

... és statisztika

Tartalom:

- ▶ Statisztikai alapfogalmak: minta, realizáció, alapstatisztikák
- ▶ Empirikus eloszlásfüggvény és konvergenciája
- ▶ Becslések tulajdonságai
- ▶ Maximum likelihood becslés, momentumbecslés
- ▶ Konfidenciaintervallumok, χ^2 - és t -eloszlás
- ▶ Hipotézisvizsgálat, paraméteres próbák

Előismeretek: differenciál- és integrálszámítás főleg egy dimenzióban,

a valszám anyagrészből leginkább: nevezetes diszkrét és folytonos eloszlások, eloszlásfüggvény, sűrűségfüggvény, súlyfüggvény, függetlenség, momentumok (főleg: várható érték és szórás), CHT és változatai, nagy számok erős törvénye.

Jegyzetkiegészítés: cs.bme.hu/~tobias/statisztika.pdf.

Idén időbeosztási okokból a maximum likelihood becslés a becslések tulajdonságai előtt hangzik el.

Emlékeztető: statisztika

Sztochasztika = a véletlen törvényszerűségének tudománya.

A **sztochasztika** két fő ága:

- ▶ **Valószínűségelmélet**: matematikai modellezés és formalizmus, elméleti alapok
- ▶ **Statisztika**: véletlen eljárásokból származó mérési adatok elemzése és interpretációja valószínűségelméleti módszerek segítségével

Statisztika: áttekintés

Véletlen eljárásból származó mérési adatok → milyen eloszlásúak?

Gyakori szituáció: a mérési adatok eloszlása valamilyen ismeretlen paraméterrel/paraméterekkel rendelkezik. (Pl. tudjuk, hogy Poisson, de nem tudjuk, hogy mi λ értéke.)

A statisztika két fő ága:

- ▶ **Becslélmélet:** ismételt mintavétel segítségével próbáljuk meg az ismeretlen paramétert megbecsülni. A becslés lehet
 - ▶ **pontbecslés** → a becslésünk értéke egy valós szám (vagy n paraméter esetén egy $\mathbb{R}^n$ -beli vektor), a paraméter valamilyen értelemben „legjobb közelítése”.
 - ▶ **intervallumbecslés/konfidenciaintervallumok** → a becslésünk értéke egy intervallum, amibe az ismeretlen paraméter legalább egy bizonyos (előre megadott) valószínűséggel beleesik.
- ▶ **Hipotézisvizsgálat:** szintén ismételt mintavétel segítségével valamilyen állítást szeretnénk tesztelni → „van-e elég bizonyítékunk arra, hogy az állítást elfogadjuk/elutasítsuk”? → **Statisztikai próbák**, ebből a tárgyból: u -próba és t -próba. (Példák ismeretlen paraméterre, mintavételre, hipotézisre: 13. fejele)

Minta és realizáció

(13.1.1. Def.) Legyenek $X_1, \dots, X_n$ független, azonos eloszlású változók, amelyek peremeloszlása nem feltétlenül ismert. Ekkor az $\mathbf{X} = (X_1, \dots, X_n)$ val. vektorváltozót **n elemű független, azonos eloszlású mintának** (röviden **n elemű fae. mintának**) nevezzük.

Jelölések: Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása valamilyen ϑ paramétertől függ (azaz az eloszlásuk a ϑ ismeretében meghatározható), ahol a ϑ lehetséges értékeinek halmaza $\Theta \subseteq \mathbb{R}^d$ valamely $d \geq 1$ -re. Ekkor $\vartheta \in \Theta$ -ra,

- (1) abban az esetben, **ha ϑ a valódi paraméter**, egy A esemény valószínűségét $\mathbb{P}_{\vartheta}(A)$ -val jelöljük,
- (2) az X_1 (vagy X_2, X_3 stb.) valószínűségi változó ϑ paraméterhez tartozó eloszlásfüggvényét F_{ϑ} -val jelöljük, azaz $F_{\vartheta}(x) = \mathbb{P}_{\vartheta}(X_1 < x)$, $x \in \mathbb{R}$,
- (3) ha X_1 -nek ezen paraméter esetén létezik sűrűségfüggvénye, akkor ezt f_{ϑ} -val jelöljük. Vagyis:

$$f_{\vartheta}(x) = \begin{cases} F'_{\vartheta}(x), & \text{ha } F_{\vartheta} \text{ differenciálható } x\text{-ben,} \\ 0, & \text{különben,} \end{cases}$$

Minta és realizáció

(13.1.1. Def.) Legyenek $X_1, \dots, X_n$ független, azonos eloszlású változók, amelyek peremeloszlása nem feltétlenül ismert. Ekkor az $\mathbf{X} = (X_1, \dots, X_n)$ val. vektorváltozót **n elemű független, azonos eloszlású mintának** (röviden **n elemű fae. mintának**) nevezzük.

Jelölések: Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása valamilyen ϑ paramétertől függ (azaz az eloszlásuk a ϑ ismeretében meghatározható), ahol a ϑ lehetséges értékeinek halmaza $\Theta \subseteq \mathbb{R}^d$ valamely $d \geq 1$ -re. Ekkor $\vartheta \in \Theta$ -ra,

- (1) abban az esetben, **ha ϑ a valódi paraméter**, egy A esemény valószínűségét $\mathbb{P}_{\vartheta}(A)$ -val jelöljük,
- (2) az X_1 (vagy X_2, X_3 stb.) valószínűségi változó ϑ paraméterhez tartozó eloszlásfüggvényét F_{ϑ} -val jelöljük, azaz $F_{\vartheta}(x) = \mathbb{P}_{\vartheta}(X_1 < x)$, $x \in \mathbb{R}$,
- (3) ha X_1 -nek ezen paraméter esetén létezik sűrűségfüggvénye, akkor ezt f_{ϑ} -val jelöljük,
- (4) illetve ha X_1 -nek ezen paraméter esetén diszkrét az eloszlása, akkor p_{ϑ} -val jelöljük az ehhez az eloszláshoz tartozó súlyfüggvényt. Vagyis $x \in \mathbb{R}$ esetén $p_{\vartheta}(x) = \mathbb{P}_{\vartheta}(X = x)$.

Minta és realizáció

- ▶ (13.1.3. Példa: f_ϑ ill. p_ϑ az előbbi példákban)
- ▶ (13.1.4. Példa: az ismeretlen paraméterrel rendelkező eloszlású val. változó értékkészlete függhet a paramétértől)

Célunk a *realizáció* precíz definiálása $\rightarrow$ ehhez először az *adott paraméterhez tartozó lényeges értékek halmazát* definiáljuk.

(13.1.5. Def. + jelölés.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétértől függ. Legyen $\vartheta \in \Theta$ és $i = 1, \dots, n$. Ekkor ha létezik az f_ϑ sűrűségfüggvény, akkor az $S_{X_i}^{(\vartheta)}$ halmazt így definiáljuk:

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid f_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

Ha pedig létezik a p_ϑ súlyfüggvény, akkor definiáljuk az $S_{X_i}^{(\vartheta)}$ halmazt az alábbi képlettel:

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid p_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

$S_{X_i}^{(\vartheta)}$ -t mindkét esetben az X_i valószínűségi változó ϑ paraméterhez tartozó **lényeges értékei halmazának** nevezzük.

Minta és realizáció

- ▶ (13.1.3. Példa: f_ϑ ill. p_ϑ az előbbi példákban)
- ▶ (13.1.4. Példa: az ismeretlen paraméterrel rendelkező eloszlású val. változó értékkészlete függhet a paramétértől)

Célunk a *realizáció* precíz definiálása $\rightarrow$ ehhez először az **adott paraméterhez tartozó lényeges értékek halmazát** definiáljuk.

(13.1.5. Def. + jelölés.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétértől függ. Legyen $\vartheta \in \Theta$ és $i = 1, \dots, n$. Ekkor ha létezik az f_ϑ sűrűségfüggvény, akkor az $S_{X_i}^{(\vartheta)}$ halmazt így definiáljuk:

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid f_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

Ha pedig létezik a p_ϑ súlyfüggvény, akkor definiáljuk az $S_{X_i}^{(\vartheta)}$ halmazt az alábbi képlettel:

$$S_{X_i}^{(\vartheta)} = \{x \in \mathbb{R} \mid p_\vartheta(x) > 0\} \subseteq \mathbb{R}.$$

$S_{X_i}^{(\vartheta)}$ -t mindkét esetben az X_i valószínűségi változó ϑ paraméterhez tartozó **lényeges értékei halmazának** nevezzük.

Minta és realizáció

(13.1.6. Példa: lényeges értékek halmaza a korábbi példákban)

(13.1.7. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függ.

Az $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$ az $\mathbf{X} = (X_1, \dots, X_n)$ minta egy (lehetséges) **realizációja** a $\vartheta \in \Theta$ paraméter esetén, ha $x_i \in S_{X_i}^{(\vartheta)}$ teljesül minden $i \in \{1, \dots, n\}$ -re.

(13.1.8. Példa: realizációk az 13.1.6. Példa esetén)

Statisztikák

Egy n elemű fae. minta egy **statisztikájának** a mintaelemek egy olyan függvényét nevezzük, amely szimmetrikus, azaz „minden mintaelemről ugyanúgy függ”:

(13.2.1. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta. Ha $T: \mathbb{R}^n \rightarrow \mathbb{R}$ szimmetrikus függvény, azaz

$$T(x_1, \dots, x_n) = T(\pi(x_1), \dots, \pi(x_n))$$

minden $x_1, \dots, x_n \in \mathbb{R}$ és $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ permutáció esetén teljesül, akkor a $T(\mathbf{X}) = T(X_1, \dots, X_n)$ valószínűségi vektorváltozót az $X_1, \dots, X_n$ egy **statisztikájának** nevezzük.

Emlékeztető: $\pi: \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ permutáció, ha kölcsönösen egyértelmű függvény (azaz bijekció), lásd 1. diászor.

Megjegyzés: Tavaly a nagy Θ helyett kis θ -t használtam, de ez zavarba ejtő jelölésnek bizonyult, ezért lecseréltem.

Régi vizsgákban, feladatsorokon a θ felel meg a Θ -nak.

Alapstatisztikák: mintaátlag

Az első alapstatisztika mintaátlag, amit már a középiskolából is ismerünk:

Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta. Ekkor az $\overline{X}_n = \frac{X_1 + \dots + X_n}{n}$ **mintaátlag** az $\mathbf{X}$ egy statisztikája. Ha $\mathbf{x} = (x_1, \dots, x_n)$ az $\mathbf{X} = (X_1, \dots, X_n)$ egy realizációja, akkor realizáció átlagát $\overline{x}_n$ -nel fogjuk jelölni: $\overline{x}_n = \frac{x_1 + \dots + x_n}{n}$.

Az eddigiek alapján nem meglepő, hogy a mintaátlag a várható érték egyfajta becslése a minta alapján.

Az világos, hogy ha $\mathbb{E}(X_i)$ létezik (vagyis ha $\mathbb{E}[|X_i|] < \infty$), akkor a mintaátlag várható értéke megegyezik az egyes mintaelemek várható értékével:

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n}(\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)) = \mathbb{E}(X_1).$$

További tulajdonságai $\rightarrow$ lásd később.

Alapstatisztikák: korrigált empirikus szórásnégyszet

(13.2.3. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta.
Ekkor az

$$(S_n^*)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X_n})^2 \quad (1)$$

menyiséget az **$\mathbf{X}$ korrigált empirikus szórásnégyszetének** nevezzük, $S_n^* = \sqrt{(S_n^*)^2}$ -et pedig az **$\mathbf{X}$ korrigált empirikus szórásnak**.

- ▶ “empirikus” = “tapasztalati”.
- ▶ Világos: $(S_n^*)^2$ és S_n^* statisztikai $\mathbf{X}$ -nek.
- ▶ Miért $n-1$ -gyel osztunk (és nem pl. n -nel)? $\rightarrow$ Köv. állítás.

(13.2.4. Állítás.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, amelyre teljesül, hogy $\mathbb{E}(X_i^2) < \infty$. Ekkor

$$\mathbb{E}((S_n^*)^2) = \mathbb{D}^2(X_1).$$

(Bizonyítás)

(13.2.5. Megjegyzés: miért nevezzük korrigáltnak)

Rendezett minták, empirikus medián, módusz

(13.2.6. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta. Legyen $(X_1^*, X_2^*, \dots, X_n^*)$ az $X_1, \dots, X_n$ mintaelemek egy olyan felsorolása, hogy

$$X_1^* \leq X_2^* \leq \dots \leq X_n^*.$$

Ekkor $(X_1^*, X_2^*, \dots, X_n^*)$ -ot **rendezett mintának** nevezzük.

(13.2.7. Példa: nyolcszori kockadobás; ismétlődő elemek kezelése)

Vigyázat: $X_1^*, \dots, X_n^*$ általában nem függetlenek és nem is azonos eloszlásúak!

Az **empirikus medián** a rendezett minta középső eleme, ha páratlan elemszámú a minta, ha pedig páros elemszámú, akkor a két középső átlaga:

(13.2.8. Def.) Az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta **empirikus mediánján** az $m_{n,\mathbf{X}} = X_{k+1}^*$ rendezett mintaelemet értjük, ha $n = 2k + 1$, és $m_{n,\mathbf{X}} = \frac{X_k^* + X_{k+1}^*}{2}$ -et, ha $n = 2k$. (13.2.9. Példa)

Mi a (nem empirikus) *medián*? → Lásd hamarosan.

Módusz: a leggyakoribb érték a mintában. Ha több ilyen érték van, akkor mindegyiket módusznak tekintjük.

Rendezett minták, empirikus medián, módusz

(13.2.6. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta. Legyen $(X_1^*, X_2^*, \dots, X_n^*)$ az $X_1, \dots, X_n$ mintaelemek egy olyan felsorolása, hogy

$$X_1^* \leq X_2^* \leq \dots \leq X_n^*.$$

Ekkor $(X_1^*, X_2^*, \dots, X_n^*)$ -ot **rendezett mintának** nevezzük.

(13.2.7. Példa: nyolcszori kockadobás; ismétlődő elemek kezelése)

Vigyázat: $X_1^*, \dots, X_n^*$ általában nem függetlenek és nem is azonos eloszlásúak!

Az **empirikus medián** a rendezett minta középső eleme, ha páratlan elemszámú a minta, ha pedig páros elemszámú, akkor a két középső átlaga:

(13.2.8. Def.) Az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta **empirikus mediánján** az $m_{n,\mathbf{X}} = X_{k+1}^*$ rendezett mintaelemet értjük, ha $n = 2k + 1$, és $m_{n,\mathbf{X}} = \frac{X_k^* + X_{k+1}^*}{2}$ -et, ha $n = 2k$. (13.2.9. Példa)

Mi a (nem empirikus) *medián*? → Lásd hamarosan.

Módusz: a leggyakoribb érték a mintában. Ha több ilyen érték van, akkor mindegyiket módusznak tekintjük.

Empirikus eloszlásfüggvény

Emlékeztető: egy (egydimenziós) val. változó eloszlását az eloszlásfüggvénye egyértelműen meghatározza.

Tegyük fel, hogy nem tudjuk, hogy a fae. mérési adataink milyen típusú eloszláshoz tartoznak (azaz nemcsak egy paraméter, hanem a teljes eloszlás ismeretlen), viszont képesek vagyunk nagy elemszámú mintát venni.

Kérdés: Hogyan tudjuk a mintaelemek **eloszlásfüggvényét** a minta alapján közelíteni/becsülni? Erre szolgál az **empirikus eloszlásfüggvény**.

Empirikus eloszlásfüggvény

(13.3.1. Def.) Legyen $n \in \mathbb{N}$ és legyenek $X_1, \dots, X_n$ fae. valószínűségi változók. Az $\mathbb{R} \rightarrow \mathbb{R}$,

$$x \mapsto F_n^*(x) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i < x\}} = \frac{1}{n} |\{i \in \{1, \dots, n\} \mid X_i < x\}|$$

függvényt az $\underline{X} = (X_1, \dots, X_n)$ n elemű fae. mintához tartozó **empirikus eloszlásfüggvény**nek nevezzük.

Alternatív (és talán emészthetőbb) megfogalmazás az $X_1^*, \dots, X_n^*$ **rendezett minta** segítségével:

$$F_n^*(x) = \begin{cases} 0, & \text{ha } x \leq X_1^*, \\ \frac{k}{n}, & \text{ha } X_k^* < x \leq X_{k+1}^*, \quad k = 1, \dots, n-1, \\ 1, & \text{ha } x > X_n^*. \end{cases}$$

F_n^* egy *véletlen* függvény: a függvény értékei az $X_1, \dots, X_n$ valószínűségi változók értékeitől függenek.

(Össznépi gyakorlat: 12.4. feladat – az empirikus eloszlásfüggvény maga is egy (diszkrét) eloszlásfüggvény)

Az empirikus eloszlásfüggvény tulajdonságai

(13.3.2. Állítás.) Legyen $n \in \mathbb{N}$ és legyenek $X_1, \dots, X_n$ független, azonos eloszlású valószínűségi változók F eloszlásfüggvénnyel.

Ekkor minden $x \in \mathbb{R}$ esetén teljesül, hogy

- (1) $\mathbb{E}(F_n^*(x)) = F(x)$,
- (2) $\mathbb{D}^2(F_n^*(x)) = \frac{F(x)(1-F(x))}{n}$, és
- (3) $\mathbb{P}(\lim_{n \rightarrow \infty} F_n^*(x) = F(x)) = 1$.

Azaz az empirikus eloszlásfüggvény „jó becslése a valódi eloszlásfüggvénynek”: várható értéke a valódi eloszlásfüggvény, szórása 0-hoz tart, sőt az emp. eloszlásfüggvény 1 valószínűséggel pontonként a valódi eloszlásfüggvényhez tart.

Erről többet – más szavakkal – később.

(Állítás bizonyítása: (1) és (2) azért igaz, mert $nF_n^*(x) \sim B(F(x), 1/n)$; (3) a nagy számok erős törvénye alapján, lásd előadás/jegyzetkiegészítés)

(Példa: 12.3. feladat – lásd gyakorlat/mintamegoldás)

Glivenko–Cantelli-tétel (kiegészítő anyag)

Láttuk: $\mathbb{P}(\lim_{n \rightarrow \infty} F_n^*(x) = F(x)) = 1, \forall x \in \mathbb{R}$.

Ez még nem zárja ki azt, hogy $x \rightarrow \infty$ vagy $x \rightarrow -\infty$ esetén a konvergencia egyre lassabb legyen. (Ha ez így lenne, akkor az empirikus eloszlásfüggvényt nehezen lehetne a „valódi eloszlásfüggvény kirajzolására” használni.)

Azonban ez nincs így, mert az alábbi tétel szerint F_n^* 1 valószínűséggel az egész számegyenesen egyenletesen konvergál F -hez.

Glivenko–Cantelli-tétel: Legyen $n \in \mathbb{N}$ és legyenek $X_1, \dots, X_n$ független, azonos eloszlású valószínűségi változók F eloszlásfüggvénnyel. Ekkor az $x \mapsto F_n^*(x)$ függvény 1 valószínűséggel egyenletesen konvergál $x \mapsto F(x)$ -hez, azaz

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

(Megjegyzés: sűrűségfüggvény empirikus közelítése, empirikus sűrűséghisztogram)

Glivenko–Cantelli-tétel (kiegészítő anyag)

Láttuk: $\mathbb{P}(\lim_{n \rightarrow \infty} F_n^*(x) = F(x)) = 1, \forall x \in \mathbb{R}$.

Ez még nem zárja ki azt, hogy $x \rightarrow \infty$ vagy $x \rightarrow -\infty$ esetén a konvergencia egyre lassabb legyen. (Ha ez így lenne, akkor az empirikus eloszlásfüggvényt nehezen lehetne a „valódi eloszlásfüggvény kirajzolására” használni.)

Azonban ez nincs így, mert az alábbi tétel szerint F_n^* 1 valószínűséggel az egész számegyenesen egyenletesen konvergál F -hez.

Glivenko–Cantelli-tétel: Legyen $n \in \mathbb{N}$ és legyenek $X_1, \dots, X_n$ független, azonos eloszlású valószínűségi változók F eloszlásfüggvénnyel. Ekkor az $x \mapsto F_n^*(x)$ függvény 1 valószínűséggel egyenletesen konvergál $x \mapsto F(x)$ -hez, azaz

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

(Megjegyzés: sűrűségfüggvény empirikus közelítése, empirikus sűrűséghisztogram)

Becslésmélet

Szituáció: $X_1, \dots, X_n$ fae. mintaelemek, eloszlásuk egy $\vartheta \in \Theta$ paramétertől függ, és az ismeretlen ϑ paraméterre vagy annak valamilyen $\psi(\vartheta)$ függvényére a mintaelemek valamilyen statisztikája alapján minél jobb becslést akarunk adni.

Két fő változat:

- ▶ pontbecslések $\rightarrow$ a becslés egy konkrét valós szám/vektor lesz („a legjobb tippünk a paraméterre”)
- ▶ intervallumbecslések, konfidenciaintervallumok $\rightarrow$ a becslés egy intervallum lesz, amibe az ismeretlen paraméter legalább egy adott valószínűséggel (pl. 0,95, 0,99) beleesik.

Idő hiányában a pontbecslésekről szóló rész csak ízelítő lesz, egy-két módszert ismerünk csak meg részletesebben $\rightarrow$ bővebb ismeretekért lásd a SZIT MSc-s Matematikai statisztika tárgyat.

Maximum likelihood becslés

A paraméterbecslés egyik módszere a *maximum likelihood módszer*.

A módszer lényege, hogy az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta $\mathbf{x} = (x_1, \dots, x_n)$ realizációja alapján meghatározzuk azt a $\vartheta_* = \vartheta_*(x_1, \dots, x_n)$ -gal jelölt paramétert, amely esetén a legvalószínűbb, hogy ezt a realizációt kapjuk (diszkrét esetben pontosan, folytonos esetben közelítőleg).

Miután meghatároztuk a $T(x_1, \dots, x_n) = \vartheta_*(x_1, \dots, x_n)$ függvényt, a becslés eredménye a

$$T(X_1, \dots, X_n) = \vartheta_*(X_1, \dots, X_n)$$

statisztika lesz.

A becslés eredménye tehát a véletlen minta egy függvénye, nem pedig a realizációé – ezt érdemes mindig észben tartani!

Likelihood-függvény

Hogy írjuk le, hogy egy adott $\vartheta \in \Theta$ paraméter esetén milyen valószínűséggel kapunk (közelítőleg) egy adott realizációt?

→ A **likelihood-függvény** segítségével.

Emlékeztető: p_ϑ súlyfüggvény, f_ϑ sűrűv. ϑ paraméter esetén.

(14.2.1. Definíció.) Legyenek $X_1, \dots, X_n$ fae. valószínűségi változók, amelyek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függ.

- (1) Ha X_1 eloszlása bármely $\vartheta \in \Theta$ esetén diszkrét és p_ϑ jelöli a ϑ paraméterhez tartozó súlyfüggvényt, akkor az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta bármely $(x_1, \dots, x_n)$ realizációja esetén legyen

$$L_\vartheta(x_1, \dots, x_n) = \prod_{i=1}^n p_\vartheta(x_i).$$

- (2) Ha X_1 eloszlása bármely $\vartheta \in \Theta$ esetén folytonos és $x \mapsto f_\vartheta(x)$ jelöli a ϑ paraméterhez tartozó sűrűségfüggvényt, akkor az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta bármely $(x_1, \dots, x_n)$ realizációja esetén legyen

$$L_\vartheta(x_1, \dots, x_n) = \prod_{i=1}^n f_\vartheta(x_i).$$

Likelihood-függvény

(14.2.1. Definíció.) Legyenek $X_1, \dots, X_n$ fae. valószínűségi változók, amelyek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függ.

- (1) Ha X_1 eloszlása bármely $\vartheta \in \Theta$ esetén diszkrét és p_ϑ jelöli a ϑ paraméterhez tartozó súlyfüggvényt, akkor az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta bármely $(x_1, \dots, x_n)$ realizációja esetén legyen

$$L_\vartheta(x_1, \dots, x_n) = \prod_{i=1}^n p_\vartheta(x_i).$$

- (2) Ha X_1 eloszlása bármely $\vartheta \in \Theta$ esetén folytonos és $x \mapsto f_\vartheta(x)$ jelöli a ϑ paraméterhez tartozó sűrűségfüggvényt, akkor az $\mathbf{X} = (X_1, \dots, X_n)$ n elemű fae. minta bármely $(x_1, \dots, x_n)$ realizációja esetén legyen

$$L_\vartheta(x_1, \dots, x_n) = \prod_{i=1}^n f_\vartheta(x_i).$$

Az $\mathbb{R}^n \rightarrow \mathbb{R}$, $(x_1, \dots, x_n) \mapsto L_\vartheta(x_1, \dots, x_n)$ függvényt mindkét esetben (a ϑ paraméterhez tartozó) **likelihood-függvénynek** nevezzük.

Likelihood-függvény: interpretáció

Diszkrét eset $\rightarrow$ az L_{ϑ} likelihood-függvény a mintaelemek együttes eloszlását írja le abban az esetben, ha ϑ a valódi paraméter.

Valóban, ha A egy esemény, $\mathbb{P}_{\vartheta}(A)$ jelöli az A valószínűségét, feltéve, hogy ϑ a valódi paraméter (4. dia). $X_1, \dots, X_n$ függetlensége miatt

$$L_{\vartheta}(x_1, \dots, x_n) = \prod_{i=1}^n p_{\vartheta}(x_i) = \prod_{i=1}^n \mathbb{P}_{\vartheta}(X_i = x_i) = \mathbb{P}_{\vartheta}(X_1 = x_1, \dots, X_n = x_n).$$

Folytonos eset $\rightarrow$ ha ϑ a valódi paraméter, akkor mivel $f_{\vartheta}(x_i)$ az X_i valószínűségi változó sűrűségfüggvényének értéke az x_i helyen,

$$L_{\vartheta}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\vartheta}(x_i)$$

a függetlenség miatt megegyezik az $X_1, \dots, X_n$ valószínűségi változók együttes sűrűségfüggvényével.

A likelihood-függvény szemléletes jelentése tehát ebben az esetben

$$\begin{aligned} L_{\vartheta}(x_1, \dots, x_n) &= \prod_{i=1}^n f_{\vartheta}(x_i) = \prod_{i=1}^n \lim_{\Delta x_i \downarrow 0} \frac{1}{2\Delta x_i} \mathbb{P}_{\vartheta}(X_i \in (x_i - \Delta x_i, x_i + \Delta x_i)) \\ &= \lim_{\Delta x_1 \downarrow 0, \dots, \Delta x_n \downarrow 0} \frac{1}{2\Delta x_1 \cdot \dots \cdot 2\Delta x_n} \mathbb{P}_{\vartheta}(X_1 \in (x_1 - \Delta x_1, x_1 + \Delta x_1), \dots, X_n \in (x_n - \Delta x_n, x_n + \Delta x_n)). \end{aligned}$$

Maximum likelihood becslés adott realizáció esetén

(14.2.3. Definíció.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függő eloszlással rendelkeznek, és tegyük fel, hogy (minden $\vartheta \in \Theta$ esetén) létezik az L_{ϑ} likelihood-függvény. Legyen $\mathbf{x} = (x_1, \dots, x_n)$ az $\mathbf{X} = (X_1, \dots, X_n)$ egy realizációja.

A **maximum likelihood becslés** ϑ -ra az $(x_1, \dots, x_n)$ **realizáció alapján** az a $\vartheta_*(x_1, \dots, x_n) \in \Theta$ paraméter, amely esetén a likelihood-függvény $L_{\vartheta}(x_1, \dots, x_n)$ értéke maximális.

Formálisan kifejezve,

$$\vartheta_*(x_1, \dots, x_n) = \arg \max_{\vartheta \in \Theta} L_{\vartheta}(x_1, \dots, x_n).$$

Más szavakkal, $\vartheta_*(x_1, \dots, x_n)$ az a $\vartheta \in \Theta$ paraméter, amelyre teljesül, hogy bármely $\vartheta' \in \Theta$ esetén

$$L_{\vartheta_*(x_1, \dots, x_n)}(x_1, \dots, x_n) \geq L_{\vartheta'}(x_1, \dots, x_n).$$

A gyakorlatban a maximum helye szerencsére gyakran egyértelmű.

Maximum likelihood becslés fogalma

Amint már utaltunk rá, a maximum likelihood becslés végeredményeként egy statisztikát kapunk, vagyis nem az $(x_1, \dots, x_n)$ realizáció, hanem az $\mathbf{X} = (X_1, \dots, X_n)$ valószínűségi vektorváltozó egy függvényét. Erről szól a következő definíció:

(14.2.4. Definíció.) Az előző definícióban leírt helyzetben vezessük be a

$$T: \mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto \vartheta_*(x_1, \dots, x_n)$$

függvényt. Ekkor a

$$T(X_1, \dots, X_n) = \vartheta_*(X_1, \dots, X_n)$$

statisztika neve: **maximum likelihood becslés** ϑ -ra.

Kérdés: hogy számoljuk ki a maximum likelihood becslést?

Most megismerünk erre egy standard módszert, ami sokfajta eloszlásnál használható.

Utána lesz olyan példa is, ahol más ötlettel könnyebb kiszámolni a maximum likelihood becslést (de ilyesmire a vizsgán legfeljebb a csillagos feladatban lehet számítani).

Log-likelihood-függvény

$$L_{\vartheta}(x_1, \dots, x_n) = \prod_{i=1}^n p_{\vartheta}(x_i) \text{ (diszkrét)}, \quad L_{\vartheta}(x_1, \dots, x_n) = \prod_{i=1}^n f_{\vartheta}(x_i) \text{ (folyt.)}.$$

Ezt szeretnénk ϑ -ban maximalizálni. $\rightarrow$ Lehet deriválni, de csúnya (n tagú produktum, szorzatszabály, láncszabály...).

Vegyük a **logaritmusát** $\rightarrow$ maximuma ugyanott lesz.

A produktumból szumma lesz és könnyebb lesz deriválni.

(14.2.5. Def.) Ha $(x_1, \dots, x_n) \mapsto L_{\vartheta}(x_1, \dots, x_n)$ egy likelihood-függvény, akkor az

$$\mathbb{R}^n \rightarrow \mathbb{R}, \quad (x_1, \dots, x_n) \mapsto l_{\vartheta}(x_1, \dots, x_n) = \ln L_{\vartheta}(x_1, \dots, x_n)$$

függvényt **log-likelihood-függvénynek** nevezzük.

(14.2.6. Állítás.) Ha $(x_1, \dots, x_n) \mapsto L_{\vartheta}(x_1, \dots, x_n)$ egy likelihood-függvény és $(x_1, \dots, x_n) \mapsto l_{\vartheta}(x_1, \dots, x_n) = \ln L_{\vartheta}(x_1, \dots, x_n)$ az ehhez tartozó log-likelihood-függvény, akkor $\vartheta \in \Theta$ esetén $L_{\vartheta}(x_1, \dots, x_n)$ pontosan akkor maximális ϑ -ban, ha $l_{\vartheta}(x_1, \dots, x_n)$ maximális ϑ -ban.

Maximum likelihood becslés: eljárás log-likelihood függvényes módszerrel

1. Rögzített $x_1, \dots, x_n$ realizáció és $\vartheta \in \Theta$ paraméter esetén meghatározzuk a likelihood-függvény $L_{\vartheta}(x_1, \dots, x_n)$ értékét.
2. Ennek (e alapú) logaritmusa $l_{\vartheta}(x_1, \dots, x_n)$ értéke.
3. Az a $\vartheta \mapsto l_{\vartheta}(x_1, \dots, x_n)$ függvényt deriváljuk (ϑ szerint).
4. Megvizsgáljuk, hogy a derivált hol egyenlő 0-val.
5. Megvizsgáljuk, hogy a derivált zérushelyei lokális maximumhelyek-e. (+ $\rightarrow$ - előjelet vált-e a derivált? Avagy: 2. derivált negatív-e?)
Ha több lokális maximumhely is van, ezek közül az lesz a globális maximumhely, ahol a likelihood-függvény értéke a legnagyobb.
6. Tegyük fel, hogy egy egyértelmű $\vartheta_*(x_1, \dots, x_n)$ globális maximumhelyet kaptunk. Ekkor $\vartheta_*(X_1, \dots, X_n)$ lesz a keresett maximum likelihood becslés.

Amennyiben a Θ paramétertartomány nem az egész $\mathbb{R}^d$, akkor vizsgálni kell azt is, hogy a tartomány határán nincs-e maximumhely.

Maximum likelihood becslés: példák és megjegyzések

- ▶ Diszkrét példa: 14.2.7. Példa, Poisson-eloszlás
- ▶ Folytonos példa: 14.2.8. Példa, exponenciális eloszlás
- ▶ Példa L_{ϑ} maximalizálására l_{ϑ} kiszámolása nélkül: 14.2.10. Példa, egyenletes eloszlás
- ▶ További példák gyakorlaton

Hány hal van a tóban? – avagy a hipergeometriai eloszlás maximum likelihood becslése (14.2.11. Példa)

Bolla Marianna példája (ML-becslés egy $n = 1$ elemű mintánál, egy olyan esetben, amikor sok paraméter van, köztük egy ismeretlen).

Egy tóban élő halak számát szeretnénk humánus módon (az összes hal kifogása és a tó leengedése nélkül) megbecsülni.

Kifogunk néhány halat és megjelöljük egy chippel, ami nem ártalmas a halra.

Jelölje j a megjelölt és k a jelöletlen halak számát, összesen tehát $j + k$ hal van a tóban.

Kifogunk (**visszadobás nélkül**) n halat és ezek közt x jelöltet találunk.

Ez alapján mi a **maximum likelihood becslésünk** k -ra?

Ismeretlen paraméter: k . A likelihood-függvény

$$L_k(x) = \mathbb{P}_k(n \text{ kifogottból } x \text{ jelölt}) = \frac{\binom{j}{x} \binom{k}{n-x}}{\binom{j+k}{n}}.$$

Hogyan maximalizáljuk?

k csak egész értéket vehet fel, nem lehet k szerint deriválni.

Hány hal van a tóban? – avagy a hipergeometriai eloszlás maximum likelihood becslése (14.2.11. Példa)

Jelölje j a megjelölt és k a jelöletlen halak számát, összesen tehát $j + k$ hal van a tóban.

Kifogunk (**visszadobás nélkül**) n halat és ezek közt x jelöltet találunk.

Ez alapján mi a **maximum likelihood becslésünk** k -ra?

Ismeretlen paraméter: k . A likelihood-függvény

$$L_k(x) = \mathbb{P}_k(n \text{ kifogottból } x \text{ jelölt}) = \frac{\binom{j}{x} \binom{k}{n-x}}{\binom{j+k}{n}}.$$

Hogyan maximalizáljuk?

k csak egész értéket vehet fel, nem lehet k szerint deriválni.

Nézzük meg, hogy $L_k(x)$ milyen k -kra lesz nagyobb, mint $L_{k-1}(x)$.

Remény: az eloszlás **unimodális („egymódusú”)**, vagyis $L_k(x)$ egy ideig nő k -ban, aztán pedig csökken.

Hány hal van a tóban? – folytatás

A likelihood-függvény

$$L_k(x) = \mathbb{P}_k(n \text{ kifogottból } x \text{ jelölt}) = \frac{\binom{j}{x} \binom{k}{n-x}}{\binom{j+k}{n}}.$$

Ha $k \geq 1$,

$$\frac{L_k(x)}{L_{k-1}(x)} = \frac{\binom{k}{n-x} \binom{j+k-1}{n}}{\binom{k-1}{n-x} \binom{j+k}{n}} = \frac{k(j+k-n)}{(k-(n-x))(j+k)} \geq 1$$

pontosan akkor teljesül, ha

$$k \leq \frac{jn}{x} - j.$$

Itt vált növekvőről csökkenőre a likelihood-függvény.

Ha $\frac{jn}{x} \notin \mathbb{N}$, akkor $k = \lfloor \frac{jn}{x} \rfloor - j$ az ML-becslés.

Ha $\frac{jn}{x} \in \mathbb{N}$: két ML-becslésünk is van (azonos maximum-értékkel):

$$k = \frac{jn}{x} - j \text{ és } k = \frac{jn}{x} - j - 1.$$

Az összes hal számára ML-becslés: $\lfloor \frac{jn}{x} \rfloor$, illetve $\frac{jn}{x}$ és $\frac{jn}{x} - 1$.

Hipergeometriai eloszlás általában (kiegészítő anyag)

Van N darab objektum, ebből M kedvező, $N - M$ kedvezőtlen. Húzzunk $n \leq N$ darabot **visszatevés nélkül** $\rightarrow$ jelölje X a húzott kedvező objektumok számát. Ekkor

$$\mathbb{P}(X = m) = \frac{\binom{M}{m} \binom{N-M}{n-m}}{\binom{N}{n}}, \quad m = 0, 1, \dots, n.$$

- ▶ Rögzített n esetén, ha N végtelenhez tart és $M/N = p$ állandó, tart az (n, p) paraméterű binomiális eloszláshoz.
- ▶ Várható értéke mindig $\frac{nM}{N} = np$.
- ▶ Hogyan tér el a binomiálistól?

Ha már húztam egy kedvező objektumot, a következő lépésben **kevésbé** valószínű, hogy megint kedvezőt húzok (mivel visszatevés nélkül húzok) $\rightarrow$ a kedvező objektumok húzásainak indikátorai a különböző kísérletekben **negatíván korreláltak**.

A korrelációk nullához tartanak a fenti határértékben.

Becslések tulajdonságai

(14.1.1. Def.) Legyen adott egy $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása egy ismeretlen $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függ, és legyen $\psi: \mathbb{R}^d \rightarrow \mathbb{R}$ egy függvény. Azt mondjuk, hogy a minta egy $T(\mathbf{X}) = T(X_1, \dots, X_n)$ statisztikája

(1) a $\psi(\vartheta)$ paraméterfüggvény egy **torzítatlan** becslése, ha

$$\mathbb{E}_{\vartheta}(T(X_1, \dots, X_n)) = \psi(\vartheta), \quad \forall \vartheta \in \Theta,$$

ahol $\mathbb{E}_{\vartheta}$ jelöli a $\mathbb{P}_{\vartheta}$ valószínűség szerinti várható értéket,

(2) a $\psi(\vartheta)$ egy **aszimptotikusan torzítatlan** becslése, ha

$$\lim_{n \rightarrow \infty} \mathbb{E}_{\vartheta}(T(X_1, \dots, X_n)) = \psi(\vartheta), \quad \forall \vartheta \in \Theta,$$

(3) a $\psi(\vartheta)$ egy **erősen konzisztens** becslése, ha

$$\mathbb{P}_{\vartheta}\left(\lim_{n \rightarrow \infty} T(X_1, \dots, X_n) = \psi(\vartheta)\right) = 1, \quad \forall \vartheta \in \Theta.$$

(Megjegyzés: *gyengén konzisztensnek*, avagy egyszerűen *konzisztensnek* akkor nevezzük a becslést, ha ugyanez az 1 valószínűségű/majdnem biztos konvergencia helyett valószínűségben való/sztochasztikus konvergenciával teljesül. Láttuk: az 1 valószínűségű konvergenciából következik a valószínűségben való konvergencia is.)
Tehát: erősen konzisztens $\Rightarrow$ konzisztens is.)

Becslések tulajdonságai

(14.1.2. Def.)

Ha $T(\mathbf{X})$ és a minta egy másik $T'(\mathbf{X}) = T'(X_1, \dots, X_n)$ statisztikája is torzítatlan becslése $\psi(\vartheta)$ -nak, akkor azt mondjuk, hogy $T(\mathbf{X})$ **legalább annyira hatásos, mint** $T'(\mathbf{X})$, ha $\mathbb{D}_{\vartheta}^2(T(\mathbf{X})) \leq \mathbb{D}_{\vartheta}^2(T'(\mathbf{X}))$, ahol $\mathbb{D}_{\vartheta}^2$ jelöli a $\mathbb{P}_{\vartheta}$ szerinti szórásnégyzetet.

Ha $T(\mathbf{X})$ legalább annyira hatásos, mint $\psi(\vartheta)$ bármely torzítatlan becslése, akkor $T(\mathbf{X})$ a $\psi(\vartheta)$ egy **hatásos** becslése.

Becslések tulajdonságai

Megjegyzések:

- ▶ Torzítatlan $\Rightarrow$ aszimptotikusan torzítatlan.
- ▶ Vannak más konzisztenciafogalmak is $\rightarrow$ lásd a Matematikai statisztika tárgyat.

(14.1.3. Példa.) Egy lehetséges ψ függvény a

$$\Theta \ni \vartheta \mapsto \psi(\vartheta) = \mathbb{E}_{\vartheta}(X_1),$$

ha az $\mathbb{E}_{\vartheta}(X_1)$ minden $\vartheta \in \Theta$ esetén létezik (azaz ha $\mathbb{E}_{\vartheta}(|X_1|) < \infty$ minden $\vartheta \in \Theta$ esetén teljesül). Ebben az esetben a

$$T(X_1, \dots, X_n) = \bar{X}_n = \frac{X_1 + \dots + X_n}{n}$$

mintaátlag a $\psi(\vartheta) = \mathbb{E}_{\vartheta}(X_1)$ paraméterfüggvény egy torzítatlan és erősen konzisztens becslése.

(Bizonyítás: lásd előadás/jegyzet)

Becslések tulajdonságai

(14.1.4. Példa.) Egy másik lehetséges ψ függvény a

$$\Theta \ni \vartheta \mapsto \psi(\vartheta) = \mathbb{D}_{\vartheta}^2(X_1),$$

ha $\mathbb{E}_{\vartheta}(X_1^2)$ minden $\vartheta \in \Theta$ esetén véges. Ebben az esetben a

$$T(X_1, \dots, X_n) = (S_n^*)^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \overline{X}_n)^2$$

korrigált empirikus szórásnégyzet torzítatlan becslése a

$$\psi(\vartheta) = \mathbb{D}_{\vartheta}^2(X_1)$$

paraméterfüggvénynek. A

$$T'(X_1, \dots, X_n) = S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X}_n)^2 \quad (2)$$

statisztika nem torzítatlan becslése $\mathbb{D}_{\vartheta}^2(X_1)$ -nek, de aszimptotikusan torzítatlan.

(Bizonyítás: lásd előadás/jegyzet)

Becslések tulajdonságai: példák

- ▶ 14.1.5. Példa (vö. 5.16. feladat): a torzítatlanság önmagában nem garantálja a becslés jóságát, pl. a hatásosságát.
- ▶ A maximum likelihood becslés lehet torzítatlan, de nem mindig az: 14.2.9. Megjegyzés.
- ▶ Az empirikus eloszlásfüggvény értéke a valódi eloszlásfüggvény egy becslése $\rightarrow$ rögzített $x \in \mathbb{R}$ esetén $F_n^*(x)$ torzítatlan és erősen konzisztens becslése $F(x)$ -nek: össznépi gyakorlat, 12.5. feladat.

Torzítatlanság és konzisztencia: össznépi gyakorlat, 12.7. feladat: Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy $n \geq 1$ elemű fae. minta, amelynek elemei $U(\vartheta; \vartheta + 1)$ eloszlásúak, ahol $\vartheta \in \Theta$ ismeretlen paraméter és $\Theta = (0, \infty)$.

1. Számoljuk ki $\mathbb{E}_{\vartheta}(1/X_1)$ -et és $\mathbb{D}_{\vartheta}^2(1/X_1)$ -et,
2. ennek segítségével adjunk torzítatlan és erősen konzisztens becslést a $\psi(\vartheta) = \ln \frac{\vartheta+1}{\vartheta}$ paraméterfüggvényre!

Momentumbecslés/momentumok módszere

Emlékeztető: (14.3.1. Def. – a 3. diason már volt.) Legyen k egy pozitív egész szám és X egy valószínűségi változó. Ha $\mathbb{E}[|X|^k] < \infty$, akkor az X k . **momentumán** az $\mathbb{E}[X^k] \in \mathbb{R}$ mennyiséget értjük.

Kérdés: Milyen statisztikával becsülhetjük a momentumokat, ha az eloszlás ismeretlen paraméter(ek)től függ?

(14.3.2. Def.) Legyen k egy pozitív egész szám és $X_1, \dots, X_n$ független, azonos eloszlású valószínűségi változók. Az $(X_1, \dots, X_n)$ n -elemű fae. minta k -**adik empirikus momentumán** az

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^n X_i^k$$

statisztikát értjük.

$k = 1$: az 1. empirikus momentum a mintaátlag: $\hat{m}_1 = \overline{X_n}$.

Momentumok módszere: motiváció

A momentumok módszere különösen akkor lesz hasznos, ha egynél több ismeretlen paraméterünk van. Pl.:

- ▶ normális eloszlás ismeretlen μ várható értékkel és ismeretlen σ^2 szórásnégyzettel $\rightarrow$ az ismeretlen paraméterek vektora $\underline{\vartheta} = (\mu, \sigma^2) \in \mathbb{R}^2$, ahol $\mu \in \mathbb{R}$ és $\sigma^2 > 0$,
- ▶ egyenletes eloszlás egy ismeretlen kezdő- és végpontú (a, b) intervallumon $\rightarrow$ a paramétervektor $\underline{\vartheta} = (a, b) \in \mathbb{R}^2$ valamely $-\infty < a < b < \infty$ valós számokra.

Paraméterek és momentumok

Általában tegyük fel, hogy az $\mathbf{X} = (X_1, \dots, X_n)$ fae. minta elemeinek eloszlását meghatározó ismeretlen paraméterek vektora $\underline{\vartheta} = (\vartheta_1, \dots, \vartheta_k)$, amely egy $\Theta \subseteq \mathbb{R}^k$ halmaz egy eleme.

Tegyük fel, hogy az $X_1, \dots, X_n$ mintaelemeknek a paraméter bármely választása esetén létezik az első k momentuma. Ezek a momentumok természetesen a $\vartheta_1, \dots, \vartheta_k$ paraméterek függvényei:

$$m_1 = g_1(\vartheta_1, \dots, \vartheta_k), m_2 = g_2(\vartheta_1, \dots, \vartheta_k), \dots, m_k = g_k(\vartheta_1, \dots, \vartheta_k)$$

valamely $g_1, g_2, \dots, g_k: \mathbb{R}^k \rightarrow \mathbb{R}$ függvényre. A

$$(\vartheta_1, \dots, \vartheta_k) \mapsto \underline{g}(\vartheta_1, \dots, \vartheta_k) := (g_1(\vartheta_1, \dots, \vartheta_k), \dots, g_k(\vartheta_1, \dots, \vartheta_k))$$

leképezés tehát egy $\underline{g}: \mathbb{R}^k \rightarrow \mathbb{R}^k$ függvény, vagyis egy vektormező.

Ha ennek a vektormezőnek létezik inverze, akkor jelöljük ezt $\underline{h} = (h_1, \dots, h_k)$ -val. Ezen inverz segítségével a momentumok ismeretében megkaphatjuk a $\underline{\vartheta} = (\vartheta_1, \dots, \vartheta_k)$ paramétert:

$$\vartheta_1 = h_1(m_1, \dots, m_k), \vartheta_2 = h_2(m_1, \dots, m_k), \dots, \vartheta_k = h_k(m_1, \dots, m_k).$$

Momentumok módszere

$$\vartheta_1 = h_1(m_1, \dots, m_k), \vartheta_2 = h_2(m_1, \dots, m_k), \dots, \vartheta_k = h_k(m_1, \dots, m_k).$$

(14.3.3. Def.) Ekkor az $\underline{\vartheta} = (\vartheta_1, \dots, \vartheta_n)$ paramétervektor **momentumbecslésén** az $\underline{\hat{\vartheta}} = (\hat{\vartheta}_1, \dots, \hat{\vartheta}_k)$ vektort értjük, ahol

$$\hat{\vartheta}_i = h_i(\hat{m}_1, \dots, \hat{m}_k), \quad i = 1, \dots, k,$$

és $\hat{m}_i$ az i -edik empirikus momentumot jelöli.

Megjegyzés: differenciálható $\underline{g}$ leképezés esetén az $\underline{h} = (h_1, \dots, h_k)$ inverz létezésének szükséges feltétele, hogy $\underline{g}$ Jacobi-determinánsa a Θ halmazon sehol se legyen 0.

Példák momentumbecslésre:

- ▶ 14.3.4. Példa: normális eloszlás
- ▶ 14.3.5. Példa: egyenletes eloszlás

Konfidenciaintervallumok

Eddig: *pontbecslések* (átlag/szórásnégyzet becslései, maximum likelihood, momentumbecslés stb) $\rightarrow$ a „legjobb tippünk” az ismeretlen paraméter(ek)re egy valós szám vagy vektor.

Most: *intervallumbecslés/konfidenciaintervallumok* $\rightarrow$ nem egyetlen szám/vektor lesz a becslésünk, hanem egy **intervallum**,

- ▶ amelybe a keresett paraméter egy bizonyos előre megadott valószínűséggel (pl. legalább 95%, pontosan 99% stb.) beleesik,
- ▶ és amely az előbbi tulajdonságokat teljesítő összes intervallum közül valamilyen értelemben optimális (pl. a legrövidebb).

Informatikai példa: szerverválasztó?

Konfidenciaintervallumok

(15.1.1. Definíció.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy n elemű fae. minta, ahol a mintaelemek eloszlása egy $\vartheta \in \Theta \subseteq \mathbb{R}^d$ paramétertől függ, legyen $\psi: \mathbb{R}^d \rightarrow \mathbb{R}$ egy függvény, és legyen $0 < \varepsilon < 1$. Azt mondjuk, hogy a $[T_1(\mathbf{X}), T_2(\mathbf{X})]$ intervallum (pontosan) $1 - \varepsilon$ **szintű konfidenciaintervallum** a $\psi(\vartheta)$ paraméterfüggvényre, ha

$$\mathbb{P}_{\vartheta}(T_1(\mathbf{X}) \leq \psi(\vartheta) \leq T_2(\mathbf{X})) = 1 - \varepsilon,$$

minden $\vartheta \in \Theta$ esetén.

Megjegyzések:

1. A definícióban szereplő $T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ intervallum-végpontok az $\mathbf{X}$ minta **statisztikái**, tehát a véletlen mintától függenek, de a realizáció ismeretében már nem véletlenek.
2. Pontosán $1 - \varepsilon$ szintű konfidenciaintervallum általában csak akkor adható meg, ha $X_1, \dots, X_n$ peremeloszlása folytonos. Diszkrét esetben általában *legalább* $1 - \varepsilon$ *szintű konfidenciaintervallumokról* szokás beszélni, ahol a definícióban „ $= 1 - \varepsilon$ ” helyett „ $\geq 1 - \varepsilon$ ” szerepel.

Kvantilisek: bevezetés

Azt mondjuk, hogy a $[T_1(\mathbf{X}), T_2(\mathbf{X})]$ intervallum (pontosan) $1 - \varepsilon$ **szintű konfidenciaintervallum** a $\psi(\vartheta)$ paraméterfüggvényre, ha

$$\mathbb{P}_{\vartheta}(T_1(\mathbf{X}) \leq \psi(\vartheta) \leq T_2(\mathbf{X})) = 1 - \varepsilon, \quad \forall \vartheta \in \Theta.$$

$T_1(\mathbf{X})$ és $T_2(\mathbf{X})$ meghatározásához szükséges fogalom: folytonos eloszlások *kvantilisei*.

Legyen X folytonos val. változó $\rightarrow F_X$ eloszlásfv., f_X sűrűv.

Tudjuk: F_X folytonos függvény.

Az egyszerűség kedvéért tegyük fel, hogy f_X pontosan egy I nyílt intervallumon nem 0. Ilyen az összes nevezetes folytonos eloszlás:

- ▶ az egyenletes eloszlás, ahol $I = (a, b)$, $-\infty < a < b < \infty$,
- ▶ az exponenciális eloszlás, ahol $I = (0, \infty)$,
- ▶ és a normális eloszlás, ahol $I = (-\infty, \infty) = \mathbb{R}$.

Ilyen esetben $\text{Ran}(X) = I$. F_X szigorúan monoton nő I -n.

„ I alsó végpontjáig” (ha az véges) konstans 0, „ I felső végpontjától” (ha az véges) konstans 1.

Létezik az $F_X^{-1}: (0, 1) \rightarrow I$ inverzfüggvény, szintén folytonos és szig. mon. növvő.

Kvantilisok és medián

(15.1.3. Definíció.) Legyen X folytonos valószínűségi változó F_X eloszlásfüggvénnyel és f_X sűrűségfüggvénnyel, és tegyük fel, hogy létezik egy olyan I nyílt intervallum, hogy minden $x \in \mathbb{R}$ esetén $f_X(x) \neq 0$ pontosan akkor teljesül, ha $x \in I$.

Legyen $0 < y < 1$. Ekkor az $F_X^{-1}(y) \in I$ pontot az X y -kvantilisének, az $\frac{1}{2}$ -kvantilist ($y = 1/2$) **mediánnak** nevezzük.

- ▶ Tehát ha x az X y -kvantilise, akkor $\mathbb{P}(X < x) = y$.
- ▶ Miért nem jó ez a definíció diszkrét X esetén?
- ▶ Mi a probléma folytonos esetben, ha a sűrűségfüggvény pl. két távoli diszjunkt intervallum unióján nem nulla?
- ▶ A medián az 5. feladatsor 7. e*) feladatában már szerepelt.
- ▶ Emlékeztető: az **empirikus medián** páratlan elemszám esetén a minta középső eleme, párosnál a két középső elem átlaga.

Láttuk: az F_n^* empirikus eloszlásfüggvény maga is egy eloszlásfüggvény, a hozzá tartozó eloszlás diszkrét.

Az empirikus eloszlásfüggvény az empirikus medián környékén $1/2$ körül van (a részletekbe itt nem megyünk bele). Ilyen értelemben analógja az empirikus medián a mediánnak.

Kvantilisok és medián

(15.1.3. Definíció.) Legyen X folytonos valószínűségi változó F_X eloszlásfüggvénnyel és f_X sűrűségfüggvénnyel, és tegyük fel, hogy létezik egy olyan I nyílt intervallum, hogy minden $x \in \mathbb{R}$ esetén $f_X(x) \neq 0$ pontosan akkor teljesül, ha $x \in I$.

Legyen $0 < y < 1$. Ekkor az $F_X^{-1}(y) \in I$ pontot az X y -kvantilisének, az $\frac{1}{2}$ -kvantilist ($y = 1/2$) **mediánnak** nevezzük.

- ▶ Tehát ha x az X y -kvantilise, akkor $\mathbb{P}(X < x) = y$.
- ▶ Miért nem jó ez a definíció diszkrét X esetén?
- ▶ Mi a probléma folytonos esetben, ha a sűrűségfüggvény pl. két távoli diszjunkt intervallum unióján nem nulla?
- ▶ A medián az 5. feladatsor 7. e*) feladatában már szerepelt.
- ▶ **Emlékeztető:** az **empirikus medián** páratlan elemszám esetén a minta középső eleme, párosnál a két középső elem átlaga.

Láttuk: az F_n^* empirikus eloszlásfüggvény maga is egy eloszlásfüggvény, a hozzá tartozó eloszlás diszkrét.

Az empirikus eloszlásfüggvény az empirikus medián környékén $1/2$ körül van (a részletekbe itt nem megyünk bele). Ilyen értelemben analógja az empirikus medián a mediánnak.

Konfidenciaintervallum normális eloszlás várható értékére ismert szórásnégyzet esetén

A konfidenciaintervallumok legalapvetőbb példája:

Feladat. Legyen $X \sim N(\mu; \sigma^2)$, ahol a $\mu \in \mathbb{R}$ várható érték ismeretlen, de a $\sigma > 0$ szórás ismert. Adott $\varepsilon \in (0, 1)$ esetén szerkesszünk $1 - \varepsilon$ szintű konfidenciaintervallumot μ -re úgy, hogy annak hossza a lehető legkisebb legyen!

Mivel az

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

sűrűségfüggvény az x pont μ -től való távolsága függvényében szigorúan monoton csökken, ezért adott ε esetén a legrövidebb $1 - \varepsilon$ szintű konfidenciaintervallumot akkor fogjuk kapni, ha azt az $\overline{X}_n$ mintaátlag körül szimmetrikusan választjuk, vagyis

$$[\overline{X}_n - r_{\varepsilon}, \overline{X}_n + r_{\varepsilon}]$$

$$\text{alakban, úgy, hogy } \mathbb{P}_{\mu}(\mu \in [\overline{X}_n - r_{\varepsilon}, \overline{X}_n + r_{\varepsilon}]) = 1 - \varepsilon \quad (3)$$

minden $\mu \in \mathbb{R}$ esetén teljesüljön. Cél: r_{ε} (ε -tól függő) értékének meghatározása.

Konfidenciaintervallum μ -re ismert σ^2 esetén

$$\mathbb{P}_\mu(\mu \in [\overline{X}_n - r_\varepsilon, \overline{X}_n + r_\varepsilon]) = \mathbb{P}_\mu(\overline{X}_n - r_\varepsilon \leq \mu \leq \overline{X}_n + r_\varepsilon) = 1 - \varepsilon, \quad r_\varepsilon = ?$$

A mintaátlag standardizálása. A mintaátlag eloszlása a normális eloszlás ismert tulajdonságai alapján:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n \underbrace{X_i}_{\sim N(\mu; \sigma^2) \text{ fae.}} \sim N(\mu; \sigma^2/n).$$

$\underbrace{\hspace{10em}}_{\sim N(n\mu; n\sigma^2)}$

Ebből következik, hogy

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0; 1). \quad (4)$$

Itt álljunk meg egy pillanatra. A CHT pont azt mondja ki, hogy

$$\frac{\overline{X}_n - \mathbb{E}(\overline{X}_n)}{\mathbb{D}(\overline{X}_n)} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma} \sqrt{n} \xrightarrow[n \rightarrow \infty]{d} N(0; 1). \quad (5)$$

Normális eloszlású val. változók esetén tehát a CHT nemcsak eloszlásbeli konvergenciával, hanem **egyenlőséggel** teljesül!

Konfidenciaintervallum μ -re ismert σ^2 esetén

$$\mathbb{P}_\mu(\mu \in [\overline{X}_n - r_\varepsilon, \overline{X}_n + r_\varepsilon]) = \mathbb{P}_\mu(\overline{X}_n - r_\varepsilon \leq \mu \leq \overline{X}_n + r_\varepsilon) = 1 - \varepsilon, \quad r_\varepsilon = ?$$

A mintaátlag standardizálása. A mintaátlag eloszlása a normális eloszlás ismert tulajdonságai alapján:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^n \underbrace{X_i}_{\sim N(\mu; \sigma^2) \text{ fae.}} \sim N(\mu; \sigma^2/n).$$

$\underbrace{\hspace{10em}}_{\sim N(n\mu; n\sigma^2)}$

Ebből következik, hogy

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0; 1). \quad (4)$$

Fejezzük ki most a fenti egyenletet az $\frac{\overline{X}_n - \mu}{\sigma} \sqrt{n}$ valószínűségi változó segítségével:

$$\begin{aligned} \mathbb{P}_\mu(\overline{X}_n - r_\varepsilon \leq \mu \leq \overline{X}_n + r_\varepsilon) &= 1 - \varepsilon \Leftrightarrow \mathbb{P}_\mu(-r_\varepsilon \leq \overline{X}_n - \mu \leq r_\varepsilon) = 1 - \varepsilon \\ &\Leftrightarrow \mathbb{P}_\mu\left(-\frac{r_\varepsilon}{\sigma} \sqrt{n} \leq \frac{\overline{X}_n - \mu}{\sigma} \sqrt{n} \leq \frac{r_\varepsilon}{\sigma} \sqrt{n}\right) = 1 - \varepsilon. \end{aligned}$$

Konfidenciaintervallum μ -re ismert σ^2 esetén

Cél: $\mathbb{P}_\mu\left(-\frac{r_\varepsilon}{\sigma}\sqrt{n} \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq \frac{r_\varepsilon}{\sigma}\sqrt{n}\right) = 1 - \varepsilon$, $r_\varepsilon = ?$.

Mivel $\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \sim N(0; 1)$, ezért [emlékeztető: $\Phi(-x) = 1 - \Phi(x)$]

$$\mathbb{P}_\mu\left(-\frac{r_\varepsilon}{\sigma}\sqrt{n} \leq \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \leq \frac{r_\varepsilon}{\sigma}\sqrt{n}\right) = \Phi\left(\frac{r_\varepsilon}{\sigma}\sqrt{n}\right) - \Phi\left(-\frac{r_\varepsilon}{\sigma}\sqrt{n}\right) = 2\Phi\left(\frac{r_\varepsilon}{\sigma}\sqrt{n}\right) - 1.$$

Vagyis $\Phi\left(\frac{r_\varepsilon}{\sigma}\sqrt{n}\right) = 1 - \varepsilon/2$, azaz

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

(ahol $\Phi^{-1}(1 - \frac{\varepsilon}{2})$ -t a szokásos módon kiolvashatjuk a normális eloszlás táblázatából), tehát

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

Visszahelyettesítve a konfidenciaintervallum:

$$\left[\overline{X}_n - \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) \frac{\sigma}{\sqrt{n}}, \overline{X}_n + \Phi^{-1}\left(1 - \frac{\varepsilon}{2}\right) \frac{\sigma}{\sqrt{n}}\right].$$

Konfidenciaintervallum μ -re ismert σ^2 esetén

A konfidenciaintervallum:

$$[\overline{X}_n - \Phi^{-1}(1 - \varepsilon/2) \frac{\sigma}{\sqrt{n}}, \overline{X}_n + \Phi^{-1}(1 - \varepsilon/2) \frac{\sigma}{\sqrt{n}}].$$

Ezzel készen vagyunk, ezt még kicsit cicomázzuk – bevezetünk egy olyan jelölést, ami később a hipotézisvizsgálatnál is hasznos lesz:

$$u_\delta := \Phi^{-1}(1 - \delta), \quad \delta \in (0, 1).$$

Vegyük észre, hogy u_δ az $N(0; 1)$ eloszlás $1 - \delta$ -kvantilise, hiszen az eloszlásfüggvény Φ .

A konfidenciaintervallum tehát $[\overline{X}_n - u_{\varepsilon/2} \frac{\sigma}{\sqrt{n}}, \overline{X}_n + u_{\varepsilon/2} \frac{\sigma}{\sqrt{n}}]$.

Összefoglalva:

(15.2.2. Állítás.) Legyenek $X_1, \dots, X_n$ fae. $N(\mu; \sigma^2)$ eloszlású valószínűségi változók, ahol a μ várható érték ismeretlen és a $\sigma^2 > 0$ szórásnégyzet ismert, és jelölje $\overline{X}_n$ a mintához tartozó mintaátlagot. Ekkor $\varepsilon \in (0, 1)$ esetén a μ várható értékre egy $1 - \varepsilon$ szintű konfidenciaintervallum a következő:

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = [\overline{X}_n - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}],$$

ahol $u_{\varepsilon/2}$ a standard normális eloszlás $1 - \varepsilon/2$ -kvantilise.

Konfidenciaintervallum μ -re ismert σ^2 esetén

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = \left[\overline{X}_n - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \right].$$

Figyeljük meg, hogy

1. minél nagyobb a minta elemszáma, annál kevésbé hosszú konfidenciaintervallum szükséges (vö. nagy számok törvénye),
2. minél nagyobb a szórás, annál hosszabb konfidenciaintervallum szükséges (nagy szórás esetén „laposabb”, jobban szétterülő a sűrűségfüggvény),
3. minél kisebb az ε hibaszint, annál hosszabb konfidenciaintervallum szükséges (mert $1 - \varepsilon$ valószínűséggel kell μ -nek a konfidenciaintervallumba esnie, ha μ a valódi paraméter).

További kérdések: konfidenciaintervallum σ^2 -re ismeretlen μ esetén? És ismert μ esetén? És μ -re ismeretlen σ^2 esetén?

Ehhez most: t - és χ^2 -eloszlás, nagyrészt biz. nélkül.

2024-ben amihez kell a χ^2 -eloszlás, az csak kiegészítő anyag.

Amihez viszont csak a t -eloszlás kell, az vizsgaanyag.

Konfidenciaintervallum μ -re ismert σ^2 esetén

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = \left[\overline{X}_n - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \right].$$

Figyeljük meg, hogy

1. minél nagyobb a minta elemszáma, annál kevésbé hosszú konfidenciaintervallum szükséges (vö. nagy számok törvénye),
2. minél nagyobb a szórás, annál hosszabb konfidenciaintervallum szükséges (nagy szórás esetén „laposabb”, jobban szétterülő a sűrűségfüggvény),
3. minél kisebb az ε hibaszint, annál hosszabb konfidenciaintervallum szükséges (mert $1 - \varepsilon$ valószínűséggel kell μ -nek a konfidenciaintervallumba esnie, ha μ a valódi paraméter).

További kérdések: konfidenciaintervallum σ^2 -re ismeretlen μ esetén? És ismert μ esetén? És μ -re ismeretlen σ^2 esetén?

Ehhez most: t - és χ^2 -eloszlás, nagyrészt biz. nélkül.

2024-ben amihez kell a χ^2 -eloszlás, az csak kiegészítő anyag.

Amihez viszont csak a t -eloszlás kell, az vizsgaanyag.

χ^2 -eloszlás

(15.3.1. Definíció.) Legyenek $X_1, \dots, X_n$ fae. $N(0; 1)$ eloszlású valószínűségi változók. Ekkor $\sum_{i=1}^n X_i^2$ eloszlását **n szabadságfokú χ^2 -eloszlásnak** (ejtsd: khí-négyzet-eloszlásnak) nevezzük. Jelölés: $X \sim \chi^2(n)$.

(15.3.2. Definíció.) Legyen egy n pozitív egész szám. Ha az X valószínűségi változó n szabadságfokú χ^2 -eloszlású, akkor a sűrűségfüggvénye

$$f_n(x) = \begin{cases} \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, & \text{ha } x > 0, \\ 0, & \text{ha } x \leq 0. \end{cases}$$

Itt Γ az úgynevezett **Euler-féle Gamma-függvény**, amelyre $\Gamma(n/2)$ rekurzívan a $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$ és $\Gamma(r+1) = r\Gamma(r)$, $r \in [0, \infty)$ egyenletekből számítható ki.

Eloszlásfüggvény: $F_n(x) = \int_{-\infty}^x f_n(t) dt \rightarrow$ erre is van táblázat.

χ^2 -eloszlás

$X_1, \dots, X_n \sim N(0; 1)$ fae., $\sum_{i=1}^n X_i^2 \sim \chi^2(n)$.

$$f_n(x) = \begin{cases} \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, & \text{ha } x > 0, \\ 0, & \text{ha } x \leq 0. \end{cases}$$

$\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$ és $\Gamma(r+1) = r\Gamma(r)$, $r \in [0, \infty)$.

► Ha n egy pozitív egész szám:

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = \dots = (n-1)(n-2)\dots 1\Gamma(1) = (n-1)!$$

A Γ függvény folytonos (sőt analitikus) módon kiterjeszthető az $\mathbb{R} \setminus \{0, -1, -2, \dots\}$ halmazra, ez tehát a faktoriális egyfajta kiterjesztése majdnem minden valós számra.

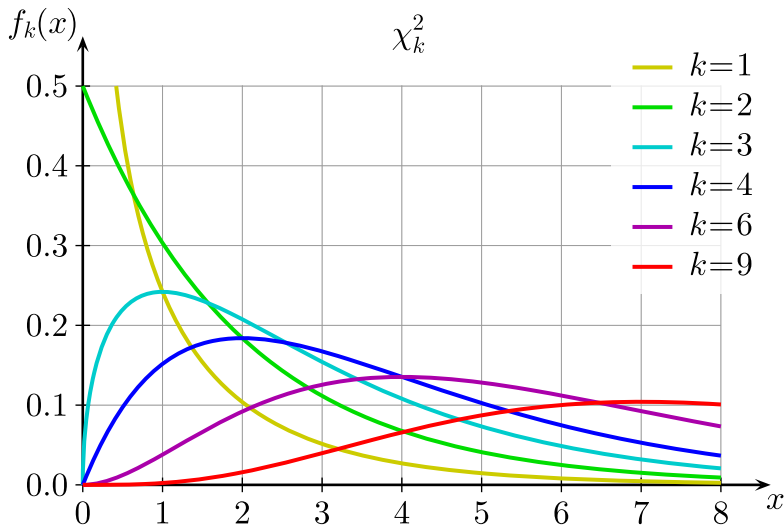
► *Kapcsolat az exponenciális eloszlással.* $n = 2$ esetén

$$f_n(x) = \begin{cases} \frac{e^{-x/2}}{2}, & \text{ha } x > 0, \\ 0, & \text{ha } x \leq 0, \end{cases}$$

vagyis $\chi^2(2) = \text{Exp}(1/2)$. Általában igaz: $\chi^2(2n)$ megegyezik n darab fae. $\text{Exp}(1/2)$ v.v. összegének eloszlásával.

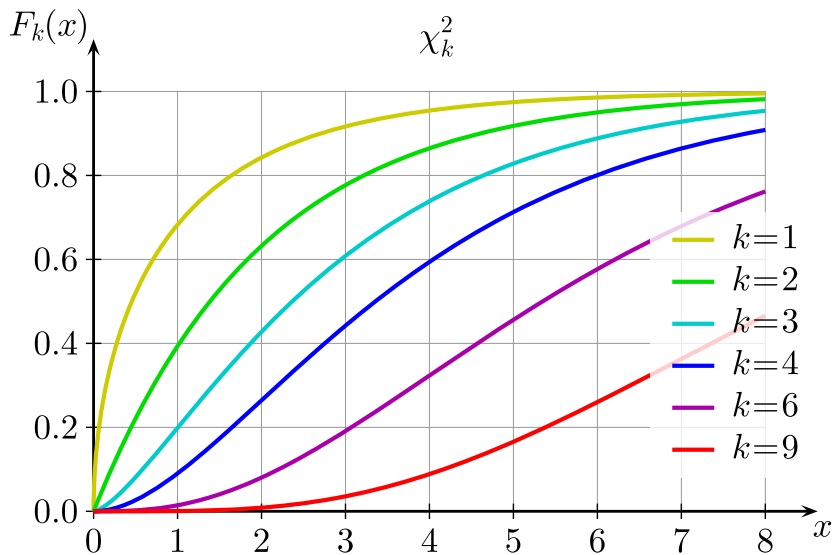
► A 7.5. feladatban (név nélkül) szerepelt a $\chi^2(1)$ eloszlás.

χ^2 -eloszlás sűrűségfüggvénye



"Chi-square pdf" by Geek3 – Own work. Licensed under CC BY 3.0 via Wikimedia Commons

χ^2 -eloszlás eloszlásfüggvénye



t -eloszlás (Student-eloszlás)

(15.3.4. Definíció.) Legyen n egy pozitív egész szám. Az X valószínűségi változó eloszlását n **szabadságfokú t -eloszlásnak**, más szóval n **szabadságfokú Student-eloszlásnak** nevezzük, ha a sűrűségfüggvénye

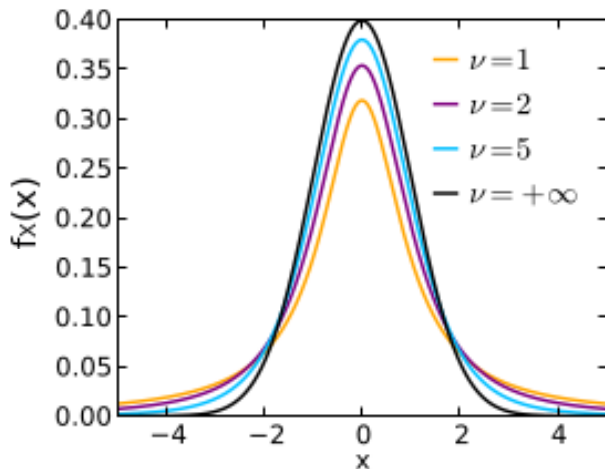
$$f_n(x) = \frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)} \left(1 + \frac{x^2}{n}\right)^{-(n+1)/2}, \quad x \in \mathbb{R},$$

ahol Γ ismét az Euler-féle Gamma-függvény. Jelölés: $X \sim t(n)$.

1. A „Student” név az eloszlást felfedező William Sealy Gosset angol statisztikus álneve volt.
2. Az 1 szabadságfokú, $f_1(x) = \frac{1}{\pi} \frac{1}{1+x^2}$ sűrűségfüggvényű t -eloszlás ismertebb nevén a *standard Cauchy-eloszlás*. Az ilyen eloszlású X valószínűségi változó érdekes tulajdonsága, hogy $\mathbb{E}[|X|] = \infty$, és emiatt az $\mathbb{E}[X]$ várható érték nem létezik.

Általában az $n \geq 2$ szabadságfokú t -eloszlásnak az első $n - 1$ momentuma létezik, az n -edik már nem.

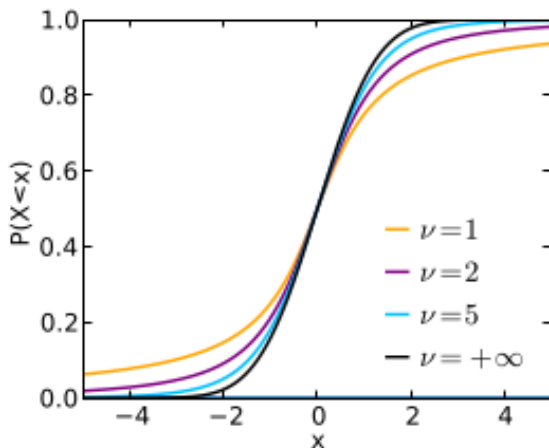
t -eloszlás (Student-eloszlás) sűrűségfüggvénye



"SStudent t cdf" by Skbkakas – Own work. Licensed under CC BY 3.0 via Wikimedia Commons

Ugyanúgy szimmetrikus a 0-ra, mint a standard normális eloszlás sűrűségfüggvénye. „ $\nu = \infty$ ”: **standard normális**.

t -eloszlás (Student-eloszlás) eloszlásfüggvénye



"SStudent t pdf" by Skbkekas – Own work. Licensed under CC BY 3.0 via Wikimedia Commons

A szimmetria és folytonosság miatt $\mathbb{P}(X < x) = 1 - \mathbb{P}(X < -x)$ és $\mathbb{P}(-x < X < x) = 2\mathbb{P}(X < x) - 1$, $x > 0$, a standard normális eloszláshoz hasonlóan.

t-eloszlás: táblázat

A t -(Student-)próba kritikus értékei
1-y értéke, ahol az y-kvantilist nézzük: 2. sor

szabadsági fok:
1. oszlop

f	0,1 0,05	0,05 0,025	0,02 0,01
1	6,314	12,71	31,82
2	2,920	4,303	6,965
3	2,353	3,182	4,541
4	2,132	2,776	3,747
5	2,015	2,571	3,365
6	1,943	2,447	3,143
7	1,895	2,365	2,998
8	1,860	2,306	2,896
9	1,833	2,262	2,821
10	1,812	2,228	2,764
11	1,796	2,201	2,718
12	1,782	2,179	2,681
13	1,771	2,160	2,650
14	1,761	2,145	2,624
15	1,753	2,131	2,602
16	1,746	2,120	2,583
17	1,740	2,110	2,567
18	1,734	2,101	2,552
19	1,729	2,093	2,539
20	1,725	2,086	2,528

f	0,1 0,05	0,05 0,025	0,02 0,01
21	1,721	2,080	2,518
22	1,717	2,074	2,508
23	1,714	2,069	2,500
24	1,711	2,064	2,492
25	1,708	2,060	2,485
26	1,706	2,056	2,479
27	1,703	2,052	2,473
28	1,701	2,048	2,467
29	1,699	2,045	2,462
30	1,697	2,042	2,457
40	1,684	2,021	2,423
50	1,676	2,009	2,403
60	1,671	2,000	2,390
70	1,667	1,994	2,381
80	1,664	1,990	2,374
90	1,662	1,987	2,369
100	1,660	1,984	2,364
200	1,653	1,972	2,345
500	1,648	1,965	2,334
∞	1,645	1,960	2,326

← $N(0;1)$

A gyakorlatban, ha f nagy (ököl szabály: $f \geq 30$), az f szabadsági fokú t -eloszlás közelíthető standard normálissal.

Hol fordul elő a t -eloszlás?

(15.3.6. Állítás.) Legyenek $X_1, \dots, X_n$ független és $N(\mu; \sigma^2)$ eloszlású valószínűségi változók. Jelölje $\bar{X}_n$ a hozzájuk tartozó mintaátlagot és $S_n^* = \sqrt{(S_n^*)^2}$ a korrigált empirikus szórásukat. Ekkor a

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n^*}$$

valószínűségi változó t -eloszlású $n - 1$ szabadságfokkal.

Konfidenciaintervallum σ^2 -re ismeretlen μ esetén

(15.4.1. Tétel: Lukács Jenő tétele.) Legyenek $X_1, \dots, X_n \sim N(\mu; \sigma^2)$ fae. valószínűségi változók. Ekkor a

$$\frac{(n-1)(S_n^*)^2}{\sigma^2}$$

valószínűségi változó χ^2 -eloszlású $n-1$ szabadságfokkal, továbbá $\overline{X}_n$ és $(S_n^*)^2$ függetlenek.

Ennek felhasználásával már le tudjuk vezetni a normális eloszlás ismeretlen σ^2 szórásnégyzetére vonatkozó konfidenciaintervallum képletét, ha a μ várható érték ismert.

Fő különbség a μ -re vonatkozó konfidenciaintervallumokhoz képest: σ^2 csak pozitív lehet, tehát a konfidenciaintervallum is $(0, \infty)$ részhalmaza lesz.

A χ^2 -sűrűségfüggvény nem szimmetrikus az y tengelyre.

Konfidenciaintervallum σ^2 -re ismeretlen μ esetén

Az $1 - \varepsilon$ szintű konfidenciaintervallumok közül a lehető legrövidebb $[a_\varepsilon, b_\varepsilon]$ ($0 < a_\varepsilon < b_\varepsilon < \infty$) konfidenciaintervallumot keressük.

Bizonyítás nélkül felhasználjuk, hogy a legrövidebb konfidenciaintervallumot akkor kapjuk, ha

$$\mathbb{P}_{\sigma^2}(0 < \sigma^2 < a_\varepsilon) = \mathbb{P}_{\sigma^2}(\sigma^2 > b_\varepsilon) = \frac{\varepsilon}{2}.$$

(Nem az intervallum szimmetrikus, hanem a konfidenciaintervallum alá és felé eső részre eső valószínűségek egyenlők.)

Ezt átírhatjuk úgy is, hogy

$$\mathbb{P}_{\sigma^2}\left(\frac{(n-1)(S_n^*)^2}{\sigma^2} > \frac{(n-1)(S_n^*)^2}{a_\varepsilon}\right) = \mathbb{P}_{\sigma^2}\left(\frac{(n-1)(S_n^*)^2}{\sigma^2} < \frac{(n-1)(S_n^*)^2}{b_\varepsilon}\right) = \frac{\varepsilon}{2}.$$

Jelölje $\alpha \in (0, 1)$ esetén $\chi^2_{n-1, \alpha}$ az $n - 1$ szabadságfokú χ^2 -eloszlás $1 - \alpha$ -kvantilisét.

Lukács tétele szerint $\frac{(n-1)(S_n^*)^2}{\sigma^2} \sim \chi^2(n-1)$.

Az a szám, amelynél ez a val. változó pontosan $\varepsilon/2$ valószínűséggel nagyobb, vagyis $1 - \varepsilon/2$ valószínűséggel kisebb, éppen $\chi^2_{n-1, \varepsilon/2}$.

Az a szám pedig, amelynél pont $\varepsilon/2$ valószínűséggel kisebb, vagyis

Konfidenciaintervallum σ^2 -re ismeretlen μ esetén

$$\mathbb{P}_{\sigma^2}(0 < \sigma^2 < a_\varepsilon) = \mathbb{P}_{\sigma^2}(\sigma^2 > b_\varepsilon) = \frac{\varepsilon}{2}.$$

Ezt átírhatjuk úgy is, hogy

$$\mathbb{P}_{\sigma^2}\left(\frac{(n-1)(S_n^*)^2}{\sigma^2} > \frac{(n-1)(S_n^*)^2}{a_\varepsilon}\right) = \mathbb{P}_{\sigma^2}\left(\frac{(n-1)(S_n^*)^2}{\sigma^2} < \frac{(n-1)(S_n^*)^2}{b_\varepsilon}\right) = \frac{\varepsilon}{2}.$$

Jelölje $\alpha \in (0, 1)$ esetén $\chi_{n-1, \alpha}^2$ az $n-1$ szabadságfokú χ^2 -eloszlás $1 - \alpha$ -kvantilisét.

Lukács tétele szerint $\frac{(n-1)(S_n^*)^2}{\sigma^2} \sim \chi^2(n-1)$.

Az a szám, amelynél ez a val. változó pontosan $\varepsilon/2$ valószínűséggel nagyobb, vagyis $1 - \varepsilon/2$ valószínűséggel kisebb, éppen $\chi_{n-1, \varepsilon/2}^2$.

Az a szám pedig, amelynél pont $\varepsilon/2$ valószínűséggel kisebb, vagyis pont $1 - \varepsilon/2$ valószínűséggel nagyobb, éppen $\chi_{n-1, 1-\varepsilon/2}^2$.

Tehát a fenti egyenlet pontosan akkor igaz, ha

$$\frac{(n-1)(S_n^*)^2}{a_\varepsilon} = \chi_{n-1, \varepsilon/2}^2 \quad \text{és} \quad \frac{(n-1)(S_n^*)^2}{b_\varepsilon} = \chi_{n-1, 1-\varepsilon/2}^2.$$

Konfidenciaintervallum σ^2 -re ismeretlen μ esetén

$$\frac{(n-1)(S_n^*)^2}{a_\varepsilon} = \chi_{n-1, \varepsilon/2}^2 \quad \text{és} \quad \frac{(n-1)(S_n^*)^2}{b_\varepsilon} = \chi_{n-1, 1-\varepsilon/2}^2.$$

Nagy örömünkre a kapott egyenletek **már nem függenek az ismeretlen σ^2 paramétertől.**

Átrendezés után azt kapjuk, hogy $(a_\varepsilon, b_\varepsilon)$ csak a mintától függ, így valóban egy konfidenciaintervallum:

$$[a_\varepsilon, b_\varepsilon] = \left[\frac{(n-1)(S_n^*)^2}{\chi_{n-1, \varepsilon/2}^2}, \frac{(n-1)(S_n^*)^2}{\chi_{n-1, 1-\varepsilon/2}^2} \right].$$

Összefoglalva:

(15.4.3. Állítás.) Legyenek $X_1, \dots, X_n$ fae. $N(\mu; \sigma^2)$ eloszlású val. változók, ahol a μ várható érték ismert és a σ^2 szórás ismeretlen. Ekkor $\varepsilon \in (0, 1)$ esetén a σ^2 szórásnégyzetre egy $1 - \varepsilon$ szintű konfidenciaintervallum a következő:

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = \left[\frac{(n-1)(S_n^*)^2}{\chi_{n-1, \varepsilon/2}^2}, \frac{(n-1)(S_n^*)^2}{\chi_{n-1, 1-\varepsilon/2}^2} \right],$$

ahol $\alpha \in (0, 1)$ esetén $\chi_{n-1, \alpha}^2$ az $n-1$ szabadságfokú χ^2 -eloszlás $1 - \alpha$ -kvantilise.

Konfidenciaintervallum σ^2 -re ismert μ esetén

Mi a helyzet az **ismert** várható érték esetével?

- ▶ Az eddigiek szerint az előző dián szereplő képlet ismert μ esetén is egy $1 - \varepsilon$ szintű konfidenciaintervallumot ad σ^2 -re.
- ▶ Azonban ez a konfidenciaintervallum nem optimális, mivel van nála rövidebb ugyanilyen szintű.
- ▶ Ugyanis a χ^2 -eloszlás definíciója (standard normálisok összegének eloszlása) és a standardizálás ismert módja szerint ha $X_1, \dots, X_n \sim N(\mu; \sigma^2)$ fae. valószínűségi változók, akkor $\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$ pontosan n darab független standard normális eloszlású valószínűségi változó négyzetének összege, ezért $\chi^2(n)$ -eloszlású.
- ▶ Így ismert μ esetén az előző érveléssel analóg módon új konfidenciaintervallumot kapunk σ^2 -re a következő állítás szerint.

Konfidenciaintervallum σ^2 -re ismert μ esetén

(15.4.4. Állítás.) Legyenek $X_1, \dots, X_n$ fae. $N(\mu; \sigma^2)$ eloszlású valószínűségi változók, ahol a μ várható érték ismert és a $\sigma^2 > 0$ szórásnégyzet ismeretlen. Ekkor $\varepsilon \in (0, 1)$ esetén a σ^2 szórásnégyzetre egy $1 - \varepsilon$ szintű konfidenciaintervallum a következő:

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = \left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, \varepsilon/2}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n, 1-\varepsilon/2}^2} \right],$$

ahol $\alpha \in (0, 1)$ esetén $\chi_{n, \alpha}^2$ az n szabadságfokú χ^2 -eloszlás $1 - \alpha$ -kvantilise.

Az ismert és az ismeretlen várható érték esete között tehát nem csupán a szabadsági fokban van különbség, hanem abban is, hogy míg ismeretlen várható érték esetén

$(n-1)(S_n^*)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2$ állt az intervallum-végpontok számlálóiiban, ezek helyett itt $\sum_{i=1}^n (X_i - \mu)^2$ szerepel.

Vagyis a várható érték becslését a valódi várható értékre cserélve pontosabb becslést (rövidebb konfidenciaintervallumot) kapunk.

Konfidenciaintervallum normális eloszlás várható értékére ismeretlen szórásnégyzet esetén

Hátravan még: konfidenciaintervallum μ -re, ha σ^2 ismeretlen.

Az ismert σ^2 szórásnégyzetű normális eloszlás μ várható értékére vett konfidenciaintervallum meghatározásakor láttuk, hogy

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \sim N(0; 1),$$

ami fontos volt a konfidenciaintervallum meghatározásához.

15.3.6. Állítás $\Rightarrow$ ha a σ szórást az S_n^* empirikus szórásra cseréljük, akkor $N(0; 1)$ helyett $n - 1$ szabadságfokú t -eloszlást kapunk:

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n^*} \sim t(n - 1).$$

Miért nagyszerű ez? Azért, mert az utóbbi egyenletben egyáltalán nem szerepel az ismeretlen σ szórás.

$\Rightarrow$ Ugyanaz lesz a képlet a konfidenciaintervallumra, mint ismert σ esetén, de σ helyett S_n^* -gal és az $N(0; 1)$ helyett a $t(n - 1)$ -eloszlás $1 - \varepsilon/2$ -kvantilisével.

Konfidenciaintervallum μ -re ismeretlen σ^2 esetén

(15.5.1. Állítás.) Legyenek $X_1, \dots, X_n$ fae. $N(\mu; \sigma^2)$ eloszlású valószínűségi változók, ahol a μ várható érték és a $\sigma^2 > 0$ szórásnégyzet ismeretlen, és jelölje $\overline{X}_n$ a mintához tartozó mintaátlagot, S_n^* pedig a mintához tartozó korrigált empirikus szórást.

Ekkor $\varepsilon \in (0, 1)$ esetén a μ várható értékre egy $1 - \varepsilon$ szintű konfidenciaintervallum a következő:

$$[T_1(\mathbf{X}), T_2(\mathbf{X})] = \left[\overline{X}_n - \frac{S_n^* t_{n-1, \varepsilon/2}}{\sqrt{n}}, \overline{X}_n + \frac{S_n^* t_{n-1, \varepsilon/2}}{\sqrt{n}} \right],$$

ahol $\alpha \in (0, 1)$ esetén $t_{n-1, \alpha}$ az $n - 1$ szabadságfokú t -eloszlás $1 - \alpha$ -kvantilisét jelöli.

Emlékeztető: nagy f szabadságfok ($\approx$ legalább 30) esetén a $t(f)$ -eloszlás közelíthető a standard normálissal.

Hipotézisvizsgálat

Hipotézisvizsgálat $\rightarrow$ valamilyen előzetes **hipotézisünket** (azaz feltevésünket) akarjuk valamilyen mérés, illetve adatelemzés alapján igazolni vagy cáfolni.

Ilyen hipotézis lehet például:

- ▶ A mintaelemek várható értéke 2.
- ▶ A mintaelemek várható értéke legfeljebb 2.
- ▶ A mintaelemek várható értéke megegyezik egy adott másik minta elemeinek várható értékével.
- ▶ A mintaelemek szórása 8.
- ▶ A mintaelemek normális eloszlásúak.
- ▶ A realizációk függetlenek valamely más, kapcsolódó mérés eredményeitől.

Ebben a bevezető statisztikai részben főleg az első háromról lesz szó, az MSc-s Matematikai statisztikai tárgybán a többiről is.

Statisztikai próbák

A mintavétel után számítások segítségével megvizsgáljuk, hogy a mérési eredmények a hipotézisünknek ellentmondanak-e, azaz hogy a hipotézisünket feltéve a mérési eredmények nagyon valószínűtlenek-e. Ha igen, akkor a hipotézist elutasítjuk.

Általános esetben a hipotézisvizsgálati eljárás (röviden **statisztikai próba**) az alábbi lépéssorozat végrehajtását jelenti:

1. A H_0 -lal jelölt *nullhipotézis* formalizálása.
2. A mintára vonatkozó *elfogadási tartomány* és *kritikus tartomány* megalkotása.
3. A kísérlet végrehajtása, amelyben az $x_1, \dots, x_n$ realizációt (adatpontokat) kapjuk.
4. Annak megvizsgálása, hogy az adatpontok az elfogadási tartományba esnek-e. Ha igen, H_0 -t elfogadjuk. Ellenkező esetben H_0 -t elutasítjuk.*

* Léteznek ún. *szekvenciális próbák* is, ahol a 4. pontnak „nem meggyőző köztes eredmény” esetén lehet olyan kimenetele is, hogy újabb mintát kell venni
→ ciklus indul újra 3.-tól. Pl.: Wald–Wolfowitz-próba.

Paraméteres és nemparaméteres próbák

Ebből a tárgyból alapvetően **paraméteres próbák**kal foglalkozunk
→ az eloszlás típusa ismert (pl. normális), van egy vagy több ismeretlen paraméter, ami(k)re a hipotézis vonatkozik.

Pl. u -próba, több változatban → normális eloszlás várható értékére vonatkozó próbák ismert szórás esetén,
 t -próba: ugyanez ismeretlen szórás esetén
(+más példák, amiket itt nem tanulunk, pl. F -próba).

MSc-s Matematikai statisztika tárgy → **nemparaméteres próbák** is, amik a konkrét eloszlástípus ismerete nélkül is használhatóak:

- ▶ függetlenségvizsgálat,
- ▶ illeszkedésvizsgálat (egy bizonyos adott eloszlást követ-e a minta?),
- ▶ homogenitásvizsgálat (két minta peremeloszlásai azonosak-e?).

Pl. χ^2 -próba, Kolmogorov–Szmirnov-próba.

Hipotézisvizsgálati alapfogalmak

(16.1.1. Def.) Legyen $\mathbf{X} = (X_1, \dots, X_n)$ egy fae. minta az $(\Omega, \mathcal{F}, \mathbb{P})$ val. mezőn, ahol a mintaelemek peremeloszlása ismeretlen. Legyen $\mathbf{x} = (x_1, \dots, x_n)$ az $\mathbf{X} = (X_1, \dots, X_n)$ egy realizációja.

- (1) Egy H_0 **nullhipotézis** egy olyan kijelentés, amelynek igazsága csak az $X_1, \dots, X_n$ együttes eloszlásától függ.

A H_1 **ellenhipotézis** a H_0 tagadása: $H_1 = \neg H_0$.

Az adott próbához tartozó **alternatíva**¹ a

$$H_0 \text{ vs. } H_1.$$

- (2) Egy $\mathfrak{X}_e$ **elfogadási tartomány** az $\mathbb{R}^n$ egy részhalmaza. Az $\mathfrak{X}_e$ elfogadási tartományhoz tartozó $\mathfrak{X}_k$ **kritikus tartomány** az $\mathfrak{X}_e$ komplementere, azaz $\overline{\mathfrak{X}_e} = \mathbb{R}^n \setminus \mathfrak{X}_e$.

A $\mathfrak{X}_k$ kritikus tartománnyal definiált **(statisztikai) próba** az az eljárás, amely során H_0 -t elfogadjuk, ha $\mathbf{X} \in \mathfrak{X}_e$ és H_0 -t elutasítjuk (azaz H_1 -et fogadjuk el), ha $\mathbf{X} \in \mathfrak{X}_k$.

¹vs.: ejtsd „versus” (latinul „ellen”).

Hipotézisvizsgálati alapfogalmak (folytatás)

- (3) A $\mathfrak{X}_k$ kritikus tartománnyal definiált próba **terjedelme** a $0 < \alpha < 1$ szám, ha

$$\mathbb{P}((X_1, \dots, X_n) \in \mathfrak{X}_k | H_0 \text{ igaz}) = \alpha.$$

- (4) **Elsőfajú hibának** nevezzük azt, amikor egy próba eljárása során H_0 -t elutasítjuk, pedig H_0 igaz.

Másodfajú hibának nevezzük azt, amikor az eljárás során H_0 -t elfogadjuk, pedig H_1 igaz (azaz H_0 hamis).

Egy α terjedelmű próba elsőfajú hibájának valószínűsége éppen a terjedelem:

$$\mathbb{P}(H_0\text{-t elutasítjuk} | H_0 \text{ igaz}) = \mathbb{P}((X_1, \dots, X_n) \in \mathfrak{X}_k | H_0 \text{ igaz}) = \alpha.$$

A másodfajú hiba valószínűségének kiszámítására viszont nem mindig áll rendelkezésre explicit módszer, amint azt konkrét példákban látni fogjuk.

Interpretáció – Bolla Marianna nyomán

Matematikailag nem precíz, de a szemléltetéshez segíthet:

(16.1.3. Megjegyzés.) A H_0 nullhipotézist gyakran úgy fogalmazzuk meg, hogy annak teljesülése esetén ismert legyen az $X_1, \dots, X_n$ mintaelemek eloszlása.

Pl. utcán talált pénzérme $\rightarrow$ **jó nullhipotézis** a H_0 : „a pénzérme szabályos” kijelentés, mert ekkor annak ismeretében, hogy H_0 igaz, hogy a mintaelemek fae. $1/2$ paraméterű indikátorváltozók.

Nem jó nullhipotézis például a H'_0 : „a pénzérme nem szabályos” kijelentés. Mi ekkor a fae. indikátorok paramétere? Terjedelem?

Interpretáció 2.

(16.1.4. Megjegyzés.) A nullhipotézist sokszor „az ártatlanság vélelmében” fogalmazzuk meg.

Pl. a pénzértmét tipikusan azért vizsgáljuk meg, mert arra gyanakodunk, hogy cinkelt, de mégis ezen állítás tagadását tesszük meg H_0 -nak. Ekkor

- ▶ az elsőfajú hiba úgy interpretálható, mint az ártatlan elítélése,
- ▶ a másodfajú hiba pedig úgy, mint a bűnös felmentése.

A legtöbb ismert próbában csak az elsőfajú hiba valószínűségét, vagyis a terjedelmet tudjuk kontrollálni: előre megadunk egy α értéket $\rightarrow$ a próba konstrukciója alapján a terjedelem éppen α lesz. Érezhető (a próbák ismerete nélkül is): ha az előírányzott terjedelmet nagyon kicsinek választjuk, akkor a másodfajú hiba valószínűsége megnő.

(„Ha biztosra akarunk menni, hogy ártatlanokat ne ítéljünk el, akkor bűnösöket is gyakrabban fogunk felmenteni.”)

A gyakorlatban $\varepsilon = 0.05$ és $\varepsilon = 0.01$ a leggyakrabban választott terjedelemértékek.

Elsőfajú és másodfajú hiba: példa

16.1.5. Példa: Egy új gyógyszert szeretnénk forgalomba hozni.

Ekkor H_0 : „a gyógyszer hatástalan vagy káros” vs. H_1 : „a gyógyszer hatásos” jó alternatíva.

Itt: elsőfajú hiba $\Leftrightarrow$ a mintavétel alapján egy hatástalan vagy káros gyógyszert hatásosnak mondunk (és ez alapján vélhetően forgalomba hozzuk).

Mivel ez az elsőfajú hiba, ennek valószínűségét tudjuk kontrollálni, tetszőlegesen alacsony $1 - \alpha$ szint alatt tarthatjuk, a másodfajú hiba valószínűségének megnövekedésének árán.

Másodfajú hiba $\Leftrightarrow$ a gyógyszer hatásos, de a mintavétel alapján mégis hatástalannak vagy károsnak nyilvánítjuk (és ezért bizonyára nem vezetjük be).

Természetesen a másodfajú hiba bekövetkezése sem kedvező (pl. gazdasági szempontból), de nem jár olyan fatális következményekkel, mint az elsőfajú hiba.

Ha csak az elsőfajú hiba valószínűsége kontrollálható, akkor H_0 -t a fenti módon kell megválasztani.

Kétoldali, egymintás μ -próba: példa/levezetés

(Bolla Marianna példája nyomán.)

A sarki péknél mindennap 1 db egykilós kenyeret veszünk. Az utóbbi időben mintha kisebbek lennének a kenyerek $\rightarrow$ tényleg 1 kilósak várható értékben?

Feltesszük: a kenyér tömege (kg-ban mérve) normális eloszlású $\sigma = 0,02$ kg szórással, tehát $N(\mu; 0,02^2)$ eloszlású. Kérdés: $\mu \stackrel{?}{=} \mu_0$.

Alternatíva („a H_0 nullhipotézis megválasztása az ártatlanság vélelmében”):

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

ahol esetünkben $\mu_0 = 1$ kg.

Előírányzott **terjedelem**: $\varepsilon = 0,05$,

vagyis azt szeretnénk, hogy a próba során legfeljebb 0,05 valószínűséggel utasítsuk el H_0 -t, amennyiben H_0 igaz.

Kétoldali, egymintás μ -próba: példa/levezetés

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

Mintát veszünk: $n = 25$ napon keresztül az aznap vásárolt kenyeret mindig mérlegre tesszük.

Kapott $\mathbf{x} = (x_1, \dots, x_{25})$ realizáció átlaga: $\bar{x}_n = \bar{x}_{25} = 0,98$ kg.

Cél: olyan próbát konstruálni, aminek a terjedelme pontosan $\varepsilon = 0,05$, azaz olyan $\mathfrak{X}_e$ elfogadási tartományt és $\mathfrak{X}_k$ kritikus tartományt keresünk, amelyekre

$$\mathbb{P}(H_0\text{-t elutasítjuk} | H_0) = \varepsilon = 0,05.$$

Az $\mathfrak{X}_e$ elfogadási tartományt keressük **olyan alakban, hogy H_0 -t akkor fogadjuk el, ha μ_0 egy $[\bar{X}_n - h, \bar{X}_n + h]$ alakú intervallumba esik**, ahol $h > 0$. Ekkor h -t úgy kívánjuk megválasztani, hogy $\mu = \mu_0 = 1$ esetén éppen ε valószínűséggel essen **a minta** a kritikus tartományba, azaz $1 - \varepsilon$ valószínűséggel essen az elfogadási tartományba, vagyis teljesüljön, hogy

$$\mathbb{P}_{\mu_0}(\mu_0 \in [\bar{X}_n - h, \bar{X}_n + h]) = 1 - \varepsilon.$$

Kétoldali, egymintás u -próba: példa/levezetés

Az eddigiek alapján $[\overline{X}_n - h, \overline{X}_n + h]$ -re teljesül, hogy

$$\mathbb{P}_{\mu_0}(\mu_0 \in [\overline{X}_n - h, \overline{X}_n + h]) = 1 - \varepsilon.$$

Azaz $[\overline{X}_n - h, \overline{X}_n + h]$ egy $\overline{X}_n$ körül szimmetrikus, $1 - \varepsilon$ szintű **konfidenciaintervallum** kell, hogy legyen a μ_0 paraméterre.

Láttuk: egy ilyen konfidenciaintervallum

$$\left[\overline{X}_n - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}, \overline{X}_n + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \right],$$

amiből tehát azt kapjuk, hogy $h = \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}}$. Vagyis H_0 -t pontosan akkor fogadjuk el, ha

$$\begin{aligned} \overline{X}_n - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \leq \mu_0 \leq \overline{X}_n + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} &\Leftrightarrow \mu_0 - \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \leq \overline{X}_n \leq \mu_0 + \frac{\sigma u_{\varepsilon/2}}{\sqrt{n}} \\ &\Leftrightarrow -u_{\varepsilon/2} \leq \frac{\overline{X}_n - \mu_0}{\sigma} \sqrt{n} \leq u_{\varepsilon/2}. \end{aligned}$$

Az elfogadási tartományunk emiatt

$$\mathfrak{X}_e = \left\{ \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n : -u_{\varepsilon/2} \leq \frac{\overline{X}_n - \mu_0}{\sigma} \sqrt{n} \leq u_{\varepsilon/2} \right\}.$$

Kétoldali, egymintás u -próba: példa/levezetés

Az elfogadási tartományunk

$$\mathfrak{X}_e = \left\{ \mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n : -u_{\varepsilon/2} \leq \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n} \leq u_{\varepsilon/2} \right\}.$$

Ha tehát az $u(\mathbf{X}) = \frac{\bar{X}_n - \mu_0}{\sigma} \sqrt{n}$ próbastatisztika $u(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$ realizációja $-u_{\varepsilon/2}$ és $u_{\varepsilon/2}$ közé esik, akkor H_0 -t elfogadjuk, különben pedig elutasítjuk (azaz H_1 -et fogadjuk el).

A konkrét esetünkben $\varepsilon = 0,05$, $n = 25$, $\bar{x}_n = 0.98$, $\mu_0 = 1$ és $\sigma = 0.03$. Ez alapján tehát $u(\mathbf{x}) = \frac{0.98-1}{0.02} \sqrt{25} = -5$.

Normális eloszlás táblázata $\Rightarrow u_{\varepsilon/2} = \Phi^{-1}(0,975) \approx 1,9600$.

Tehát H_0 -t akkor fogadnánk el, ha az $u(\mathbf{X})$ próbastatisztika értéke $-1,96$ és $1,96$ közé esne.

Mivel ez nem áll fenn, H_0 -t elutasítjuk, vagyis kellő bizonyítékot találtunk arra, hogy a kenyerek várható értékben nem egykilósak.

Kétoldali, egymintás u -próba: eljárás

Feltevés: az $x_1, \dots, x_n$ adatpontok fae. normális eloszlású $X_1, \dots, X_n$ valószínűségi változók realizációi, amelyek μ várható értéke ismeretlen, de $\sigma > 0$ szórásuk ismert.

Általános eljárás (nem feltétlenül egyoldali) egymintás u -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az elfogadási tartomány meghatározása (függ a próba altípusától).
5. A próbastatisztika $u(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$ értékének meghatározása.
6. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk (tehát H_1 -et fogadjuk el).

Kétoldali eset. Nullhipotézis: $H_0: \mu = \mu_0$.

Releváns kvantilis: $u_{\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$.

Elfogadási tartomány: $\mathfrak{X}_e = \{(x_1, \dots, x_n): |u(\mathbf{x})| \leq u_{\varepsilon/2}\}$, tehát:

- ▶ ha $|u(\mathbf{x})| > u_{\varepsilon/2}$, H_0 -t elutasítjuk,
- ▶ különben H_0 -t elfogadjuk.

Egyoldali, egymintás μ -próba: példa/levezetés

Térjünk vissza a kenyérvásárlás példájához.

Jogos kérdés: miért azt tettük meg H_1 -nek, hogy $\mu \neq \mu_0$?

Ha a mintaátlag kisebb, mint 1 (pl. 0,98 kg), akkor józan észszel arra gyanakszunk, hogy a kenyér súlyának várható értéke **kisebb**, mint 1 kg. Arra nem, hogy nagyobb.

Egy másik lehetséges alternatíva:

$$H_0: \mu \geq \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0, \quad (5)$$

ahol μ_0 változatlanul 1-gyel egyenlő.

Kérdés: ha tudjuk, hogy H_0 teljesül, akkor a mintaelemek eloszlása is ismert?

Szigorú értelemben nem. H_0 teljesülése esetén lehetnek a mintaelemek például $N(\mu_0; \sigma^2)$ és $N(\mu_0 + 113; \sigma^2)$ eloszlásúak is. Azonban a H_0 teljesülése esetén egy μ_0 -nál kisebb átlagú mintát akkor kapunk a legnagyobb valószínűséggel, ha μ megegyezik μ_0 -lal (nem pedig nagyobb nála).

Egyoldali, egymintás μ -próba: példa/levezetés

Térjünk vissza a kenyérvásárlás példájához.

Jogos kérdés: miért azt tettük meg H_1 -nek, hogy $\mu \neq \mu_0$?

Ha a mintaátlag kisebb, mint 1 (pl. 0,98 kg), akkor józan észszel arra gyanakszunk, hogy a kenyér súlyának várható értéke **kisebb**, mint 1 kg. Arra nem, hogy nagyobb.

Egy másik lehetséges alternatíva:

$$H_0: \mu \geq \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0, \quad (5)$$

ahol μ_0 változatlanul 1-gyel egyenlő.

Kérdés: ha tudjuk, hogy H_0 teljesül, akkor a mintaelemek eloszlása is ismert?

Szigorú értelemben nem. H_0 teljesülése esetén lehetnek a mintaelemek például $N(\mu_0; \sigma^2)$ és $N(\mu_0 + 113; \sigma^2)$ eloszlásúak is. Azonban a H_0 teljesülése esetén egy μ_0 -nál kisebb átlagú mintát akkor kapunk a legnagyobb valószínűséggel, ha μ megegyezik μ_0 -lal (nem pedig nagyobb nála).

Egyoldali, egymintás u -próba: példa/levezetés

$$H_0: \mu \geq \mu_0 \quad \text{vs.} \quad H_1: \mu < \mu_0,$$

Ha egy tetszőleges A eseményre – a jelöléssel enyhén visszaélve – bevezetjük a

$$\mathbb{P}(A|H_0 \text{ igaz}) := \mathbb{P}(A|\text{a mintaelemek várható értéke } \mu_0)$$

jelölést, vagyis úgy fogjuk fel, hogy H_0 teljesülése esetén a várható érték egyenlő μ_0 -al, akkor a kétoldali, egymintás u -próbához hasonlóan járhatunk el.

A próbastatisztika most is $u = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$ lesz, de H_0 -t most akkor utasítjuk el, ha u értéke **nagyon kicsi**.

Elfogadási tartomány: jobbról végtelen, és $\mu = \mu_0$ esetén $1 - \varepsilon$ valószínűséggel esik μ_0 az $\mathfrak{X}_e$ elfogadási tartományba, tehát ε valószínűséggel a tőle balra eső $\mathfrak{X}_k$ kritikus tartományba:

$$\mathfrak{X}_k = \{\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n : u(\mathbf{x}) < -u_\varepsilon\},$$

ahol $u_\varepsilon = \Phi^{-1}(1 - \varepsilon)$, ahogy eddig.

Egyoldali, egymintás u -próba: példa, előnyök, változatok

Példánkban $u(\mathbf{x}) = -5$ -öt kaptunk. A standard normális eloszlás $1 - \varepsilon = 0,95$ -kvantilise $\Phi^{-1}(1 - \varepsilon) = \Phi^{-1}(0,95) \approx 1,6449$.

Mivel $u(\mathbf{x}) = -5 < -u_\varepsilon$, H_0 -t ismét elutasítjuk, vagyis azt kapjuk, hogy a kenyerek tömege várható értékben kisebb, mint 1 kg.

Az egyoldali, egymintás u -próba tehát μ_0 -nál nagyobb átlagú minta esetén soha nem utasítja el H_0 -t, viszont μ_0 -nál kisebb átlagú minták esetén **szigorúbb**, mint a kétoldali, egymintás próba. A példánkban ha $u(\mathbf{x})$ a -1.9600 és a -1.6449 számok közé esik, akkor az egyoldali próba már elutasítja H_0 -t, de a kétoldali még elfogadja.

Ez az egyoldali próba fő előnye.

Ha viszont a $H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$

alternatívára kívánunk egyoldali, egymintás u -próbát szerkeszteni, akkor a szimmetria miatt az

$$\mathfrak{X}_k = \{\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n: u(\mathbf{x}) > u_\varepsilon\}$$

kritikus tartományt kapjuk. H_0 -t elutasítjuk $\Leftrightarrow u(\mathbf{x})$ nagyon nagy.

Kétoldali, egymintás u -próba: eljárás

Feltevés: az $x_1, \dots, x_n$ adatpontok fae. normális eloszlású $X_1, \dots, X_n$ valószínűségi változók realizációi, amelyek μ várható értéke ismeretlen, de $\sigma > 0$ szórásuk ismert.

Általános eljárás (nem feltétlenül egyoldali) egymintás u -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az elfogadási tartomány meghatározása (függ a próba altípusától).
5. A próbastatisztika $u(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$ értékének meghatározása.
6. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk (tehát H_1 -et fogadjuk el).

Egyoldali eset ($H_0: \mu \geq \mu_0$ változat). Nullhipotézis: $H_0: \mu \geq \mu_0$.

Releváns kvantilis: $-u_\varepsilon = -\Phi^{-1}(1 - \varepsilon)$.

Elfogadási tartomány: $\mathfrak{X}_\varepsilon = \{(x_1, \dots, x_n): u(\mathbf{x}) \geq -u_\varepsilon\}$, tehát:

- ▶ ha $u(\mathbf{x}) < -u_\varepsilon$, H_0 -t elutasítjuk,
- ▶ különben H_0 -t elfogadjuk.

Kétoldali, egymintás u -próba: eljárás

Feltevés: az $x_1, \dots, x_n$ adatpontok fae. normális eloszlású $X_1, \dots, X_n$ valószínűségi változók realizációi, amelyek μ várható értéke ismeretlen, de $\sigma > 0$ szórásuk ismert.

Általános eljárás (nem feltétlenül egyoldali) egymintás u -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az elfogadási tartomány meghatározása (függ a próba altípusától).
5. A próbastatisztika $u(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{\sigma} \sqrt{n}$ értékének meghatározása.
6. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk (tehát H_1 -et fogadjuk el).

Egyoldali eset ($H_0: \mu \leq \mu_0$ változat). Nullhipotézis: $H_0: \mu \leq \mu_0$.

Releváns kvantilis: $u_\varepsilon = \Phi^{-1}(1 - \varepsilon)$.

Elfogadási tartomány: $\mathfrak{X}_e = \{(x_1, \dots, x_n): u(\mathbf{x}) \leq u_\varepsilon\}$, tehát:

- ▶ ha $u(\mathbf{x}) > u_\varepsilon$, H_0 -t elutasítjuk,
- ▶ különben H_0 -t elfogadjuk.

Próba ereje, az u -próba optimalitása (kieg. anyag)

Láttuk: a kétoldali, egymintás u -próba terjedelme, vagyis elsőfajú hibájának valószínűsége legfeljebb az előzetesen választott ε szám lesz.

Kérdés: Mit mondhatunk a másodfajú hibáról?

Ha tudjuk, hogy H_1 teljesül $\Rightarrow$ mintaelemek eloszlása??

Vegyünk most egy $\mu_1 \neq \mu_0$ paramétert, ahol μ_0 a nullhipotézis szerinti várható érték. Tehát ha μ_1 a valódi paraméter, akkor H_1 igaz. Ha tudnánk, hogy μ_1 a valódi paraméter, akkor a másodfajú hiba képlete ez lenne:

$$\mathbb{P}(H_0\text{-t elfogadjuk} \mid \mu = \mu_1) = \mathbb{P}(u(\mathbf{X}) \in (-u_{\varepsilon/2}, u_{\varepsilon/2}) \mid X_1 \sim N(\mu_1; \sigma^2)).$$

Ebben az esetben a próba **erejének** 1 mínusz a másodfajú hiba valószínűségét nevezzük.

Az erő tehát itt csak konkrét H_1 -beli paraméter esetén definiálható. Minél távolabb μ_1 a μ_0 értéktől, annál kisebb lesz a másodfajú hiba valószínűsége, vagyis annál nagyobb lesz a próba ereje.

Egyenletesen legerősebb próbák (kieg. anyag)

Miért a (kétoldali, egymintás) u -próbát alkalmazzuk a gyakorlatban és nem valamilyen más, szintén ε terjedelmű próbát ugyanerre az alternatívára (feltételezve, hogy a mintaelemek eloszlása normális és a σ szórásuk adott)?

Azért, mert bizonyítható, hogy az u -próba az **egyenletesen legerősebb** próba valamennyi ilyen próba közül.

Vagyis: bárhogy is választunk egy H_1 -beli μ_1 paramétert, $\mu = \mu_1$ esetén a másodfajú hiba akkor lesz a legkisebb, ha az u -próbát alkalmazzuk.

(Nem bizonyítjuk, sőt ki sem mondjuk formálisan.

Megjegyzés: ez a **Neyman–Pearson-lemma** következménye.

A Neyman–Pearson-lemma az egyenletesen legerősebb próba létezését mondja ki és konstrukcióját adja meg általános esetben → lásd BME-TTK-s matematikai statisztika tárgyak.)

Ilyen értelemben egyenletesen legerősebb próba az összes u - és t -próba, amit az előadáson tárgyalunk, mindegyik a maga alternatívájára vonatkozóan, a maga feltevései mellett.

Az egymintás u -próba alkalmazásai

Az u -próba esetenként olyan esetekben is alkalmazható, ahol a mintaelemek ugyan nem normális eloszlásúak, de a CHT/de Moivre–Laplace-tétel miatt azzal közelíthetők.

(Példa: u -próba binomiális eloszlásra, 16.2.4. Példa)

Megjegyzés: p -érték

Ha ε -t csökkentjük, akkor H_0 -t gyakrabban fogjuk elfogadni; ez nem meglepő, hiszen ε a próba terjedelmével, vagyis az elsőfajú hiba valószínűségével egyezik meg.

A pénzérmés 16.2.4. példában $\varepsilon = 0.01$ és $\varepsilon = 0.05$ között van az a legnagyobb terjedelem, amelyre H_0 -t az adott mintának megfelelő $u(\mathbf{x})$ próbastatisztika-érték esetén még elfogadjuk.

Ezt szokás a próba p -értékének (angolul p -value) nevezni. Minél kisebb a p -érték, annál jogosabb a H_0 nullhipotézis elutasítása.

Ebben a példában (egyoldali, egymintás u -próba) a p -érték éppen 1 mínusz Φ -nek a próbastatisztikában vett értéke:

$$u(\mathbf{x}) = u_\varepsilon \Leftrightarrow \mathbb{P}(u(\mathbf{X}) > u(\mathbf{x})) = \varepsilon \Leftrightarrow \varepsilon = 1 - \Phi(u(\mathbf{x})) = 1 - \Phi(2) \approx 0.0228.$$

Vagyis a próba p -értéke kb. 2,28%. Maximum ekkora terjedelemnél H_0 -t elfogadjuk, ennél nagyobb terjedelem esetén elutasítjuk.

Kétoldali u -próba esetén, például a kenyeres példában:

$$|u(\mathbf{x})| = u_{\varepsilon/2} \Leftrightarrow \varepsilon = 2(1 - \Phi(|u(\mathbf{x})|)) = 2(1 - \Phi(5)) \approx 5.733 \times 10^{-7}.$$

Vagyis a próba p -értéke rendkívül alacsony.

Megjegyzés: p -érték

p -érték használata $\rightarrow$ alternatív szemlélet a terjedelemhez képest. Itt nem rögzítjük a mintavétel előtt a terjedelmet, hanem a minta alapján adjuk meg azt az ε terjedelmet, amely pont a H_0 elfogadása és elutasítása határán áll, azaz a p -értéket.

Gyakori szóhasználat (nemcsak u -próbánál): hogy valamilyen mintára vonatkozó állítás „szignifikáns”, ha a p -érték legfeljebb 0,05, és „nagyon szignifikáns”, ha legfeljebb 0,01. Ha elfogadjuk a nullhipotézist, max. 0,05, ill. 0,01 lehet az elsőfajú hiba.

t -próba

Kérdés: egyoldali/kétoldali próba a normális eloszlás várható értékére ismeretlen szórás esetén?

A gyakorlatban sokszor nehéz megtippelni, hogy mi a szórás, míg a minta (korrigált) empirikus szórása könnyen kiszámítható.

Ahogy említettük: a t -próba hasonló nullhipotézisekre vonatkozik, mint az u -próba (a normális eloszlás várható értékével kapcsolatban), de ismeretlen szórás esetén.

- ▶ Az u - és a t -próba között hasonló a viszony, mint a konfidenciaintervallumok ismert és ismeretlen szórás esetére vonatkozó képletei között.
- ▶ Egymintás próbánál a szórást az empirikus szórás helyettesíti. A próbastatisztika most t -eloszlású lesz standard normális helyett, szabadsági foka: mintaelemszám -1 .
- ▶ Kétmintás eset $\rightarrow$ lásd később.

Kétoldali, egymintás t -próba

Legyen $\mathbf{X} = (X_1, \dots, X_n)$ fae. minta $N(\mu; \sigma^2)$ eloszlású mintaelemekkel, ahol ezúttal mind a μ várható érték, mind a σ^2 szórásnégyszet ismeretlen. Kétoldali, egymintás esetben az alternatívánk továbbra is $H_0: \mu = \mu_0$ vs. $\mu \neq \mu_0$.

A kétoldali, egymintás, ε terjedelmű u -próba alapja az volt, hogy

$$u(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \sim N(0; 1),$$

és ez alapján ha $|u(\mathbf{X})| \leq u_{\varepsilon/2}$, akkor H_0 -t elfogadjuk, különben elutasítjuk. Ugyanakkor a konf.intervallumoknál említettük, hogy

$$t(\mathbf{X}) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{S_n^*} \sim t(n-1),$$

ahol $S_n^* = \sqrt{(S_n^*)^2}$ a mintához tartozó korrigált empirikus szórást jelöli. u -próba levezetésével analóg érvelés $\Rightarrow$

akkor kapunk ε terjedelmű próbát, ha H_0 -t akkor fogadjuk el, ha $|t(\mathbf{X})| \leq t_{n-1, \varepsilon/2}$, különben pedig elutasítjuk, ahol $t_{n-1, \varepsilon/2}$ az $n-1$ szabadságfokú t -eloszlás $1 - \varepsilon/2$ -kvantilise.

Egyoldali változat: hasonlóan.

Egy- vagy kétoldali, egymintás t -próba: eljárás

Feltevés: az $x_1, \dots, x_n$ adatpontok $X_1, \dots, X_n$ fae. normális eloszlású val. változók realizációi, amelyek szórása és μ várható értéke ismeretlen. Eljárás egymintás t -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az $f = n - 1$ szabadsági fok meghatározása.
5. Az elfogadási tartomány meghatározása, **függ a próba altípusától** $\rightarrow n - 1$ szabadságfokú t -eloszlás táblázata (hasonló a normális eloszláséhoz, 0-ra szimmetrikus sűrűv.).
6. A próbastatisztika $t(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{s_n^*} \sqrt{n}$ értékének meghatározása.
7. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk.

Egy- vagy kétoldali, egymintás t -próba: eljárás

Feltevések: az $x_1, \dots, x_n$ adatpontok $X_1, \dots, X_n$ fae. normális eloszlású val. változók realizációi, amelyek szórása és μ várható értéke ismeretlen.

Eljárás egymintás t -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az $f = n - 1$ szabadsági fok meghatározása.
5. Az elfogadási tartomány meghatározása, **függ a próba altípusától** $\rightarrow$ $n - 1$ szabadságfokú t -eloszlás táblázata (hasonló a normális eloszláséhoz, 0-ra szimmetrikus sűrűv.).
6. A próbastatisztika $t(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{s_n^*} \sqrt{n}$ értékének meghatározása.
7. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk.

Kétoldali, egymintás t -próba. Alternatíva: $H_0: \mu = \mu_0$ vs. $H_1: \mu \neq \mu_0$.

Releváns kvantilis: $t_{n-1, \varepsilon/2}$ az $n - 1$ szabadságfokú t -eloszlás $1 - \varepsilon/2$ -kvantilise. Elfogadási tartomány: $\mathfrak{X}_e = \{(x_1, \dots, x_n): |t(\mathbf{x})| \leq t_{n-1, \varepsilon/2}\}$.

Tehát: ha $|t(\mathbf{x})| > t_{n-1, \varepsilon/2}$, H_0 -t elutasítjuk, különben elfogadjuk.

Egy- vagy kétoldali, egymintás t -próba: eljárás

Feltevések: az $x_1, \dots, x_n$ adatpontok $X_1, \dots, X_n$ fae. normális eloszlású val. változók realizációi, amelyek szórása és μ várható értéke ismeretlen.

Eljárás egymintás t -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az $f = n - 1$ szabadsági fok meghatározása.
5. Az elfogadási tartomány meghatározása, **függ a próba altípusától** $\rightarrow$ $n - 1$ szabadságfokú t -eloszlás táblázata (hasonló a normális eloszláséhez, 0-ra szimmetrikus sűrűségfüggvény).
6. A próbastatisztika $t(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{s_n^*} \sqrt{n}$ értékének meghatározása.
7. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk.

Egyoldali, egymintás t -próba $H_0: \mu \geq \mu_0$ vs. $H_1: \mu < \mu_0$ alternatívával.

Releváns kvantilis: $-t_{n-1, \varepsilon}$ az $n - 1$ szabadságfokú t -eloszlás $1 - \varepsilon$ -kvantilisének (-1) . Elfogadási tart.: $\mathcal{X}_e = \{(\mathbf{x}_1, \dots, \mathbf{x}_n): t(\mathbf{x}) \geq -t_{n-1, \varepsilon}\}$.

Tehát: ha $t(\mathbf{x}) < -t_{n-1, \varepsilon}$, H_0 -t elutasítjuk, különben elfogadjuk.

Egy- vagy kétoldali, egymintás t -próba: eljárás

Feltevések: az $x_1, \dots, x_n$ adatpontok $X_1, \dots, X_n$ fae. normális eloszlású val. változók realizációi, amelyek szórása és μ várható értéke ismeretlen.

Eljárás egymintás t -próbánál:

1. A H_0 nullhipotézis felállítása.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_n$ adatpontok kinyerése.
4. Az $f = n - 1$ szabadsági fok meghatározása.
5. Az elfogadási tartomány meghatározása, **függ a próba altípusától** $\rightarrow$ $n - 1$ szabadságfokú t -eloszlás táblázata (hasonló a normális eloszláséhez, 0-ra szimmetrikus sűrűségf.)).
6. A próbastatisztika $t(\mathbf{x}) = \frac{\bar{x}_n - \mu_0}{s_n^*} \sqrt{n}$ értékének meghatározása.
7. Annak eldöntése, hogy a próbastatisztika értéke az elfogadási tartományba esik-e. Ha igen, H_0 -t elfogadjuk. Ha nem, H_0 -t elutasítjuk.

Egyoldali, egymintás t -próba $H_0: \mu \leq \mu_0$ vs. $H_1: \mu > \mu_0$ alternatívával.

Releváns kvantilis: $t_{n-1, \varepsilon}$ az $n - 1$ szabadságfokú t -eloszlás $1 - \varepsilon$ -kvantilise. Elfogadási tartomány: $\mathfrak{X}_e = \{(x_1, \dots, x_n): t(\mathbf{x}) \leq t_{n-1, \varepsilon}\}$.

Tehát: ha $t(\mathbf{x}) > t_{n-1, \varepsilon}$, H_0 -t elutasítjuk, különben elfogadjuk.

Egymintás t -próba: példa – kenyérvásárlás

Korábbi példa $\rightarrow$ a péknél vásárolt kenyerek (nagyon) szignifikánsan kisebbek voltak 1 kg-nál.

Az u -próba-hoz feltettük, hogy a normális eloszlású kenyértömeg σ szórása megegyezik a 0,02 értékkel (honnan tudjuk, hogy annyi?)

Kérdés: lehet, hogy azért utasítottuk el azt a nullhipotézist, hogy a kenyerek súlya 1 kg, mert hibás feltételezésünk volt a szórásról?

Szerencsére mindennap felírtuk, hogy aznap milyen súlyú kenyeret vettünk, így ki tudjuk számolni a korrigált empirikus szórásokat is. Péknél vett kenyerek: $n=25$ elemű $(x_1, \dots, x_n)$ mintarealizáció, ahol $\overline{x_n} = 0.98$ kg teljesült $\rightarrow$ azt kapjuk, hogy a korrigált empirikus szórás $s_{25}^* = 0.05$ kg (vagyis lényegesen nagyobb a valódi szórás korábban feltételezett 0.02 értékénél).

Futtassuk le az előbbi alternatívára u -próba helyett a t -próbát is $\rightarrow$ más döntéshez vezet-e?

Egymintás t -próba: példa – kenyérvásárlás

Kétoldali, egymintás t -próba esetén:

$$H_0: \mu = 1 \quad \text{vs.} \quad H_1: \mu \neq 1,$$

terjedelem: $\varepsilon = 0,05$. A próbastatisztika értéke:

$$t(\mathbf{x}) = \frac{\sqrt{25}(0.98 - 1)}{0.05} = -\frac{10}{5} = -2.$$

Ennek abszolút értékét kell az $n - 1 = 24$ szabadságfokú t -eloszlás $1 - \varepsilon/2 = 0,975$ -kvantilisével összehasonlítani. A kvantilis értéke

$$t_{24,0,025} \approx 2,0639.$$

Ez nagyobb, mint $u_{0,025} \approx 1.9600$, és némileg nagyobb $|t(\mathbf{x})|$ -nél is. Ezért H_0 -t elfogadjuk, vagyis a kenyerek átlagosan mégis 1 kilósok. Egyoldali, egymintás t -próba esetén:

$$H_0: \mu \geq 1 \quad \text{vs.} \quad H_1: \mu < 1,$$

terjedelem: $\varepsilon = 0,05$. $t(\mathbf{x}) = -2$, változatlanul. Ezt kell $-t_{24,0,05} \approx -1.7109$ -cel összehasonlítani. Mivel $t(\mathbf{x}) < -t_{24,0,05}$, H_0 -t a „szigorúbb” egyoldali próba alapján még mindig elutasítjuk.

t-próba R-ben

t-próba R-ben (dokumentáció): <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test>.

Kétmintás próbák

Következő téma: **kétmintás u - és t -próba** $\rightarrow$ időtakarékosági okokból csak a kétoldali verziót tárgyaljuk.

Ebből az egyoldali verziók nagyon hasonlóan vezethetők le, mint az egymintás esetben.

Itt **két független** (nem feltétlenül azonos elemszámú) fae. normális eloszlású minta lesz adott. Szeretnénk tesztelni, hogy a két minta várható értéke megegyezik-e.

- ▶ kétoldali, kétmintás u -próba: **ismert szórásnégyzetek esetén** tudunk konstruálni egy normális eloszlású próbastatisztikát,
- ▶ kétoldali, kétmintás t -próba: ha a két szórás ismeretlen, **de egyenlő**, akkor „ki tudunk keverni” egy t -eloszlású próbastatisztikát $\rightarrow$ ennek pontos alakját és szabadsági fokát nehéz a kétoldali, kétmintás u -próba esetéből levezetni.

Kétoldali, kétmintás u -próba

Kenyeres példa $\rightarrow \varepsilon = 0.05$ terjedelem mellett el kell utasítanunk azt a nullhipotézist, hogy a péknél vásárolt kenyerek várható értékben egykilósak $\rightarrow$ alternatív kenyérforrás után nézünk. Van egy másik pék, aki nagyon hasonló kinézetű, ízű és árú egykilós kenyereket gyárt, mint az első.

Reklámja szerint képes olyan kenyeret gyártani, amelyek tömegének szórása csak 1 dkg (azaz $\sigma_1 = 0.01$ kg, tehát $\sigma_1^2 = 0.0001$ kg² a szórásnégyzet).

Most $n_1 = 10$ napon keresztül ennél a péknél vásárolunk napi egy kenyeret, és megmérjük a tömegüket.

Egy olyan $\mathbf{x} = (x_1, \dots, x_{n_1})$ realizációt kapunk, amelynek átlaga $\overline{x_{n_1}} = 0.9825$ kg.

A későbbiekben hasznos **jelölés** a régi péknél vásárolt kenyerek tömegeire: $\mathbf{y} = (y_1, \dots, y_{n_2})$, ahol a mintaelemszám $n_2 = 25$, a mintaátlag $\overline{y_{n_2}} = 0.98$, a szóráss $\sigma_2 = 0.02$ volt. (Feltételezzük, hogy az – akárki által gyártott – kenyerek tömegei együttesen függetlenek.)

Kérdés: ugyanaz marad-e a helyzetünk, ha az új pékhez járunk?

Kétoldali, kétmintás u -próba

Vagyis: az új péknél vásárolt kenyerek tömegének (szintén ismeretlen) μ_1 várható értéke ugyanaz-e, mint a régi péknél vásároltak tömegének μ_2 várható értéke? Ezért a

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2$$

alternatívára kívánunk egy $\varepsilon = 0,05$ terjedelmű próbát konstruálni. Feltételezve, hogy az $(X_1, \dots, X_{n_1})$ minta független az $(Y_1, \dots, Y_{n_2})$ mintától, az $\overline{X_{n_1}}$ mintaátlag is független az $\overline{Y_{n_2}}$ mintaátlagtól. Ezért $\overline{X_{n_1}}$ és $\overline{Y_{n_2}}$ lineáris kombinációi is normális eloszlásúak. Elemi számításokkal megmutatható, hogy

$$u(\mathbf{X}, \mathbf{Y}) = \frac{\overline{X_{n_1}} - \overline{Y_{n_2}}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0; 1).$$

Így a kétoldali, egymintás u -próbához hasonlóan, ha

$$u(\mathbf{X}, \mathbf{Y}) \in [-u_{\varepsilon/2}, u_{\varepsilon/2}] = [-\Phi^{-1}(1 - \varepsilon/2), \Phi^{-1}(1 - \varepsilon/2)],$$

akkor H_0 -t elfogadjuk, különben elutasítjuk. Kritikus tartományunk:

$$\mathfrak{X}_k = \{(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2} : |u(\mathbf{x}, \mathbf{y})| > u_{\varepsilon/2}\}.$$

Kétoldali, kétmintás u -próba

Konkrét példánkban

$$u(\mathbf{x}, \mathbf{y}) = \frac{\overline{x_{n_1}} - \overline{y_{n_2}}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \frac{0.9825 - 0.98}{\sqrt{\frac{0.0001}{10} + \frac{0.0004}{25}}} \approx 0,4903.$$

Mivel $|u(\mathbf{x}, \mathbf{y})| < u_{\varepsilon/2} = \Phi^{-1}(0.975) = 1.9600$, H_0 -t elfogadjuk.
Vagyis várható értékben az új péknél is ugyanolyan tömegűek a kenyerek, mint a réginél, tehát a pékváltás nem változtat szignifikánsan a helyzetünkön...

(Törölve: "De kis híján szignifikánsan változtat.")

Kétoldali, kétmintás u -próba: eljárás

Feltevés: az $x_1, \dots, x_{n_1}$ adatpontok fae. $N(\mu_1, \sigma_1^2)$ eloszlású $X_1, \dots, X_{n_1}$ valószínűségi változók realizációi, az $y_1, \dots, y_{n_2}$ adatpontok pedig az $X_1, \dots, X_{n_1}$ -től (együttesen) független, egymás között fae. $N(\mu_2, \sigma_2^2)$ eloszlású $Y_1, \dots, Y_{n_2}$ valószínűségi változók realizációi. A $\sigma_1, \sigma_2 > 0$ szórások ismertek, a μ_1, μ_2 várható értékek ismeretlenek. Eljárás:

1. Nullhipotézis: $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$ adatpontok kinyerése.
4. Próbat statisztika:

$$u(\mathbf{x}, \mathbf{y}) = \frac{\overline{x_{n_1}} - \overline{y_{n_2}}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

5. Elfogadási tartomány: $\mathfrak{X}_e = \{(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2} : |u(\mathbf{x}, \mathbf{y})| \leq u_{\varepsilon/2}\}$, ahol $u_{\varepsilon/2} = \Phi^{-1}(1 - \varepsilon/2)$.
6. Döntés: ha $|u(\mathbf{x}, \mathbf{y})| > u_{\varepsilon/2}$, H_0 -t elutasítjuk, kül. elfogadjuk.

Megjegyzés: egyoldali, kétmintás u -próba

Itt nem részletezzük, de a kétmintás u -próbának ugyanúgy vannak egyoldali változatai, mint az egymintásnak. Az egyoldali és a kétoldali próba közötti különbség is analóg az egymintás esethez.

- ▶ $H_0: \mu_1 \geq \mu_2$ vs. $H_1: \mu_1 < \mu_2$: akkor utasítjuk el a nullhipotézist, ha $u(\mathbf{x}, \mathbf{y}) = \frac{\bar{x}_{n_1} - \bar{y}_{n_2}}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} < -u_\epsilon$.
- ▶ $H_0: \mu_1 \leq \mu_2$ vs. $H_1: \mu_1 > \mu_2$: akkor utasítjuk el H_0 -t, ha $u(\mathbf{x}, \mathbf{y}) > u_\epsilon$.

Kenyeres példa: $H_0: \mu_1 \geq \mu_2$ vs. $H_1: \mu_1 < \mu_2$ is jó alternatíva. Ha H_0 -t elutasítjuk, akkor az új pék választásával javul a helyzetünk.

Alkalmazás pl.: bevezessünk-e egy új gyógyszert, jobb-e, mint a régi?

A kétmintás t -próbánál (lásd lejjebb) is csak a kétoldali változattal foglalkozunk, de annak is ugyanúgy vannak egyoldali változatai is.

Kétoldali, kétmintás u -próba: példa

(13.6. a) feladat)

Véletlenszerűen kiválasztottuk a BME 10 gépészmérnök férfi hallgatóját, az ő átlagos testmagasságuk 179 cm volt, 2 cm korrigált empirikus szórással, valamint a BME 20 építőmérnök férfi hallgatóját, az ő esetükben az átlag 178 cm, a korrigált empirikus szórás 1 cm volt. Feltéve, hogy az egyes hallgatók testmagasságai függetlenek, normális eloszlásúak és azonos szak esetén azonos eloszlásúak is, vizsgáljuk meg a minta alapján 5% és 1% terjedelmű próba segítségével is, hogy a két szakra járók testmagasságának várható értéke megegyezik-e, ha a gépészmérnök férfi hallgatók testmagasságának szórása $\sqrt{2}$ cm, az építőmérnök férfi hallgatóké pedig 1 cm.

Adjuk meg H_0 -t, H_1 -et, a próbastatisztika elfogadási tartománynak megfelelő értékei halmazát, a próbastatisztika értékét és a döntést.

Kétoldali, kétmintás t -próba

Emlékeztető: kétoldali, kétmintás u -próba

Legyenek $X_1, \dots, X_{n_1} \sim N(\mu_1; \sigma_1^2)$ és $Y_1, \dots, Y_{n_2} \sim N(\mu_2; \sigma_2^2)$ független fae. minták, ahol a μ_1 és μ_2 várható értékek ismeretlenek, és tekintsük a

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2$$

alternatívát. Láttuk (bár nem bizonyítottuk), hogy

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

$\Rightarrow$ Ha a σ_1^2 és σ_2^2 szórásnégyzetek ismertek, ezzel a próbastatisztikával mehet a kétoldali, kétmintás u -próba. Hasonlítsuk a próbastatisztikát az $N(0; 1)$ eloszlás megfelelő kvantiliséhez.

Kérdés: mi legyen a próbastatisztika, ha a szórásek ismeretlenek?

Kétoldali, kétmintás t -próba

Emlékeztető: kétoldali, kétmintás u -próba

Legyenek $X_1, \dots, X_{n_1} \sim N(\mu_1; \sigma_1^2)$ és $Y_1, \dots, Y_{n_2} \sim N(\mu_2; \sigma_2^2)$ független fae. minták, ahol a μ_1 és μ_2 várható értékek ismeretlenek, és tekintsük a

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2$$

alternatívát. Láttuk (bár nem bizonyítottuk), hogy

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

$\Rightarrow$ Ha a σ_1^2 és σ_2^2 szórásnégyzetek ismertek, ezzel a próbastatisztikával mehet a kétoldali, kétmintás u -próba. Hasonlítsuk a próbastatisztikát az $N(0; 1)$ eloszlás megfelelő kvantiliséhez.

Kérdés: mi legyen a próbastatisztika, ha a szórásek ismeretlenek?

Kétoldali, kétmintás t -próba

Kérdés: ha a szórások ismeretlenek, mi egy olyan próbastatisztika, aminek az eloszlását ismerjük?

Biz. nélkül: **ha $\mathbb{D}^2(X_i) = \mathbb{D}^2(Y_i)$, akkor**

$$t(\mathbf{X}, \mathbf{Y}) = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\frac{(n_1-1)(s_X^*)^2 + (n_2-1)(s_Y^*)^2}{n_1+n_2-2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

t -eloszlású $f = n_1 + n_2 - 2$ szabadsági fokkal.

Ez jó lesz próbastatisztikának. Kétoldali esetben ennek abszolút értékét a $t(n_1 + n_2 - 2)$ -eloszlás $1 - \varepsilon/2$ -kvantilisével hasonlítjuk össze, hasonlóan a kétoldali, egymintás t -próbához.

Megjegyzés: a szabadsági fok általában a teljes mintaelemszám mínusz a **becsült paraméterek száma**.

- ▶ Egymintás t -próba: mintaelemszám n , becsült paraméter (a szórás helyett) s_X^* , tehát a szabadsági fok $n - 1$.
- ▶ Kétmintás t -próba: teljes mintaelemszám $n_1 + n_2$, becsült paraméterek (a szórások helyett) s_X^* és s_Y^* , tehát a szabadsági fok $n_1 + n_2 - 2$.

Kétoldali, kétmintás t -próba: eljárás

Feltevés: az $x_1, \dots, x_{n_1}$ adatpontok fae. normális eloszlású $X_1, \dots, X_{n_1}$ valószínűségi változók realizációi, amelyek μ_1 várható értéke ismeretlen is, az $y_1, \dots, y_{n_2}$ adatpontok pedig az $X_1, \dots, X_{n_1}$ -től (együttesen) független, egymás között fae. normális eloszlású $Y_1, \dots, Y_{n_2}$ valószínűségi változók realizációi, ezek μ_2 várható értéke sem ismert, és $\mathbb{D}^2(X_1) = \mathbb{D}^2(Y_1)$ ismeretlen. Eljárás:

1. Nullhipotézis: $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$.
2. Az előírányzott $\varepsilon \in (0, 1)$ terjedelem kiválasztása.
3. Az $x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}$ adatpontok kinyerése.
4. Az $f = n_1 + n_2 - 2$ szabadsági fok meghatározása.
5. Próbat statisztika:
$$t(\mathbf{x}, \mathbf{y}) = \frac{\overline{x_{n_1}} - \overline{y_{n_2}}}{\sqrt{\frac{(n_1-1)(s_x^*)^2 + (n_2-1)(s_y^*)^2}{n_1+n_2-2}}} \sqrt{\frac{n_1 n_2}{n_1+n_2}}.$$
6. Elfogadási tartomány: $\mathfrak{X}_e = \{(\mathbf{x}, \mathbf{y}) = (x_1, \dots, x_{n_1}, y_1, \dots, y_{n_2}) \in \mathbb{R}^{n_1+n_2} : |t(\mathbf{x}, \mathbf{y})| \leq t_{n_1+n_2-2, \varepsilon/2}\},$ ahol $t_{n_1+n_2-2, \varepsilon/2}$ a $t(n_1 + n_2 - 2)$ -eloszlás $1 - \varepsilon/2$ -kvantilise.
7. Döntés: ha $|t(\mathbf{x}, \mathbf{y})| > t_{\varepsilon/2}$, akkor H_0 -t elutasítjuk, különben elfogadjuk.

Kétoldali, kétmintás t -próba: példa – kenyérvásárlás

Az új péknél vásárolt 10 db kenyér adatai: $n_1 = 10$, $\overline{x_{n_1}} = 0,9825$ kg. Szerencsére gondosan felírtuk az új péknél vásárolt kenyerek tömegadatait is, ezek korrigált empirikus szórására pedig $s_x^* = 0,015$ kg adódott.

A régi péknél vásárolt 25 kenyérhez tartozó adatok: $n_2 = 25$, $\overline{y_{n_2}} = 1,98$ kg, $s_y^* = 0,05$ kg.

Így a szabadsági fok $f = n_1 + n_2 - 2 = 33$. Alternatíva:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2$$

Erre ismeretlen szórások mellett a kétoldali, kétmintás t -próbát alkalmazzuk. A próbastatisztika értéke:

$$t(\mathbf{x}, \mathbf{y}) = \frac{\overline{x_{n_1}} - \overline{y_{n_2}}}{\sqrt{\frac{(n_1-1)(s_x^*)^2 + (n_2-1)(s_y^*)^2}{n_1 + n_2 - 2}}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = \frac{0.9825 - 0.98}{\sqrt{\frac{9 \cdot 0.015^2 + 24 \cdot 0.05^2}{33}}} \sqrt{\frac{10 \cdot 25}{10 + 25}} \approx 0.1541.$$

Ez (abszolút értékben) jóval kisebb, mint az $n_1 + n_2 - 2 = 33$ szabadságfokú t -eloszlás $1 - \varepsilon/2 = 0,975$ -kvantilise, $t_{33,0,975} \approx 2.0345$. Ezért H_0 -t elfogadjuk ($\Leftrightarrow$ a régi és az új péknél vásárolt kenyér tömegének várható értéke megegyezik).

Kétoldali, kétmintás t -próba: példa

(13.6. b) feladat)

Véletlenszerűen kiválasztottuk a BME 10 gépészmérnök férfi hallgatóját, az ő átlagos testmagasságuk 179 cm volt, 2 cm korrigált empirikus szórással, valamint a BME 20 építőmérnök férfi hallgatóját, az ő esetükben az átlag 178 cm, a korrigált empirikus szórás 1 cm volt. Feltéve, hogy az egyes hallgatók testmagasságai függetlenek, normális eloszlásúak és azonos szak esetén azonos eloszlásúak is, vizsgáljuk meg a minta alapján 5% és 1% terjedelmű próba segítségével is, hogy a két szakra járók testmagasságának várható értéke megegyezik-e, ha mindkét szakra járó hallgatók testmagasságának szórása ismeretlen, de egyenlő!

Mindkét esetben adjuk meg H_0 -t, H_1 -et, a próbastatisztika elfogadási tartománynak megfelelő értékei halmazát, a próbastatisztika értékét és a döntést.

t-próba: kitekintés

Akik felveszik a mérnök- és gazdaságinformatikus MSc képzés Matematikai statisztikai tárgyát, további esetekkel is találkozhatnak, ahol *t*-eloszlású a próbastatisztika. Pl.:

- ▶ *Egyoldali, kétmintás verziók (u-próbánál is):* értelemszerűen, ahogy már említettük, részletesen lásd a statisztika tárgyból.
- ▶ Kétmintás *t*-próbánál: **mi a teendő, ha a szórások nem egyenlők**, vagy nem tudjuk, hogy egyenlők-e?
 - ▶ Szórások egyenlőségének ellenőrzésére is van próba: *F*-próba, $H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$.
 - ▶ Ha az *F*-próbánál H_0 -t elfogadjuk, mehet a kétmintás *t*-próba.
 - ▶ Ha az *F*-próbánál H_0 -t elutasítjuk: **Welch-próba**.
A próbastatisztika még mindig *t*-eloszlású, de máshogy néz ki, a szabadsági foka általában nem egész szám.
- ▶ **Összetartozó mintás *t*-próba:** $(x_1, y_1), \dots, (x_n, y_n)$ adatpontok, egymás közt $x_1, \dots, x_n$ és $y_1, \dots, y_n$ is fae. normális eloszlású val. változók realizációi, de x_i és y_i nem függetlenek, hanem ugyanabból a mérésből származnak.
Pl: n csavart mérünk, x_i az i -edik csavar hossza, y_i az átmérője. Erre is van *t* próbastatisztikájú próba.