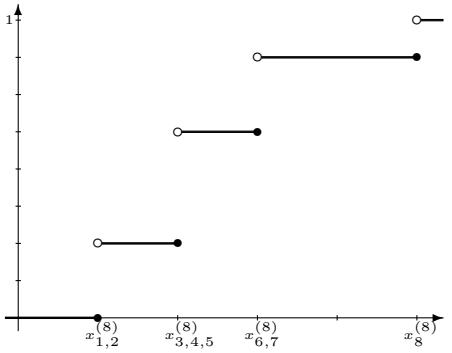
Mo: Ha például a minta a következő:

	1	2	3	3	5	1	2	2
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

A tapasztalati eloszlás:

érték	1	2	3	4	5
valószínűség	$\frac{2}{8}$	$\frac{3}{8}$	$\frac{2}{8}$	0	$\frac{1}{8}$

mintaátlag:	$19/8 = 2.38$
tapasztalati szórásnégyzet:	$95/64 = 1.48$
mintaterjedelem:	4
tapasztalati medián:	2



■

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

2. Elégségesség

Az X minta információt tartalmaz arról, hogy melyik $\vartheta \in \Theta$ az igazi paraméter. A $P_\vartheta(X = x)$ valószínűség függ ϑ-tól (bizonyos ϑ-kra nagy a valószínűsége, hogy ezt a mintát kapjuk, másokra kisebb). A $T(X)$ statisztika is hordoz információt, hiszen a $P_\vartheta(T(X) = t)$ valószínűség is függ ϑ-tól. Az eredeti minta általában több információt tartalmaz a paraméterről, mint a belőle kiszámolt statisztika. Pl. ha diszkrét mintát veszünk, akkor a $P_\vartheta(X = x)$ valószínűségek sokfélesége hordozza a ϑ-ra vonatkozó információt. A $T(X)$-ben rejlő információ pedig a $P_\vartheta(T(X) = t)$ valószínűségek sokféleségéből származik. A $\{T(X) = t\}$ esemény felbomlik $\{X = x\}$ eseményekre, olyan x-ekre, melyekre $T(x) = t$. Az X-ben tartalmazott pluszinformáció tehát a $P_\vartheta(X = x|T(X) = t)$ valószínűségek sokféleségéből adódik, olyan x-ekre, melyekre $T(x) = t$. Ha a $P_\vartheta(X = x|T(X) = t)$ feltételes valószínűségek már nem függnek ϑ-tól, akkor, mivel $P_\vartheta(X = x) = P(X = x|T(X) = T(x))P_\vartheta(T(X) = T(x))$, a minta nem tartalmaz több információt, mint a statisztika.

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

2.1. Példa. Legyen $X_i \sim Ind(p)$, ahol $0 \le p \le 1$ ismeretlen paraméter. Pl. 100 kockadobás mindegyikéről feljegyezzük, hogy hatos-e. Belátjuk, hogy $\sum_i X_i$ elégséges statisztika p-re, azaz nem kell a

dobásokról egyesével feljegyezni, hogy melyik volt hatos, melyik nem, hanem elég feljegyezni, hogy összesen hány hatos volt. Ezzel nem veszítünk információt p-ről. A definíció alapján számolunk:

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

Azaz a kapott feltételes valószínűség tényleg nem függ p-től. ■

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

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

$$P_\vartheta(X=x) = P_\vartheta(X=x|T(X)=T(x))\cdot P_\vartheta(T(X)=T(x)) = h(x)\cdot g_\vartheta(T(x)),$$

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

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

$$\frac{h(x)}{\sum_{y:T(y)=t} h(y)}, \text{ ha } T(x)=t,$$

egyébként pedig a feltételes valószínűség nulla. ■

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

2.2. Példa. Legyen $X_1,...,X_n \sim Poisson(\lambda)$, keressünk elégséges statisztikát λ-ra!

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

azaz a mintaelemek összege elégséges statisztika. ■

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

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

2.3. Példa. Legyen $X_i \sim \mathrm{E}(0,b)$, b a paraméter. Próbáljunk faktorizálni!

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

tehát $X_n^{(n)}$ elégséges statisztika. ■

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

2.1. Feladat. Keressünk elégséges statisztikát a paraméter(ek)re a következő mintákból:
1) $X_i \sim Geo(p)$
2) $X_i \sim Bin(m,p)$ $\quad m$ ismert
3) $X_i \sim Exp(\lambda)$
4) $X_i \sim N(m,\sigma)$
5) $X_i \sim E(-a,a)$
Mo: 1) $\sum X_i$ elégséges:

$$p_{n;p}(x) = \prod_{i=1}^{n}(1-p)^{x_i-1}\cdot p = (1-p)^{\sum x_i - n}p^n.$$

2) $\sum X_i$ elégséges:

$$p_{n;p}(x) = \prod_{i=1}^{n}\binom{m}{x_i}p^{x_i}(1-p)^{m-x_i} = p^{\sum x_i}(1-p)^{m\cdot n-\sum x_i}\cdot\prod_{i=1}^{n}\binom{m}{x_i}.$$

3) $\sum X_i$ elégséges:

$$f_{n;\lambda}(x) = \prod_{i=1}^{n}\lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda\sum x_i}.$$

4) Ha (m,σ) ismeretlen, akkor $(\sum X_i, \sum X_i^2)$ elégséges:

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

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

Ha σ ismert, akkor $\sum X_i$ elégséges:

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

Ha m ismert, akkor $\sum(X_i-m)^2$ elégséges.
5) $\max|X_i|$ elégséges:

$$f_{n;a}(x) = \prod_{i=1}^{n}\frac{1}{2a}I(-a<x<+a) = \frac{1}{(2a)^n}\cdot I(|x_i|<a \quad i=1,...,n) = \frac{1}{(2a)^n}\cdot I(|x|_n^{(n)}<a).$$

■

3. Becslések és jóságuk

Legyen $X_1,...,X_n$ (független, azonos eloszlású) minta az F_ϑ eloszlásfüggvényű eloszlásból. Szeretnénk a paraméter $\psi(\vartheta)$ függvényét becsülni (gyakran magát a paramétert kell becsülni, de nem mindig). Nem remélhetjük, hogy a pontos értéket eltaláljuk, de azt igen, hogy jól meg tudjuk közelíteni.

3.1. Definíció. A $\psi(\vartheta)$ mennyiség *becslése* valamely $T(X)$ statisztika.

Azért vezettünk be egy új elnevezést a $T(X)$ statisztikára, mert most úgy gondolunk rá, mint a $\psi(\vartheta)$ mennyiséget jól közelítő becslésre. Mit várhatunk el a $T(X)$ becsléstől?
1) A $T(X)$ becslés nagyjából $\psi(\vartheta)$ körül ingadozzék.
2) A $T(X)$ minél kevésbé ingadozzék $\psi(\vartheta)$ körül, azaz a becslés legyen minél pontosabb.
3) Tegyük fel, hogy minden n mintaelemszámra van egy T_n becslésünk. Megkövetelhetjük a $T_n(X_1,...,X_n)\to\psi(\vartheta)$ sztochasztikus konvergenciát.

3.2. Definíció. A $T(X)$ becslés *torzítatlan* $\psi(\vartheta)$-ra, ha

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

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

3.1. Példa. Legyen $X_1, ..., X_n \sim F_\vartheta$ és $\psi(\vartheta) = E_\vartheta(X_i)$. Ekkor $\overline{X}$ torzítatlan becslés, mivel

$$E_\vartheta(\overline{X}) = E_\vartheta(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{E_\vartheta(X_i)}_{\psi(\vartheta)} = \psi(\vartheta).$$

Ez alkalmazható például a Poisson vagy az indikátor eloszlás paraméterének becslésére. ■

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

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

ez pedig legfeljebb n-edfokú polinomja p-nek. Másrészt p^k egy torzítatlan becslése: $I(X_1 = 1) \cdot I(X_2 = 1) \cdots I(X_k = 1) \quad (k \leq n)$. ■

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Azt kell tehát belátni, hogy

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

A számtani és négyzetes közép közötti egyenlőtlenségből

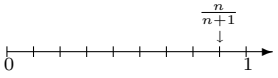
$$\sqrt{\frac{\sum c_i^2}{n}} \ge \frac{\sum c_i}{n} = \frac{1}{n}, \text{ azaz } \sum c_i^2 \ge \frac{n}{n^2} = \frac{1}{n}.$$

■

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

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

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



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

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

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

Másrészt

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

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

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

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

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

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

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

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

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

$$D^2_\vartheta(\frac{n+1}{n} \cdot X^{(n)}_n) = \frac{\vartheta^2}{n(n+1)} \longrightarrow 0.$$

■

4. Fisher-információ

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

4.2. Definíció. Az n elemű, $F_{n;\vartheta}$ eloszlásfüggvényű minta *Fisher-információja* az

$$I_n(\vartheta) = E_\vartheta([\frac{\partial}{\partial\vartheta} \ln L_n(X;\vartheta)]^2)$$

érték, ha a derivált létezik, és ez a mennyiség véges.

4.1. Tétel. (*) Legyenek $X_1, \ldots, X_n$ független, azonos eloszlású mintaelemek, és jelölje egy mintaelem likelihood függvényét $L_1(x_i;\vartheta)$. Tegyük fel, hogy $I_1(\vartheta) < \infty$, továbbá hogy

$$E_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_1(X_1;\vartheta)) = 0.$$

Ekkor $I_n(\vartheta) = n \cdot I_1(\vartheta)$.

Bizonyítás. A feltétel miatt $I_1(\vartheta) = D^2_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_1(X_1;\vartheta))$, továbbá

$$E_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_n(X;\vartheta)) = E_\vartheta(\frac{\partial}{\partial\vartheta} \ln \prod_{i=1}^n L_1(X_i;\vartheta)) = \sum_{i=1}^n \underbrace{E_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_1(X_i;\vartheta))}_{0} = 0.$$

Ebből következik, hogy

$$I_n(\vartheta) = D^2_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_n(X;\vartheta)) = D^2_\vartheta(\sum_{i=1}^n \frac{\partial}{\partial\vartheta} \ln L_1(X_i;\vartheta)) = \sum_{i=1}^n \underbrace{D^2_\vartheta(\frac{\partial}{\partial\vartheta} \ln L_1(X_i;\vartheta))}_{I_1(\vartheta)} = n \cdot I_1(\vartheta).$$

■

Mj: A feltétel gyakran teljesül, mivel ez egy bederiválhatóság álruhába bújtatva. Tegyük fel mondjuk, hogy a minta abszolút folytonos. Tudjuk, hogy

$$\int_{-\infty}^{+\infty} f_{1;\vartheta}(x)\,\mathrm{d}x = 1 \quad \Rightarrow \quad \frac{\partial}{\partial\vartheta} \int_{-\infty}^{+\infty} f_{1;\vartheta}(x)\,\mathrm{d}x = 0.$$

Ha a deriválás és integrálás felcserélhető, azaz be lehet deriválni, akkor kapjuk, hogy

$$0 = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\vartheta} f_{1;\vartheta}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{\frac{\partial}{\partial\vartheta} f_{1;\vartheta}(x)}{f_{1;\vartheta}(x)} \cdot f_{1;\vartheta}(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\vartheta} \ln f_{1;\vartheta}(x) \cdot f_{1;\vartheta}(x)\,\mathrm{d}x = E_\vartheta(\frac{\partial}{\partial\vartheta} \ln f_{1;\vartheta}(X_1)).$$

Diszkrét esetben, hasonlóan, a tétel feltétele azzal ekvivalens, hogy

$$\frac{\partial}{\partial\vartheta}\sum_{x_1} p_{1;\vartheta}(x_1) = \sum_{x_1}\frac{\partial}{\partial\vartheta}p_{1;\vartheta}(x_1).$$

Ez biztosan teljesül, ha a valószínűségi változó értékkészlete véges.

4.1. Példa. Legyen $X_i \sim Ind(p)$. Ekkor $L_1(x,p) = p^x(1-p)^{1-x}$ $(x \in \{0,1\})$, amiből

$$\frac{\partial}{\partial p}\ln L_1(x,p) = \frac{x}{p} - \frac{1-x}{1-p}.$$

Ellenőrizzük a **(*)** Tétel feltételét!

$$E_p\left(\frac{\partial}{\partial p}\ln L_1(X_i,p)\right) = \frac{p}{p} - \frac{1-p}{1-p} = 0.$$

Tehát az egyelemű minta Fisher-információja:

$$I_1(p) = D_p^2\left(\frac{\partial}{\partial p}\ln L_1(X_i,p)\right) = D_p^2\left(\frac{X_i}{p(1-p)}\right) = \frac{1}{p(1-p)},$$

és $I_n(p) = n\cdot I_1(p) = \dfrac{n}{p(1-p)}$. ■

4.2. Példa. Legyen $X_i \sim N(\vartheta,\sigma_0)$, ahol σ_0 ismert. Ekkor

$$L_1(x,\vartheta) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\cdot e^{-\frac{(x-\vartheta)^2}{2\sigma_0^2}}.$$

Logaritmust véve

$$\ln L_1(x,\vartheta) = \ln\frac{1}{\sqrt{2\pi\sigma_0^2}} - \frac{(x-\vartheta)^2}{2\sigma_0^2}.$$

Ellenőrizzük a **(*)** Tétel feltételét!

$$E_\vartheta\left[(\frac{\partial}{\partial\vartheta}\ln L_1(X_i,\vartheta)\right] = E_\vartheta\left[\frac{2(X_i-\vartheta)}{2\sigma_0^2}\right] = 0.$$

Ebből

$$I_1(\vartheta) = D_\vartheta^2\left(\frac{X_i-\vartheta}{\sigma_0^2}\right) = \frac{D_\vartheta^2(X_i)}{\sigma_0^4} = \frac{1}{\sigma_0^2},$$

és $I_n(\vartheta) = \dfrac{n}{\sigma_0^2}$. ■

Mj: Nézhetnénk azt is, hogy a minta mennyi információt tartalmaz a paraméter valamely $\psi(\vartheta)$ függvényére, illetve mennyi információt tartalmaz egy paramétervektorra nézve. Ezekkel most nem foglalkozunk.

Bizonyítás nélkül megjegyezzük a következőt. Legyen X minta, a benne rejlő információ $I_X(\vartheta)$. Továbbá legyen $T = T(X)$ egy statisztika, a benne rejlő információt jelölje $I_T(\vartheta)$. Belátható, hogy (bizonyos regularitási feltételek mellett) $I_T(\vartheta) \le I_X(\vartheta)$, és egyenlőség akkor és csak akkor van, ha T elégséges statisztika. Például a következő regularitási feltételeket szokták tenni:
1) $\sqrt{L_1(x,\vartheta)}$ folytonosan differenciálható ϑ szerint,
2) $\infty > I_1(\vartheta) > 0$, és $I_1(\vartheta)$ folytonos ϑ-ban.
Ha ezek teljesülnek, akkor minden *szép és jó*, pl. ekkor teljesül a **(*)** tétel bederiválhatósági feltétele.

A következő tétel arról szól, hogy egy torzítatlan becslés nem lehet tetszőlegesen pontos, a szórásnégyzetre adható egy alsó korlát, mely a mintában található információ mennyiségtől függ (azonban ez az alsó korlát általában nem éles).

4.2. Tétel. (Cramér-Rao egyenlőtlenség) Legyen $X = (X_1, ..., X_n)$ minta, és tegyük fel, hogy teljesülnek a **(*)** tétel feltételei. Legyen még $T(X)$ olyan torzítatlan becslése $\psi(\vartheta)$-nak, melyre $D^2_\vartheta(T) < \infty$ minden ϑ-ra. Feltesszük még a következő bederiválhatósági feltételt:

$$\psi'(\vartheta) = E_\vartheta\left(T(X)\frac{\partial}{\partial\vartheta}\ln L_n(X;\vartheta)\right).$$

Ekkor $D^2_\vartheta(T) \geq \frac{(\psi'(\vartheta))^2}{I_n(\vartheta)}$ teljesül minden ϑ-ra.

Bizonyítás. Az ötlet: tudjuk, hogy minden S-re

$$\mathrm{cov}_\vartheta(T,S)^2 \leq D^2_\vartheta(T)D^2_\vartheta(S) \quad \Rightarrow \quad D^2_\vartheta(T) \geq \frac{\mathrm{cov}_\vartheta(T,S)^2}{D^2_\vartheta(S)}.$$

Ez olyan S választással ad használható korlátot, melyre

$$\mathrm{cov}_\vartheta(T,S) = E_\vartheta(TS) - E_\vartheta(T)E_\vartheta(S) = E_\vartheta(TS) - \psi(\vartheta)E_\vartheta(S)$$

nem függ T-től. Legyen $S = \frac{\partial}{\partial\vartheta}\ln L_n(X;\vartheta)$, ez a feltétel szerint jó lesz. Egyrészt $E_\vartheta(S) = 0$ (ez volt a **(*)** Tétel egyik feltétele), tehát

$$\mathrm{cov}_\vartheta(T,S) = E_\vartheta(T\cdot S) = \psi'(\vartheta),$$

másrészt $D^2_\vartheta(S) = I_n(\vartheta)$. ■

Mj: Nézzük a feltételt!

$$\psi(\vartheta) = E_\vartheta(T) = \int T(x)f_{n;\vartheta}(x)\,\mathrm{d}x\,.$$

Bederiválhatóság esetén ebből

$$\psi'(\vartheta) = \int T(x)\frac{\partial}{\partial\vartheta}f_{n;\vartheta}(x)\,\mathrm{d}x = E_\vartheta\left(T(X)\frac{\partial}{\partial\vartheta}\ln L_n(X;\vartheta)\right).$$

Az alsó korlát, azaz a $\frac{(\psi'(\vartheta))^2}{I_n(\vartheta)}$ mennyiség neve *információs határ*. Ha $D^2_\vartheta(T) = \frac{(\psi'(\vartheta))^2}{I_n(\vartheta)}$ teljesül minden ϑ-ra, akkor azt mondjuk, hogy a T becslés eléri az információs határt. Nyilván, ha a T torzítatlan becslés eléri az információs határt, akkor hatásos.

Speciálisan, ha magát a paramétert akarjuk becsülni, akkor $\psi(\vartheta) = \vartheta$, azaz az információs határ $1/I_n(\vartheta)$.

4.1. Feladat. Legyen $X_i \sim N(\vartheta, \sigma_0)$, ϑ-ra keressünk hatásos becslést! Az információs határ $1/I_n(\vartheta) = \sigma_0^2/n$, ugyanakkor $D^2_\vartheta(\overline{X}) = \sigma_0^2/n$, ezért $\overline{X}$ hatásos becslés ϑ-ra. ■

4.2. Feladat. Legyen $X_i \sim Ind(p)$, p-re keressünk hatásos becslést! Az információs határ $1/I_n(p) = p(1-p)/n$, ugyanakkor $D^2_p(\overline{X}) = p(1-p)/n$, ezért $\overline{X}$ hatásos becslés p-re. ■

4.3. Feladat. Legyen $X_i \sim Exp(\lambda)$, $\psi(\lambda) = \frac{1}{\lambda}$-t szeretnénk becsülni. Tudjuk, hogy $\overline{X}$ torzítatlan becslés. Szórásnégyzete

$$D^2_\lambda(\overline{X}) = \frac{\frac{1}{\lambda^2}}{n} = \frac{1}{\lambda^2\cdot n}.$$

A Fisher-információ:

$$E_\lambda\left[\frac{\partial}{\partial\lambda}\ln f_{1;\lambda}(X_i)\right] = E_\lambda\left[\frac{\partial}{\partial\lambda}(\ln\lambda - \lambda X_i)\right] = E_\lambda\left[\frac{1}{\lambda} - X_i\right] = 0.$$

Ebből $I_1(\lambda) = D^2_\lambda\left(\frac{1}{\lambda} - X_i\right) = D^2_\lambda(X_i) = \frac{1}{\lambda^2}$, $I_n(\lambda) = \frac{n}{\lambda^2}$. Most $\psi(\lambda) = 1/\lambda$, $\psi'(\lambda) = -1/\lambda^2$. Az információs határ tehát

$$\frac{\left(-\frac{1}{\lambda^2}\right)^2}{\frac{n}{\lambda^2}} = \frac{1}{n\cdot\lambda^2}.$$

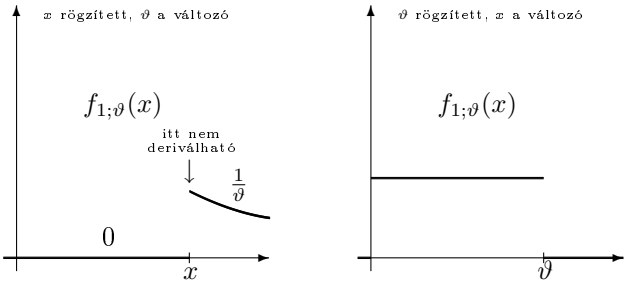
Kaptuk, hogy $\overline{X}$ hatásos becslés $1/\lambda$-ra. ■

4.4. Feladat. Legyen $X_i \sim \text{Poisson}(\lambda)$. Mutassuk meg, hogy $\overline{X}$ hatásos becslés λ-ra! ■

4.5. Feladat. Legyen $X_i \sim E(0,\vartheta)$. Most nem teljesülnek a regularitási feltételek, rögzített x mellett $f_{1;\vartheta}(x)$ nem folytonosan deriválható ϑ szerint.
$f_{1;\vartheta}(x) = \begin{cases} \frac{1}{\vartheta} & \text{ha } 0 \le x \le \vartheta \\ 0 & \text{egyébként} \end{cases}$

$$I_1(\vartheta) = E_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\ln L_1(X_1;\vartheta)\right)^2\right] = E_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\ln\frac{1}{\vartheta}\right)^2\right] = \left(-\frac{1}{\vartheta}\right)^2 = \frac{1}{\vartheta^2},$$

felhasználva, hogy 1 valószínűséggel $X_1 < \vartheta$, azaz 1 valószínűséggel $L_1(X_1;\vartheta)$ deriválható. Ugyanakkor $I_n(\vartheta) = E_\vartheta\left[\left(\frac{\partial}{\partial\vartheta}\ln\left(\frac{1}{\vartheta}\right)^n\right)^2\right] = \left(-\frac{n}{\vartheta}\right)^2 = \frac{n^2}{\vartheta^2} = n^2\cdot I_1(\vartheta)$.



■

5. Becslési módszerek

5.1. Blackwellizálás

Először nézzünk egy kis ismétlést! Ha X, Y diszkrét valószínűségi változók, akkor X feltételes várható értéke az $Y = y$ feltétel mellett $E(X|Y=y) = \sum_x x\cdot P(X=x|Y=y) = V(y)$. Az X feltételes várható értéke az Y változóra nézve pedig $E(X|Y) = V(Y)$. Ez tehát nem egy szám, hanem egy valószínűségi változó. Bizonyítás nélkül elevenítsük fel a következő tulajdonságokat:
1) $E(c|Y) = c$
2) $E(X_1 + X_2|Y) = E(X_1|Y) + E(X_2|Y)$
3) $E(cX|Y) = cE(X|Y)$
4) $X_1 \le X_2 \;\Rightarrow\; E(X_1|Y) \le E(X_2|Y)$.
Ezeken kívül szükségünk lesz még néhány tulajdonságra:
5) Teljes várható érték tétele:

$$E\left(E(X|Y)\right) = E(V(Y)) = \sum_y V(y)\cdot P(Y=y) = \sum_y\sum_x x\cdot P(X=x|Y=y)\cdot P(Y=y) =$$
$$\sum_x x\underbrace{\sum_y P(X=x|Y=y)\cdot P(Y=y)}_{P(X=x)\text{ teljes valószínűség tétele}} = \sum_x x\cdot P(X=x) = E(X).$$

5) Kiemelés: $E(XW(Y)|Y) = W(Y)E(X|Y)$, magyarázat: Y rögzítése mellett $W(Y)$ konstans, azaz kiemelhető a várható értékből
6) Teljes szórásnégyzet tétele: $D^2(X) = E\left(D^2(X|Y)\right) + D^2\left(E(X|Y)\right)$, ahol $D^2(X|Y) = E\left((X - E(X|Y))^2|Y\right)$.
Bizonyítás:

$$D^2(X|Y) = E\left(X^2 - 2XE(X|Y) + E(X|Y)^2|Y\right) = E(X^2|Y) - E\left(2XE(X|Y)|Y\right) + E\left(E(X|Y)^2|Y\right) =$$
$$E(X^2|Y) - 2E(X|Y)E(X|Y) + E(X|Y)^2 = E(X^2|Y) - E(X|Y)^2.$$

Várható értéket véve:

$$E\left(D^2(X|Y)\right) = \underbrace{E\left(E(X^2|Y)\right)}_{E(X^2)} - E\left(E(X|Y)^2\right) = \underbrace{E(X^2) - E(X)^2}_{D^2(X)} - \underbrace{\left[E\left(E(X|Y)^2\right) - E(X)^2\right]}_{D^2(E(X|Y))}.$$

Következményként kapjuk, hogy $D^2\left(E(X|Y)\right) \leq D^2(X)$.

5.1. Példa. Dobjunk fel egy szabályos kockát n-szer. Legyen X a hatosok száma, Y a páratlanok száma. Az $\{Y = y\}$ feltétel mellett X eloszlása $Bin(n-y, 1/3)$. Tehát $E(X|Y) = (n-Y)/3$. Erre

$$E(E(X|Y)) = E((n-Y)/3) = \frac{n - E(Y)}{3} = \frac{n}{6}, \quad E(X) = \frac{n}{6}.$$

Másrészt

$$D^2(X|Y) = (n-Y) \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2(n-Y)}{9}, \quad E(D^2(X|Y)) = \frac{n}{9}.$$

Végül

$$D^2(X) = \frac{5n}{36}, \quad D^2(E(X|Y)) = \frac{n}{36}.$$

■

5.1. Tétel. (Rao-Blackwell tétel) Legyen X diszkrét minta, S torzítatlan becslése $\psi(\vartheta)$-nak, T pedig elégséges statisztika ϑ-ra. Ekkor $U = E(S|T)$ is torzítatlan becslés $\psi(\vartheta)$-ra, és hatásosabb S-nél.

Bizonyítás. A feltételes várható érték tulajdonságaiból rögtön kötetkezik:

$$E_\vartheta(U) = E_\vartheta\left(E_\vartheta(S|T)\right) = E_\vartheta(S) = \psi(\vartheta), \quad D^2_\vartheta(U) = D^2_\vartheta\left(E_\vartheta(S|T)\right) \leq D^2_\vartheta(S).$$

■

Mj: Azt, hogy T elégséges statisztika, ott használtuk fel, hogy az $E(S|T)$ várható érték nem függ ϑ-tól, azaz a mintából kiszámolható mennyiség. Az eljárásnak lényege tehát az, hogy egy akármilyen torzítatlan becslés hatásosságát javíthatjuk azzal, ha egy elégséges statisztikára vett feltételes várható értékét képezzük. Ennek az eljárásnak elnevezése *blackwellizálás*. A tételből következik, hogy ha van hatásos becslés, akkor az tetszőleges elégséges statisztikának függvénye.

5.2. Példa. Legyen $X_i \sim Geo(p)$, p-t szeretnénk becsülni. Elégséges statisztika: $T = \sum_{i=1}^{n} X_i$. Egy egyszerű torzítatlan becslés: $S = I(X_1 = 1)$. Blackwellizáljunk, azaz számoljuk ki az $U = E(S|T) = V(T)$ becslést!

$$V(\ell) = E_p(S|T=\ell) = 0 \cdot P_p(S=0|T=\ell) + 1 \cdot P_p(S=1|T=\ell) = P_p(X_1 = 1|\sum_{i=1}^{n} X_i = \ell) = \frac{P_p(X_1 = 1\,,\, \sum_{i=1}^{n} X_i = \ell)}{P(\sum_{i=1}^{n} X_i = \ell)}.$$

A számláló

$$P_p(X_1 = 1, \sum_{i=1}^{n} X_1 = \ell) = P_p(X_1 = 1, \sum_{i=2}^{n} X_i = \ell - 1) = P_p(X_1 = 1) P_p(\sum_{i=2}^{n} X_i = \ell - 1).$$

A nevezőnél pedig azt használjuk, hogy $\sum_{i=1}^{n} X_i \sim NegBin(n, p)$. Tehát

$$V(\ell) = \frac{P_p(X_1 = 1) P_p(\sum_{i=2}^{n} X_i = \ell - 1)}{P(\sum_{i=1}^{n} X_i = \ell)} = \frac{p\binom{\ell-2}{n-2} p^{n-1}(1-p)^{\ell-n}}{\binom{\ell-1}{n-1} p^n (1-p)^{\ell-n}} = \frac{n-1}{\ell-1}.$$

Azaz $U = V(\sum_{i=1}^{n} X_i) = \dfrac{n-1}{\sum_{i=1}^{n} X_i - 1}$ lesz az S blackwellizáltja (megmutatható, hogy ez hatásos becslés). ■

5.1. Feladat. Legyen $X_i \sim Poisson(\lambda)$, keressünk jó torzítatlan becslést $\psi(\lambda) = e^{-\lambda}$-ra blackwellizálással!
Első ötlet: $e^{-\overline{X}}$, de belátható, hogy ez nem torzítatlan.
Vegyük észre, hogy $e^{-\lambda} = P_\lambda(X_1 = 0)$, azaz $S = I(X_1 = 0)$ jó lesz. Továbbá $T = \sum_{i=1}^n X_i$ elégséges.

$$E_\lambda(S|T = \ell) = P_\lambda(X_1 = 0|\sum_{i=1}^n X_i = \ell) = \frac{e^{-\lambda} \cdot e^{-(n-1)\lambda} \cdot \frac{((n-1)\lambda)^\ell}{\ell!}}{e^{-n\lambda} \cdot \frac{(n\lambda)^\ell}{\ell!}} = \left(1 - \frac{1}{n}\right)^\ell.$$

Felhasználtuk, hogy $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ és $\sum_{i=2}^n X_i \sim \text{Poisson}((n-1)\lambda)$. Azaz a megjavított becslés:

$$V = \left(1 - \frac{1}{n}\right)^{\sum X_i} = \left[\underbrace{\left(1 - \frac{1}{n}\right)^n}_{\longrightarrow e^{-1}}\right]^{\overline{X}}.$$

■

5.2. Feladat. Legyen $X_i \sim \text{Ind}(p)$, keressünk jó torzítatlan becslést $\psi(p) = p^2$-re blackwellizálással!
Első ötlet: $\overline{X}^2$, de ez nem torzítatlan.
$S = I(X_1 = 1, X_2 = 1)$, $T = \sum_{i=1}^n X_i$,

$$E_p(S|T = \ell) = P_p(X_1 = 1, X_2 = 1 \,|\, \sum_{i=1}^n X_i = \ell) = \frac{p \cdot p \cdot \binom{n-2}{\ell-2} p^{\ell-2}(1-p)^n}{\binom{n}{\ell} p^\ell (1-p)^{n-\ell}} = \frac{\ell(\ell-1)}{n(n-1)},$$

azaz a megjavított becslés $V = \frac{\sum X_i}{n} \cdot \frac{\sum X_i - 1}{n-1}$. ■

5.2. Tapasztalati becslés

A tapasztalati becslés azt jelenti, hogy az elméleti eloszlás valamely jellemzőjét a tapasztalati eloszlás megfelelő jellemzőjével becsüljük. Ezek a becslések általában konzisztensek, mivel a tapasztalati eloszlásfüggvény tart az elméletihez. A következő táblázatban tapasztalati becslésre látunk néhány példát.

Jellemző	**Elméleti eloszlásé**	**Tapasztalati eloszlásé**
várható érték	$E(X)$	$\frac{1}{n}\sum_{i=1}^n X_i = \overline{X}$
szórásnégyzet	$D^2(X)$	$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = S_n^2$
terjedelem	$\sup\{\, x \,|\, F(x) < 1 \,\} - \inf\{\, x \,|\, F(x) > 0 \,\}$	$X_n^{(n)} - X_1^{(n)}$
eloszlásfüggvény	$F(x) = P(X < x)$	$\frac{1}{n}\sum_{i=1}^n I(X_i < x) = \hat{F}_n(x)$

5.3. Maximum likelihood becslés

5.1. Definíció. Legyen $X_1, ..., X_n$ minta F_ϑ eloszlásból, $\vartheta \in \Theta$. Ekkor a
ϑ maximum likelihood (ML) becslése $\hat{\vartheta}$, ha $L_n(X; \hat{\vartheta}) = \max\{L_n(X; \vartheta) : \vartheta \in \Theta\}$.

Lehet, hogy ilyen nem létezik, vagy nem egyértelmű. Ha $L_n(X; \vartheta)$ "elég sima", akkor a következőt szokás csinálni: $\frac{\partial}{\partial \vartheta} \ln L_n(X; \vartheta) = 0$ *likelihood egyenlet* megoldását keressük.

5.3. Példa. $X_i \sim Exp(\lambda)$. $L_n(X; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum X_i}$, $\ln L_n(X; \lambda) = n \cdot \ln \lambda - \lambda \sum X_i$, $\frac{\partial}{\partial \lambda} \ln L_n(X; \lambda) = \frac{n}{\lambda} - \sum X_i$, ez akkor 0, ha $\lambda = \frac{n}{\sum X_i} = \frac{1}{\overline{X}}$, azaz $\hat{\lambda} = \frac{1}{\overline{X}}$. ■

5.4. Példa. $X_i \sim \text{Egyenletes}(\vartheta, \vartheta + 1)$. $L_n(X; \vartheta) = I(\vartheta \le X_i \le \vartheta + 1, \forall i) \left(\frac{1}{1}\right)^n = I(\vartheta \le X_1^{(n)}, X_n^{(n)} \le \vartheta + 1)$. Ennek a függvénynek a maximumhelye nem egyértelmű, minden $\hat{\vartheta} \in \left[X_n^{(n)} - 1, X_1^{(n)}\right]$ érték ML becslés. ■

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

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

5.3. Tétel. Legyen $X_1,...,X_n$ minta az F_ϑ eloszlásból, ϑ ML becslése pedig $\hat\vartheta$. Ekkor $\psi(\hat\vartheta)$ ML becslés $\psi(\vartheta)$-ra.

Bizonyítás. Legyen $\tilde L_n(x;\psi) = \sup\{\, L_n(x;\vartheta) \,|\, \psi(\vartheta)=\psi \,\}$ az *indukált likelihood függvény.* Definíció szerint $\psi(\vartheta)$ ML becslése $\tilde L_n(x;\psi)$ maximumhelye ψ-ben. $\tilde L_n(x;\psi) \le L_n(x;\hat\vartheta)$ és $\tilde L_n(x;\psi(\hat\vartheta)) = L_n(x;\hat\vartheta)$. Tehát $\tilde L_n(x;\psi(\hat\vartheta)) = \max_\psi\left\{\tilde L_n(x;\psi)\right\}$ ■

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

5.4. Tétel. Bizonyos (erős) regularitási feltételek mellett elég nagy n-re a $\hat\vartheta_n$ ML becslés létezik, konzisztens,

a) aszimptotikusan normális eloszlású: $\sqrt{n}\cdot(\hat\vartheta_n - \vartheta) \longrightarrow N(m(\vartheta),\sigma(\vartheta))$ eloszlásban $(n\to\infty)$.
b) aszimptotikusan torzítatlan: $m(\vartheta)=0$.
c) aszimptotikusan optimális: $\sigma^2(\vartheta) = \frac{1}{I_1(\vartheta)}$.

5.5. Példa. $X_i \sim E(a,b)$, adjuk meg a paraméterek ML becslését!
$L_n(X;a,b) = \left(\frac{1}{b-a}\right)^n \cdot I(a\le X_i\le b \quad \forall i)$, azaz a legkisebb olyan $[a,b]$ intervallumot keressük, amely mindegyik megfigyelést tartalmazza. Azaz $\hat a = X_1^{(n)}$, $\hat b = X_n^{(n)}$. ■

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

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

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

A likelihood egyenlet most két egyenletből áll:

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

$$\frac{\partial}{\partial \sigma}\ln L_n(X;m,\sigma) = -\frac{n}{\sigma} - \frac{2\sum(X_i-m)^2}{2\sigma^3} = 0$$

(1) $\sum X_i = n\cdot m \quad \Rightarrow \quad \hat m = \frac{\sum X_i}{n} = \overline{X}$

(2) $\sum(X_i-m)^2 = n\cdot\sigma^2 \quad \Rightarrow \quad \sigma^2 = \frac{\sum(X_i-m)^2}{n}$

$\hat\sigma^2 = \frac{1}{n}\sum(X_i-\overline{X})^2 = S_n^2$ ■

5.4. Momentumok módszere

Ezt a módszert akkor szokás alkalmazni, ha sok ismeretlen paraméter van, és a ML becslést nehéz kiszámítani. Az eljárás a következő.

Eljárás

1) Az eloszlás ℓ. (elméleti) momentuma: $\mu_\ell = E_\vartheta(X_i^\ell)$
Felírunk annyi momentumot, amennyi meghatározza az eloszlást.
2) Kifejezzük az ismeretlen paramétereket a momentumok segítségével.
3) $E_\vartheta(X_i^\ell)$ helyébe $\frac{1}{n}\sum_{i=1}^n X_i^\ell$-et írunk (ezek a tapasztalati momentumok).

5.7. Példa. Legyen $X_i \sim E(a,b)$.

$$\mu_1 = E_{a,b}(X_i) = \frac{a+b}{2},\ \mu_2 = E_{a,b}(X_i^2) = D^2_{a,b}(X_i) + E_{a,b}(X_i)^2 = \frac{(b-a)^2}{12} + \left(\frac{a+b}{2}\right)^2$$

$$\frac{(b-a)^2}{12} = \mu_2 - \mu_1^2 \quad \Rightarrow \quad \begin{aligned} b-a &= \sqrt{12(\mu_2-\mu_1^2)} \\ b &= \mu_1 + \sqrt{3(\mu_2-\mu_1^2)} \\ a &= \mu_1 - \sqrt{3(\mu_2-\mu_1^2)} \end{aligned}$$

$\hat{a} = \overline{X} - \sqrt{3(\frac{1}{n}\sum X_i^2 - \overline{X}^2)} = \overline{X} - \sqrt{3}S_n$, $\hat{b} = \overline{X} + \sqrt{3(\frac{1}{n}\sum X_i^2 - \overline{X}^2)}\overline{X} + \sqrt{3}S_n$. ■

5.8. Példa. Legyen $X_i \sim E(-a,a)$.

$\mu_1 = E_a(X_i) = 0$

$$\mu_2 = E_a(X_i^2) = \frac{a^2}{3} \quad \Rightarrow \quad a = \sqrt{3\mu_2}$$

$$\hat{a} = \sqrt{3 \cdot \tfrac{1}{n} \cdot \sum X_i^2}.$$

5.5. Intervallum becslések

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

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

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

Mj: A KI pontos megbízhatósági szintje

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

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

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

■

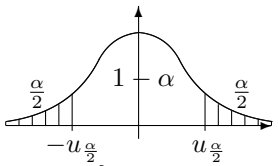
Az egyik legfontosabb (és legszebb) eset a normális eloszlás várható értékére KI konstruálása, ismert vagy ismeretlen szórás mellett. Nézzük meg ezeket!

5.9. Példa. Legyen $X_i \sim N(\mu, \sigma)$, ahol σ ismert, μ ismeretlen. Adjunk μ-re $KI(1-\alpha)$-t!

Kiindulásként vegyük észre, hogy $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$, azaz $\sqrt{n} \cdot \dfrac{\overline{X}-\mu}{\sigma} \sim N(0,1)$. Ezért

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

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



Ebből megkaphatjuk a KI alsó és felső határát:

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

Azaz kaptuk, hogy

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

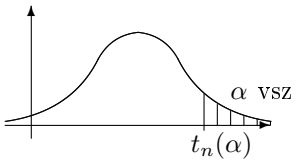
Ha a σ szórás nem ismert, akkor nehezebb dolgunk van. A megoldáshoz meg kell ismernünk két új eloszlást, a χ^2-eloszlást és a t-eloszlást.

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

Könnyű látni, hogy ha $Y \sim \chi_n^2$, akkor $E(Y) = n$ és $D^2(Y) = 2n$. A khi-négyzet eloszlás sűrűségfüggvénye is kiszámolható, erre azonban nem lesz szükségünk. Csak megjegyezzük, hogy $n = 1$-re a sűrűségfüggvény $\frac{1}{\sqrt{2\pi x}} e^{-x/2}$, $n = 2$-re pedig az $Exp(1/2)$ eloszlást kapjuk. Ha $n \geq 3$, akkor a sűrűségfüggvény egy darabig monoton nő, majd monoton csökken.

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

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



5.6. Tétel (Fisher-Bartlett). Legyen $X_i \sim N(\mu, \sigma)$. Ekkor $\overline{X}$ és S_n^2 függetlenek, és $\frac{n \cdot S_n^2}{\sigma^2} \sim \chi_{n-1}^2$.

Mj: Legyen $X_i \sim N(\mu, \sigma)$, és $Y_i = (X_i - \mu)/\sigma$. Ekkor $\frac{n \cdot S_n^2}{\sigma^2} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$. Mivel az $(Y_i - \overline{Y})$ valószínűségi változók nem függetlenek, csökken a szabadsági fok.
Mj: A mintaátlag és a tapasztalati szórásnégyzet függetlensége karakterizálja a normális eloszlást.

5.10. Példa. Legyen $X_i \sim N(\mu, \sigma)$, és most μ, σ ismeretlenek. Adjunk μ-re $KI(1-\alpha)$-t!
Legyen $S_n^* = \sqrt{\frac{1}{n-1} \sum (X_i - \overline{X})^2}$ a korrigált tapasztalati szórás. Ekkor a Fisher-Bartlett tétel szerint

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

Azaz

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

ebből a KI

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

■

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

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

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

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

egyenlőtlenséget rendezzük át. ■

5.3. Feladat. Legyen $X_i \sim Exp(\lambda)$. Adjunk hozzávetőleges $KI(1-\alpha)$-t λ-ra!

5.4. Feladat. Egy cukorgyárban kockacukrokat gyártanak. Tegyük fel, hogy a cukrok milliméterben kifejezett élhosszúsága közelítőleg normális eloszlású. 16 darab cukor élhosszúságát megmérve, a következő adatokat kaptuk:

10.10	8.94	9.61	10.00	10.42	10.33	10.68	9.25
10.32	10.26	10.32	10.74	9.98	10.23	9.88	9.89

Az adatok átlaga 10.06, tapasztalati szórása 0.46.
(a) Adjunk az élhossz m várható értékére $KI(0.95)$-t, ha tudjuk, hogy a szórás $\sigma = 0.5$.
(b) Adjunk KI-t m-re, ha a szórás ismeretlen!
(c) Ismert és ismeretlen szórás mellett is adjunk $KI(0.9)$-t m^3-re, azaz egy „átlagos" kockacukor térfogatára (ami nem egyenlő a kockacukrok átlagos térfogatával)!

6. Statisztikai próbák és jóságuk

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

Tegyük fel, hogy egy gyárban minőségellenőrzést végzünk, azaz megvizsgáljuk, hogy a gyártott termékek minősége megfelelő-e. Előzetes feltételezésünk szerint a gyártási folyamat rendben van, azaz a termékek legfeljebb 5%-a selejtes. A feltételezés ellenőrzéséhez 25 véletlenszerűen választott terméket megvizsgálunk, és ha legfeljebb 3 selejtes van köztük, akkor a feltételezést elfogadjuk. Ellenkező esetben a feltételezést elvetjük. Kérdés, hogy jó-e ez az eljárás?

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

Először tegyük fel, hogy a feltételezés igaz, azaz $p \leq 0.05$.

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

Ez a valószínűség akkor a legnagyobb, ha $p = 0.05$, így az elsőfajú hiba valószínűsége legfeljebb

$$\alpha = \sup_{p \leq 0.05} P_p(\geq 4 \text{ selejt}) = P_{0.05}(\geq 4 \text{ selejt}) = 0.034.$$

Most tegyük fel, hogy a feltételezés hamis, azaz $p > 0.05$. A másodfajú hiba valószínűsége

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

például ha a selejtarány $p = 0.1$, akkor 0.763 valószínűséggel fogjuk a feltételezést tévesen elfogadni.

Definiáljuk most formálisan az alapfogalmakat!

6.1. Definíció. Legyen $(\Omega, \mathcal{A}, \mathcal{P})$ paraméteres statisztikai mező, tehát $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, ahol Θ a paramétertér.
Hipotézisek:
nullhipotézis: $H_0 : \vartheta \in \Theta_0$
ellenhipotézis: $H_1 : \vartheta \in \Theta_1$,
ahol $\Theta = \Theta_0 \cup \Theta_1$ a paramétertér diszjunkt felbontása ($\Theta_0 \cap \Theta_1 = \emptyset$).
Minta:
$X = (X_1, ..., X_n)$ vektorváltozó (legtöbbször független, azonos eloszlású),
lehetséges értékeinek halmaza az $\mathcal{X}$ mintatér.
Statisztikai próba:
elfogadási tartomány: $\mathcal{X}_e$
kritikus (elutasítási) tartomány: $\mathcal{X}_k$,
ahol $\mathcal{X} = \mathcal{X}_e \cup \mathcal{X}_k$ a mintatér diszjunkt felbontása ($\mathcal{X}_e \cap \mathcal{X}_k = \emptyset$).
Ha $X \in \mathcal{X}_e$, akkor H_0-t elfogadjuk, ha $X \in \mathcal{X}_k$, akkor H_0-t elutasítjuk.

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

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

Nagyon fontos, hogy a két hipotézis szerepe általában nem egyenrangú. Az alapfeltevést csak nagyon indokolt esetben szeretnénk elutasítani, ezért az elsőfajú hiba súlyosabbnak számít, mint a másodfajú. Az elsőfajú hiba maximális valószínűségét szokás megadni, emellett természetesen a másodfajú hiba esélyének minimalizására törekszünk. Ebből kifolyólag a döntések értelmezése is különböző:
H_0-t elfogadjuk: nem jelenti azt, hogy igaz, csak azt, hogy nincs okunk elutasítani.
H_0-t elutasítjuk: komoly bizonyítékot találtunk arra, hogy H_0 nem igaz.
Például egy új gyógyszer vizsgálatnál a gyögyszer hatásosságára keresünk bizonyítékot, ezért a hipotézisek:
H_0: a gyógyszer nem hatásos,
H_1: a gyógyszer hatásos.
Folytassuk a definíciókat!

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

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

Másodfajú hiba: H_0 hamis, mégis elfogadjuk. Ennek valószínűsége: $P_\vartheta(X \in \mathcal{X}_e) \quad \vartheta \in \Theta_1$. A próba erőfüggvénye:

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

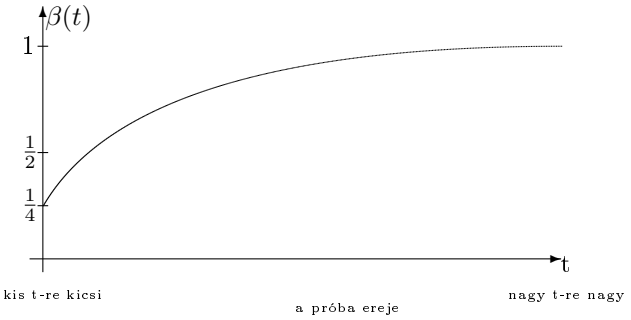
Másképp $\beta(\vartheta_1)$ a próba ereje a $H_1 : \vartheta = \vartheta_1$ ellenhipotézissel szemben.

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

6.1. Példa. Legyen $X \sim E(-t, 1 + 2t)$ egyetlen megfigyelés, ahol $t > 0$ ismeretlen.
$H_0 : t = 0$ (egyszerű hipotézis – Θ_0 egyelemű halmaz – nem kell sup a terjedelemben)
$H_1 : t > 0$ (összetett hipotézis – Θ_1 többelemű halmaz).
A próba: $\mathcal{X}_e = (0.1, 0.85)$, $(\mathcal{X} = \mathbb{R})$.

$$\alpha = P_0(X \in \mathcal{X}_k) = 1 - P_0(X \in \mathcal{X}_e) = 1 - P_0(0.1 < X < 0.85) = 1 - 0.75 = \underset{\sim}{0.}\underset{\sim}{25}.$$

erőfüggvény: $\beta(t) = P_t(X \in \mathcal{X}_k) = 1 - P_t(0.1 < X < 0.85) = 1 - \frac{0.75}{1+3t}$.



■

Mikor jó a próba?

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

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

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

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

Az egyenletesen legerősebb próba olyasmi a hipotézisvizsgálatban, mint a hatásos becslés a becslés-elméletben. Ilyen próbát általában nem könnyű konstruálni. Arra az esetre viszont tudunk egyenletesen legerősebb próbát adni, ha két egyszerű hipotézis között kell dönteni. Ehhez szükségünk lesz a véletlenített próba fogalmára. Eddigi próbák másképp megfogalmazva: Legyen $\Psi : \mathcal{X} \to \{0, 1\}$ egy függvény. Ha $x \in \mathcal{X}$-et figyeljük meg, akkor $\Psi(x)$ valószínűséggel utasítjuk el H_0-t. Ekkor:
$\mathcal{X}_e = \{\, x \,|\, \Psi(x) = 0 \,\}$
$\mathcal{X}_k = \{\, x \,|\, \Psi(x) = 1 \,\}$

6.4. Definíció. Véletlenített próba: Legyen $\Psi : \mathcal{X} \to [0, 1]$ egy függvény (ennek neve próbafüggvény). Ha $x \in \mathcal{X}$-et figyeljük meg, akkor $\Psi(x)$ valószínűséggel utasítjuk el H_0-t.

Számítsuk ki véletlenített próba esetén a hibavalószínűségeket (mondjuk diszkrét esetet véve)!

$$P_\vartheta(H_0\text{-t elutasítjuk}) = \sum_x \Psi(x) P_\vartheta(X = x) = E_\vartheta(\Psi(X)).$$

Így $\alpha = \sup_{\vartheta \in \Theta_0} E_\vartheta(\Psi(X))$, és $\beta(\vartheta) = E_\vartheta(\Psi(X)) \quad (\vartheta \in \Theta_1)$.

6.1. Tétel. (Neyman-Pearson lemma) Legyen H_0 és H_1 két egyszerű hipotézis:
$H_0 : \vartheta = \vartheta_0$, vagy másképp: a minta likelihood-függvénye $L_n(0; x)$
$H_1 : \vartheta = \vartheta_1$, vagy másképp: a minta likelihood-függvénye $L_n(1; x)$,
ahol X egy n elemű minta. Tekintsük a következő próbafüggvényt: $\Psi(x) = \begin{cases} 1 & \text{ha } \frac{L_n(1,x)}{L_n(0;x)} > c \\ \gamma & \text{ha } \frac{L_n(1,x)}{L_n(0;x)} = c \\ 0 & \text{ha } \frac{L_n(1,x)}{L_n(0;x)} < c \end{cases}$. Ekkor
1) Minden $0 < \alpha \leq 1$ esetén létezik c és γ, hogy a Ψ próba terjedelme pontosan α
2) A Ψ próba egyenletesen legerősebb az összes $\leq \alpha$ terjedelmű próba között (és lényegében egyértelmű).

Bizonyítás.
1) Legyen $Y = \frac{L_n(1;X)}{L_n(0;X)}$. Ekkor a próba terjedelme

$$\alpha = P_0(Y > c) + \gamma \cdot P_0(Y = c) = 1 - P_0(Y \leq c) + \gamma \cdot P_0(Y = c),$$

azaz $1-\alpha = P_0(Y \le c) - \gamma \cdot P_0(Y = c)$. Ha van olyan c_0, melyre $P_0(Y \le c_0) = 1-\alpha$ akkor a $c = c_0$, $\gamma = 0$ választás jó lesz. Ha nem létezik ilyen c_0, akkor is van olyan c_0, melyre $P_0(Y < c_0) \le 1-\alpha < P_0(Y \le c_0)$. Ekkor legyen

$$c = c_0, \quad \gamma = \frac{P_0(Y \le c_0) - (1-\alpha)}{P_0(Y = c_0)} \quad (0 < \gamma \le 1).$$

2) Legyen Ψ' egy másik próba próbafüggvénye. A feltevés szerint $E_0(\Psi'(X)) \le \alpha = E_0(\Psi(X))$. Azt kell belátni, hogy az erőkre $E_1(\Psi'(X)) \le E_1(\Psi(X))$ áll fenn. Tegyük fel, hogy a minta abszolút folytonos eloszlású, ekkor $L_n(1;x) = f_{n;1}(x)$, ahol $f_{n;1}(x)$ a minta együttes sűrűségfüggvénye a H_1 mellett, hasonlóan $L_n(0;x) = f_{n;0}(x)$, ahol $f_{n;0}(x)$ a minta együttes sűrűségfüggvénye a H_0 mellett. Megmutatjuk, hogy

$$\int_{\mathbb{R}^n} (\Psi(x) - \Psi'(x)) \cdot (f_{n;1}(x) - c \cdot f_{n;0}(x))\,\mathrm{d}x \ge 0. \tag{1}$$

Ugyanis, ha $f_{n;1}(x) - c \cdot f_{n;0}(x) = 0$, akkor az integrandus nulla. Ha $f_{n;1}(x) - c \cdot f_{n;0}(x) > 0$, akkor $\Psi(x) = 1$, viszont $\Psi'(x) \le 1$, így $\Psi(x) - \Psi'(x) \ge 0$, tehát az integrandus ≥ 0. Ha $f_{n;1}(x) - c \cdot f_{n;0}(x) < 0$, akkor $\Psi(x) = 0$, így $\Psi(x) - \Psi'(x) \le 0$, tehát az integrandus megint ≥ 0. Az (1) integrált szétbontva,

$$0 \le \int \Psi f_{n;1} - \int \Psi' f_{n;1} - c \cdot \int \Psi f_{n;0} + c \cdot \int \Psi' f_{n;0} = E_1(\Psi(X)) - E_1(\Psi'(X)) - c \cdot E_0(\Psi(X)) + c \cdot E_0(\Psi'(X)).$$

Átrendezve kapjuk, hogy

$$E_1(\Psi(X)) - E_1(\Psi'(X)) \ge c \cdot [E_0(\Psi(X)) - E_0(\Psi'(X))] \ge 0.$$

■

A próba neve: likelihood-hányados próba. Véletlenítésre általában csak diszkrét minták esetén van szükség, hogy a teljes terjedelmet ki tudjuk használni, és egyúttal növeljük az erőt. Ennek inkább elméleti, mint gyakorlati jelentősége van. Megmutatható, hogy (független, azonos elemű mintákra) a próba konzisztens, mivel $\beta_n \ge 1 - c^n$, ahol c a két likelihood függvénytől függő (egynél kisebb) konstans.

6.2. Példa. Egy érmével kétszer dobunk. H_0 : $P(\text{fej}) = \frac{1}{2}$, H_1 : $P(\text{fej}) = \frac{1}{6}$. A mintatér: $\mathcal{X} = \{FF, FI, IF, II\}$, és a rövidség kedvéért jelölje a likelihoodokat L_1 és L_0. A likelihoodok, illetve likelihood-hányadosok táblázata:

	FF	FI	IF	II
L_1	$\frac{1}{36}$	$\frac{5}{36}$	$\frac{5}{36}$	$\frac{25}{36}$
L_0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$\frac{L_1}{L_0}$	$\frac{1}{9}$	$\frac{5}{9}$	$\frac{5}{9}$	$\frac{25}{9}$

Legyen először $\alpha = 0.25$. A Neyman-Pearson lemma alapján olyan γ, c párt keresünk, melyre a $\Psi(x) = \begin{cases} 1 & \text{ha } \frac{L_1}{L_0} > c \\ \gamma & \text{ha } \frac{L_1}{L_0} = c \\ 0 & \text{ha } \frac{L_1}{L_0} < c \end{cases}$ próba terjedelme α, azaz $0.25 = \alpha = 1 \cdot P_0(\frac{L_1}{L_0} > c) + \gamma \cdot P_0(\frac{L_1}{L_0} = c)$. A terjedelmet úgy „szedjük össze," hogy a legnagyobb likelihood-hányadostól, 25/9-től indulunk. Látszik, hogy $c \in (\frac{5}{9}, \frac{25}{9})$ jó lesz, és véletlenítésre nincs szükség ($\gamma = 0$), tehát II esetén H_0-t elutasítjuk, minden más esetben H_0-t elfogadjuk.

Ha most $\alpha = 0.3$, akkor tovább kell menni. Ha $\frac{5}{9} < c < \frac{25}{9}$, akkor az elsőfajú hiba $0.25 < 0.3$ ha viszont $\frac{1}{9} < c < \frac{5}{9}$, akkor az elsőfajú hiba $0.25 + 0.5 > 0.3$. Tehát $c = \frac{5}{9}$, és a megoldandó egyenlet: $0.25 + \gamma \cdot 0.5 = 0.3$, amiből $\gamma = 0.1$. Tehát II esetén H_0-t elutasítjuk, FF esetén H_0-t elfogadjuk, IF vagy FI esetén véletlenítünk: H_0-t 0.9 valószínűséggel elfogadjuk, 0.1 valószínűséggel elvetjük. Mekkora a próba ereje?

$$\beta = P_1\left(\frac{L_1}{L_0} > c\right) + \gamma \cdot P_1\left(\frac{L_1}{L_0} = c\right) = \frac{25}{36} + 0.1 \cdot \frac{10}{36} = \frac{26}{36} = 0.72.$$

■

6.3. Példa. Egy urnában 7 db golyóból k piros, $7-k$ zöld. $H_0: k=3$, $H_1: k=4$. Három golyót kihúzunk visszatevés nélkül, jelölje X a pirosak számát a kihúzottak között ($\mathcal{X}=\{0,1,2,3\}$). Adjuk meg az $\alpha=0.2$ terjedelmű likelihood-hányados próbát!

$$L_0(x)=\frac{\binom{3}{x}\binom{4}{3-x}}{\binom{7}{3}},\quad L_1(x)=\frac{\binom{4}{x}\binom{3}{3-x}}{\binom{7}{3}}.$$

Ezért a táblázat:

	0	1	2	3
L_1	$\frac{1}{35}$	$\frac{12}{35}$	$\frac{18}{35}$	$\frac{4}{35}$
L_0	$\frac{4}{35}$	$\frac{18}{35}$	$\frac{12}{35}$	$\frac{1}{35}$
$\frac{L_1}{L_0}$	$\frac{1}{4}$	$\frac{2}{3}$	$\frac{3}{2}$	4

Mivel $1/35<0.2$, de $1/35+12/35>0.2$, így $c=\frac{3}{2}$. Továbbá a $0.2=\frac{1}{35}+\gamma\cdot\frac{12}{35}$ egyenletből $\gamma=0.5$. Tehát ha legfeljebb egy piros golyót húzunk, akkor elfogadjuk a nullhipotézist, ha három piros golyót húzunk, akkor elvetjük a nullhipotézist, ha pedig két piros golyót húzunk, akkor véletlenítünk: 1/2 valószínűséggel elvetjük, 1/2 valószínűséggel elfogadjuk a nullhipotézist.
A próba másodfajú hibája:

$$1-\beta=P_1(H_0\text{-t elfogadjuk})=\frac{1}{35}+\frac{12}{35}+\frac{1}{2}\cdot\frac{18}{35}=\frac{22}{35}.$$

■

6.4. Példa. Legyen $X\sim\mathrm{Exp}(\lambda)$ egy villanykörte élettartama (években kifejezve). $H_0:\lambda=\frac{1}{2}$, $H_1:\lambda=\frac{1}{3}$. Adjuk meg az $\alpha=1/8$ terjedelmű likelihood-hányados próbát!
Most $\mathcal{X}=(0,\infty)$. A sűrűségfüggvények és hányadosuk:

$$L_0(x)=\frac{1}{2}e^{-\frac{1}{2}x},\quad L_1(x)=\frac{1}{3}e^{-\frac{1}{3}x},\quad \frac{L_1(x)}{L_0(x)}=\frac{2}{3}e^{\frac{x}{2}-\frac{x}{3}}=\frac{2}{3}e^{\frac{x}{6}}.$$

Mivel a minta folytonos eloszlású, véletlenítésre nem lesz szükség. Kell tehát:

$$\alpha=\frac{1}{8}=P_0\left(\frac{2}{3}e^{\frac{X}{6}}>c\right)=P_0\left(X>6\ln\frac{3c}{2}\right)=1-F_0\left(6\ln\frac{3c}{2}\right)=e^{-3\ln\frac{3c}{2}}=\left(\frac{2}{3c}\right)^3.$$

Ebből pedig $c=\frac{4}{3}$. Megkaptuk tehát c-t, de igazából nem ez a fontos, hanem az, hogy mikor utasítjuk el a nullhipotézist. A számítás másképp:

$$\alpha=\frac{1}{8}=P_0\left(\frac{2}{3}e^{\frac{X}{6}}>c\right)=P_0\left(X>d\right)=1-F_0\left(d\right)=e^{-\frac{d}{2}},$$

amiből $d=6\ln 2=4.16$. Tehát akkor utasítjuk el a nullhipotézist, ha a villanykörte tovább él, mint 4.16 év.
Mj.: a kapott próba egyenletesen legerősebb a $H_0:\lambda=\frac{1}{2}$, $H_1:\lambda<\frac{1}{2}$ hipotézisvizsgálati feladatra, mivel minden $H_1':\lambda=\lambda_1(<\frac{1}{2})$ egyszerű ellenhipotézisre ugyanez a próba a legerősebb. ■

6.1. Feladat. Legyen öt elemű mintánk λ paraméterű Poisson eloszlásból. $H_0:\lambda=2$, $H_1:\lambda=1$. Adjuk meg az $\alpha=0.05$ terjedelmű likelihood-hányados próbát!

6.2. Feladat. Legyen n elemű mintánk az $N(m,1)$ normális eloszlásból. $H_0:m=0$, $H_1:m=1$. Adjuk meg az $\alpha=0.05$ terjedelmű likelihood-hányados próbát!

7. A normális eloszlás paramétereire vonatkozó próbák

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

1. A várható értékre vonatkozó próba, ha a szórás ismert $\rightarrow$ u-próba.

2. A várható értékre vonatkozó próba, ha a szórás ismeretlen $\rightarrow$ t-próba.

3. A szórásra vonatkozó próba, ha a várható érték ismeretlen (vagy akár ismert) $\rightarrow$ F-próba.

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

Egymintás u-próba

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

$$a)\ \begin{matrix} H_0: & m = m_0 \\ H_1: & m \neq m_0 \end{matrix} \qquad b)\ \begin{matrix} H_0: & m \leq m_0 \\ H_1: & m > m_0 \end{matrix} \qquad c)\ \begin{matrix} H_0: & m \geq m_0 \\ H_1: & m < m_0 \end{matrix} \tag{2}$$

A próbastatisztika:

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

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

$$a)\, \mathcal{X}_k = \{|u| > u_{\frac{\alpha}{2}}\} \quad b)\, \mathcal{X}_k = \{u > u_\alpha\} \quad c)\, \mathcal{X}_k = \{u < -u_\alpha\} \tag{3}$$

ahol az u_δ kritikus érték olyan, hogy $\Phi(u_\delta) = 1 - \delta$, ezt a $\Phi(x)$ függvény táblázatából keressük ki.

Mj.: (2)-ben az a) esetben kétoldali, a b) és c) esetekben egyoldali ellenhipotézisről beszélünk.
Hogy néz ki a próba erőfüggvénye (kétoldali ellenhipotézisre)? Vezessük be a $\Delta_n(m) := \frac{m - m_0}{\sigma}\sqrt{n}$ jelölést.

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

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

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

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

7.1. Példa. Legyen $X_1, ..., X_{16} \sim N(m, 1)$ minta, melyre $\overline{X} = 0,1$. Hipotéziseink: $H_0: m = 0$, $H_1: m \neq 0$. $\alpha = 0,1$ terjedelem mellett szeretnénk dönteni.
Egymintás u-próbát kell végezni, a kritikus érték:

$$0.95 = 1 - \frac{\alpha}{2} = \Phi(u_{\frac{\alpha}{2}}) \overset{\text{tábl.}}{\Rightarrow} u_{\frac{\alpha}{2}} = 1.65.$$

A próbastatisztika: $u = \frac{0,1-0}{1}\sqrt{16} = \underset{\sim}{0.4}$. Mivel $|0.4| \leq 1.65$, elfogadjuk H_0-t.
Keressük most meg a legnagyobb terjedelmet, ami mellett még elfogadjuk H_0-t!

$$u_{\frac{\alpha}{2}} = 0.4 \ \Rightarrow\ \Phi(u_{\frac{\alpha}{2}}) = 0.66 = 1 - \frac{\alpha}{2} \ \Rightarrow\ \frac{\alpha}{2} = 0.34 \ \Rightarrow\ \underset{\sim}{\underset{\sim}{\alpha = 0.68}}.$$

■

Kétmintás u-próba

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

$$a)\ \begin{matrix} H_0: & m_1 = m_2 \\ H_1: & m_1 \neq m_2 \end{matrix} \qquad b)\ \begin{matrix} H_0: & m_1 \leq m_2 \\ H_1: & m_1 > m_2 \end{matrix} \qquad c)\ \begin{matrix} H_0: & m_1 \geq m_2 \\ H_1: & m_1 < m_2 \end{matrix} \tag{4}$$

A próbastatisztika:

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

Tehát a kritikus tartományok ugyanazok, mint egymintás esetben, azaz (3) adja meg őket.

Egymintás t-próba

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

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

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

$$a)\, \mathcal{X}_k = \{|t| > t_f(\tfrac{\alpha}{2})\} \quad b)\, \mathcal{X}_k = \{t > t_f(\alpha)\} \quad c)\, \mathcal{X}_k = \{t < -t_f(\alpha)\} \tag{5}$$

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

Kétmintás t-próba

Legyenek $X_1, \ldots, X_{n_1} \sim N(m_1, \sigma)$ és $Y_1, \ldots, Y_{n_2} \sim N(m_2, \sigma)$ független minták, ahol m_1, m_2 és σ ismeretlenek.
A hipotézisek: ugyanazok, mint a kétmintás u-próbánál (4).
A próbastatisztika:

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

Jelölje a szabadsági fokot f, tehát $f = n_1 + n_2 - 2$.
A kritikus tartomány ugyanaz, mint az egymintás esetben (5).

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

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

Másrészt teljesül, hogy

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

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

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

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

ahol az f szabadsági fok

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

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

A szórásra vonatkozó próbához szükségünk lesz az F eloszlásra.

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

<u>Kétmintás F-próba</u>

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

A hipotézisek:

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

A próbastatisztika:

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

Jelölje a szbadsági fokokat $f_1 = n_1 - 1$ és $f_2 = n_2 - 1$.

A kritikus tartomány:

$$a)\mathcal{X}_k = \{F < F_{f_1,f_2}(1-\tfrac{\alpha}{2}) \text{ vagy } F > F_{f_1,f_2}(\tfrac{\alpha}{2})\} \quad b)\mathcal{X}_k = \{F > F_{f_1,f_2}(\alpha)\} \quad c)\mathcal{X}_k = \{F < F_{f_1,f_2}(1-\alpha)\}, \tag{6}$$

ahol az $F_{f_1,f_2}(\delta)$ kritikus érték olyan, hogy $G(F_{f_1,f_2}(\delta)) = 1-\delta$, ha G jelöli az (f_1, f_2) szabadsági fokú F-eloszlás eloszlásfüggvényét.

A kritikus értékeket az „F-próba kritikus értékei” című táblázatból keressük ki.

A próbastatisztika H_0 melletti eloszlása:

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

A próba praktikusabb formája kétoldali ellenhipotézisre:

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

Így a kritikus tartomány ekvivalens alakja:

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

A gyakorlatban használt α-kra a kritikus érték 1-nél nagyobb, ezért elég F és $1/F$ közül a nagyobbikat összehasonlítani a megfelelő kritikus értékkel.
A próba tehát: az $F' = \max\left\{\frac{{S^*_{n_1}}^2}{{S^*_{n_2}}^2}, \frac{{S^*_{n_2}}^2}{{S^*_{n_1}}^2}\right\}$ statisztikát hasonlítjuk össze vagy az $F_{n_1-1,n_2-1}(\alpha/2)$ vagy az $F_{n_2-1,n_1-1}(\alpha/2)$ kritikus értékekkel.

Egymintás F-próba (vagy χ^2-próba)

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

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

A próbastatisztika:

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

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

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

ahol a $\chi^2_f(\delta)$ kritikus érték olyan, hogy $G(\chi^2_f(\delta)) = 1-\delta$, ha G jelöli az f szabadsági fokú χ^2-eloszlás eloszlásfüggvényét.
A kritikus értékeket a „χ^2-próba kritikus értékei” című táblázatból keressük ki.
Ehelyett végezhetünk az

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

statisztikára F-próbát, azaz ekkor a kritikus tartományt (6) adja meg (az $F_{n-1,\infty}$ eloszlás a χ^2_{n-1} eloszlás átskálázott változata).

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

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

Vajon van-e szignifikáns különbség a két gyógyszer hatásossága között ($\alpha = 0.01$ terjedelem mellett)?
Hipotézisek: $H_0: m_1 = m_2$, $H_1: m_1 \neq m_2$. A két minta azonban nem független, mert ugyanazokon a betegeken próbálták ki mind a két gyógyszert. Vegyük ezért az $A-B$ különbséget, és teszteljük egymintás t-próbával a $H_0: m = 0$, $H_1: m \neq 0$ hipotéziseket!
$S^*_n = 1.23$, $\overline{X} = 1.58$, $t = \frac{\overline{X}-m_0}{S^*_n}\sqrt{n} = 4.06$. A kritikus érték: $t_9(0.005) = 3.35$, így, mivel $|4.06| > 3.35$, a nullhipotézist elvetjük, azaz a két gyógyszer hatásossága között szignifikáns különbség van.
Tegyük most fel, hogy a két gyógyszert más-más 10 betegen tesztelték (de továbbra is a fenti táblázat adatait használjuk). Ekkor a két minta független, kétmintás t-próba végezhető. $\overline{X} = 2.33$, $\overline{Y} = 0.75$, ${S^*_A}^2 = 4$, ${S^*_B}^2 = 3.8$. A kétféle gyógyszer hatásának szórásai feltételezhetően egyenlőek, de ellenőrizhetjük

is F-próbával: $F' = 1.1$, míg a kritikus érték $F_{9,9}(0.025) = 4.03$. A t-próba próbastatisztikája:

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(10-1)S_A^{*\,2} + (10-1)S_B^{*\,2}}} \cdot \sqrt{\frac{10 \cdot 10 \cdot (10+10-2)}{10+10}} = 1.78.$$

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

8. χ^2 próbák

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

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

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

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

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

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

$\underline{\chi^2\text{-próba (tiszta eset)}}$

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

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

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

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

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

$$\mathcal{X}_k = \{\chi^2 > \chi^2_f(\alpha)\}, \tag{7}$$

ahol a $\chi^2_f(\delta)$ kritikus érték olyan, hogy $G(\chi^2_f(\delta)) = 1 - \delta$, ha G jelöli az f szabadsági fokú χ^2-eloszlás eloszlásfüggvényét.
A kritikus értékeket a „χ^2-próba kritikus értékei" című táblázatból keressük ki.

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

χ^2-próba (becsléses eset)

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

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

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

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

Jelölje a szabadsági fokot $f = r - s - 1$.
A kritikus tartomány: (7), azaz ugyanaz, mint az előbb.

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

8.1. Illeszkedésvizsgálat

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

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

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

A próbastatisztika:

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

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

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

	1	2	3	4
ν_i	324	57	14	5

A minta átlaga: $\overline{X} = \frac{1 \cdot 324 + 2 \cdot 57 + 3 \cdot 14 + 4 \cdot 5}{400} = \frac{500}{400} = \frac{5}{4}$, ebből a paraméter ML becslése $\hat{p} = \frac{1}{\overline{X}} = 0.8$. A geometriai eloszlás pozitív egész értékeket vehet fel, tehát a lehetséges értékek egy csoportosítása: $A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}, A_4 = \{4, \ldots\}$. A négy csoport becsült valószínűsége:

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

A várt értékek táblázata tehát:

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

A próbastatisztika:

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

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

Ha a feltételezett eloszlás folytonos, akkor a teljes számegyenest kell intervallumokra, illetve félegyenesekre bontani. Az intervallumok megválasztásához néhány jó tanács:
1)Az intervallumok száma ne legyen se túl kevés (nem elég erős a próba, a mintában lévő információ nagy része elveszik), se túl sok (a χ^2 közelítés sérül).
2) Az osztópontokat úgy válasszuk, hogy az intervallumok p_i valószínűségei közel egyformák legyenek.

Összefoglalva, az illeszkedésvizsgálat lépései:
1) Ha becsléses esettel van dolgunk, akkor a paramétereket megbecsüljük a mintából ML módszerrel.
2) A lehetséges értékkészletet véges sok csoportba osztjuk (TER-t hozunk létre).
3) Kiszámoljuk a csoportok várt gyakoriságát.
4) Ha vannak kicsi (pl. 5-nél kisebb) várt gyakoriságok, akkor összevonunk csoportokat.
5) Kiszámoljuk a próbastatisztikát, kikeressük a szabadsági foknak és a terjedelemnek megfelelő kritikus értéket.
6) Összehasonlítva a kettőt, levonjuk a következtetést.

8.2. Függetlenségvizsgálat

Két szempont szerint soroljuk osztályokba a megfigyeléseket:
az 1. szempont szerint r osztály van: $A_1, ..., A_r$,
a 2. szempont szerint s osztály van: $B_1, ..., B_s$.
Nullhipotézisünk: H_0 : a két szempont független egymástól, azaz $P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$ minden i, j-re. Az ellenhipotézis pedig az, hogy a két szempont összefügg.

n darab független megfigyelésből jelölje ν_{ij} az $A_i \cap B_j$ esemény gyakoriságát, valamint legyen $\nu_{i\bullet} = \sum_{j=1}^{s} \nu_{ij}$ az A_i gyakorisága és $\nu_{\bullet j} = \sum_{i=1}^{r} \nu_{ij}$ a B_j gyakorisága.

Két eset lehetséges aszerint, hogy az egyes szempontok osztályainak valószínűségét ismerjük-e.

A) eset (ritkább): $P(A_i) = p_i$ és $P(B_j) = q_j$ ismertek. Mivel az $\{A_i \cap B_j\}$ események $r \cdot s$ elemű teljes eseményrendszert alkotnak,

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

B) eset: p_i, q_j nem ismertek. Ekkor először megbecsüljük őket a mintából ML-módszerrel:

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

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

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

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

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

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

Legyen ugyanis $\hat{\nu}_{ij} = \frac{\nu_{i\bullet}\nu_{\bullet j}}{n}$. Ekkor

$$\nu_{11} + \nu_{12} = \nu_{1\bullet} \text{ és } \hat{\nu}_{11} + \hat{\nu}_{12} = \frac{\nu_{1\bullet}\nu_{\bullet 1}}{n} + \frac{\nu_{1\bullet}\nu_{\bullet 2}}{n} = \frac{\nu_{1\bullet}}{n}(\nu_{\bullet 1} + \nu_{\bullet 2}) = \nu_{1\bullet}.$$

Kaptuk, hogy $\nu_{12} - \hat{\nu}_{12} = -(\nu_{11} - \hat{\nu}_{11})$. Hasonlóan, $\nu_{21} - \hat{\nu}_{21} = -(\nu_{11} - \hat{\nu}_{11})$ és $\nu_{22} - \hat{\nu}_{22} = \nu_{11} - \hat{\nu}_{11}$. Tehát a χ^2 statisztika mind a négy tagjában ugyanaz a számláló, méghozzá (négyzetreemelés előtt):

$$\nu_{11} - \hat{\nu}_{11} = \nu_{11} - \frac{1}{n}(\nu_{11} + \nu_{12})(\nu_{11} + \nu_{21}) = \frac{\nu_{11}\nu_{22} - \nu_{12}\nu_{21}}{n}.$$

A közös számláló és n kiemelése után marad:

$$\frac{1}{\nu_{1\bullet}\nu_{\bullet 1}} + \frac{1}{\nu_{1\bullet}\nu_{\bullet 2}} + \frac{1}{\nu_{2\bullet}\nu_{\bullet 1}} + \frac{1}{\nu_{2\bullet}\nu_{\bullet 2}} = \frac{n^2}{\nu_{1\bullet}\nu_{\bullet 1}\nu_{2\bullet}\nu_{\bullet 2}}.$$

A fentieket összerakva, épp a bizonyítani kívánt képletet kapjuk.

Mj.: Gyököt vonva, $\frac{\sqrt{n}\cdot(\nu_{11}\nu_{22} - \nu_{12}\nu_{21})}{\sqrt{\nu_{1\bullet}\nu_{2\bullet}\nu_{\bullet 1}\nu_{\bullet 2}}} \overset{H_0}{\approx} N(0,1)$, így u-próba is végezhető. Akár egyoldali u-próbát is végezhetünk, ha az ellenhipotézis adott irányú összefüggésre vonatkozik. (Pl.: H_0: a szemszín és a hajszín független, H_1: a szőkek nagyobb eséllyel kékszeműek.)

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

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

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

haj/szem	kék	más
szőke	25	25
más	75	75

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

b) Mondhatjuk-e, hogy a szőkék között több a kékszemű ($\alpha = 0.05$)?
A próbastatisztika: $\frac{\sqrt{200}\cdot(30\cdot 80 - 70\cdot 20)}{\sqrt{100\cdot 100\cdot 50\cdot 150}} = 1.63$.
A kritikus érték: $u_{0.05} = 1.65$.
Tehát a szőkék között nem szignifikánsan több a kékszemű (de majdnem :).
■

8.3. Homogenitásvizsgálat

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

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

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

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

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

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

A próbastatisztika:

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

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

9. Általánosított likelihood-hányados próba

Tekintsük a $\Theta = \Theta_0 \cup \Theta_1$ hipotézisvizsgálati feladatot. Egy n elemű X mintából számolt *általánosított likelihood-hányados statisztika* alatt a következő mennyiséget értjük:

$$\frac{\sup\{L_n(X;\vartheta) : \vartheta \in \Theta\}}{\sup\{L_n(X;\vartheta) : \vartheta \in \Theta_0\}}.$$

Ha a nullhipotézis teljesül, akkor ez közel van 1-hez, míg ha nem, akkor a számláló jóval nagyobb a nevezőnél, ezért a hányados nagy. A nullhipotézis tesztelése tehát úgy történhet, hogy ha az általánosított likelihood-hányados statisztika nagyobb, mint valamely kritikus érték, akkor H_0-t elvetjük, ha kisebb, akkor H_0-t elfogadjuk, egyenlőség esetén pedig, ha szükséges, randomizálunk.

Ha mind a nullhipotézis, mind az ellenhipotézis egyszerű, azaz a hipotézisünk szerint a sűrűségfüggvény f_0, az ellenhipotézis szerint pedig f_1, akkor a statisztika a következő alakot ölti:

$$\frac{\max\{f_0(X), f_1(X)\}}{f_0(X)} = \max\Big\{1,\ \frac{f_1(X)}{f_0(X)}\Big\}.$$

Itt a második tag a Neyman–Pearson lemmából ismerős likelihood-hányados, vagyis a klasszikus likelihood-hányados próbát kapjuk, ha a kritikus érték nagyobb 1-nél.

Az általánosított likelihood-hányados statisztika H_0 melletti eloszlását – néhány kivételes esettől eltekintve – nehéz meghatározni. Bizonyos feltételek mellett azonban határeloszlástételt tudunk bizonyítani rá, amint a minta nagysága minden határon túl nő, és ezzel aszimptotikus próbát konstruálhatunk.

Az aszimptotikus vizsgálathoz a fenti statisztika logaritmusának a kétszeresét tekintjük:

$$T = 2\big[\sup_{\vartheta\in\Theta} \log L_n(X;\vartheta) - \sup_{\vartheta\in\Theta_0} \log L_n(X;\vartheta)\big].$$

Tegyük fel, hogy a paraméterek két csoportra vannak felosztva, az első csoportba tartozók vektorát jelölje $\sigma \in \mathbb{R}^q$, a többiek $(p-q)$-dimenziós vektorát pedig τ. Tekintsük a következő hipotézisvizsgálati feladatot:

$$H_0 : \sigma = \sigma_0,\ \tau \text{ tetszőleges}, \quad H_1 : \sigma \neq \sigma_0,\ \tau \text{ tetszőleges}.$$

Megmutatható, hogy bizonyos regularitási feltételek mellett $n \to \infty$ esetén a T statisztika határeloszlása χ^2_q. Így khi-négyzet próba végezhető. (Illetve a σ paraméterre – amennyiben egy dimenziós – konfidencia-intervallum adható.)

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

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

10.1. Illeszkedésvizsgálat

$X_1, \ldots, X_n$ valamilyen folytonos eloszlásból vett minta (a mintaelemek eloszlásfüggvényét jelölje F). $H_0 : F = F_0$, $H_1 : F \neq F_0$.

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

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

Megmutatható, hogy D_n H_0 melletti eloszlása nem függ az F_0 eloszlástól: mivel $X_i^{(n)}$ és $X_{i+1}^{(n)}$ között a tapasztalati eloszlásfüggvény konstans $\frac{i}{n}$, az F_0 eloszlásfüggvény pedig monoton nő, kapjuk, hogy

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

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

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

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

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

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

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

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

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

10.2. Függetlenségvizsgálat

Legyen (X_i, Y_i) $(i = 1, \ldots, n)$ egy folytonos $H(x,y)$ eloszlásfüggvényű eloszlásból származó minta. Jelölje a H-hoz tartozó marginális eloszlásokat $F(x) = \lim_{w \to \infty} H(x,w)$ és $G(y) = \lim_{z \to \infty} H(z,y)$.
H_0 : a két koordináta független, azaz $H(x,y) = F(x)G(y)\ \forall x, y$.

Kendall próba:
Tekintsük a következő próbastatisztikát:

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

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

Megmutatható az is, hogy K_n aszimptotikusan normális eloszlású, azaz ha n elég nagy, akkor u-próba végezhető, ha pedig n kicsi, akkor külön táblázat tartalmazza a kritikus értékeket. Az u-próbához

standardizálni kell K_n-et:
a) K_n várható értéke H_0 mellett:

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

Hasonlóan, $E_0(I((X_i-X_j)(Y_i-Y_j)<0)) = 0.5$, azaz $E_0(K_n)=0$.
b) K_n szórásnégyzete H_0 mellett:
legyen $\delta_{ij} = [I((X_i-X_j)(Y_i-Y_j)\geq 0) - I((X_i-X_j)(Y_i-Y_j)<0)]$.

$$D_0^2(K_n) = \operatorname{cov}_0\left(\sum_{i<j}\delta_{ij}, \sum_{k<l}\delta_{kl}\right) = \sum_{i<j}\sum_{k<l}\operatorname{cov}_0(\delta_{ij},\delta_{kl}).$$

Itt $\operatorname{cov}_0(\delta_{ij},\delta_{kl}) = \begin{cases} 0 & \text{ha} \quad \{i,j\}\cap\{k,l\}=\emptyset \\ 1 & \text{ha} \quad i=k, j=l \\ 1/9 & \text{egyébként} \end{cases}$

Azaz $D_0^2(K_n) = 1\cdot\dfrac{n(n-1)}{2} + \dfrac{1}{9}\cdot\dfrac{n(n-1)(n-2)}{6}\cdot 6 = \dfrac{n(n-1)(2n+5)}{18}$.

Mj.: a próba csak a $\tau\neq 0$ típusú ellenhipotézisekre konzisztens, ahol

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

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

Blum-Kiefer-Rosenblatt próba:
Legyen

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

$$\begin{array}{c|c} N_3(i) & N_4(i) \\ \hline & (X_i,Y_i) \\ \hline N_1(i) & N_2(i) \end{array}$$

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

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

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

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

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

10.3. Homogenitásvizsgálat

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

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

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

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

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

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

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

$$D_0^2(W_{n,m}) = \operatorname{cov}_0\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\delta_{ij}, \sum_{l=1}^{n}\sum_{k=1}^{m}\delta_{lk}\right) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{n}\sum_{k=1}^{m}\operatorname{cov}_0(\delta_{ij}, \delta_{lk}).$$

$$\text{Itt } \operatorname{cov}_0(\delta_{ij}, \delta_{lk}) = \begin{cases} 0 & \text{ha} \quad i \neq l, j \neq k \\ 1/4 & \text{ha} \quad i = l, j = k \\ 1/12 & \text{ha} \quad i = l, j \neq k \\ 1/12 & \text{ha} \quad i \neq l, j = k \end{cases}$$

Azaz $D_0^2(W_{n,m}) = nm\frac{1}{4} + nm(m-1)\frac{1}{12} + mn(n-1)\frac{1}{12} = \frac{n \cdot m(n+m+1)}{12} = \frac{n \cdot m}{4} \cdot \frac{n+m+1}{3}$.

$\frac{n \cdot m}{4}$ lenne, ha δ_{ij}-k függetlenek lennének. Mj.: a próba csak a $P(X > Y) \neq 1/2$ típusú ellenhipotézis esetén konzisztens.

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

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

Megmutatható, hogy $D_{n,m}$ H_0 melletti eloszlása nem függ a két minta közös eloszlásától, és $\sqrt{\frac{m \cdot n}{m+n}} \cdot D_{n,m}$ aszimptotikusan Kolmogorov eloszlású. Azaz ha n, m elég nagyok, akkor a Kolmogorov eloszlásból számolhatunk kritikus értéket , ha pedig n, m kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

Mj.: Vizsgálhatjuk a $H_1 : \ F(x) > G(x)\ \forall x$ egyoldali ellenhipotézist is, ekkor a próbastatisztika

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

Erre is igaz, hogy H_0 melletti eloszlása nem függ a két minta közös eloszlásától. Ebben az esetben $\sqrt{\frac{m \cdot n}{m+n}} \cdot D_{n,m}^{+}$ aszimptotikusan Szmirnov eloszlású. Azaz ha n, m elég nagyok, akkor a Szmirnov eloszlásból számolhatunk kritikus értéket, ha pedig n, m kicsi, akkor külön táblázat tartalmazza a kritikus értékeket.

11. Lineáris regresszió, lineáris modellek

Legyen (X, Y) kétdimenziós valváltozó.
Kérdés: X milyen lineáris függvénye közelíti legjobban Y-t?

Pl.: 100 kockadobásból a párosak számát szeretnénk a hatosok számának lineáris függvényével

közelíteni.
Elnevezés: X a magyarázó változó.

A közelítés hibáját mérje a $h^2 = E[(Y - (aX + b))^2]$ mennyiség, ezt szeretnénk minimalizálni.
Megoldás:

$$h^2 = E(X^2)a^2 + b^2 + 2E(X)ab - 2E(XY)a - 2E(Y)b + E(Y^2)$$

A kifejezést a szerint deriválva, akkor lesz 0, ha

$$a = \frac{E(XY) - bE(X)}{E(X^2)}.$$

A kifejezést b szerint deriválva, akkor lesz 0, ha

$$b = E(Y) - aE(X).$$

A minimumot tehát a következő paraméterek adják:

$$a^* = \frac{cov(X,Y)}{D^2(X)}, \quad b^* = E(Y) - a^*E(X).$$

A h^2 hibát Y szórásnégyzetével osztva megkapjuk, hogy a regresszió a szórásnégyzet hány százalékát magyarázza meg.
Ugyanezt a feladatot tekinthetjük több magyarázó változó esetén is: legyen $(X_1, \ldots, X_q, Y)$ valváltozó, és jelölje $X = (X_1, \ldots, X_q)$ a magyarázó változókat. Minimalizálni szeretnénk a $h^2 = E[(Y - (a^T X + b))^2]$ mennyiséget, ahol $a^T = (a_1, \ldots, a_q)$ vektor.
A feladat az egy-dimenzióshoz hasonlóan oldható meg, és

$$a^* = cov(Y,X)\Sigma^{-1}(X), \quad b^* = E(Y) - a^*E(X).$$

Itt most $cov(Y,X) = (cov(Y,X_1), \ldots, cov(Y,X_q))$ sorvektor, $\Sigma(X)$ pedig $q \times q$-as mátrix, melynek (i,j)-dik eleme $cov(X_i, X_j)$. (Tegyük fel, hogy a mátrix invertálható.)

Ez utóbbi felállás arra az esetre is alkalmazható, amikor egy magyarázó változó van, de annak polinomjával szeretnénk Y-t közelíteni.

Pl. Egy teve legel Erzsi néni hátsó kertjében.
Ebből a mondatból véletlenszerűen választva egy szót, legyen X az „e" betűk száma, Y pedig a szó hossza. A legjobb lineáris közelítés Y-ra: $0.82X + 4.06$. A legjobb másodfokú közelítés Y-ra: $1.02X^2 - 1.25X + 4.38$. Feladat: számoljuk ki a két közelítés hibáját!

Az előző statisztikai megfelelője: $(X_{i1}, \ldots, X_{iq}, Y_i)$, $i = 1, \ldots, n$ független, azonos eloszlású minta. Most a $h^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (a^T X_i + b))^2$ hibát szeretnénk minimalizálni. Ha a minta tapasztalati eloszlását tekintjük, akkor az átlag a várható értéknek felel meg, azaz a megoldás ugyanaz, mint az előbb, csak a megfelelő tapasztalati mennyiségeket kell beírni.

A fentiekben teljes általánosságban megkerestük a legjobb lineáris közelítést.

Lineáris modell: $Y_i = a^T X_i + b + \epsilon_i$, ahol X_i ugyanaz, mint fent, de most ismertnek feltételezzük, az Y_i megfigyelések pedig X_i lineáris függvényei, hibával terhelten. Itt ϵ_i független $N(0, \sigma^2)$ hibák. Tehát Y_i $N(a^T X_i + b, \sigma^2)$ eloszlású, és függetlenek.
Példa: egy ember súlya a neme, életkora, magassága valamilyen lineáris függvénye, plusz még az egyéni ingadozásból származó hiba.

Áll: Ebben a modellben a és b ML becslése épp a fenti a^* és b^*, és a becslések torzítatlanok.

Áll: Ebben a modellben σ^2 ML becslése:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (a^T X_i + b))^2,$$

ez nem torzítatlan, de

$$\frac{1}{n-q-1}\sum_{i=1}^{n}(Y_i - (a^T X_i + b))^2$$

már az.