

Statisztika mérnököknek (levelező)

Szűcs Gábor

SZTE TTIK Bolyai Intézet

2025/26 tavaszi félév

Események valószínűsége, valószínűségi változók

A valószínűesszámitás a matematika egyik ága, melynek célja a véletlen jelenségek vizsgálata. Alapfogalmak:

- **Véletlen kísérlet:** Egy véletlen jelenség megfigyelése.
- **Esemény:** A kísérlet eredményével kapcsolatos állítás.
- **Valószínűség (P):** Annak az esélye, hogy az esemény bekövetkezik.

Feladat: Magyarországon az emberek 44 százaléka esik az A vércsoportba. Véletlenszerűen kiválasztunk egy magyar embert. Mennyi az esélye, hogy a kiválasztott ember az A vércsoportba esik?

Kísérlet: véletlenszerűen kiválasztunk egy embert.

Véletlenszerű választás = mindenkinek azonos az esélye.

Esemény: a kiválasztott ember az A vércsoportba esik.

$$P(\text{a kiválasztott ember az A vércsoportba esik}) = 44\% = 0.44$$

A statisztika biostatisztikai megközelítése:

- Adott egy sokaság (egy halmaz, tipikusan egy populáció).
- Egy mennyiség (testtömeg, utódok száma, stb.) eloszlását vizsgáljuk.
- Véletlenszerűen kiválasztunk egy elemet/egyedet a sokaságból. Ekkor:
$$\text{valószínűség} = \text{arány a sokaságon belül}$$
- ξ = a vizsgált mennyiség értéke a kiválasztott elem/egyed esetében.
- ξ egy véletlen szám, egy véletlen mennyiség.

Valószínűségi változó: Egy véletlen kísérletből származó véletlen szám. Jellemzően görög betűkkel jelöljük: ξ (kszi), η (éta), stb.

Értékkészlet: A változó lehetséges értékeinek a halmaza. Jele: R_ξ, R_η

A valószínűesszámitást ennél általánosabban között is lehet alkalmazni:

- Tekintsünk egy tetszőleges véletlen jelenséget.
- A sokaság elemei a forgatókönyvek, ahogyan a jelenség bekövetkezhet.
- A valószínűség szemléletes jelentése: a kedvező forgatókönyvek aránya.

Feladat: Egy családban 4 gyerek van, a testtömegük: 50, 54, 60, 70 kg. Véletlenszerűen kiválasztunk egy gyereket.

- ξ = kiválasztott gyerek tömege
- Értékkészlet: $R_\xi = \{50, 54, 60, 70\}$
- Arány a családon belül:

$$P(50 \leq \xi \leq 60) = 75\% = 0.75 \quad P(\xi \geq 55) = 50\% = 0.5$$

Tegyük fel, hogy a sokaságban N elem található, és a vizsgált mennyiség értékei az elemeken: $x_1, \dots, x_N$. Véletlenszerűen kiválasztunk egy elemet.

- ξ = a vizsgált mennyiség értéke a kiválasztott elem esetében
- Értékkészlet: $R_\xi = \{x_1, \dots, x_N\}$
- Arány a sokaságon belül: tetszőleges a, b számok esetén

$$P(a \leq \xi \leq b) = \text{az } a \text{ és } b \text{ közé eső elemek aránya}$$

A ξ valószínűségi változóra vonatkozó fontosabb mutatószámok:

- **Várható érték (expected value)** = átlagos érték:

$$E(\xi) = \frac{x_1 + \dots + x_N}{N}$$

- **Variancia, szórásnégyzet:**

$$\text{Var}(\xi) = E\left([\xi - E(\xi)]^2\right) = \frac{[x_1 - E(\xi)]^2 + \dots + [x_N - E(\xi)]^2}{N}$$

- **Szórás (deviation):** $D(\xi) = \sqrt{\text{Var}(\xi)}$.

A szórás szemléletes jelentése: várható értéktől vett átlagos eltérés.

$$D(\xi) = \sqrt{E\left([\xi - E(\xi)]^2\right)} \approx E\left(\sqrt{[\xi - E(\xi)]^2}\right) = E(|\xi - E(\xi)|)$$

A szórás a sokaság homogenitását méri: kis szórás = homogén sokaság.

Feladat: Egy családban 4 gyerek van, a testtömegük: 50, 54, 60, 70 kg. Véletlenszerűen kiválasztunk egy gyereket.

- ξ = kiválasztott gyerek tömege
- Várható érték (átlagos érték):

$$E(\xi) = \frac{50 + 54 + 60 + 70}{4} = 58.5$$

- Variancia, szórásnégyzet:

$$\text{Var}(\xi) = \frac{[50 - 58.5]^2 + [54 - 58.5]^2 + [60 - 58.5]^2 + [70 - 58.5]^2}{4} = 56.75$$

- Szórás (várható értéktől vett átlagos eltérés): $D(\xi) = \sqrt{56.75} = 7.53$

Probléma: nagy méretű sokaságok esetén ezt kényelmetlen végigszámolni.

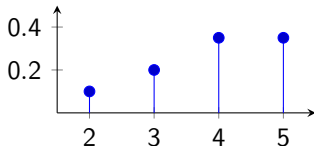
Diszkrét valószínűségi változók

Diszkrét valószínűségi változó: Olyan véletlen szám, melynek egy véges vagy végtelen sorozat az értékkészlete.

Valószínűségeloszlás: A lehetséges értékek valószínűségei (arányai).

Feladat: Egy lengyel felmérés alapján a fehér gólyák 2–5 tojást raknak az alábbi táblázatban található megoszlásban. Véletlenszerűen kiválasztunk egy gólyafészket, és jelölje ξ a fészekben található tojások számát.

x	2	3	4	5
p_x	10%	20%	35%	35%



$R_\xi = \{2, 3, 4, 5\}$, tehát a ξ diszkrét változó

$P(\xi = 2) = 0.1$, $P(\xi = 3) = 0.2$, $P(\xi = 4) = 0.35$, $P(\xi = 5) = 0.35$

$$P(\xi = 2) = 0.1, \quad P(\xi = 3) = 0.2, \quad P(\xi = 4) = 0.35, \quad P(\xi = 5) = 0.35$$

Feladat: Mennyi a ξ változó lehetséges értékeinek összvalószínűsége?

$$P(\xi = 2) + P(\xi = 3) + P(\xi = 4) + P(\xi = 5) = 0.1 + 0.2 + 0.35 + 0.35 = 1$$

Feladat: A fészkek mekkora hányadában található legfeljebb 3 tojás?

$$P(\text{legfeljebb 3 tojás}) = P(\xi \leq 3) = P(\xi = 2) + P(\xi = 3) = 0.3$$

Diszkrét változó **móduszai**: a változó legnagyobb valószínűségű értékei.
Jelentése: a változó leggyakoribb értékei a teljes sokaságon belül.

Feladat: Hány módusza van a ξ változónak, és mik ezek?

Két módusz van, a 4 és az 5. Mindkettő 35% arányban fordul elő.

Feladat: Mennyi a ξ változó várható értéke?

Tegyük fel, hogy a sokaságban összesen N egyed található! A sokaságon belüli arányok és gyakoriságok:

x	2	3	4	5	összesen
p_x	10%	20%	35%	35%	100%
k_x	$0.1N$	$0.2N$	$0.35N$	$0.35N$	N

Ekkor a várható érték (átlag):

$$\begin{aligned}
 E(\xi) &= \frac{\overbrace{2 + \dots + 2}^{0.1N \text{ db}} + \overbrace{3 + \dots + 3}^{0.2N \text{ db}} + \overbrace{4 + \dots + 4}^{0.35N \text{ db}} + \overbrace{5 + \dots + 5}^{0.35N \text{ db}}}{N} \\
 &= \frac{2 \cdot 0.1N + 3 \cdot 0.2N + 4 \cdot 0.35N + 5 \cdot 0.35N}{N} \\
 &= 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.35 + 5 \cdot 0.35 = 3.95
 \end{aligned}$$

Házi feladat: Határozzuk meg ugyanilyen módon ξ varianciáját is!

Legyen ξ diszkrét valószínűségi változó!

- **Várható érték:** $E(\xi) = \sum_{x \in R_\xi} x P(\xi = x)$
- **Variancia:** $\text{Var}(\xi) = \sum_{x \in R_\xi} [x - E(\xi)]^2 P(\xi = x)$
- **Szórás:** $D(\xi) = \sqrt{\text{Var}(\xi)}$

Feladat: Számoljuk ki a várható értéket és a szórást a gólyás feladatban!

Várható érték (átlagos érték):

$$\begin{aligned} E(\xi) &= 2 \cdot P(\xi = 2) + 3 \cdot P(\xi = 3) + 4 \cdot P(\xi = 4) + 5 \cdot P(\xi = 5) \\ &= 2 \cdot 0.1 + 3 \cdot 0.2 + 4 \cdot 0.35 + 5 \cdot 0.35 = 3.95 \end{aligned}$$

Variancia:

$$\text{Var}(\xi) = [2 - 3.95]^2 \cdot 0.1 + [3 - 3.95]^2 \cdot 0.2 + [4 - 3.95]^2 \cdot 0.35 + [5 - 3.95]^2 \cdot 0.35$$

Eredmény: $\text{Var}(\xi) = 0.95$

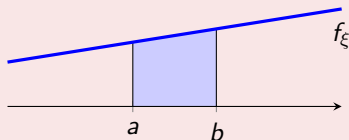
Szórás (várható értéktől vett átlagos eltérés): $D(\xi) = \sqrt{\text{Var}(\xi)} = 0.97$

Folytonos valószínűségi változók

Folytonos valószínűségi változó: az értékkészlet egy véges vagy végtelen intervallum.

Folytonos változó **sűrűségfüggvénye:** egy olyan f_ξ függvény, melyre tetszőleges $a \leq b$ számok esetén:

$$P(a \leq \xi \leq b) = \int_a^b f_\xi(x) dx$$

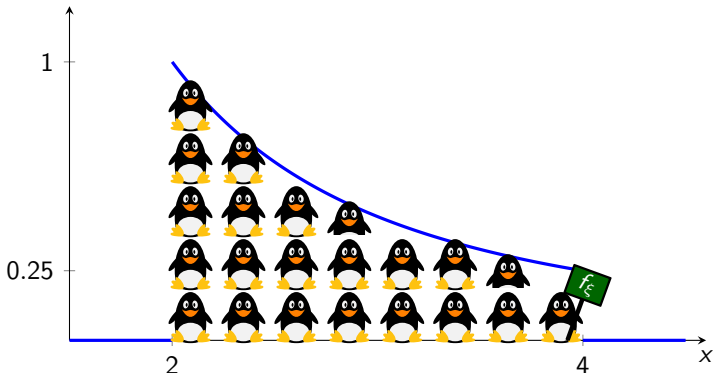


Egy mennyiség eloszlását vizsgáljuk egy sokaságon belül. Véletlenszerűen kiválasztunk egy elemet, és ξ a mennyiség értéke ezen elem esetében.

azon elemek aránya, melyeknél a vizsgált mennyiség a és b közé esik
 $= P(a \leq \xi \leq b) =$ görbe alatti terület a és b között

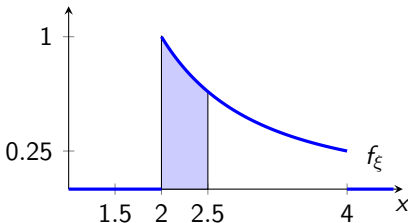
A sűrűségfüggvény szemléletes jelentése:

- Az egyes elemeket/egyedeket arra a pontra tesszük a számegyenesen, amennyi a vizsgált mennyiség értéke az adott elem/egyed esetében.
- A sűrűségfüggvény értéke a „rakás” magassága.
- Az integrál nem az elemek száma, hanem a sokaságon belüli arány.
- Folytonos változó értékkészlete: azon x valós számok, ahol $f_{\xi}(x) > 0$.



Feladat: Egy pingvinfajt vizsgálva legyen ξ egy véletlenszerűen kiválasztott egyed tömege kilogrammban. A ξ változó az alábbi f_ξ sűrűségfüggvénnyel írható le. Az egyedek mekkora hányada esik 1.5 és 2.5 kg közé?

$$f_\xi(x) = \begin{cases} 4/x^2, & \text{ha } 2 \leq x \leq 4, \\ 0, & \text{különben.} \end{cases}$$

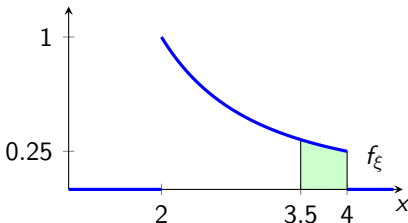


$$\begin{aligned} P(1.5 \leq \xi \leq 2.5) &= \int_{1.5}^{2.5} f_\xi(x) dx = \int_{1.5}^2 0 dx + 4 \int_2^{2.5} x^{-2} dx \\ &= 0 + 4 \left[\frac{x^{-1}}{-1} \right]_2^{2.5} = 4 \left[\frac{1}{-x} \right]_2^{2.5} = 4 \left(\frac{1}{-2.5} - \frac{1}{-2} \right) = 4 \cdot 0.1 = 0.4 \end{aligned}$$

$$\int x^c dx = \frac{x^{c+1}}{c+1} \quad \text{ha } c \neq -1, \quad \int x^{-1} dx = \ln |x|$$

Feladat: Mennyi azon egyedek aránya, melyek elérik a 3.5 kg tömeget?

$$f_{\xi}(x) = \begin{cases} 4/x^2, & \text{ha } 2 \leq x \leq 4, \\ 0, & \text{különben.} \end{cases}$$



$$\begin{aligned} P(\xi \geq 3.5) &= P(3.5 \leq \xi \leq +\infty) = \int_{3.5}^{+\infty} f_{\xi}(x) dx = 4 \int_{3.5}^4 x^{-2} dx \\ &= 4 \left[\frac{1}{-x} \right]_{3.5}^4 = 4 \left(\frac{1}{-4} - \frac{1}{-3.5} \right) \approx 0.14 = 14\% \end{aligned}$$

Feladat: Milyen testtömegek fordulnak elő a populációban?

Folytonos változó értékkészlete: azon x valós számok, ahol $f_{\xi}(x) > 0$.

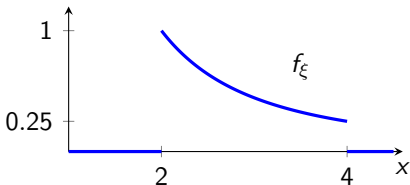
Most: $R_{\xi} = [2, 4]$

Legyen ξ folytonos valószínűségi változó!

- **Várható érték:** $E(\xi) = \int_{-\infty}^{+\infty} x f_{\xi}(x) dx$
- **Variancia:** $\text{Var}(\xi) = \int_{-\infty}^{+\infty} [x - E(\xi)]^2 f_{\xi}(x) dx$
- **Szórás:** $D(\xi) = \sqrt{\text{Var}(\xi)}$

Feladat: Mennyi a ξ változó várható értéke a pingvines feladatban?

$$f_{\xi}(x) = \begin{cases} 4/x^2, & \text{ha } 2 \leq x \leq 4, \\ 0, & \text{különben.} \end{cases}$$



$$\begin{aligned} E(\xi) &= \int_{-\infty}^{+\infty} x f_{\xi}(x) dx = \int_{-\infty}^2 x \cdot 0 dx + \int_2^4 x \cdot \frac{4}{x^2} dx + \int_4^{+\infty} x \cdot 0 dx \\ &= 0 + 4 \int_2^4 x^{-1} dx + 0 = 4 \left[\ln |x| \right]_2^4 = 4(\ln 4 - \ln 2) = 2.77 \end{aligned}$$

A varianciára adható egy könnyebben számolható formula is:

$$\begin{aligned}
 \text{Var}(\xi) &= \int_{-\infty}^{+\infty} [x - E(\xi)]^2 f_{\xi}(x) dx \\
 &= \int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx - \int_{-\infty}^{+\infty} 2E(\xi) x f_{\xi}(x) dx + \int_{-\infty}^{+\infty} (E(\xi))^2 f_{\xi}(x) dx \\
 &= \int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx - 2E(\xi) \int_{-\infty}^{+\infty} x f_{\xi}(x) dx + (E(\xi))^2 \int_{-\infty}^{+\infty} f_{\xi}(x) dx \\
 &= \int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx - 2E(\xi)E(\xi) + (E(\xi))^2 \cdot 1 = \int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx - (E(\xi))^2
 \end{aligned}$$

Feladat: Mennyi a ξ változó szórása a pingvines feladatban?

$$\int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx = \int_2^4 x^2 \frac{4}{x^2} dx = \int_2^4 4 dx = 4 \cdot 2 = 8$$

$$\text{Var}(\xi) = \int_{-\infty}^{+\infty} x^2 f_{\xi}(x) dx - (E(\xi))^2 = 8 - (2.77)^2 \approx 0.33$$

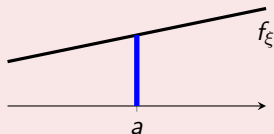
$$D(\xi) = \sqrt{\text{Var}(\xi)} = \sqrt{0.33} = 0.57$$

Folytonos változó esetén miért nem a valószínűségeloszlással számolunk?

Ha ξ folytonos változó, akkor tetszőleges a szám esetén $P(\xi = a) = 0$.

Bizonyítás:

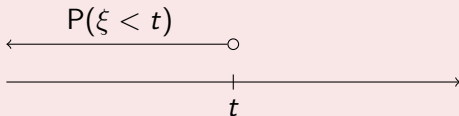
$$P(\xi = a) = P(a \leq \xi \leq a) = \int_a^a f_\xi(x) dx = 0$$



Egy ξ valószínűségi változó **eloszlásfüggvénye** a következő függvény:

$$F_\xi : \mathbb{R} \rightarrow [0, 1]$$

$$F_\xi(t) = P(\xi < t)$$

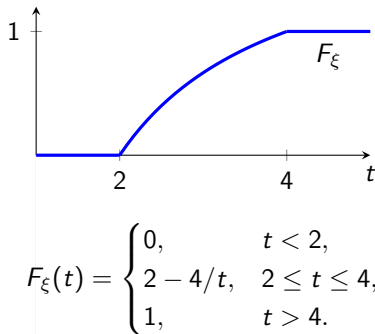
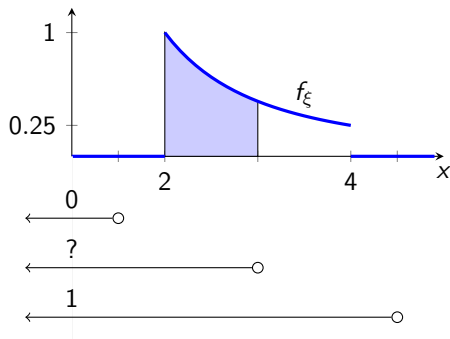


Szemléletes jelentés:

$F_\xi(t)$ = azon egyedek aránya a sokaságban, ahol ξ kisebb, mint t

Feladat: Írjuk fel az eloszlásfüggvényt a jelen feladatban!

$F_{\xi}(t) = P(\xi < t) =$ azon egyedek aránya, melyek t -nél kisebb tömegűek

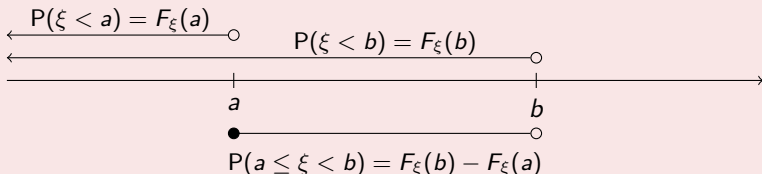


Ha $2 \leq t \leq 4$, akkor:

$$F_{\xi}(t) = P(\xi < t) = 4 \int_2^t x^{-2} dx = 4 \left[\frac{1}{-x} \right]_2^t = 4 \left(\frac{1}{-t} - \frac{1}{-2} \right) = 2 - \frac{4}{t}$$

Tetszőleges ξ valószínűségi változó és $a \leq b$ valós számok esetén:

$$P(a \leq \xi < b) = F_{\xi}(b) - F_{\xi}(a).$$



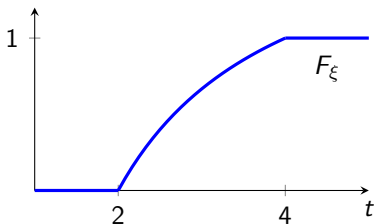
Ha a ξ változó folytonos eloszlású, akkor

$$P(a \leq \xi \leq b) = P(a < \xi < b) = P(a < \xi \leq b) = F_{\xi}(b) - F_{\xi}(a).$$

Diszkrét változók esetén ritkán használjuk az eloszlásfüggvényt, ugyanis általában egyszerűbb a valószínűségeloszlással dolgozni. De a folytonos esetben az eloszlásfüggvény nagyon hasznos eszköz.

Vissza a pingvines feladathoz...

$$F_{\xi}(t) = \begin{cases} 0, & t < 2, \\ 2 - 4/t, & 2 \leq t \leq 4, \\ 1, & t > 4. \end{cases}$$



Feladat: Az egyedek mekkora hányada esik 1.5 és 2.5 kg közé?

$$P(1.5 \leq \xi \leq 2.5) = F_{\xi}(2.5) - F_{\xi}(1.5) = \left(2 - \frac{4}{2.5}\right) - 0 = 0.4$$

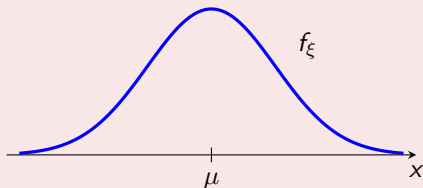
Feladat: Mennyi azon egyedek aránya, melyek elérik a 3.5 kg tömeget?

$$P(\xi \geq 3.5) = P(3.5 \leq \xi \leq 4) = F_{\xi}(4) - F_{\xi}(3.5) = \left(2 - \frac{4}{4}\right) - \left(2 - \frac{4}{3.5}\right) = 0.14$$

A normális eloszlás

A ξ valószínűségi változó **normális** (normál, Gauss-) eloszlást követ $\mu \in \mathbb{R}$ (mű) és $\sigma > 0$ (szigma) paraméterekkel, ha a sűrűségfüggvénye:

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



A sűrűségfüggvény neve: **Gauss-görbe**, **haranggörbe**.

A normális eloszlás néhány alkalmazása:

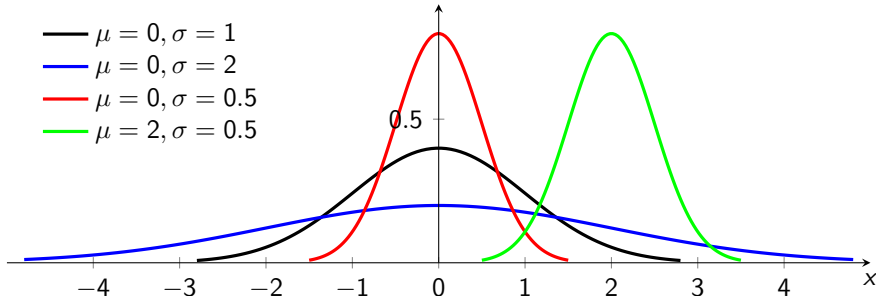
- Mérési hibák modellezése: mért érték = igazi érték + mérési hiba, ahol a mérési hiba normális eloszlást követ.
- Élettudományok: számos mennyiség (testmagasság, vérnyomás, IQ) normális vagy a normálisból származtatott eloszlást követ.

A normális eloszlás fontosabb tulajdonságai:

- A sűrűségfüggvény mindenhol pozitív, ezért $R_\xi = \mathbb{R}$.
- $\mu = E(\xi)$ és $\sigma = D(\xi)$.

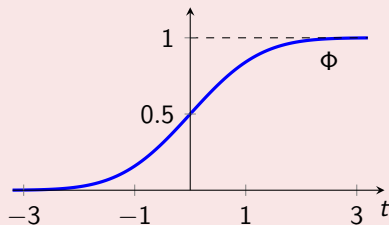
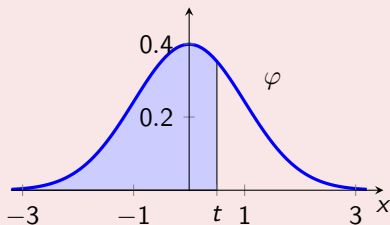
Hogyan hat a két paraméter a sűrűségfüggvényre?

- σ : a sűrűségfüggvény alakját határozza meg,
- μ : eltolás, a sűrűségfüggvény szimmetriatengelye.



Standard normális eloszlás: normális eloszlás $\mu = 0$ és $\sigma = 1$ paraméterrel. Sűrűségfüggvénye és eloszlásfüggvénye:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(t) = P(\xi < t) = \int_{-\infty}^t \varphi(x) dx.$$



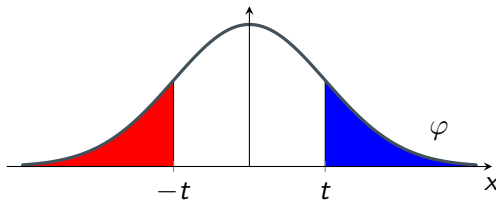
Hogyan tudjuk meghatározni a Φ függvény értékét? A formula nagyon bonyolult, táblázatot használunk.

A Φ eloszlásfüggvény rendelkezik az alábbi tulajdonságokkal.

- Szigorúan monoton növekvő és mindenhol folytonos.
- $\Phi(0) = 0.5$.
- Tetszőleges t valós szám esetén $\Phi(-t) = 1 - \Phi(t)$.

Az utolsó állítás bizonyítása:

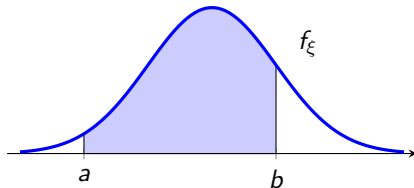
$$\begin{aligned}\Phi(-t) &= P(\xi < -t) = \int_{-\infty}^{-t} \varphi(x) dx = \int_t^{+\infty} \varphi(x) dx \\ &= P(\xi \geq t) = 1 - P(\xi < t) = 1 - \Phi(t)\end{aligned}$$



Ha ξ normális eloszlású, akkor tetszőleges $a \leq b$ valós számokra:

$$P(a \leq \xi \leq b) = \int_a^b f_{\xi}(x) dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Probléma: ezt az integrált nem tudjuk papíron kiszámolni.



Legyen ξ normális eloszlású μ várható értékkel és σ szórással! Ekkor:

$$F_{\xi}(t) = \Phi\left(\frac{t - \mu}{\sigma}\right).$$

Következmény: $P(a \leq \xi \leq b) = F_{\xi}(b) - F_{\xi}(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$

Bizonyítás:

$$F_{\xi}(t) = P(\xi < t) = \int_{-\infty}^t f_{\xi}(x) dx = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2} dx$$

Helyettesítéses integrálás, új változót vezetünk be:

$$y = \frac{x - \mu}{\sigma} = \frac{x}{\sigma} - \frac{\mu}{\sigma}, \quad dy = \frac{1}{\sigma} dx$$

Az új integrálási határok:

- alsó végpont: $x = -\infty \implies y = -\infty$
- felső végpont: $x = t \implies y = \frac{t-\mu}{\sigma}$

Mindezt visszaírjuk az integrálba:

$$F_{\xi}(t) = \int_{-\infty}^{\frac{t-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy = \int_{-\infty}^{\frac{t-\mu}{\sigma}} \varphi(y) dy = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

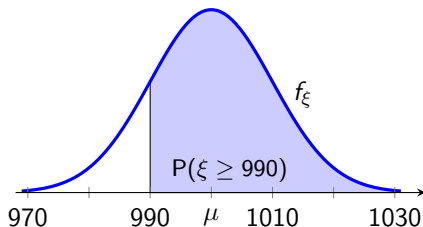
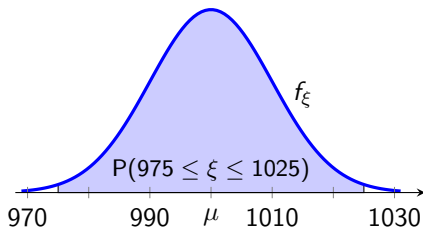
Feladat: Egy tejgyárban az 1 literes dobozos tej csomagolását automata töltőberendezés végzi. A dobozokba töltött mennyiség normális eloszlást követ, a várható érték a névleges tartalom, a szórás $\sigma = 10$ ml.

- A dobozok mekkora hányada tér el legfeljebb 2.5%-kal a névleges tartalomtól?
- A dobozok mekkora hányada tartalmaz legalább 990 ml tejet?

Véletlenszerűen kiválasztunk egy dobozt.

ξ = a kiválasztott dobozban található mennyiség

ξ normális eloszlású, $\mu = 1000$ ml, $\sigma = 10$ ml



Feladat: A dobozok mekkora hányada tér el legfeljebb 2.5%-kal a névleges tartalomtól?

$$\begin{aligned}P(975 \leq \xi \leq 1025) &= F_{\xi}(1025) - F_{\xi}(975) \\&= \Phi\left(\frac{1025 - 1000}{10}\right) - \Phi\left(\frac{975 - 1000}{10}\right) \\&= \Phi(2.5) - \Phi(-2.5) = 0.9938 - 0.0062 = 0.9876\end{aligned}$$

Itt felhasználtuk azt, hogy

$$\Phi(-2.5) = 1 - \Phi(2.5) = 1 - 0.9938 = 0.0062$$

Feladat: A dobozok mekkora hányada tartalmaz legalább 990 ml tejet?

$$\begin{aligned}P(\xi \geq 990) &= 1 - P(\xi < 990) = 1 - F_{\xi}(990) \\&= 1 - \Phi\left(\frac{990 - 1000}{10}\right) = 1 - \Phi(-1) = 1 - 0.16 = 0.84\end{aligned}$$

Itt felhasználtuk azt, hogy

$$\Phi(-1) = 1 - \Phi(1) = 1 - 0.84 = 0.16$$

Statisztikai alapfogalmak

Adott egy sokaság és egy mennyiség az elemeken. Véletlenszerűen kiválasztunk egy elemet, ξ jelöli a vizsgált mennyiséget ezen az elemen. Amire kíváncsiak vagyunk:

- $E(\xi)$ = a vizsgált mennyiség átlagos értéke a sokaságon belül;
- $D(\xi)$ = a vizsgált mennyiség szórása a sokaságon belül;
- $P(a \leq \xi \leq b)$ = arány a teljes sokaságon belül.

Valószínűségszámítás:

- ismerjük a ξ változó valószínűségeloszlását vagy sűrűségfüggvényét;
- mindent ki tudunk számolni papíron/számítógéppel.

Matematikai statisztika:

- nem ismerjük a ξ változó valószínűségeloszlását/sűrűségfüggvényét;
- megfigyeléseket végzünk a ξ változóra, és a kapott minta alapján vonunk le következtetéseket.

Mikor lehet következtetéseket levonni a minta alapján a teljes sokaságra?

- Tekintsünk úgy a megfigyelt elemekre, mint egy részsokaságra!
- **Reprezentatív minta:** a mintában minden tulajdonság hasonló arányban jelenik meg, mint a teljes sokaságban. A minta a teljes sokaság kicsinyített mása.
- Ebben az esetben a minta információt szolgáltat a teljes sokaságra.

Hogyan kaphatunk reprezentatív mintát?

- Irányított módon választjuk ki az elemeket a teljes sokaságból.
 - Jellemzően szociológiai és politikai felméréseknél alkalmazzák.
 - Előnye: kis mintaelemszám esetén is sok információt tartalmaz.
 - Hátránya: nehéz megvalósítani.
- Véletlenszerű mintavételezés: minden elemet azonos valószínűséggel választunk ki a teljes sokaságból.
 - Előnye: könnyű megvalósítani.
 - Hátránya: nagyobb mintaméret szükséges.

Mi a továbbiakban mindig véletlenszerű mintavételezést végzünk.

Statisztikai alapfogalmak:

- **Háttérváltozó:** A vizsgált ξ valószínűségi változó.
- **Statisztikai minta:** Megfigyelések a ξ háttérváltozóra, $\xi_1, \dots, \xi_n$ valószínűségi változók.
- **Mintarealizáció:** A $\xi_1, \dots, \xi_n$ változók konkrét megfigyelt értékei.
- **Mintaméret:** A megfigyelések száma (n).

Hogyan történik ez a gyakorlatban:

- Kíváncsiak vagyunk a ξ mennyiség eloszlására a sokaságban.
- Megtervezzük a mintavételezést és a statisztikai kiértékelést. Ezen a ponton a mintaelemek véletlen számok, valószínűségi változók.
- Elvégezzük a mintavételezést, kiválasztunk elemeket a sokaságból. A mintarealizáció a ξ mennyiség értéke ezeken az egyedeken.
- Elvégezzük a statisztikai kiértékelést a konkrét mintarealizáción.

Hiányzó adatnak nevezzük azt, ha a mintarealizációban egyes értékek nem állnak rendelkezésre. Lehetséges okai: sikertelen volt a megfigyelés, a megfigyelt érték elveszett, a megfigyelt értéket szándékosan töröltük, stb.

Az adathiány fontosabb típusai:

- **Teljesen véletlenszerű adathiány:** a hiányzás ténye független az adat igazi értékétől.

Ilyenkor dolgozhatunk csak a rendelkezésre álló adatokból, ugyanis ezek is reprezentálják a teljes sokaságot.

- **Nem véletlenszerű adathiány:** a hiányzás ténye valamilyen módon függ az adat igazi értékétől.

Példa: politikai közvélemény-kutatás során az egyes pártok támogatói eltérő arányban tagadják meg a válaszadást.

Ebben az esetben a rendelkezésre álló adatok nem reprezentálják a teljes sokaságot, tehát hiba csak ezeket az adatokat figyelembe venni. Valamilyen módon „rekonstruálni” kell a hiányzó adatokat.

Statisztikai becslések, alapstatisztikák

A minta alapján adjunk becslést a sokaságot jellemző mutatószámokra.
Ötlet: tekintsünk úgy a megfigyelt elemekre, mint egy részsokaságra.
Számoljuk ki a kérdéses mutatószámokat ebben a részsokaságban!

Várható érték becslése: **empirikus várható érték, mintaátlag.**

$$E(\xi) \approx \bar{\xi} = E_n(\xi) = \frac{\xi_1 + \dots + \xi_n}{n}$$

Variancia becslése: **empirikus variancia, statisztikai variancia.**

$$\text{Var}(\xi) \approx \text{Var}_n(\xi) = \frac{(\xi_1 - \bar{\xi})^2 + \dots + (\xi_n - \bar{\xi})^2}{n}$$

Szórás becslése: **empirikus szórás, statisztikai szórás.**

$$D(\xi) \approx S_\xi = D_n(\xi) = \sqrt{\text{Var}_n(\xi)}$$

Az előző oldalon felsorolt becslések **erősen konzisztensek**:

$$E_n(\xi) \rightarrow E(\xi), \quad \text{Var}_n(\xi) \rightarrow \text{Var}(\xi), \quad D_n(\xi) \rightarrow D(\xi), \quad \text{amint } n \rightarrow \infty.$$

Tehát ezek a becslések nagy n esetén pontosak lesznek.

Probléma: kis n esetén $\text{Var}_n(\xi)$ és $D_n(\xi)$ tipikusan alábecsli az igazi varianciát és szórást. Megoldás: megnöveljük ezeket a becsléseket.

Korrigált empirikus variancia és korrigált empirikus szórás:

$$\text{Var}_n^*(\xi) = \frac{n}{n-1} \text{Var}_n(\xi), \quad S_\xi^* = D_n^*(\xi) = \sqrt{\text{Var}_n^*(\xi)}$$

Tulajdonságok:

- A korrigált becslések szintén erősen konzisztensek:

$$\text{Var}_n^*(\xi) \rightarrow \text{Var}(\xi), \quad D_n^*(\xi) \rightarrow D(\xi), \quad \text{amint } n \rightarrow \infty.$$

- A korrigált becslések kis n esetén pontosabbak.

Feladat: A kar férfi hallgatóinak testmagasságát vizsgáljuk.

Háttérváltozó: ξ = egy véletlenszerűen kiválasztott hallgató magassága.

Mintarealizáció: 180, 175, 188, 168, 173, 183.

$$E(\xi) \approx \bar{\xi} = E_6(\xi) = \frac{180 + 175 + 188 + 168 + 173 + 183}{6} = 177.8$$

$$\text{Var}(\xi) \approx \text{Var}_6(\xi) = \frac{(180 - 177.8)^2 + \dots + (183 - 177.8)^2}{6} = 43.81$$

$$D(\xi) \approx D_6(\xi) = \sqrt{43.81} = 6.62$$

A kis mintaméret miatt ($n = 6$) érdemes korrigálást végezni:

$$\text{Var}(\xi) \approx \text{Var}_6^*(\xi) = \frac{6}{5} 43.81 = 52.57, \quad D(\xi) \approx D_6^*(\xi) = \sqrt{52.57} = 7.25.$$

Foglaljuk össze, hogy mit kaptunk:

- átlagos testmagasság a sokaságban = $E(\xi) \approx 177.8$,
- a testmagasság szórása a sokaságban = $D(\xi) \approx 7.25$.

Ezt a két értéket publikációkban így szokták közölni: 177.8 ± 7.25 cm.

Mennyire pontos becslés a mintaátlag a várható értékre: $|\bar{\xi} - E(\xi)| = ?$

- A mintavételezéskor egy véletlenszerű mintarealizációt kapunk.
- A mintaátlag egy valószínűségi változó, mert a mintarealizációtól függ.
- Emiatt a becslési hiba is véletlen nagyságú.

Mennyi a becslés átlagos hibája: $E(|\bar{\xi} - E(\xi)|) = ?$

- Képzeletben sorra vesszük az összes lehetséges mintarealizációt.
- Mindegyik realizációból kiszámoljuk a $\bar{\xi}$ mintaátlagot.
- Ebből kapunk egy átlagos becslési hibát.

Standard hiba (standard error of the mean, s.e.m.):

$$SE = \frac{D_n^*(\xi)}{\sqrt{n}}$$

Szemléletes jelentése: az $E(\xi) \approx \bar{\xi}$ becslés átlagos hibája adott n esetén.

Hogyan értelmezzük a standard hibát?

- Ha a standard hiba kicsi, akkor a mintaátlag minden realizáció esetén pontos becslés lesz a várható értékre.
- Ha a standard hiba nagy, akkor vannak olyan realizációk, melyekre a mintaátlag pontatlan becslést ad a várható értékre.
- Fontos: a standard hiba nem alkalmas a becslés pontosítására!

Feladat: Határozzuk meg a standard hibát a hallgatók testmagasságára!

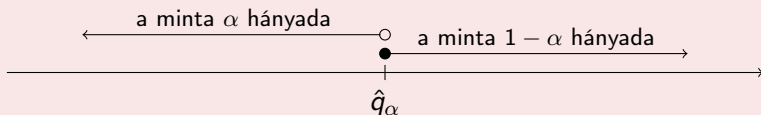
Amit tudunk: $n = 6$, $\bar{\xi} = E_6(\xi) = 177.8$, $D_6^*(\xi) = 7.25$.

Ekkor: $SE = 7.25/\sqrt{6} = 2.96$.

Foglaljuk össze, hogy mit kaptunk:

- Az ismeretlen várható értékre adott becslésünk: 177.8. Ez csak egy becslés, nem fogja pontosan telibe találni az igazi várható értéket.
- A standard hiba: 2.96. A mintaátlag várhatóan ennyivel tér el az igazi várható értéktől, átlagosan ekkora a becslés hibája.

Empirikus kvantilis, statisztikai kvantilis, percentilis: Az a $\hat{q}_\alpha$ szám, melyre teljesül, hogy mintaelemek α hányada kisebb, mint $\hat{q}_\alpha$.



Példa: empirikus medián

$$\hat{q}_{50\%} = \begin{cases} \text{a középső mintaelem,} & \text{ha } n \text{ páratlan,} \\ \text{a két középső átlaga,} & \text{ha } n \text{ páros.} \end{cases}$$

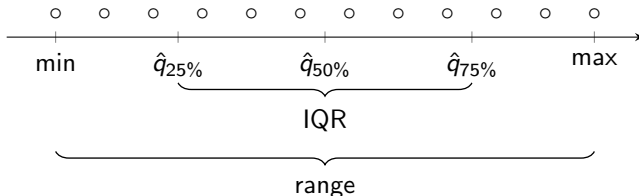
A ξ valószínűségi változó α -kvantilise az a q_α osztáspont, ami az egész sokaságot bontja fel ugyanilyen arányok szerint. Kérdés: hogyan becsülhető a sokaság kvantilise a minta alapján?

Ha a q_α elméleti kvantilis létezik és egyértelmű, akkor az empirikus kvantilis erősen konzisztens becslés: $\hat{q}_\alpha \rightarrow q_\alpha$ amint $n \rightarrow \infty$.

Néhány további statisztikai mutatószám:

- **Minimum, maximum:** a legkisebb és a legnagyobb érték a mintában.
- **Terjedelem (range):** maximum – minimum
Ilyen hosszúságú intervallumon helyezkedik el a teljes minta.
- **Empirikus alsó kvartilis, empirikus felső kvartilis:** $\hat{q}_{25\%}$, $\hat{q}_{75\%}$
- **Interkvartilis terjedeleme:** $IQR = \hat{q}_{75\%} - \hat{q}_{25\%}$
Ilyen hosszúságú intervallumon helyezkedik el a minta középső 50%-a.

Példa: a mintaméret $n = 12$

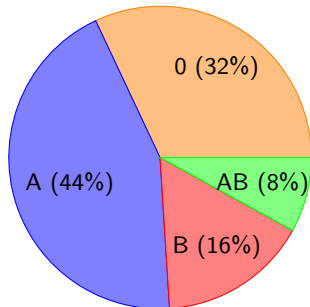
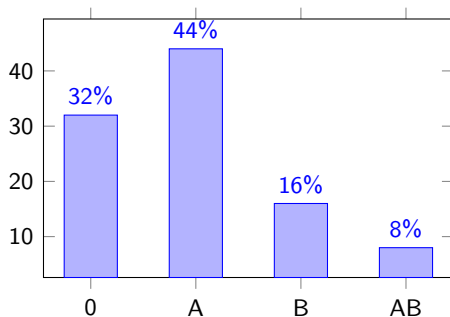


Statisztikai grafikonok

Legyen a ξ háttérváltozó diszkrét!

- **Oszlopdiaagram:** Oszlopokkal ábrázoljuk, hogy az egyes értékek hányszor (vagy milyen arányban) szerepelnek a mintában.
- **Kördiaagram:** Körcikkekkel reprezentáljuk a mintát, a középponti szögek arányosak az értékek megjelenési arányával.

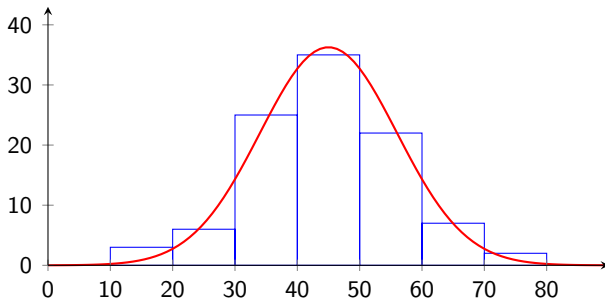
Példa: a vércsoportok aránya a magyar népesség belül.



Legyen a ξ háttérváltozó folytonos!

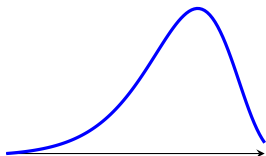
Hisztogram: Felbontjuk a számegyenest azonos hosszúságú intervallumokra. Minden intervallumra olyan magas oszlopot állítunk, ahány mintaelem az adott intervallumba esik.

Nagy elemszám (legalább $n \geq 100$) esetén a hisztogram egy jó grafikus becslés a sűrűségfüggvényre: a sűrűségfüggvény követi a hisztogram tetejét.

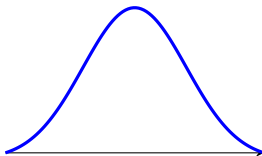


Ferdeség (skewness): egy olyan statisztikai mutatószám, ami a hisztogram és az illesztett sűrűségfüggvény szimmetriáját jellemzi.

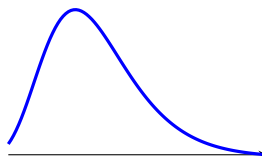
- $\text{skewness} = 0 \implies$ a hisztogram szimmetrikus
- $\text{skewness} > 0 \implies$ a hisztogram jobbra ferde
- $\text{skewness} < 0 \implies$ a hisztogram balra ferde



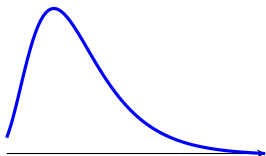
$\text{skewness} = -1$



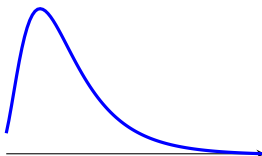
$\text{skewness} = 0$



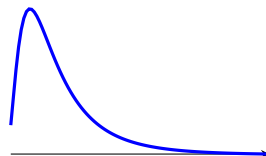
$\text{skewness} = 1$



$\text{skewness} = 1.5$

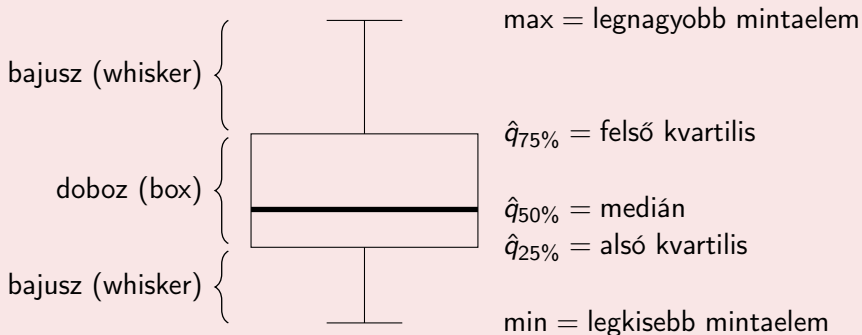


$\text{skewness} = 2$



$\text{skewness} = 3$

Boxplot outlier elemek nélkül:



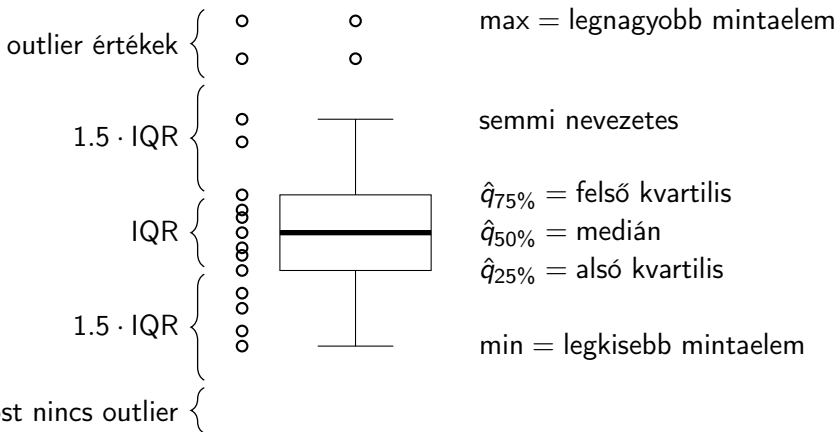
További mutatószámok az ábrán:

- Terjedelem (range) = $\max - \min$ = a boxplot magassága
- IQR = felső kvartilis – alsó kvartilis = a doboz magassága

Outlier értékek: Azok a mintaelemek, melyek

- kisebbek, mint $\hat{q}_{25\%} - 1.5 \cdot \text{IQR}$;
- vagy nagyobbak, mint $\hat{q}_{75\%} + 1.5 \cdot \text{IQR}$.

Általában a boxploton külön ábrázoljuk az outlier értékeket:



Egyes statisztikai adatelemző eljárások megkövetelik, hogy a minta normális eloszlásból származzon. Ezt előzetesen ellenőrizni kell teszteléssel vagy grafikus illeszkedésvizsgálattal.

Grafikus illeszkedésvizsgálat:

- Grafikonok alapján nyilatkozunk arról, hogy a vizsgált mennyiség milyen sűrűségfüggvénnyel írható le.
- Ez nem egzakt módszer, a döntés részben szubjektív!

Ha a minta valóban normális eloszlásból származik, akkor:

- a normális eloszlás sűrűségfüggvénye illeszkedik a hisztogramhoz;
- skewness ≈ 0 ;
- az outlier értékek aránya körülbelül 1%.

Egyes statisztikai módszereknél az is elég, ha a minta „közel normális.” Ezt tankönyvenként különböző módon definiálják, nálunk a következőt jelenti:

- a sűrűségfüggvény egymódusú;
- a minta közel szimmetrikus: $-1 \leq \text{skewness} \leq +1$.

Konfidencia intervallumok

A statisztikában egy minta alapján kétféle formában becsülhetünk meg egy ismeretlen mutatószámot (várható értéket, szórást, stb.)

- **Pontbecslés:** A mutatószámot egyetlen számmal becsüljük meg.
- **Intervallumbecslés:** Egy intervallumot adunk meg, amely nagy megbízhatósággal tartalmazza a kérdéses mutatószámot.

Legyen $\xi_1, \dots, \xi_n$ statisztikai minta egy ξ valószínűségi változóra, és legyen $\alpha \in (0, 1)$. A minta alapján felírt $[a, b]$ intervallum egy $1 - \alpha$ megbízhatóságú **konfidencia intervallum** a **várható értékre**, ha

$$P(E(\xi) \in [a, b]) = 1 - \alpha.$$

A megbízhatóság tipikus értékei: 90%, 95% vagy 99%.

A biostatisztikában jellemzően a 95%-ot használják.

A konfidencia intervallum hasonló módon definiálható tetszőleges más mutatószámra is (szórás, variancia, medián, stb.)

Megjegyzések:

- A félév folyamán számos mutatószámra kapunk majd konfidencia intervallumot. Az intervallumokra nem tanulunk elméleti formulákat, ugyanis ez túl sok időt igényelne.
- A konfidencia intervallumok általában megkövetelik, hogy a minta normális eloszlásból származzon. Ha ez teljesül, akkor az intervallum az előírt megbízhatósággal tartalmazza a kérdéses mutatószámot tetszőleges mintaméret esetén.
- Viszont az általunk tanult konfidencia intervallumok nem normális eloszlás esetén is alkalmazhatóak, ha a mintaméret elegendően nagy.

Kérdés: Mit jelent az „elegendően nagy” mintaméret?

Erre a kérdésre nincs egyszerű válasz, a szükséges mintaméret függ attól, hogy a ξ változó eloszlása mennyire hasonlít a normális eloszláshoz:

- Közel normális eloszlás esetén 20–30 mintaelem elég szokott lenni.
- Ferde eloszlás esetén jellemzően kell legalább 50 mintaelem.
- Nagyon ferde eloszlás esetén érdemesebb inkább más módszerrel felírni a konfidencia intervallumot.

Hipotézisvizsgálat

A hipotézisvizsgálat (hypothesis testing) alapfogalmai:

- Adott egy ξ háttérváltozó és egy $\xi_1, \dots, \xi_n$ statisztikai minta.
- **Nullhipotézis (H_0 , null hypothesis):** Egy állítás a ξ változóra.
- **Alternatív hipotézis (H_A , alternative hypothesis):** Egy másik állítás a ξ változóra.
- Tudjuk, hogy a két hipotézis közül valamelyik igaz.
Feladat: döntsük el a minta alapján, hogy H_0 vagy H_A igaz!

Például: $H_0 : E(\xi) = 2$, $H_A : E(\xi) = 4$.

A továbbiakban a kurzuson az alternatív hipotézis mindig a nullhipotézis tagadása lesz. Ezt kétoldali (two-sided) alternatívának nevezik. Például:

- $H_0 : E(\xi) = 2$, $H_A : E(\xi) \neq 2$.
- $H_0 : D(\xi) = D(\eta)$, $H_A : D(\xi) \neq D(\eta)$.

A hipotézisvizsgálat menete:

- Eldöntjük, hogy milyen módszerrel tesztelünk.
- Választunk egy alkalmas **szignifikancia szintet**: $\alpha \in (0, 1)$.
- A minta alapján kiszámoljuk a **p-értéket**: p-value.
- Ha $p\text{-value} \geq \alpha$, akkor **elfogadjuk** a nullhipotézist.
Ha $p\text{-value} < \alpha$, akkor **elvetjük** a nullhipotézist.

Az egész olyan, mint egy bírósági tárgyalás:

- A nullhipotézis a vádlott szava („ártatlan vagyok”).
- A statisztikai minta a bizonyítékok halmaza.
- A p-érték mindig a $[0, 1]$ intervallumba esik. Azt fejezi ki, hogy a minta alapján a nullhipotézis mennyire tűnik igaznak. A bizonyítékok milyen mértékben támasztják alá a vádlott szavát.
- Az α szignifikancia szint egy küszöbérték:
 - Ha $p\text{-value} \geq \alpha$, akkor a bíró hisz a vádlottnak, és felmenti.
 - Ha $p\text{-value} < \alpha$, akkor nem hisz neki, és elítéli.

Milyen hibákat véthetünk a hipotézisvizsgálat során?

- **Elsőfajú hiba (type I error):** Elvetjük az igaz nullhipotézist, tehát börtönbe küldünk egy ártatlant. Valószínűsége:
$$P(\text{elvetjük } H_0\text{-t, ha igaz}).$$
- **Másodfajú hiba (type II error):** Elfogadjuk a hamis nullhipotézist, tehát felmentünk egy bűnöst. Valószínűsége:
$$P(\text{elfogadjuk } H_0\text{-t, ha hamis}).$$

A hipotézisvizsgálat során a két fajta hiba egymással ellentétesen mozog, ezeket nem lehet egyszerre alacsonyan tartani. A célunk az elsőfajú hiba kontrollálása.

Szignifikancia szint (α): az elsőfajú hiba előre megadott értéke.

A szignifikancia szint kicsi szokott lenni: tipikusan 1%, 5% vagy 10%.
A biostatisztikában jellemzően: $\alpha = 5\%$.

A gyakorlatban mit jelent a nullhipotézis elfogadása?

- A nullhipotézist akkor fogadjuk el, ha a minta nem cáfolja meg. A bíróság felmenti a vádlottat, ugyanis nincs ellene bizonyíték.
- De ez még nem jelenti azt, hogy a nullhipotézis biztosan igaz. Gyakran elfogadunk hamis nullhipotéziseket is.
- A nullhipotézis elfogadása csak annyit jelent, hogy nem különbözik szignifikáns (=statisztikailag kimutatható) mértékben az igazságtól.
- Ilyenkor a nullhipotézis alkalmazható a valóság modellezésére.

Mi jelent a nullhipotézis elvetése?

- A nullhipotézist akkor vetjük el, ha a minta határozottan megcáfolja. Ebben az esetben nagy mennyiségű bizonyíték áll az ítélet mögött.
- A nullhipotézis elvetése már kellően alátámasztott tudományos eredménynek tekinthető.
- Ha elvetjük a nullhipotézist, akkor szignifikáns mértékben különbözik az igazságtól. Emiatt a valóság modellezésére sem alkalmazható.

Az egymintás t-próba

Egymintás t-próba (One sample t test)

Feltevések:

- ξ normális eloszlású változó ismeretlen várható értékkel,
- μ_0 egy tetszőleges hipotetikus érték.

Nullhipotézis: $H_0 : E(\xi) = \mu_0$.

Megjegyzések:

- A próba pontosan akkor fogadja el a μ_0 hipotetikus várható értéket, ha μ_0 az $1 - \alpha$ megbízhatóságú konfidencia intervallumba esik. A konfidencia intervallum értelmezhető olyan módon, mint a „hihető” várható értékek halmaza.
- Az egymintás t-próba **robustus** a normalitásfeltételre nézve. Ha a minta nem normális eloszlású, de a mintaméret elegendően nagy, akkor a t-próba használható a várható érték tesztelésére.

A páros t-próba

Legyenek ξ és η valószínűségi változók. Két statisztikai minta:

- $\xi_1, \dots, \xi_n$ független megfigyelések ξ -re,
- $\eta_1, \dots, \eta_m$ független megfigyelések η -ra.

A minták tipikusan két fajta kapcsolatban állhat egymással:

- **Független minták (independent samples):** A két minta egymástól független mintavételezésből származik, ezért nincs közöttük kapcsolat.
- **Összetartozó minták (paired samples, related samples):**
A ξ_i és az η_i megfigyelés minden i esetén a sokaság ugyanazon egyedére vonatkozik. Ebben az esetben mindig $n = m$. Példák:
 - $\xi_1, \dots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben
 $\eta_1, \dots, \eta_n$: ugyanezen hallgatók testmagassága egy 5 évvel ezelőtt
 - $\xi_1, \dots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben
 $\eta_1, \dots, \eta_n$: ugyanezen hallgatók édesapjának testmagassága

Páros t-próba (Paired samples t test)

Cél a várható értékek tesztelése összetartozó minták esetén.

Feltevések:

- ξ és η együttesen normális eloszlásúak,
- $\xi_1, \dots, \xi_n$ és $\eta_1, \dots, \eta_n$ összetartozó minták.

Nullhipotézis: $H_0 : E(\xi) = E(\eta)$.

Megjegyzések:

- ξ és η **együttesen normális eloszlású**, ha tetszőleges a és b valós számok esetén $a\xi + b\eta$ normális eloszlású. Ez egy kicsivel több annál, hogy ξ és η külön-külön normális eloszlású.
- A páros t-próba **robosztus** a normalitásfeltételre nézve.
- Konfidencia intervallumot is adhatunk a várható értékek különbségére:

$$P\left(E(\xi) - E(\eta) \in [a, b]\right) = 1 - \alpha.$$

Egyszempontos varianciaanalízis és a Levene-teszt

Tegyük fel, hogy a teljes sokaság $r \geq 2$ darab csoportra bontható fel! Azt vizsgáljuk, hogy egy adott mennyiség szempontjából van-e statisztikailag kimutatható különbség a csoportok között.

Véletlenszerűen kiválasztunk egy elemet a sokaságból, ξ a mennyiség értéke a kiválasztott elemen! Ekkor:

- $E(\xi) =$ átlag a teljes sokaságban,
- $D(\xi) =$ a teljes sokaságon belüli szórás,
- $E(\xi|j) =$ a j . csoporton belüli átlag,
- $D(\xi|j) =$ a j . csoporton belüli szóródás.

A cél a következő nullhipotéziseket tesztelni:

- $H_0 : \text{azonosak a csoportonkénti átlagok, } E(\xi|1) = \dots = E(\xi|r),$
- $H_0 : \text{azonosak a csoportonkénti szórások, } D(\xi|1) = \dots = D(\xi|r).$

Független mintákat veszünk az egyes csoportokból:

- minta az 1. csoportra: $\xi_{11}, \xi_{12}, \dots, \xi_{1n_1}$
- minta a 2. csoportra: $\xi_{21}, \xi_{22}, \dots, \xi_{2n_2}$
- ...
- minta az r . csoportra: $\xi_{r1}, \xi_{r2}, \dots, \xi_{rn_r}$

Teljes minta elemszáma: $n = n_1 + n_2 + \dots + n_r$

Becslések a minta alapján:

$$E(\xi|j) \approx \bar{\xi}_j = \frac{\xi_{j1} + \xi_{j2} + \dots + \xi_{jn_j}}{n_j} = \text{mintaátlag a } j. \text{ csoportban}$$

$$D(\xi|j) \approx D_{n_j}^*(\xi|j) = \text{korrigált empirikus szórás a } j. \text{ csoportban}$$

Ugyanígy becsülhető minden mutatószám csoportonkénti bontásban is.

Szórás tesztek

A cél azt tesztelni, hogy a csoportonkénti szórások azonosak, tehát:

$$H_0 : D(\xi|1) = \dots = D(\xi|r)$$

Feltevés: a ξ változó minden csoportban normális eloszlást követ.

Alkalmazható tesztek:

Teszt	Csoportok száma	Robusztus?	Ajánlott?
F-próba	$r = 2$	Nem	Nem!!!
Bartlett-teszt	$r \geq 2$	Kis mértékben	Van jobb
Levene-teszt	$r \geq 2$	Igen	Igen
Brown–Forsythe-teszt	$r \geq 2$	Igen	Igen

Egyszempontos varianciaanalízis (Analysis of Variances, ANOVA)

A cél azt tesztelni, hogy a csoportonkénti átlagok azonosak, tehát:

$$H_0 : E(\xi|1) = \dots = E(\xi|r).$$

Feltevések:

- a ξ változó minden csoporton belül normális eloszlást követ,
- azonosak a csoportonkénti szórások, tehát $D(\xi|1) = \dots = D(\xi|r)$

Megjegyzések:

- Akkor fogadjuk el a nullhipotézist, ha $\bar{\xi}_1 \approx \bar{\xi}_2 \approx \dots \approx \bar{\xi}_r$.
- Az ANOVA robusztus a normalitásfeltételekre nézve.
- Az ANOVA nem robusztus a szórásfeltételre! A szórásokat teszteléssel lehet ellenőrizni. Ha a szórások nem azonosak, akkor alkalmazzuk inkább a **Welch-féle F-próbát**.

Lineáris és nemlineáris regresszió

Két folytonos valószínűségi változó: ξ és η .

Feladat: adjunk becslést az η alapján a ξ változóra! Példák:

- Lineáris regresszió: $\xi \approx a\eta + b$
- Exponenciális regresszió: $\xi \approx e^{a\eta+b}$
- Reciprokos regresszió: $\xi \approx a/\eta + b$

A változók elnevezése:

- ξ a **függő változó** avagy **eredményváltozó**,
- η a **független változó** avagy **magyarázó változó**.

Kérdések:

- Mely a és b valós számokra lesz a legpontosabb a becslés?
- Mennyire pontos a becslés?

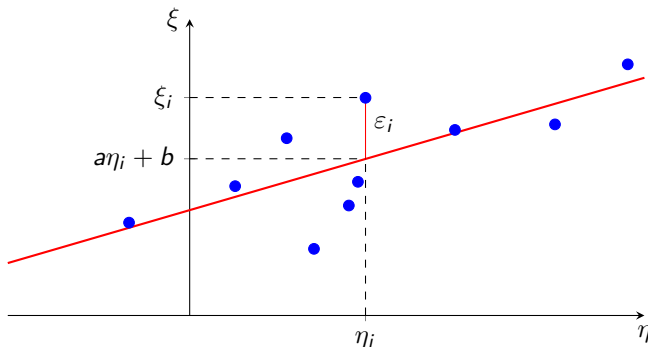
Lineáris regresszió: $\xi \approx a\eta + b$

A becslés hibája: $\varepsilon = \xi - (a\eta + b)$

$\xi = (a\eta + b) + \varepsilon = \text{predikciós tag} + \text{hibatag (egyedi hatás)}$

Összetartozó minták: $\xi_1, \dots, \xi_n$ és $\eta_1, \dots, \eta_n$

Lineáris regresszió a mintaelemeken: $\xi_i = (a\eta_i + b) + \varepsilon_i$



Hogyan határozzuk meg a két paraméter értékét?

- A mintához legjobban illeszkedő egyenest keressük, tehát minimalizálni akarjuk a hibatagokat.
- **Legkisebb négyzetes becslés (least squares estimation):**

$$S(a, b) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (\xi_i - a\eta_i - b)^2 \longrightarrow \min$$

- Deriváljuk az S függvényt a két változó szerint, és megoldjuk a következő egyenleteket:

$$\frac{dS}{da} = 0, \quad \frac{dS}{db} = 0.$$

- Az egyenletrendszer megoldása:

$$a = \frac{r_n(\xi, \eta) D_n^*(\xi)}{D_n^*(\eta)}, \quad b = \bar{\xi} - a\bar{\eta}.$$

Itt $r_n(\xi, \eta)$ a Pearson-féle korrelációs együttható. (Lásd később.)

Mennyire jó az illeszkedés a regressziós egyeneshez? Két becslés:

- Lineáris regresszió: $\xi \approx a\eta + b$

A becslés teljes hibája a mintán:

$$\text{SSE} = \text{sum of squares (errors)} = \sum_{i=1}^n \varepsilon_i^2$$

- Becslés a mintaátlaggal: $\xi \approx \bar{\xi}$

A becslés teljes hibája a mintán:

$$\text{SST} = \text{sum of squares (total)} = \sum_{i=1}^n (\xi_i - \bar{\xi})^2$$

Megmutatható, hogy $\text{SSE}/\text{SST} \in [0, 1]$.

A tört jelentése: a lineáris regresszió teljes becslési hibája ennyi százaléka a várható értékkel történő becslés teljes becslési hibájának.

Meghatározottsági együttható (coefficient of determination):

$$R^2 = 1 - \text{SSE}/\text{SST}.$$

Jelentése: a lineáris regresszió teljes becslési hibája ennyi százalékkal kisebb, mint a mintaátlaggal történő becslés teljes becslési hibájának.

Tulajdonságai:

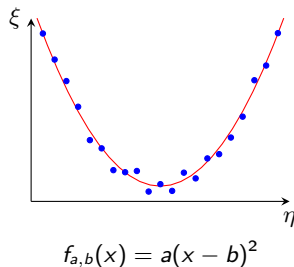
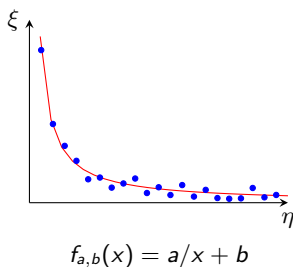
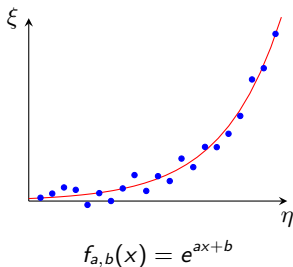
- $0 \leq R^2 \leq 1$.
- Minél nagyobb az R^2 értéke, annál pontosabb a lineáris becslés.
- Ha $R^2 \leq 0.5$, akkor a lineáris becslés nagyon pontatlan.

Megjegyzések:

- Lineáris regressziós esetén: $R^2 = r_n^2(\xi, \eta)$.
- A meghatározottsági együtthatót nemlineáris regresszió esetén is alkalmazzuk, de ekkor SSE az adott regresszió összesített becslési hibája. Viszont ekkor már nincs kapcsolat a korrelációs együtthatóval.

Előfordul, hogy a változók közötti kapcsolat nem lineáris jellegű. Ilyenkor a kapcsolatot nemlineáris függvény segítségével keressük. Például:

- Exponenciális regresszió: $\xi \approx e^{a\eta+b}$,
- Reciprokos regresszió: $\xi \approx a/\eta + b$,
- Másodfokú regresszió: $\xi \approx a(\eta - b)^2$.



A regresszió általános modellje:

$$\xi = f_{a,b,\dots}(\eta) + \varepsilon = \text{predikciós tag} + \text{hibatag}$$

A formulában:

- f egy adott típusú függvény;
- $a, b, \dots$ a függvény paraméterei.

A cél a paramétereket olyan módon meghatározni, hogy a teljes hiba minimális legyen. Legkisebb négyzetek módszere:

$$S(a, b, \dots) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n \left(\xi_i - f_{a,b,\dots}(\eta_i) \right)^2 \longrightarrow \min$$

Megjegyzések:

- A nemlineáris esetben nincsen szép formula a paraméterekre.
- Bizonyos esetekben a nemlineáris regresszió visszavezethető lineárisra.
- A becslés pontosságát most is az R^2 számszerűsíti.

Hogyan vezethető vissza a nemlineáris regresszió lineárisra? Legyen g egy adott függvény, nincsen benne ismeretlen paraméter.

1. speciális eset: $\xi \approx ag(\eta) + b$.

Példa: $\xi \approx a \ln \eta + b$.

Bevezetünk egy új változót: $\eta' = g(\eta)$.

Lineáris regressziót kapunk: $\xi \approx a\eta' + b$.

2. speciális eset: $\xi \approx g(a\eta + b)$.

Ha a g függvénynek létezik az inverze, akkor: $g^{-1}(\xi) \approx a\eta + b$

Példa: $\xi \approx \ln(a\eta + b)$, $e^\xi \approx a\eta + b$.

Bevezetünk egy új változót: $\xi' = g^{-1}(\xi)$.

Lineáris regressziót kapunk: $\xi' \approx a\eta + b$.

Korrelációs együtthatók

Két folytonos valószínűségi változó: ξ és η .

Független változók: nincs kapcsolat a változók között, nem tartalmaznak információt egymásra nézve. Ilyenkor nincs értelme regressziót végezni.

Függő változók: kapcsolat van közöttük. Az egyik változó ismeretében valamilyen(!) módszerrel becslés adható a másik változó értékére.

- Erős kapcsolat: az egyik változó ismeretében pontos becslés adható.
- Gyenge kapcsolat: csak pontatlan becslésekre van lehetőség.
- Pozitív irányú kapcsolat: a két változó értéke jellemzően azonos irányba változik. Ha ξ nő, akkor η is nő.
- Negatív irányú kapcsolat: a két változó értéke jellemzően ellentétes irányba változik. Ha ξ nő, akkor η csökken.

Két folytonos valószínűségi változó: ξ és η .

Összetartozó minták: $\xi_1, \dots, \xi_n$ és $\eta_1, \dots, \eta_n$

Korrelációs együtthatók: Olyan statisztikai mutatószámok, melyek a változók közötti kapcsolat irányát és erősségét mérik.

Empirikus kovariancia és Pearson-féle korrelációs együttható:

$$C_n(\xi, \eta) = \frac{1}{n-1} \sum_{i=1}^n (\xi_i - \bar{\xi})(\eta_i - \bar{\eta}), \quad r_n(\xi, \eta) = \frac{C_n(\xi, \eta)}{D_n^*(\xi) D_n^*(\eta)}.$$

Fontosabb tulajdonságok:

- Szimmetria: $r_n(\xi, \eta) = r_n(\eta, \xi)$
- Lehetséges értékek: $-1 \leq r_n(\xi, \eta) \leq +1$
- A Pearson-féle korrelációs együttható a két változó közötti **lineáris kapcsolat** irányát és erősségét jellemzi.

Kérdés: Mit jelent a lineáris kapcsolat?

Lineáris regressziót végzünk a minta alapján: $\xi \approx a\eta + b$.

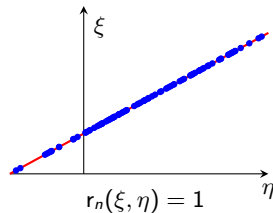
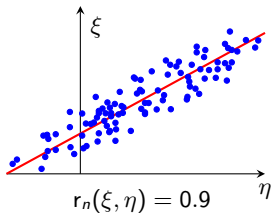
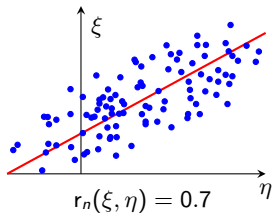
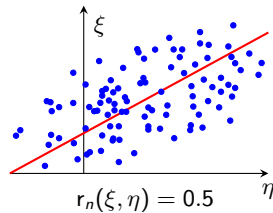
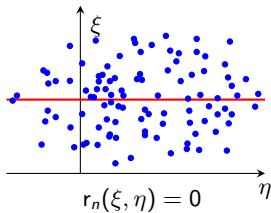
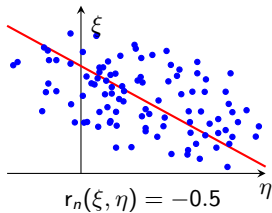
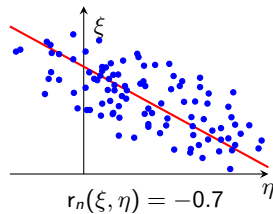
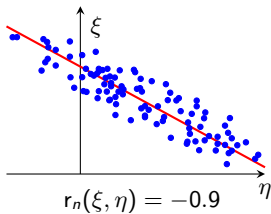
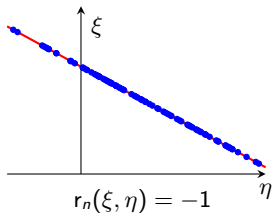
A Pearson-féle korrelációs együtthetó jellemzi a kapcsolat irányát:

- Ha $r_n(\xi, \eta) > 0$, akkor $a > 0$, tehát a változók között pozitív irányú kapcsolat van.
- Ha $r_n(\xi, \eta) < 0$, akkor $a < 0$, tehát a változók között negatív irányú kapcsolat van.

A Pearson-féle korrelációs együtthetó jellemzi a kapcsolat erősségét:

- Ha $r_n(\xi, \eta) \approx \pm 1$, akkor a változók között erős lineáris kapcsolat van, tehát a regressziós becslés pontos.
- Ha $r_n(\xi, \eta) \approx 0$, akkor a változók függetlenek vagy gyenge lineáris kapcsolat van közöttük. Emiatt a regressziós becslés pontatlan.

A Pearson-féle korrelációs együtthetó hátránya: csak a lineáris kapcsolatot képes mérni. Ha a változók között van kapcsolat, de nem lineáris jellegű, akkor ezt nem mindig detektálja.



Korrelációs teszt

Cél a függetlenség tesztelése összetartozó minták alapján.

Feltevések:

- ξ és η együttesen normális eloszlású változók.
- $\xi_1, \dots, \xi_n$ és $\eta_1, \dots, \eta_n$ összetartozó minták.

Nullhipotézis: $H_0 : \xi$ és η függetlenek.

Megjegyzések:

- A teszt robusztus a normalitásfeltételre, közel szimmetrikus folytonos eloszlású változók esetén lehet alkalmazni.
- Ha van kapcsolat a változók között, de ezt a korrelációs együttható nem tudja detektálni, akkor a teszt elfogadja a nullhipotézist.
Pearson-féle korreláció: csak a lineáris kapcsolatot érzékeli.