Biostatisztika

Szűcs Gábor

SZTE TTIK Bolyai Intézet

2023/24 tavaszi félév

Események valószínűsége, valószínűségi változók

A valószínűségszámítás 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.

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

A továbbiakban a félév folyamán:

- Adott egy véges sokaság (egy halmaz, például egy populáció).
- Egy mennyiség (testtömeg, utódok száma, stb.) eloszlását vizsgáljuk.
- Véletlenszerűen kiválasztunk egy elemet/egyedet a sokaságból: mindegyik elemnek azonos az esélye. Emiatt:

$$\text{valószínűség} = \text{arány a sokaságon belül}$$

- ξ = a vizsgált mennyiség értéke a kiválasztott elem/egyed esetében.
- ξ egy véletlen szám, 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

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

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

$$\mathrm{P}\big(50 \le \xi \le 60\big) = 75\% = 0.75 \qquad \mathrm{P}\big(\xi \ge 55\big) = 50\% = 0.5$$

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

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

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

A ξ valószínűségi változóra vonatkozó fontosabb mutatószámok:

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

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

- **Variancia, szórásnégyzet:**

$$\mathrm{Var}(\xi) = \mathrm{E}\Big(\big[\xi - \mathrm{E}(\xi)\big]^2\Big) = \frac{\big[x_1 - \mathrm{E}(\xi)\big]^2 + \ldots + \big[x_N - \mathrm{E}(\xi)\big]^2}{N}$$

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

A szórás szemléletes jelentése: várható értéktől vett átlagos eltérés.

$$\mathrm{D}(\xi) = \sqrt{\mathrm{E}\Big(\big[\xi - \mathrm{E}(\xi)\big]^2\Big)} \approx \mathrm{E}\Big(\sqrt{\big[\xi - \mathrm{E}(\xi)\big]^2}\Big) = \mathrm{E}\Big(\big|\xi - \mathrm{E}(\xi)\big|\Big)$$

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.

- $\xi =$ kiválasztott gyerek tömege
- Várható érték (átlagos érték):

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

- Variancia, szórásnégyzet:

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

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

Probléma: nagy méretű sokaságok esetén ezt kényelmetlen végigszámolni.

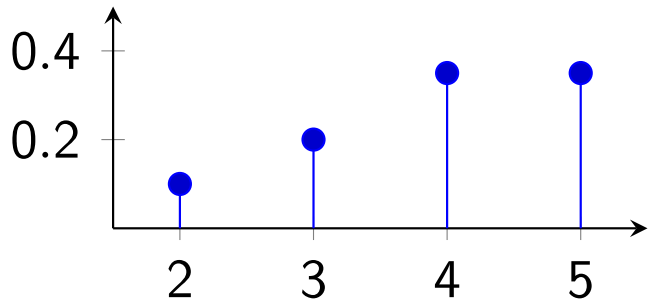
Diszkrét valószínűségi változók

Diszkrét valószínűségi változó: Olyan változó, melynek véges 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

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

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

Feladat: Mennyi a ξ változó lehetséges értékeinek összvalószínűsége?

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

Feladat: A fészkek mekkora hányadában található legfeljebb 3 tojás?

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

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):

$$\mathrm{E}(\xi) = \frac{\overbrace{2+\ldots+2}^{0.1N \text{ db}}+\overbrace{3+\ldots+3}^{0.2N \text{ db}}+\overbrace{4+\ldots+4}^{0.35N \text{ db}}+\overbrace{5+\ldots+5}^{0.35N \text{ db}}}{N}$$

$$= \frac{2\cdot 0.1N + 3\cdot 0.2N + 4\cdot 0.35N + 5\cdot 0.35N}{N}$$

$$= 2\cdot 0.1 + 3\cdot 0.2 + 4\cdot 0.35 + 5\cdot 0.35 = 3.95$$

Házi feladat: Adjuk meg ugyanilyen módon ξ varianciáját és szórását!

Legyen ξ diszkrét valószínűségi változó!

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

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

Várható érték (átlagos érték):

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

Variancia:

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

Eredmény: $\mathrm{Var}(\xi) = 0.95$

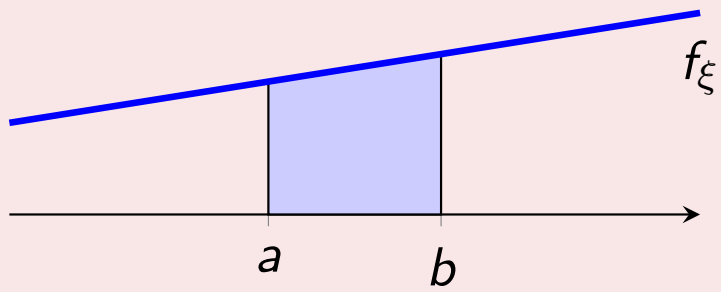
Szórás (várható értéktől vett átlagos eltérés): $\mathrm{D}(\xi) = \sqrt{\mathrm{Var}(\xi)} = 0.97$

Folytonos valószínűségi változók

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

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

$$\mathrm{P}(a \leq \xi \leq b) = \int_a^b f_\xi(x)\,dx$$



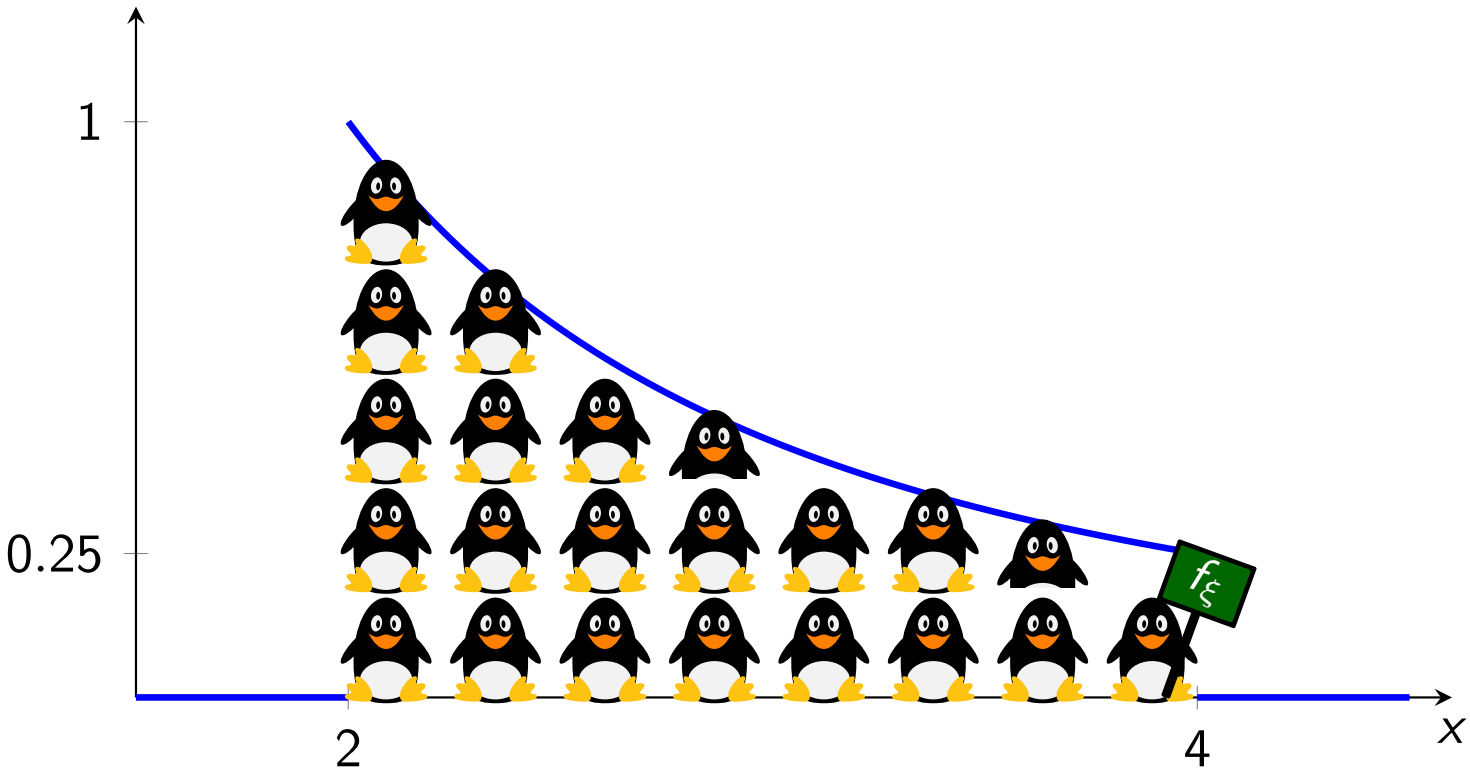
Egy mennyiség eloszlását vizsgáljuk egy sokaságon belül. Véletlenszerűen kiválasztunk egy elemet, és ξ a mennyiség értéke ezen elem esetében.

azon elemek aránya, melyeknél a vizsgált mennyiség a és b közé esik

$= \mathrm{P}(a \leq \xi \leq b) =$ görbe alatti terület a és b között

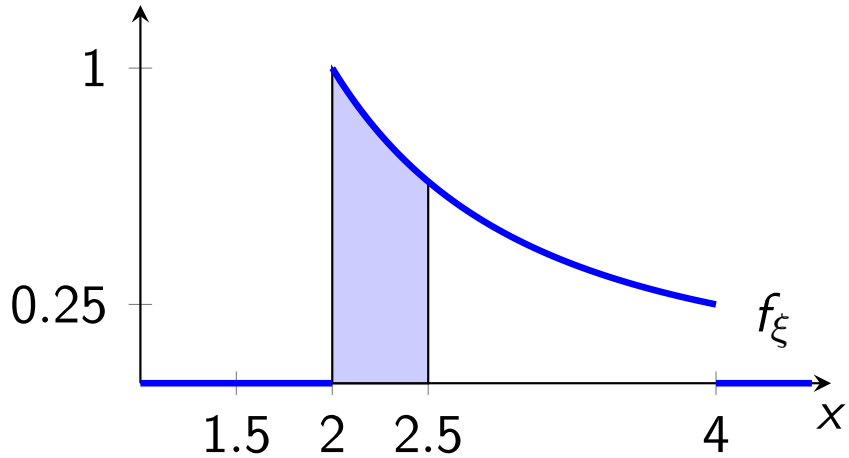
A sűrűségfüggvény szemléletes jelentése:

- 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?

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



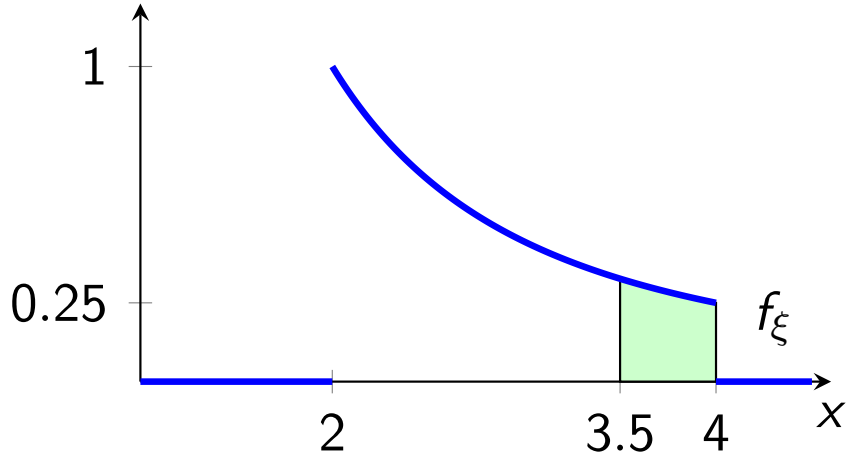
$$\mathrm{P}\big(1.5 \le \xi \le 2.5\big) = \int_{1.5}^{2.5} f_\xi(x)dx = \int_{1.5}^{2} 0\,dx + 4\int_{2}^{2.5} x^{-2}dx$$

$$= 0 + 4\left[\frac{x^{-1}}{-1}\right]_2^{2.5} = 4\left[\frac{1}{-x}\right]_2^{2.5} = 4\left(\frac{1}{-2.5} - \frac{1}{-2}\right) = 4 \cdot 0.1 = 0.4$$

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

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

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



$$\mathrm{P}(\xi \ge 3.5) = \mathrm{P}\big(3.5 \le \xi \le +\infty\big) = \int_{3.5}^{+\infty} f_\xi(x)dx = 4\int_{3.5}^{4} x^{-2}dx$$

$$= 4\left[\frac{1}{-x}\right]_{3.5}^{4} = 4\left(\frac{1}{-4} - \frac{1}{-3.5}\right) \approx 0.14 = 14\%$$

Feladat: Milyen testtömegek fordulnak elő a populációban?

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

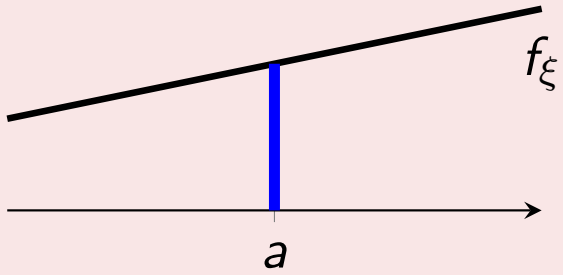
Most: $R_\xi = [2, 4]$

Folytonos változó esetén miért nem a valószínűségeloszlással számolunk?

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

Bizonyítás:

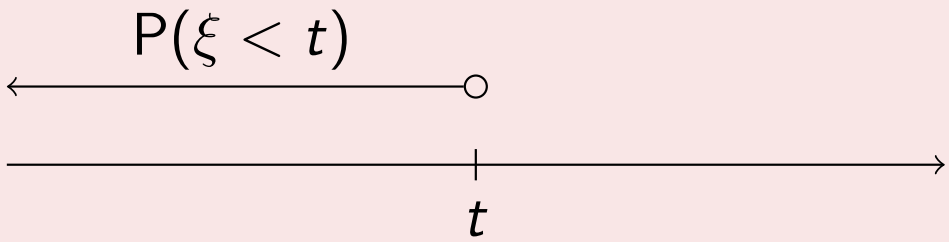
$$\mathrm{P}(\xi = a) = \mathrm{P}\big(a \leq \xi \leq a\big) = \int_a^a f_\xi(x)\,dx = 0$$



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

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

$$F_\xi(t) = \mathrm{P}(\xi < t)$$

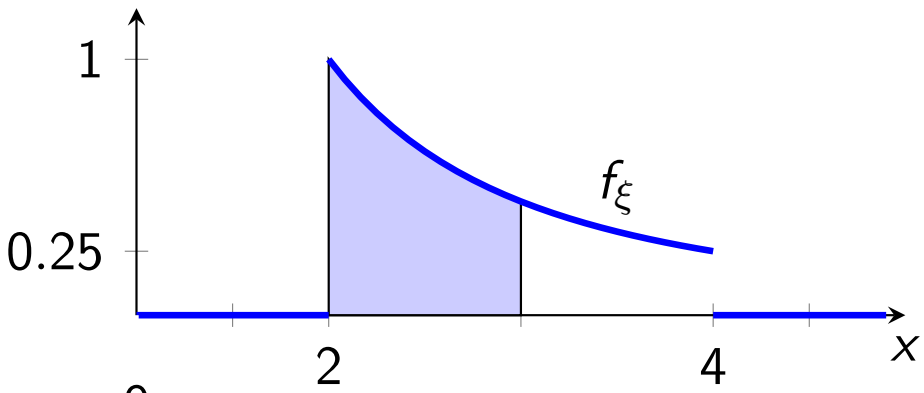


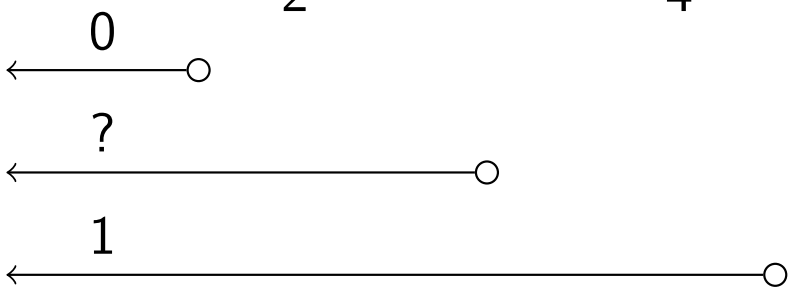
Szemléletes jelentés:

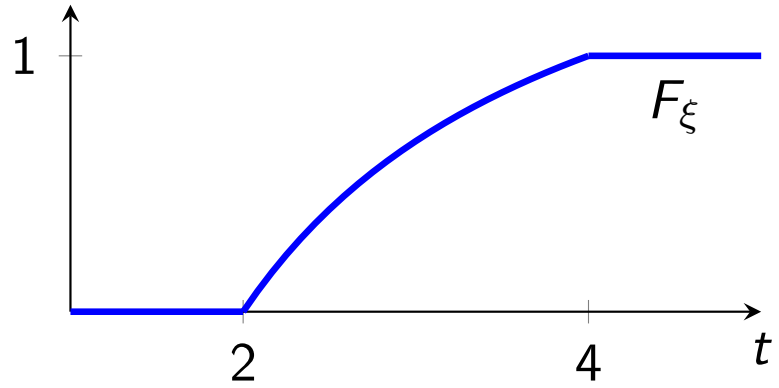
$F_\xi(t) =$ azon egyedek aránya a sokaságban, ahol ξ kisebb, mint t

Feladat: Írjuk fel az eloszlásfüggvényt a jelen feladatban!

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







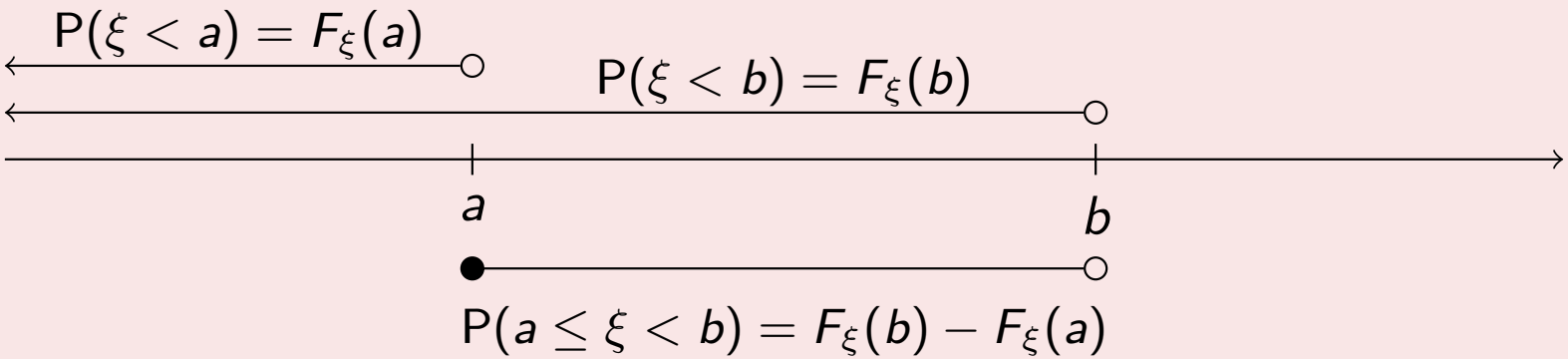
$$F_\xi(t) = \begin{cases} 0, & t < 2, \\ 2 - 4/t, & 2 \le t \le 4, \\ 1, & t > 4. \end{cases}$$

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

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

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

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

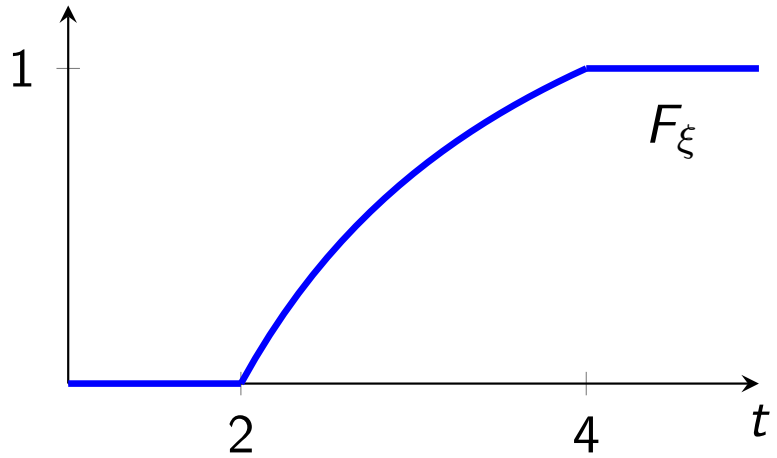


Speciálisan ha a ξ változó folytonos eloszlású, akkor

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

Vissza a pingvines feladathoz...

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



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

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

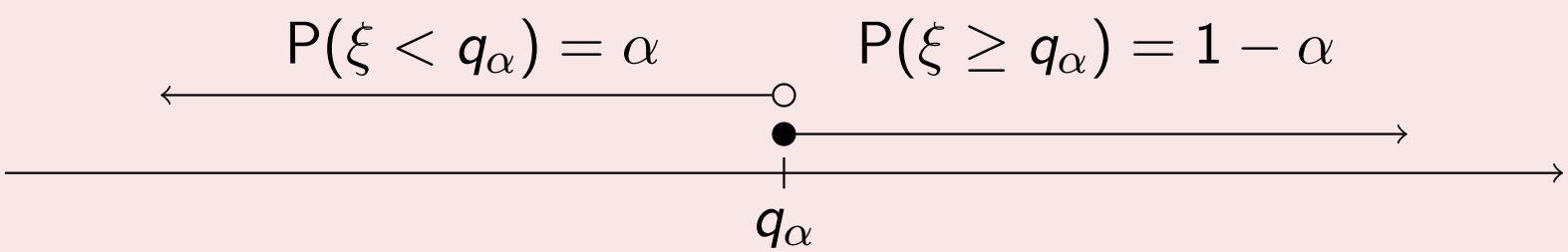
Feladat: Mennyi azon egyedek aránya, melyek elérik a 3.5 kg tömeget?

$$\mathrm{P}\big(\xi \ge 3.5\big) = \mathrm{P}\big(3.5 \le \xi \le 4\big) = F_\xi(4) - F_\xi(3.5) = \left(2 - \frac{4}{4}\right) - \left(2 - \frac{4}{3.5}\right) = 0.14$$

Legyen ξ valószínűségi változó, és legyen $0 < \alpha < 1$. Akkor mondjuk azt, hogy egy q_α szám a változó **α-kvantilise**, ha $\mathrm{P}(\xi < q_\alpha) = \alpha$.

A α-kvantilis jelentése: a ξ mennyiség a teljes sokaságon belül

- az egyedek α hányadánál kisebb, mint q_α,
- az egyedek $1 - \alpha$ hányadánál nagyobb vagy egyenlő, mint q_α.



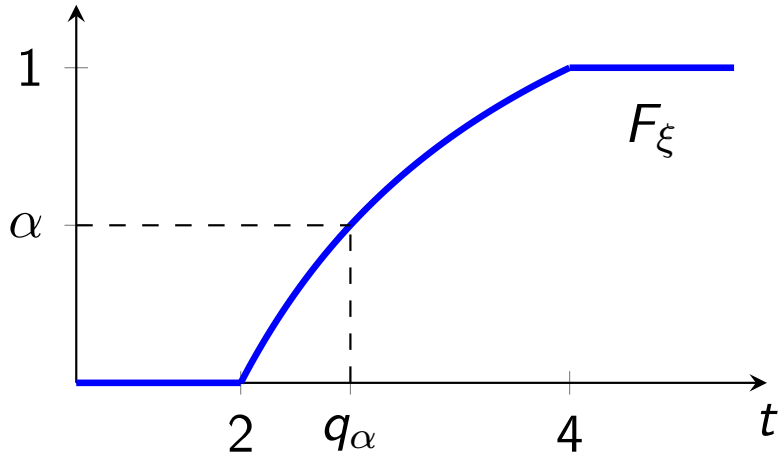
Megjegyzés: Az α-kvantilis nem mindig létezik, és ha létezik, akkor nem feltétlenül egyértelmű.

Nevezetes kvantilisek:

- **Medián:** $q_{50\%}$
- **Alsó és felső kvartilis:** $q_{25\%}$ és $q_{75\%}$
- **Decilisek:** $q_{10\%}, q_{20\%}, \ldots, q_{90\%}$

Feladat: Adjuk meg és értelmezzük a mediánt valamint az alsó és a felső kvartilist a jelen feladatban!

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

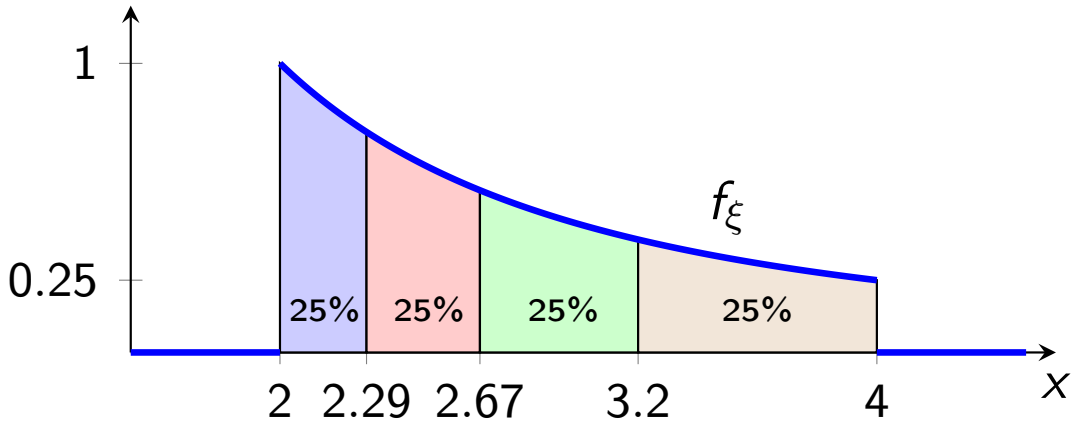


Tetszőleges $0 < \alpha < 1$ szám esetén:

$$\alpha = \mathrm{P}(\xi < q_\alpha) = F_\xi(q_\alpha) = 2 - 4/q_\alpha \implies$$

$$4/q_\alpha = 2 - \alpha \implies q_\alpha/4 = 1/(2-\alpha) \implies q_\alpha = 4/(2-\alpha)$$

α	25%	50%	75%
q_α	2.29	2.67	3.2

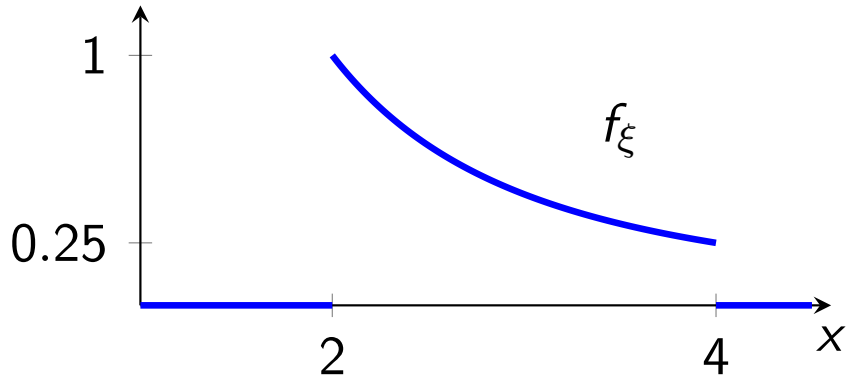


Legyen ξ folytonos valószínűségi változó!

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

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

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



$$\mathrm{E}(\xi) = \int_{-\infty}^{+\infty} x f_\xi(x) dx = \int_{-\infty}^{2} x \cdot 0 dx + \int_{2}^{4} x \cdot \frac{4}{x^2} dx + \int_{4}^{+\infty} x \cdot 0 dx$$

$$= 0 + 4 \int_{2}^{4} x^{-1} dx + 0 = 4 \Big[\ln |x| \Big]_2^4 = 4 \big(\ln 4 - \ln 2 \big) = 2.77$$

A varianciára adható egy könnyebben számolható formula is:

$$\mathrm{Var}(\xi) = \int_{-\infty}^{+\infty} \left[x - \mathrm{E}(\xi)\right]^2 f_\xi(x)dx$$

$$= \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - \int_{-\infty}^{+\infty} 2\mathrm{E}(\xi) x f_\xi(x)dx + \int_{-\infty}^{+\infty} \left(\mathrm{E}(\xi)\right)^2 f_\xi(x)dx$$

$$= \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - 2\mathrm{E}(\xi) \int_{-\infty}^{+\infty} x f_\xi(x)dx + \left(\mathrm{E}(\xi)\right)^2 \int_{-\infty}^{+\infty} f_\xi(x)dx$$

$$= \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - 2\mathrm{E}(\xi)\mathrm{E}(\xi) + \left(\mathrm{E}(\xi)\right)^2 \cdot 1 = \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - \left(\mathrm{E}(\xi)\right)^2$$

Feladat: Mennyi a ξ változó szórása a pingvines feladatban?

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

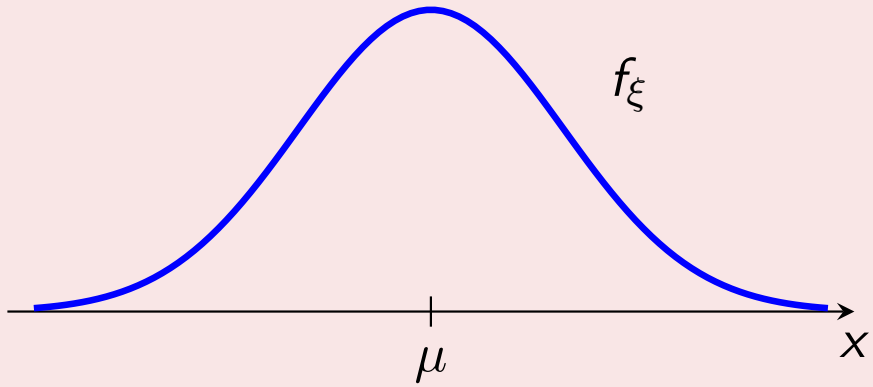
$$\mathrm{Var}(\xi) = \int_{-\infty}^{+\infty} x^2 f_\xi(x)dx - \left(\mathrm{E}(\xi)\right)^2 = 8 - (2.77)^2 \approx 0.33$$

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

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:

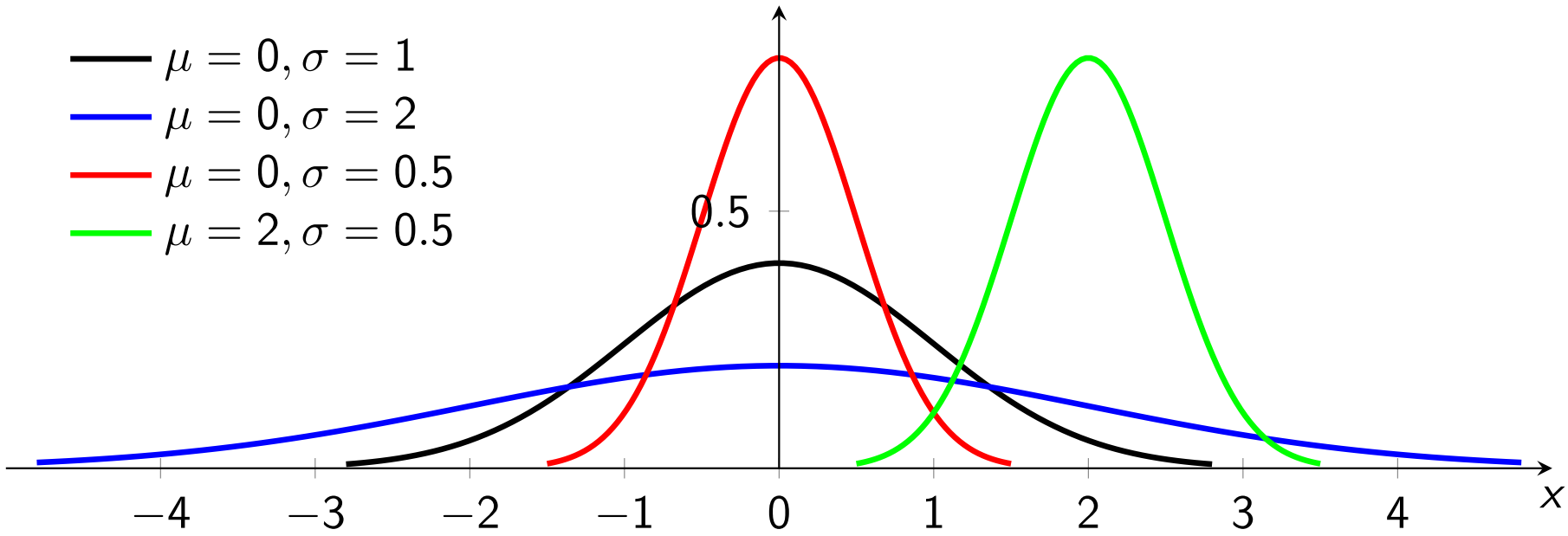
- Mérési hibák modellezése: mért érték = igazi érték + mérési hiba, ahol a mérési hiba normális eloszlást követ.
- Élettudományok: számos mennyiség (testmagasság, vérnyomás, IQ) normális vagy a normálisból származtatott eloszlást követ.

A normális eloszlás fontosabb tulajdonságai:

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

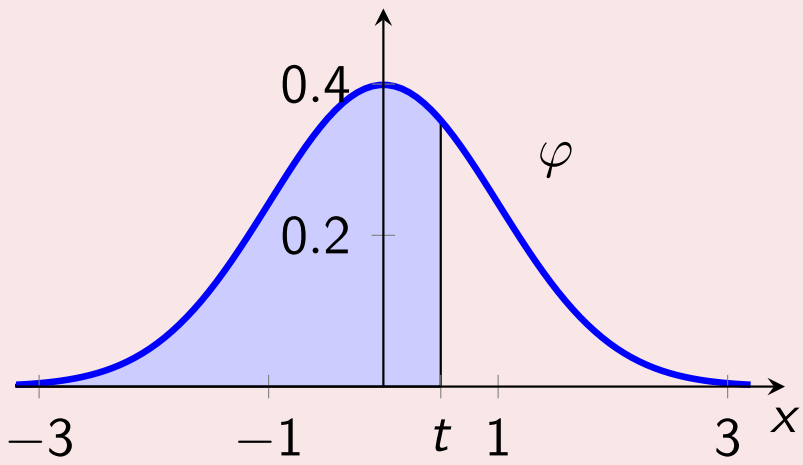
Hogyan hat a két paraméter a sűrűségfüggvényre?

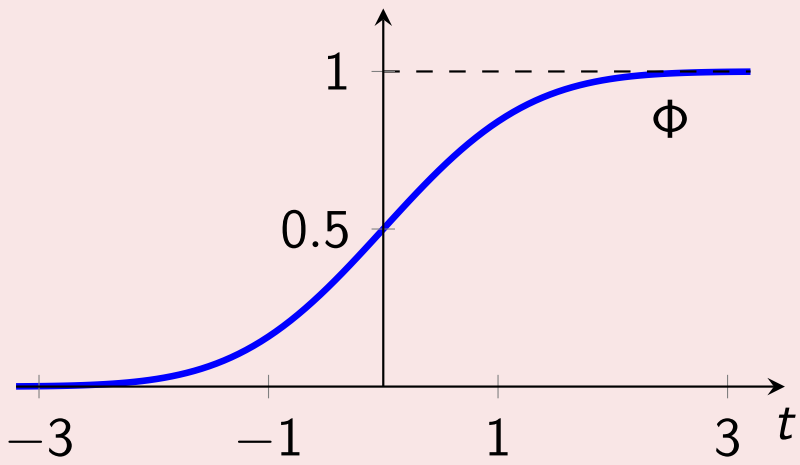
- σ: a sűrűségfüggvény alakját határozza meg,
- μ: eltolás, a sűrűségfüggvény szimmetriatengelye.



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

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





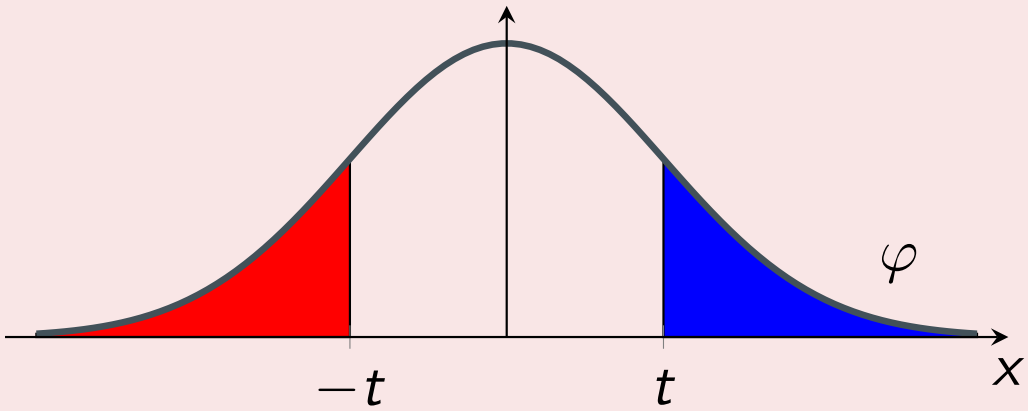
Hogyan tudjuk meghatározni a Φ függvény értékét? A formula nagyon bonyolult, táblázatot használunk.

A Φ eloszlásfüggvény rendelkezik az alábbi tulajdonságokkal.

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

Bizonyítás:

$$\Phi(-t) = \mathrm{P}(\xi < -t) = \int_{-\infty}^{-t} \varphi(x)\,dx = \int_{t}^{+\infty} \varphi(x)\,dx$$
$$= \mathrm{P}(\xi \geq t) = 1 - \mathrm{P}(\xi < t) = 1 - \Phi(t)$$



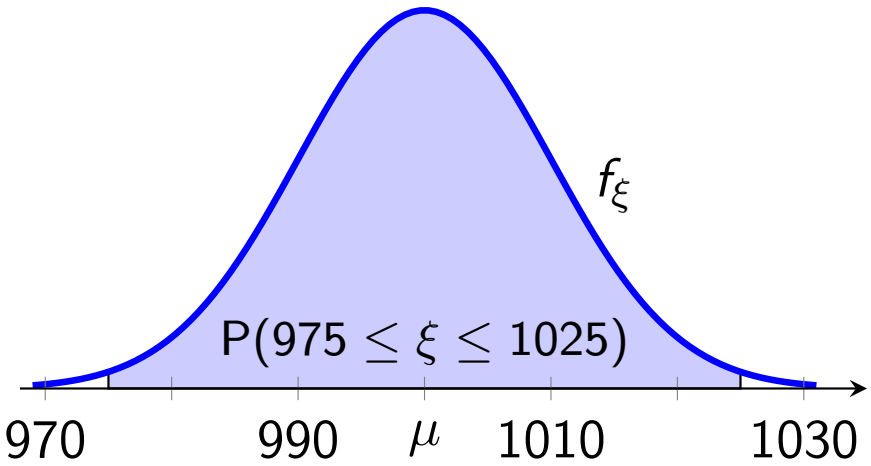
Feladat: Egy tejgyárban az 1 literes dobozos tej csomagolását automata töltőberendezés végzi. A dobozokba töltött mennyiség normális eloszlást követ, a várható érték a névleges tartalom, a szórás $\sigma = 10$ ml.

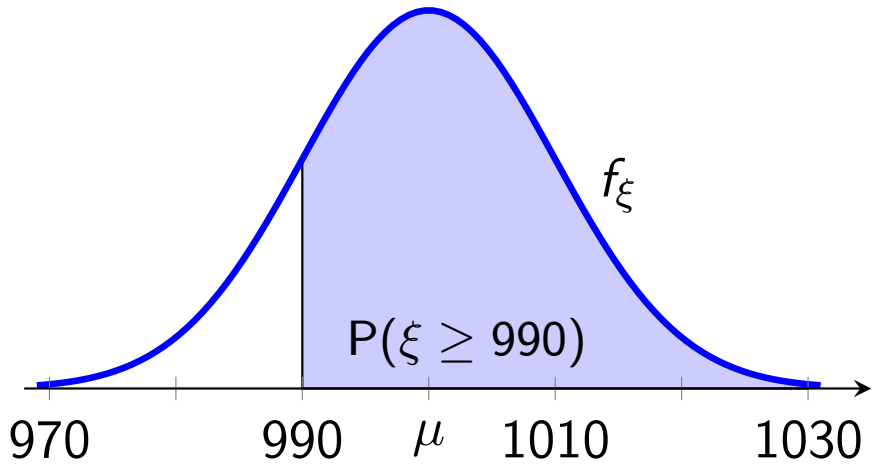
- A dobozok mekkora hányada tér el legfeljebb 2.5%-kal a névleges tartalomtól?
- A dobozok mekkora hányada tartalmaz legalább 990 ml tejet?

Véletlenszerűen kiválasztunk egy dobozt.

$\xi =$ a kiválasztott dobozban található mennyiség

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

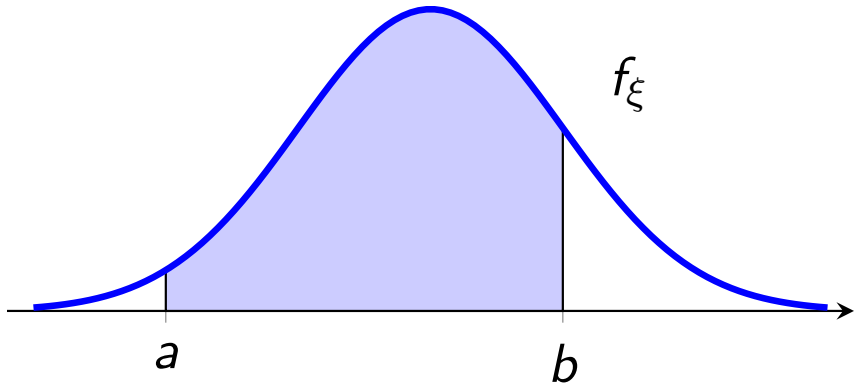




Ha ξ normális eloszlású, akkor tetszőleges $a \leq b$ valós számokra:

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

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

Bizonyítás:

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

Helyettesítéses integrálás, új változót vezetünk be:

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

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

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

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?

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

Itt felhasználtuk azt, hogy

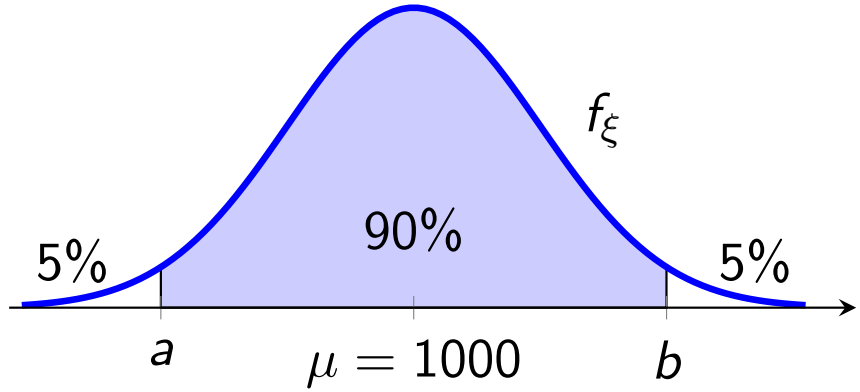
$$\Phi(-1) = 1 - \Phi(1) = 1 - 0.84 = 0.16$$

Feladat. Adjunk meg egy olyan $[a, b]$ intervallumot, amire teljesül, hogy a tejesdobozok 90%-a ebbe az intervallumba esik!

Cél: $\mathrm{P}(a \le \xi \le b) = 0.9$

Kérdés: $a, b = ?$

Ötlet: $a = q_{5\%}, \quad b = q_{95\%}$



$$0.95 = \mathrm{P}(\xi < b) = F_\xi(b) = \Phi\left(\frac{b - 1000}{10}\right)$$

$$0.95 = \Phi(1.65) \quad \Longrightarrow \quad \frac{b - 1000}{10} = 1.65 \quad \Longrightarrow \quad b = 1016.5$$

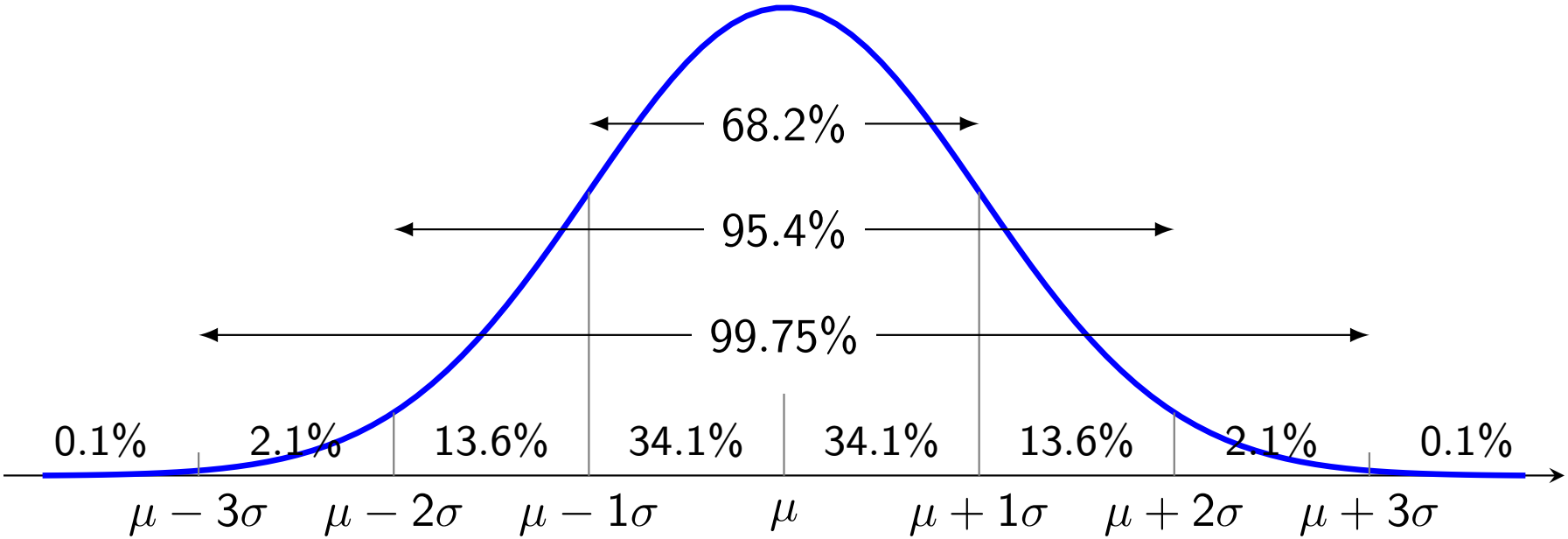
$$0.05 = \mathrm{P}(\xi < a) = F_\xi(a) = \Phi\left(\frac{a - 1000}{10}\right)$$

$$0.05 = 1 - 0.95 = 1 - \Phi(1.65) = \Phi(-1.65)$$

$$\Longrightarrow \quad \frac{a - 1000}{10} = -1.65 \quad \Longrightarrow \quad a = 983.5$$

Megoldás: $[983.5, 1016.5]$

Az alábbi ábra azt mutatja meg, hogy egy ξ normális eloszlású változó mekkora eséllyel esik a várható érték két oldalára felmért intervallumokba:

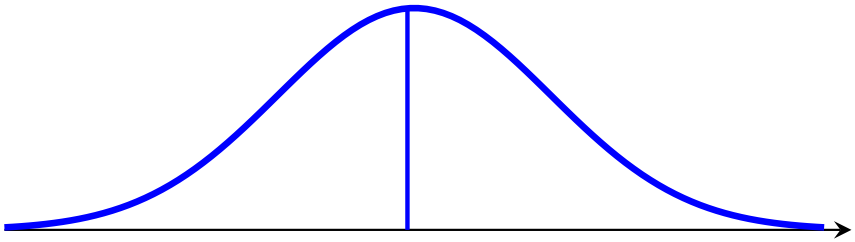


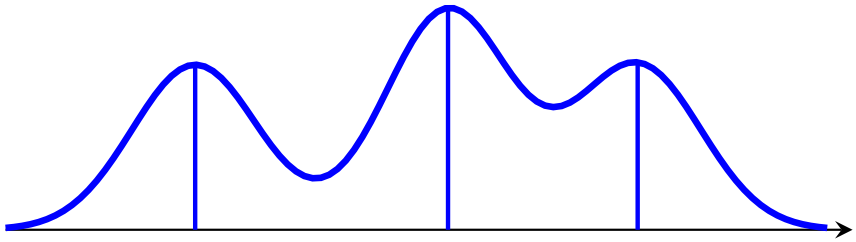
Legyen ξ normális eloszlású változó! Ekkor:

- 1σ-szabály: $\mathrm{P}\big(\mu - \sigma \leq \xi \leq \mu + \sigma\big) \approx 68\%$,
- 2σ-szabály: $\mathrm{P}\big(\mu - 2\sigma \leq \xi \leq \mu + 2\sigma\big) \approx 95\%$,
- 3σ-szabály: $\mathrm{P}\big(\mu - 3\sigma \leq \xi \leq \mu + 3\sigma\big) \approx 99.75\%$.

Sűrűségfüggvények módusza és ferdesége

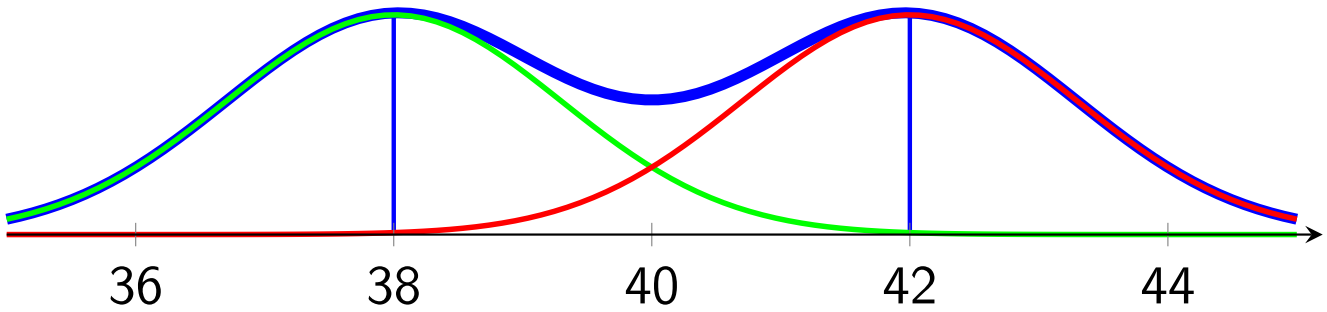
Folytonos változó **móduszai**: a sűrűségfüggvény lokális maximumhelyei.
A sűrűségfüggvény lehet **egymóduszú** vagy **többmóduszú.**





A több módusz gyakran arra utal, hogy a sokaságot több részre lehet felbontani, melyeken belül a vizsgált mennyiség már egymóduszú. Példa:

- kék görbe: a lábméret sűrűségfüggvénye a felnőtt népességen belül,
- zöld görbe: a lábméret sűrűségfüggvénye a nők körében,
- piros görbe: a lábméret sűrűségfüggvénye a férfiak körében.



Tegyük fel, hogy a sűrűségfüggvénynek csak egyetlen módusza van!

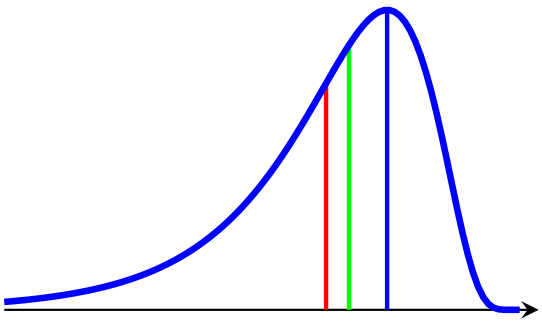
- Módusz: A változó ezen érték közelébe esik a legnagyobb eséllyel.
- Medián: A változó „középső” értéke.
- Várható érték: A változó átlagos értéke.

Ha a sűrűségfüggvény szimmetrikus (például normális eloszlás), akkor:

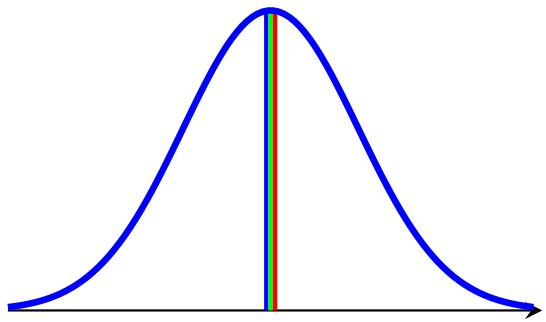
$$\text{várható érték} = \text{medián} = \text{módusz}$$

Ha a sűrűségfüggvény nem szimmetrikus, akkor jellemzően(!):

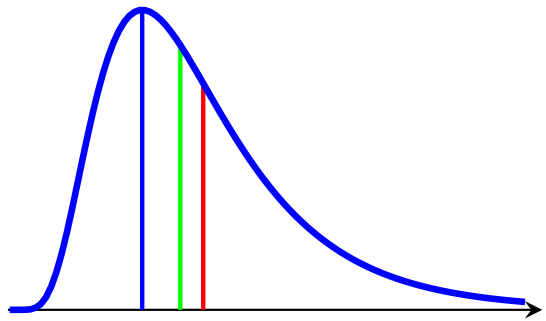
- Balra ferde sűrűségfüggvény esetén: $\text{várható érték} < \text{medián} < \text{módusz}$
- Jobbra ferde sűrűségfüggvény esetén: $\text{módusz} < \text{medián} < \text{várható érték}$



balra ferde függvény



szimmetrikus függvény



jobbra ferde függvény

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:

- $\mathrm{E}(\xi) =$ a vizsgált mennyiség átlagos értéke a sokaságon belül;
- $\mathrm{D}(\xi) =$ a vizsgált mennyiség szórása a sokaságon belül;
- $\mathrm{P}(a \leq \xi \leq b) =$ 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, $\xi_1, \ldots, \xi_n$ valószínűségi változók.
- **Mintarealizáció:** A $\xi_1, \ldots, \xi_n$ 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: várható érték, szórás, medián, stb.

Ötlet: tekintsünk úgy a megfigyelt elemekre, mint egy részsokaságra. Számoljuk ki a kérdéses mutatószámokat ebben a részsokaságban!

Empirikus várható érték, mintaátlag:

$$\overline{\xi} = \mathsf{E}_n(\xi) = \frac{\xi_1 + \ldots + \xi_n}{n}$$

Empirikus variancia, statisztikai variancia:

$$\mathrm{Var}_n(\xi) = \frac{\left(\xi_1 - \overline{\xi}\right)^2 + \ldots + \left(\xi_n - \overline{\xi}\right)^2}{n}$$

Empirikus szórás, statisztikai szórás: $S_\xi = \mathsf{D}_n(\xi) = \sqrt{\mathrm{Var}_n(\xi)}$

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

$$\mathrm{E}_n(\xi) \to \mathrm{E}(\xi), \quad \mathrm{Var}_n(\xi) \to \mathrm{Var}(\xi), \quad \mathrm{D}_n(\xi) \to \mathrm{D}(\xi), \qquad \text{amint} \quad n \to \infty.$$

Tehát ezek a becslések nagy n esetén pontosak lesznek.

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

Korrigált empirikus variancia és **korrigált empirikus szórás**:

$$\mathrm{Var}_n^*(\xi) = \frac{n}{n-1}\,\mathrm{Var}_n(\xi), \qquad S_\xi^* = \mathrm{D}_n^*(\xi) = \sqrt{\mathrm{Var}_n^*(\xi)}$$

Tulajdonságok:

- A korrigált becslések szintén erősen konzisztensek:

$$\mathrm{Var}_n^*(\xi) \to \mathrm{Var}(\xi), \qquad \mathrm{D}_n^*(\xi) \to \mathrm{D}(\xi), \qquad \text{amint} \quad n \to \infty.$$

- A korrigált becslések kis n esetén pontosabbak.

Feladat: A kar férfi hallgatóinak testmagasságát vizsgáljuk.
Háttérváltozó: $\xi =$ egy véletlenszerűen kiválasztott hallgató magassága.
Mintarealizáció: 180, 175, 188, 168, 173, 183.

$$\mathrm{E}(\xi) \approx \overline{\xi} = \mathrm{E}_6(\xi) = \frac{180+175+188+168+173+183}{6} = 177.8$$

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

$$\mathrm{D}(\xi) \approx \mathrm{D}_6(\xi) = \sqrt{43.81} = 6.62$$

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

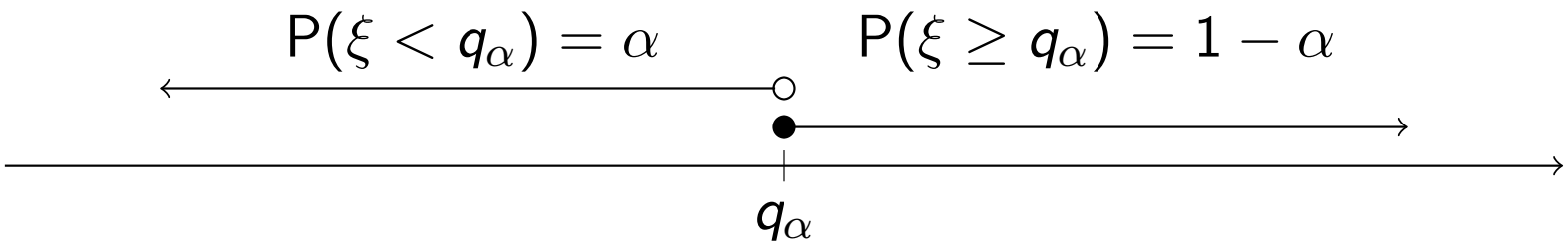
$$\mathrm{Var}(\xi) \approx \mathrm{Var}_6^*(\xi) = \frac{6}{5}\,43.81 = 52.57, \quad \mathrm{D}(\xi) \approx \mathrm{D}_6^*(\xi) = \sqrt{52.57} = 7.25.$$

Foglaljuk össze, hogy mit kaptunk:

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

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

A ξ valószínűségi változó α-kvantilise: $\mathrm{P}(\xi < q_\alpha) = \alpha$.
Jelentése: a változó az egyedek α hányadánál kisebb, mint q_α.



Hogyan becsülhető a sokaság kvantilise a minta alapján?

Empirikus kvantilis, statisztikai kvantilis, percentilis: Az a $\hat{q}_\alpha$ szám, melyre teljesül, hogy a $\xi_1, \ldots, \xi_n$ értékek α hányada kisebb, mint $\hat{q}_\alpha$.

Példa: **empirikus medián**

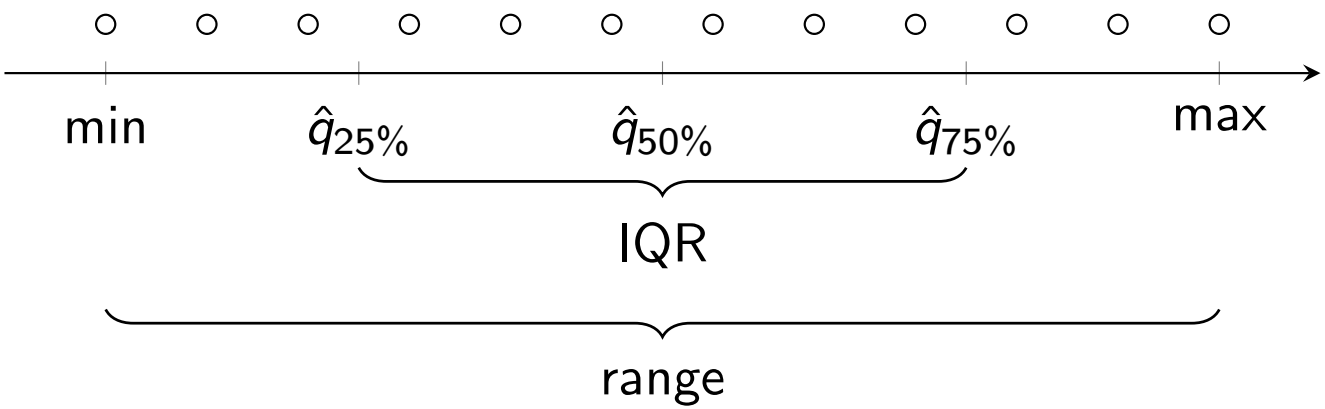
$$\hat{q}_{50\%} = \begin{cases} \text{a középső mintaelem,} & \text{ha } n \text{ páratlan,} \\ \text{a két középső átlaga,} & \text{ha } n \text{ páros.} \end{cases}$$

Ha ξ folytonos változó, továbbá ha q_α létezik és egyértelmű, akkor az empirikus kvantilis erősen konzisztens becslés: $\hat{q}_\alpha \to q_\alpha$ amint $n \to \infty$.

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 távolság:** $\mathrm{IQR} = \hat{q}_{75\%} - \hat{q}_{25\%}$
  Ilyen hosszúságú intervallumon helyezkedik el a minta középső 50%-a.

Példa: a mintaméret $n = 12$

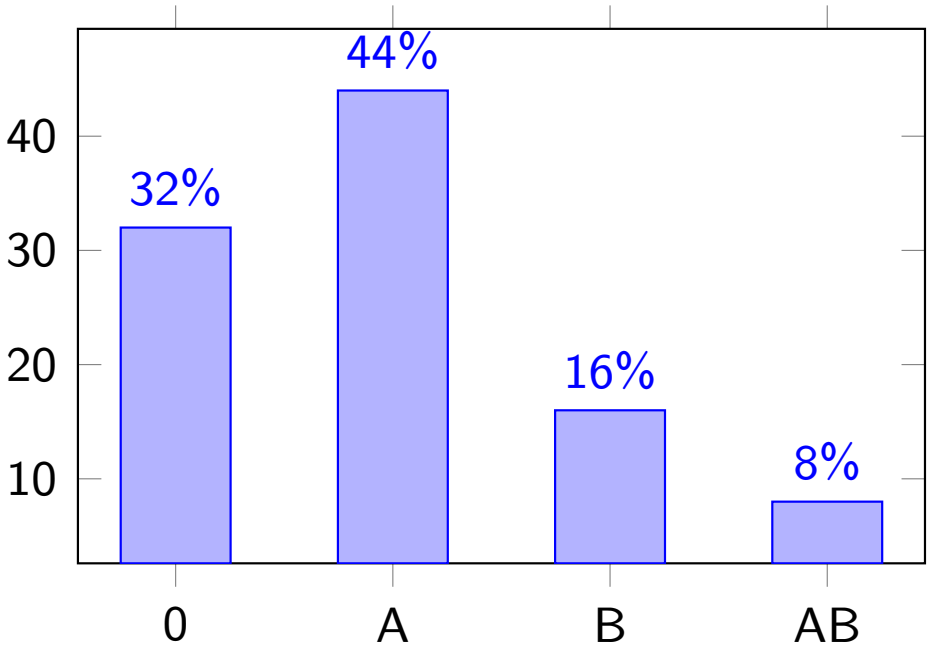


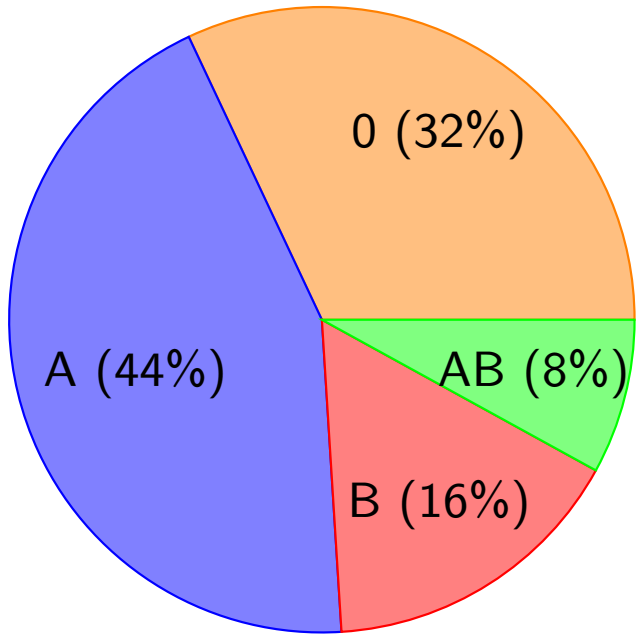
Statisztikai grafikonok

Legyen a ξ háttérváltozó diszkrét!

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

Példa: a vércsoportok aránya a magyar népességen belül.

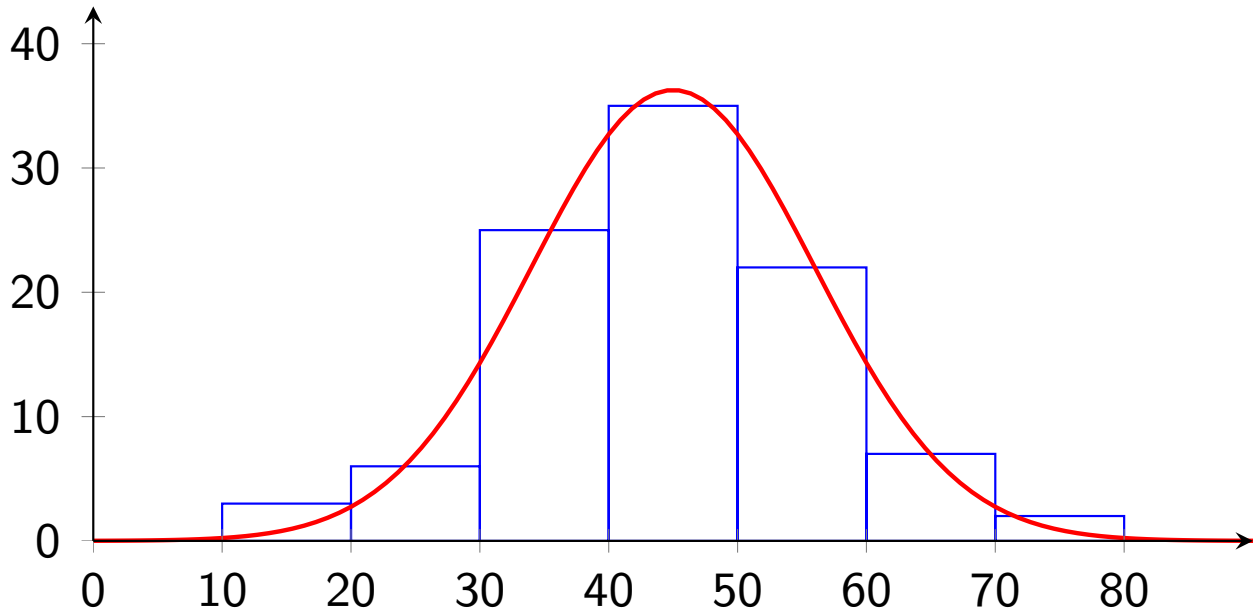




Legyen a ξ háttérváltozó folytonos!

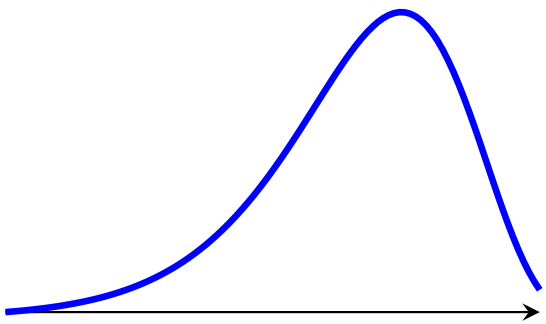
Hisztogram: Felbontjuk a számegyenest azonos hosszúságú intervallumokra. Minden intervallumra olyan magas oszlopot állítunk, ahány mintaelem esik az adott intervallumba.

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

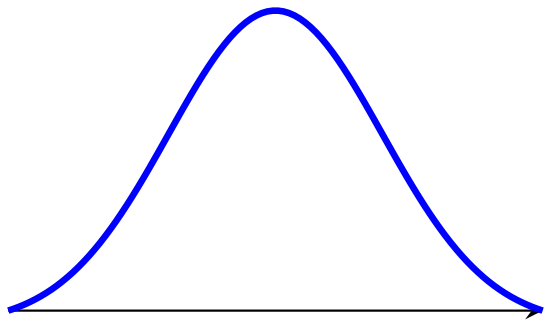


Ferdeség (skewness): egy olyan statisztikai mutatószám, ami a hisztogram és az illesztett sűrűségfüggvény szimmetriáját jellemzi.

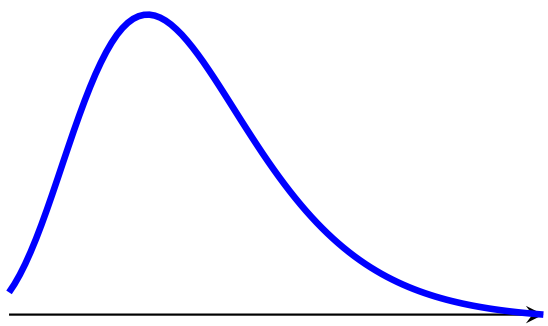
- skewness $= 0 \implies$ a hisztogram szimmetrikus
- skewness $> 0 \implies$ a hisztogram jobbra ferde
- skewness $< 0 \implies$ a hisztogram balra ferde



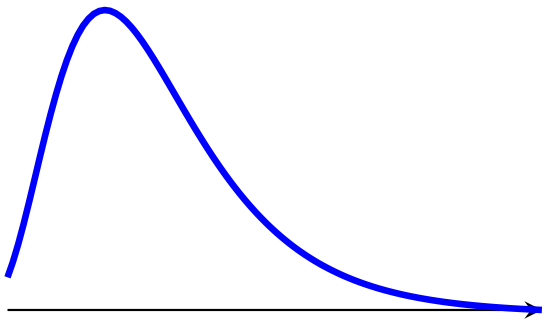
skewness $= -1$



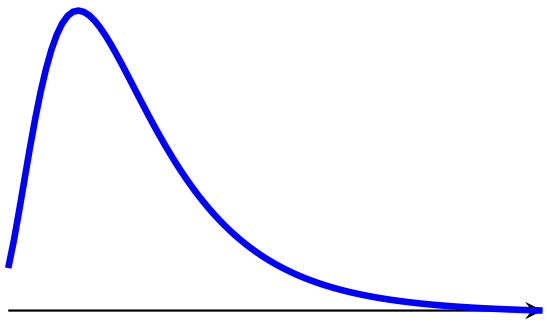
skewness $= 0$



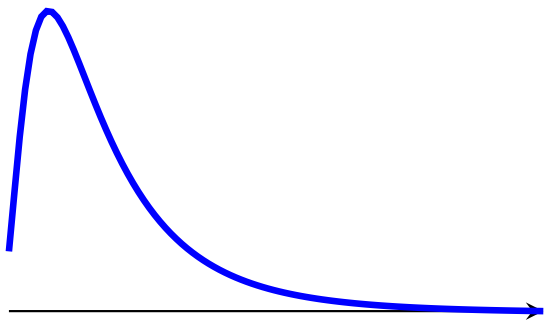
skewness $= 1$



skewness $= 1.5$

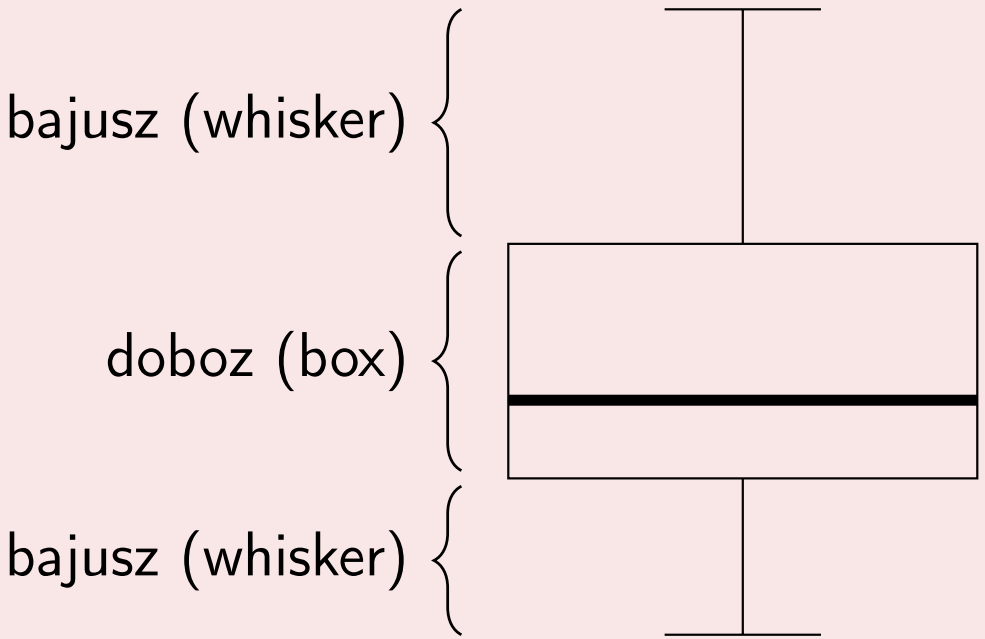


skewness $= 2$



skewness $= 3$

Boxplot outlier elemek nélkül:



max = legnagyobb mintaelem

$\hat{q}_{75\%}$ = felső kvartilis

$\hat{q}_{50\%}$ = medián

$\hat{q}_{25\%}$ = alsó kvartilis

min = legkisebb mintaelem

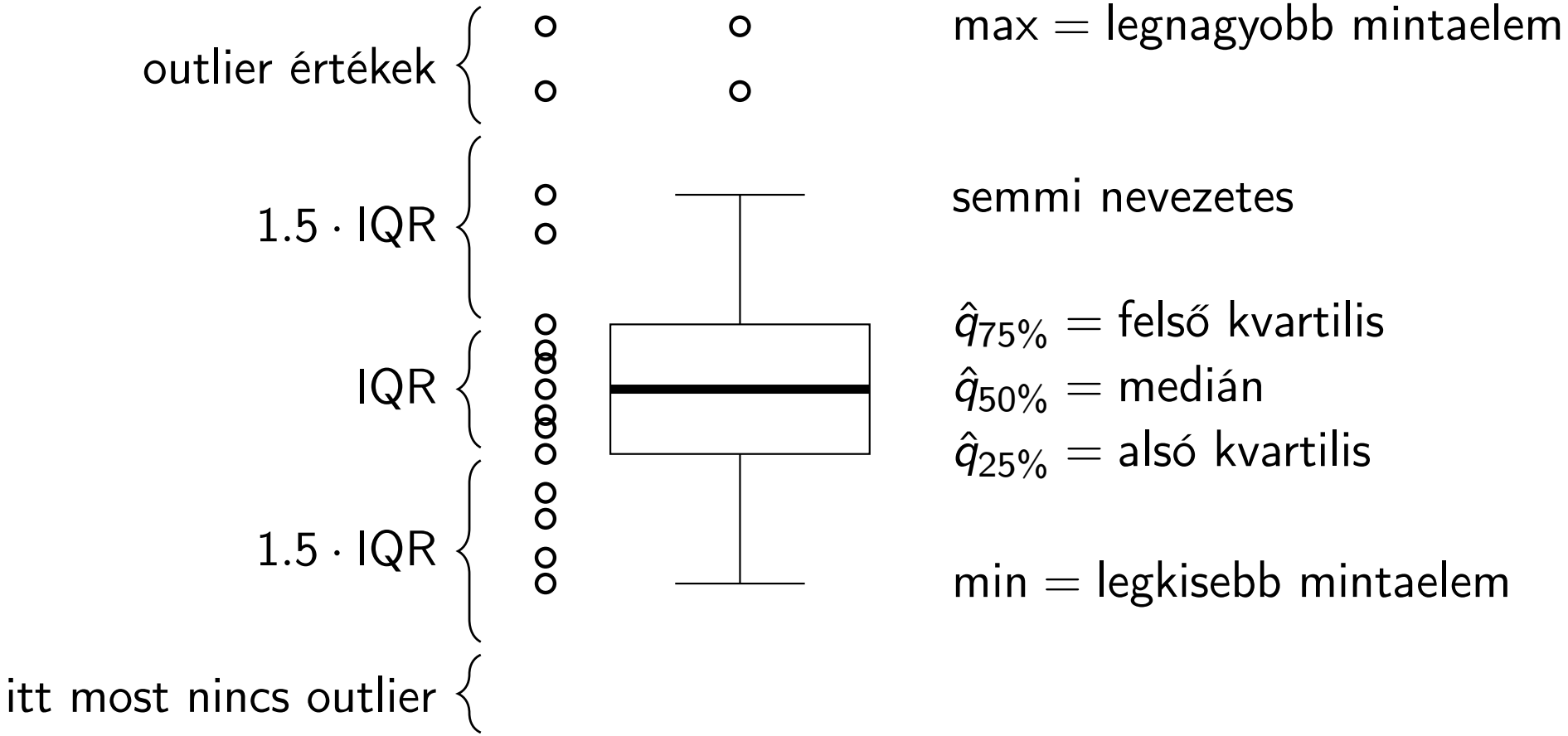
További mutatószámok az ábrán:

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

Outlier értékek: Azok a mintaelemek, melyek

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

Általában a boxploton külön ábrázoljuk az outlier értékeket:



Grafikus illeszkedésvizsgálat:

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

Normalitásvizsgálat grafikonok és a skewness segítségével:

- Hisztogram:
  - Nagy elemszám esetén a sűrűségfüggvény illeszkedik a hisztogramhoz.
  - Ha a hisztogram nem követi a haranggörbe alakját, akkor a minta nem normális eloszlásból származik.
  - Főleg a hisztogram ferdeségét érdemes figyelni.
- Skewness:
  - Normális eloszlás esetén a minta közel szimmetrikus.
  - Ha a minta nagyon ferde, akkor nem normális eloszlásból származik.
- Boxplot:
  - Normális eloszlás esetén az outlier értékek aránya körülbelül 1%.
  - Ha a mintában magas (legalább 5%) az outlier értékek aránya, akkor a minta nem normális eloszlásból származik.

Standard hiba, konfidencia intervallumok

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

- 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: $\mathsf{E}\big(|\overline{\xi} - \mathsf{E}(\xi)|\big) = ?$

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

Standard hiba (standard error of the mean, s.e.m.): $\mathsf{SE} = \mathsf{D}^*_n(\xi)/\sqrt{n}$

Szemléletes jelentése: az $\mathsf{E}(\xi) \approx \overline{\xi}$ 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: $n = 6$, $\overline{\xi} = \mathsf{E}_6(\xi) = 177.8$, $\mathsf{D}^*_6(\xi) = 7.25$.

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

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.

A statisztikában egy minta alapján kétféle formában becsülhetjük meg az ismeretlen mutatószámot (várható érték, szórást, stb.):

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

Legyen $\xi_1, \ldots, \xi_n$ 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

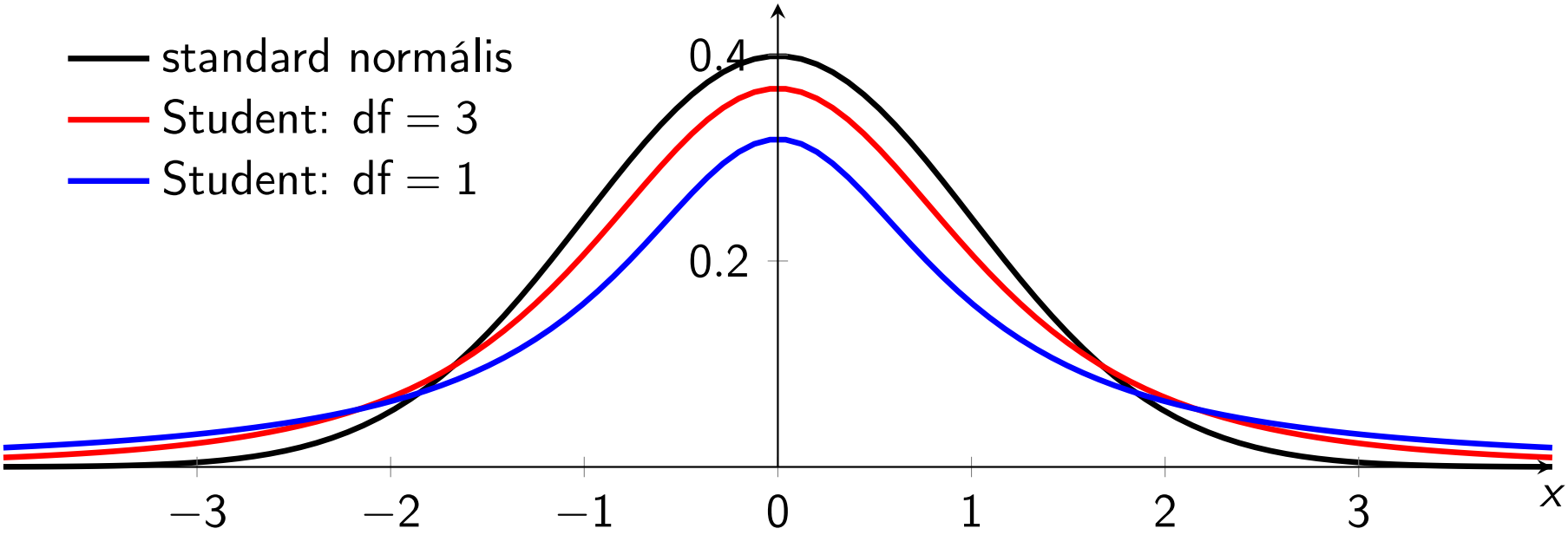
$$\mathrm{P}\Big(\mathrm{E}(\xi) \in [a, b]\Big) = 1 - \alpha.$$

Megjegyzések:

- 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.)

Student-eloszlás:

- Folytonos eloszlás, a sűrűségfüggvénye hasonlít a haranggörbéhez:



- **Szabadsági fok (degrees of freedom, df):** az eloszlás paraméterre. Ettől függ a sűrűségfüggvény alakja.
- A Student-eloszlás eloszlásfüggvénye: $\Phi_{\mathrm{df}}(t) = \mathrm{P}(\xi < t)$.
- A Student-eloszlás kvantilise:

$$\alpha = \mathrm{P}(\xi < q_\alpha) = \Phi_{\mathrm{df}}(q_\alpha), \qquad \text{tehát} \qquad q_\alpha = \Phi_{\mathrm{df}}^{-1}(\alpha).$$

Legyen a ξ háttérváltozó normális eloszlású ismeretlen szórással!
$1-\alpha$ megbízhatóságú konfidencia intervallum a változó várható értékére:

$$\left[\overline{\xi} - c \cdot \mathrm{SE}\,,\, \overline{\xi} + c \cdot \mathrm{SE}\right], \qquad c = \Phi_{n-1}^{-1}\left(1-\alpha/2\right).$$

Itt Φ_{n-1} az $n-1$ szabadsági fokú Student-eloszlás eloszlásfüggvénye.

Feladat: Tegyük fel, hogy a kar férfi hallgatóinak a testmagassága normális eloszlású! Adjunk 95% megbízhatóságú konfidencia intervallumot az átlagos testmagasságra!

Korábban láttuk: $n = 6,\ \overline{\xi} = 177.8,\ \mathrm{SE} = 2.96$

A szabadsági fok: $n - 1 = 5$

Most $\alpha = 0.05$, tehát $1 - \alpha/2 = 0.975$

A Student-eloszlás táblázatából: $c = \Phi_5^{-1}(0.975) = 2.571$

A minta alapján felírt konfidencia intervallum:

$$\left[177.8 - 2.571 \cdot 2.96\,,\, 177.8 + 2.571 \cdot 2.96\right] = \left[170.2\,,\, 185.4\right]$$

Kérdés: Hogyan értelmezhető a kapott eredmény?

A mintavételezés során számos mintarealizációt kaphatunk:

- „Jó" mintarealizációk: az ezekből számolt konfidencia intervallum tartalmazza az ismeretlen várható értéket. Ezek teszik ki az összes lehetséges mintarealizáció $1 - \alpha = 0.95$ hányadát.
- „Rossz" mintarealizációk: ezek félrevezetőek, ugyanis a belőlük számolt konfidencia intervallum nem tartalmazza a várható értéket. Ezek alkotják az összes realizáció $\alpha = 0.05$ hányadát.

Kérdés: Ebben a feladatban jó vagy rossz mintarealizációt kaptunk?

Ezt nem tudjuk eldönteni. Csak reménykedhetünk benne, hogy a jók közül kaptunk egyet, ugyanis ezek vannak többségben.

Kérdés: Miért nem számolunk 99.99%-os megbízhatósággal?

A magasabb megbízhatóság szélesebb intervallumot jelent. A túl széles intervallum viszont nehezíti az eredmény alkalmazhatóságát.

Kérdés: Mi a helyzet akkor, ha a ξ nem normális eloszlású?

Ha a minta nem normális eloszlásból jön, akkor a felírt intervallum nem lesz pontosan $1-\alpha$ megbízhatóságú. Helyette:

$$\mathsf{P}\Big(\mathsf{E}(\xi) \in \big[\overline{\xi} - c \cdot \mathsf{SE}\,,\, \overline{\xi} + c \cdot \mathsf{SE}\big]\Big) \to 1-\alpha, \quad \text{amint} \quad n \to \infty.$$

Tehát ha a mintaméret nagy, akkor az intervallum jó közelítéssel $1-\alpha$ megbízhatóságú.

Kérdés: Mit jelent ebben az esetben a „nagy” mintaméret?

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

- (közel) szimmetrikus eloszlás esetén 20–30 mintaelem tipikusan elég szokott lenni a pontos közelítéshez,
- ferde eloszlás esetén jellemzően kell legalább 50, vagy akár még annál is több mintaelem.

Hipotézisvizsgálat

A **hipotézisvizsgálat** (hypothesis testing) alapfogalmai:

- Adott egy ξ háttérváltozó és egy $\xi_1, \ldots, \xi_n$ 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:

- $H_0 : P(\xi = 5) = 1/2$, $H_A : P(\xi = 5) \neq 1/2$.
- $H_0 : \xi$ normális eloszlású, $H_A : \xi$ nem normális eloszlású.

A hipotézisvizsgálat menete:

- Eldöntjük, hogy milyen módszerrel tesztelünk.
- A minta alapján kiszámoljuk a **próbastatisztika** értékét: t_n.
- Meghatározzuk a **kritikus értéket**: c.
- Ha $|t_n| \leq c$, akkor **elfogadjuk** a nullhipotézist.
  Ha $|t_n| > c$, 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 próbastatisztika (t_n) azt fejezi ki, hogy a vádlott szava mennyire van ellentmondásban a bizonyítékokkal.
- A c kritikus érték egy küszöbérték:
  - Ha $|t_n| \leq c$, akkor a bíró hisz a vádlottnak, és felmenti.
  - Ha $|t_n| > c$, akkor nem hisz neki, és elítéli.

Feladat: Teszteljük le azt a nullhipotézist, hogy a kar férfi hallgatóinak az átlagos testmagassága 175 cm! Feltehető, hogy a testmagasság normális eloszlású a teljes sokaságban.

Amit tudunk:

- $\xi =$ egy véletlenszerűen kiválasztott hallgató testmagassága.
- Megfigyelt értékek: 180, 175, 188, 168, 173, 183.
- Nullhipotézis: $H_0 : E(\xi) = 175$.

A várható értéket az **egymintás t-próba** segítségével tesztelhetjük:

- Hipotetikus várható érték: $\mu_0 = 175$.
- Próbastatisztika:

$$t_n = \frac{\overline{\xi} - \mu_0}{SE} = \frac{177.8 - 175}{2.96} = 0.946$$

- A kritikus érték: $c = 2.571$. (Erre majd még visszatérünk.)
- Döntés: $|t_n| \leq c$, tehát a nullhipotézist elfogadjuk.

Milyen hibákat véthetünk a hipotézisvizsgálat során?

- **Elsőfajú hiba (type I error):** Elvetjük az igaz nullhipotézist, tehát börtönbe küldünk egy ártatlant. Valószínűsége:

$$\alpha = \mathrm{P}(\text{elvetjük } \mathrm{H}_0\text{-t, ha igaz}).$$

- **Másodfajú hiba (type II error):** Elfogadjuk a hamis nullhipotézist, tehát felmentünk egy bűnöst. Valószínűsége:

$$\beta = \mathrm{P}(\text{elfogadjuk } \mathrm{H}_0\text{-t, ha hamis}).$$

Még egy fogalom:

$$\textbf{erő (power)} = \mathrm{P}(\text{elvetjük } \mathrm{H}_0\text{-t, ha hamis}) = 1 - \beta.$$

A lehetőségeket az alábbi táblázatban foglalhatjuk össze:

	elfogadjuk	elvetjük
H_0 igaz	helyes döntés	elsőfajú hiba
H_0 hamis	másodfajú hiba	helyes döntés

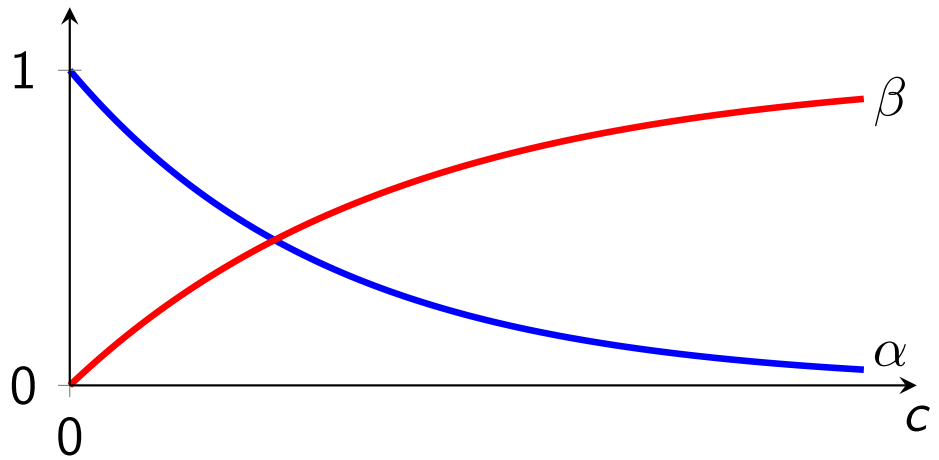
Mire hathatunk és mire nem a hipotézisvizsgálat során?

- Akkor vetjük el a nullhipotézist, ha $|t_n| > c$.
- A t_n próbastatisztika értékét nem tudjuk befolyásolni.
- A c kritikus értéket (=mennyire szigorú a bíró) mi választjuk.

Meg lehet választani úgy a kritikus értéket, hogy mindkét hiba alacsony maradjon? Erre sajnos nincs lehetőség:

magas kritikus érték $\Rightarrow$ alacsony elsőfajú hiba, de magas másodfajú hiba

alacsony kritikus érték $\Rightarrow$ alacsony másodfajú hiba, de magas elsőfajú hiba

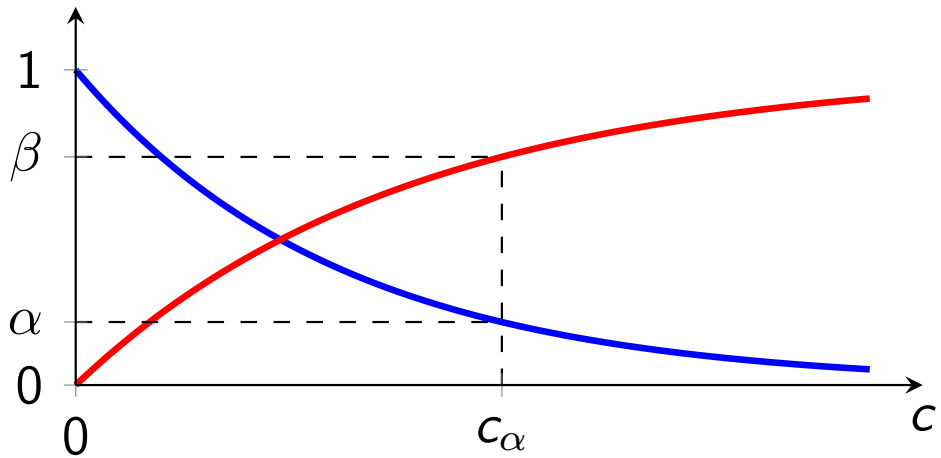


Szignifikancia szint: az α elsőfajú hiba előre megadott értéke.

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

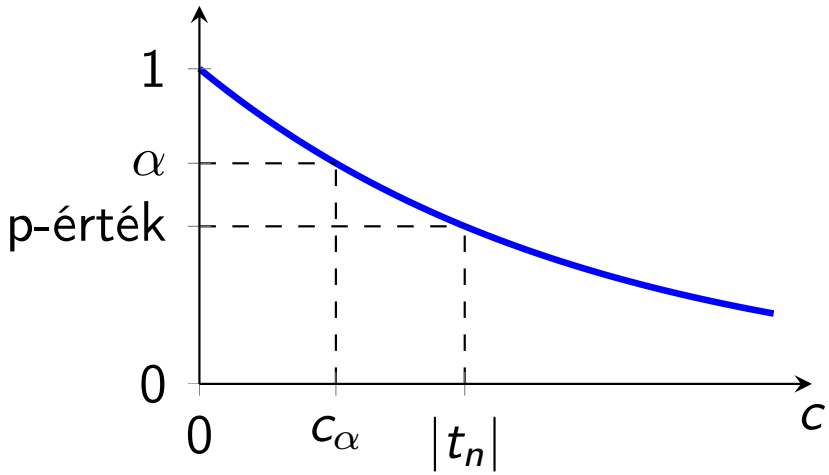
Első- és másodfajú hiba a gyakorlatban:

- A feladat megadja az α szignifikancia szintet (=elsőfajú hiba).
- Meghatározzuk a hozzá tartozó kritikus értéket (c_α) és tesztelünk.
- A β másodfajú hibára nincsen ráhatásunk.

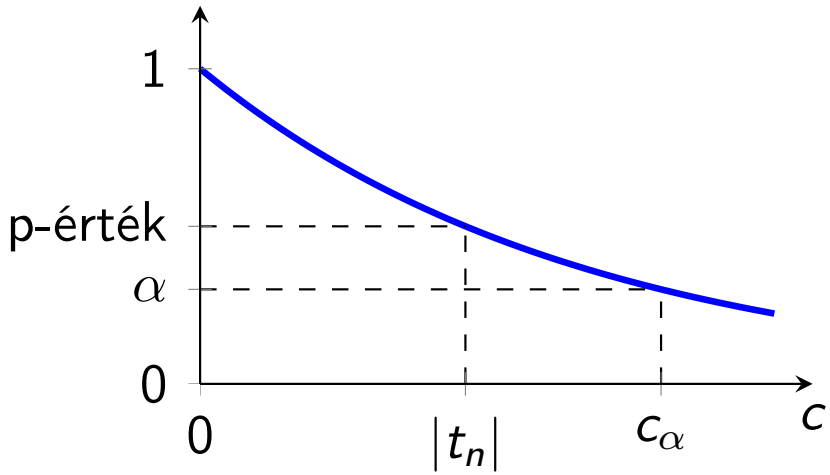


p-érték (p-value): az a szignifikancia szint, melyre $c_{\text{p-érték}} = |t_n|$.
Döntés a p-érték segítségével:

$$\text{elvetjük } H_0\text{-t} \iff |t_n| > c_\alpha \iff \text{p-érték} < \alpha.$$



elvetjük H_0-t



elfogadjuk H_0-t

Megjegyzések:

- A p-érték mindig 0 és 1 közé esik.
- A p-érték szintén értelmezhető olyan módon, hogy mennyire „hihető” a nullhipotézis. A nullhipotézist akkor vetjük el, ha a p-érték alacsony.

Az egymintás t-próba

Egymintás t-próba (One sample t test)

Cél a ξ valószínűségi változó várható értékének a tesztelése.

Feltevések:

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

Nullhipotézis: $\mathrm{H}_0 : \mathrm{E}(\xi) = \mu_0$.

Próbastatisztika:

$$t_n = \frac{\overline{\xi} - \mu_0}{\mathrm{SE}}.$$

Kritikus érték: $c_\alpha = \Phi^{-1}_{n-1}(1 - \alpha/2)$.

Döntés: pontosan akkor fogadjuk el a nullhipotézist, ha $|t_n| \leq c_\alpha$.

Mi a gondolat a t-próba mögött?

- Nullhipotézis: $H_0 : E(\xi) = \mu_0$.
- Ha H_0 igaz, akkor

$$t_n = \frac{\overline{\xi} - \mu_0}{SE} \approx \frac{E(\xi) - \mu_0}{SE} = 0.$$

- Ha H_0 nem igaz, akkor

$$t_n = \frac{\overline{\xi} - \mu_0}{SE} \approx \frac{E(\xi) - \mu_0}{SE} \neq 0.$$

A nullhipotézist akkor fogadjuk el, ha t_n nullához közeli szám. Ez egy logikus ötlet, hiszen:

- ha $t_n \approx 0$, akkor az arra utal, hogy H_0 igaz,
- ha $t_n \not\approx 0$, akkor az arra utal, hogy H_0 nem igaz.

Megjegyzés: Algebrai átalakításokkal bebizonyítható, hogy

$$\text{elfogadjuk } H_0\text{-t} \iff |t_n| \leq c_\alpha \iff \mu_0 \in \left[\overline{\xi} - c_\alpha \,\mathrm{SE}\,,\, \overline{\xi} + c_\alpha \,\mathrm{SE}\right]$$

Ez azt jelenti, hogy:

- 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.

Megjegyzés: Az egymintás t-próba **robusztus** a normalitásfeltételre nézve. Ha a minta nem normális eloszlású, de a mintaméret nagy, akkor a t-próba használható a várható érték tesztelésére. A szükséges mintaméret:

- (közel) szimmetrikus eloszlás esetén 20–30 mintaelem;
- ferde eloszlás esetén legalább 50.

A páros t-próba

Legyenek ξ és η valószínűségi változók. Két statisztikai minta:

- $\xi_1, \ldots, \xi_n$ független megfigyelések ξ-re,
- $\eta_1, \ldots, \eta_m$ 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:
  - $\xi_1, \ldots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben
    $\eta_1, \ldots, \eta_n$: ugyanezen hallgatók testmagassága egy 5 évvel ezelőtt
  - $\xi_1, \ldots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben
    $\eta_1, \ldots, \eta_n$: 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,
- $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_n$ 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 **robusztus** a normalitásfeltételre nézve.
- Konfidencia intervallumot is adhatunk a várhat értékek különbségére:

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

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:

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

A cél a következő nullhipotéziseket tesztelni:

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

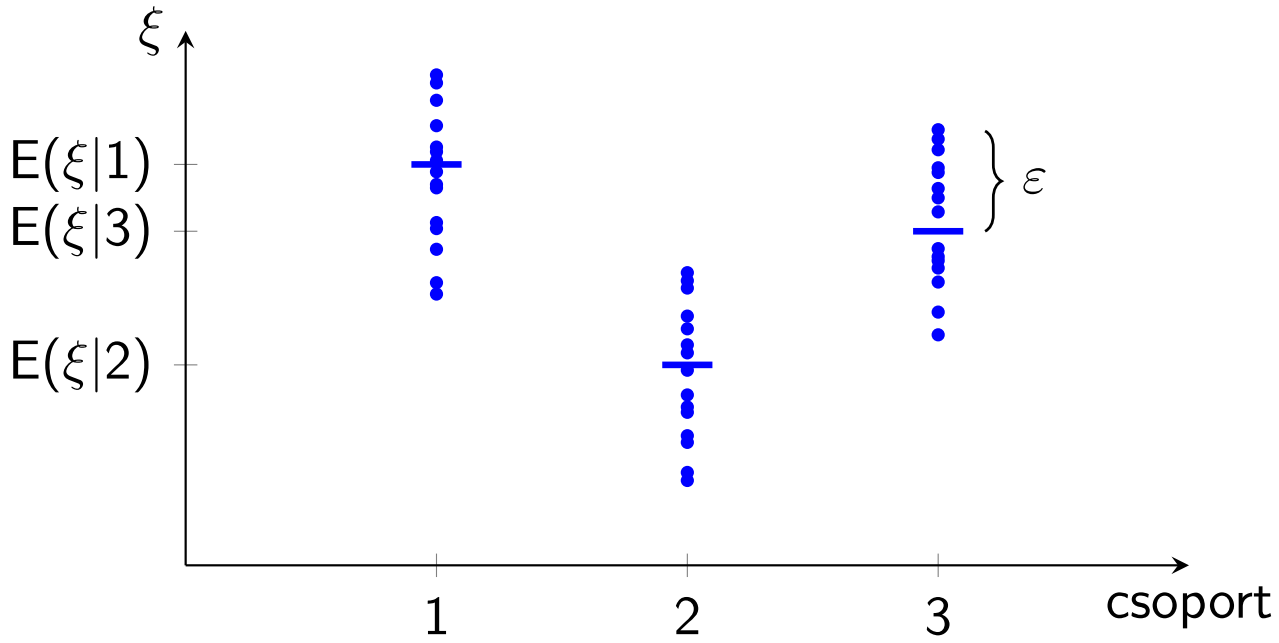
Tegyük fel, hogy a kiválasztott elem a j. csoportból származik! Ekkor:

$$\text{egyedi hatás} = \varepsilon = \xi - \mathrm{E}(\xi|j) = \text{mért érték} - \text{csoportátlag}$$

Tehát a mért érték:

$$\xi = \mathrm{E}(\xi|j) + \varepsilon = \text{csoportátlag} + \text{egyedi hatás}$$

Példa a teljes sokaságra $r = 3$ csoport esetén:



Független mintákat veszünk az egyes csoportokból:

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

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

Becslések a minta alapján:

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

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

Ugyanígy becsülhető minden mutatószám csoportonkénti bontásban is.

Szórástesztek

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

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

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:

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

Feltevések:

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

Megjegyzések:

- Akkor fogadjuk el a nullhipotézist, ha $\overline{\xi}_1 \approx \overline{\xi}_2 \approx \ldots \approx \overline{\xi}_r$.
- 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**.

További próbák a csoporthatások tesztelésére

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

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

Formálisan mindegyik teszthez szükséges a csoportonkénti normalitás, de mindegyik módszer robusztus erre nézve.

- **Kétmintás t-próba:** Az ANOVA speciális esete $r = 2$ csoportra.
- **Welch-féle F-próba:** Az ANOVA általánosítása arra a esetre, amikor a csoportonkénti szórások nem azonosak.
- **Welch-próba:** A Welch-féle F-próba speciális esete $r = 2$ csoportra.

Csoportok száma	Csoportonként azonos szórás	Tetszőleges szórás
$r = 2$	kétmintás t-próba	Welch-próba
$r \geq 2$	ANOVA	Welch-féle F-próba

Normalitásvizsgálat

Számos statisztikai módszernél meg van követelve, hogy a háttérváltozó normális eloszlású legyen a teljes sokaságban.

- Ezek a módszerek normális eloszlás esetén bármilyen n mintaméretre alkalmazhatóak.
- De hogyan döntsük el, hogy ez a feltétel teljesül-e?

Ha a próba robusztus a normalitásfeltételre nézve, akkor:

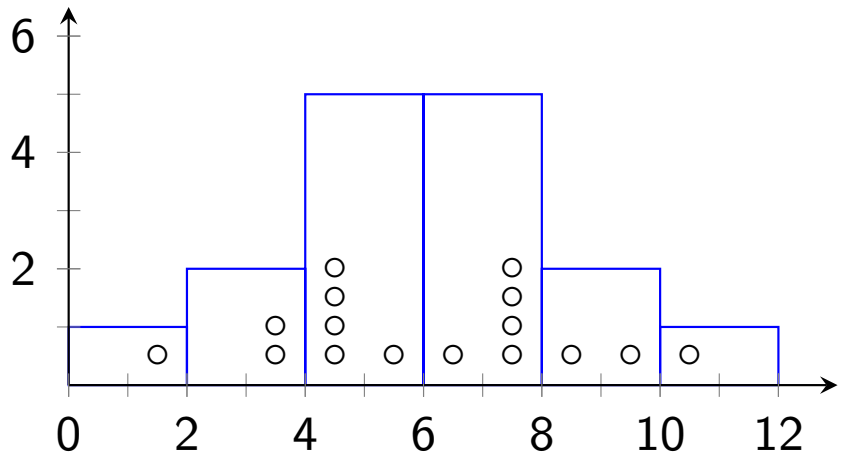
- a próba „közel normális” eloszlás esetén is alkalmazható, cserébe nagyobb mintára van szükség;
- a „közel normális” eloszlás ellenőrizhető grafikus illeszkedésvizsgálattal.

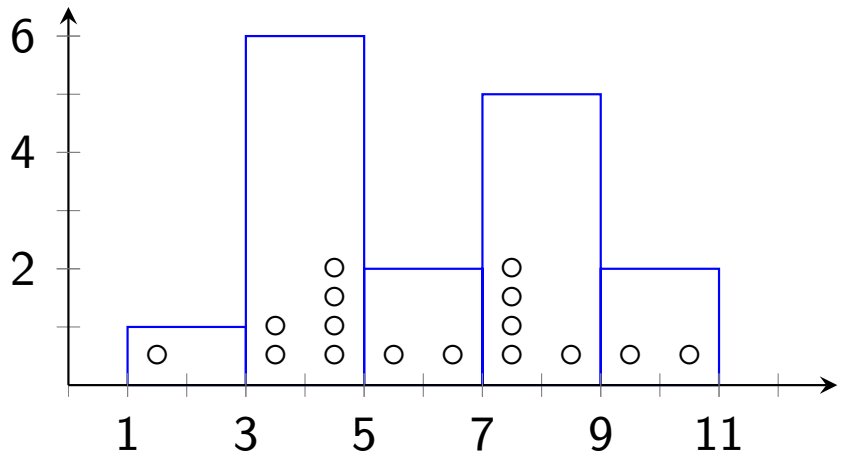
Ha a próba nem robusztus, akkor:

- nem elég, ha a minta „közel normális” eloszlású;
- teszteléssel kell ellenőrizni, hogy a háttérváltozó normális eloszlású.

Eddig hisztogram alkalmazásával végeztünk grafikus normalitásvizsgálatot. Probléma:

- A hisztogram alakját nagymértékben befolyásolja a beosztáspontok választása. A beosztáspontok megváltoztatásával ugyanazon mintára teljesen más hisztogramot kaphatunk.
- Emiatt a grafikus normalitásvizsgálat megbízhatatlan lehet.
- Ez különösen igaz kis mintaelemszám esetén.



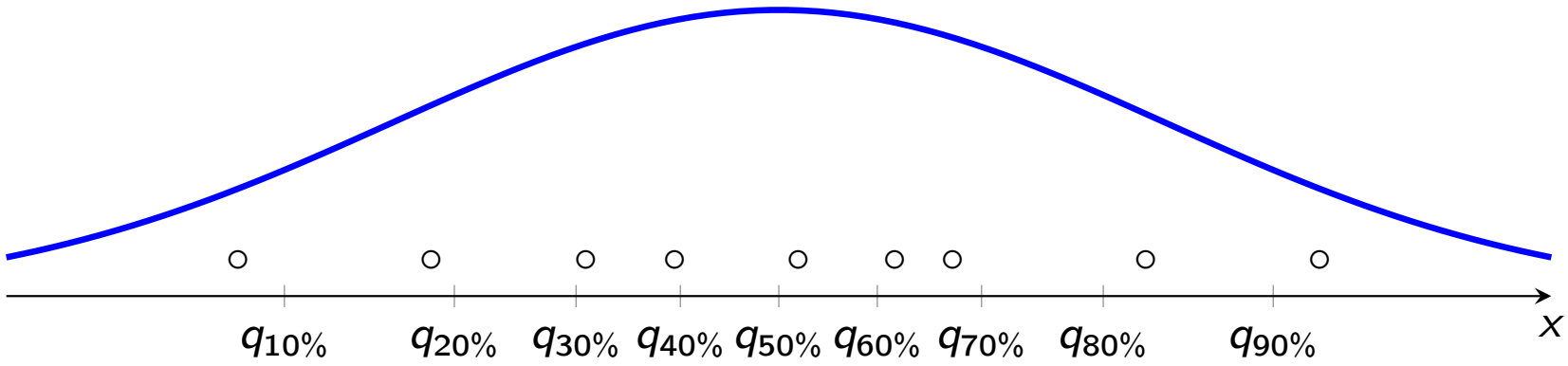


A hisztogram csak nagy minta ($n \geq 100$) esetén ad jó becslést az igazi sűrűségfüggvényre!

Tegyük fel, hogy a ξ háttérváltozó folytonos!

- Statisztikai minta: $\xi_1, \ldots, \xi_n$
- Rendezett minta: $\xi_1^* \leq \ldots \leq \xi_n^*$
- Elméleti α-kvantilis: $P(\xi < q_\alpha) = \alpha$
- Megmutatható, hogy ekkor $\xi_i^* \approx q_{\frac{i}{n+1}}$ minden i esetén.
- A közelítés a mintaméret növelésével egyre pontosabb.

Példa: ha ξ normális eloszlású, akkor $n = 9$ esetén



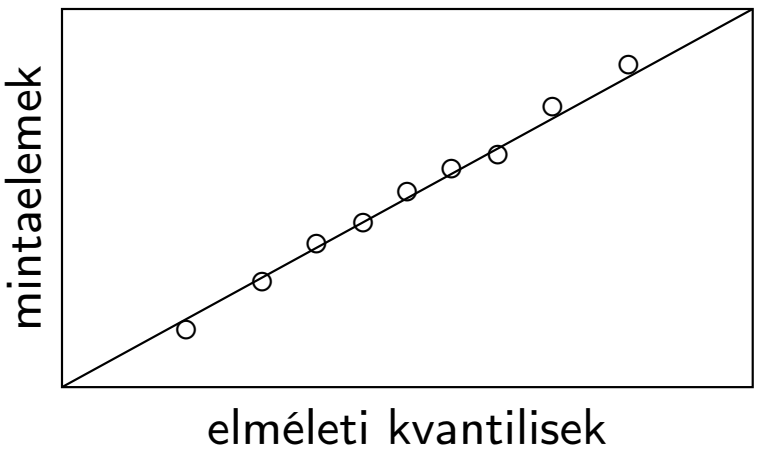
Ellenőrizni akarjuk, hogy a ξ változó normális eloszlású-e:

- rendezett minta: $\xi_1^* \leq \ldots \leq \xi_n^*$
- a normális eloszlás elméleti kvantilisei: $q_{\frac{1}{n+1}}, \ldots, q_{\frac{n}{n+1}}$

Q-Q plot: Koordináta-rendszerben ábrázoljuk a következő pontokat:

$$\left(q_{\frac{1}{n+1}}, \xi_1^*\right), \ldots, \left(q_{\frac{n}{n+1}}, \xi_n^*\right)$$

- Ha a ξ változó normális eloszlású, akkor $\xi_i^* \approx q_{\frac{i}{n+1}}$ minden i-re.
- Ekkor az ábrázolt pontok az $x = y$ egyeneshez közel helyezkednek el.
- Ha a pontok az egyenestől távol esnek, akkor a változó nem normális.



Shapiro–Wilk-próba

Feltevés: ξ folytonos változó ismeretlen eloszlásfüggvénnyel.

Nullhipotézis: $H_0 : \xi$ normális eloszlású.

Megjegyzések:

- A Shapiro–Wilk-próba a Q-Q plot alapján dönt a nullhipotézisről: azt vizsgálja, hogy a rendezett minta milyen mértékben tér el az elméleti kvantilisektől.
- A tapasztalatok szerint Shapiro–Wilk-próba a legjobb normalitásteszt: ez veti el a legnagyobb arányban a hamis nullhipotéziseket.
- A normalitás tesztelésére számos további próba is létezik:
  - Lilliefors-próba
  - Cramer–von Mises-próba
  - Anderson–Darling-próba
  - stb.

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?

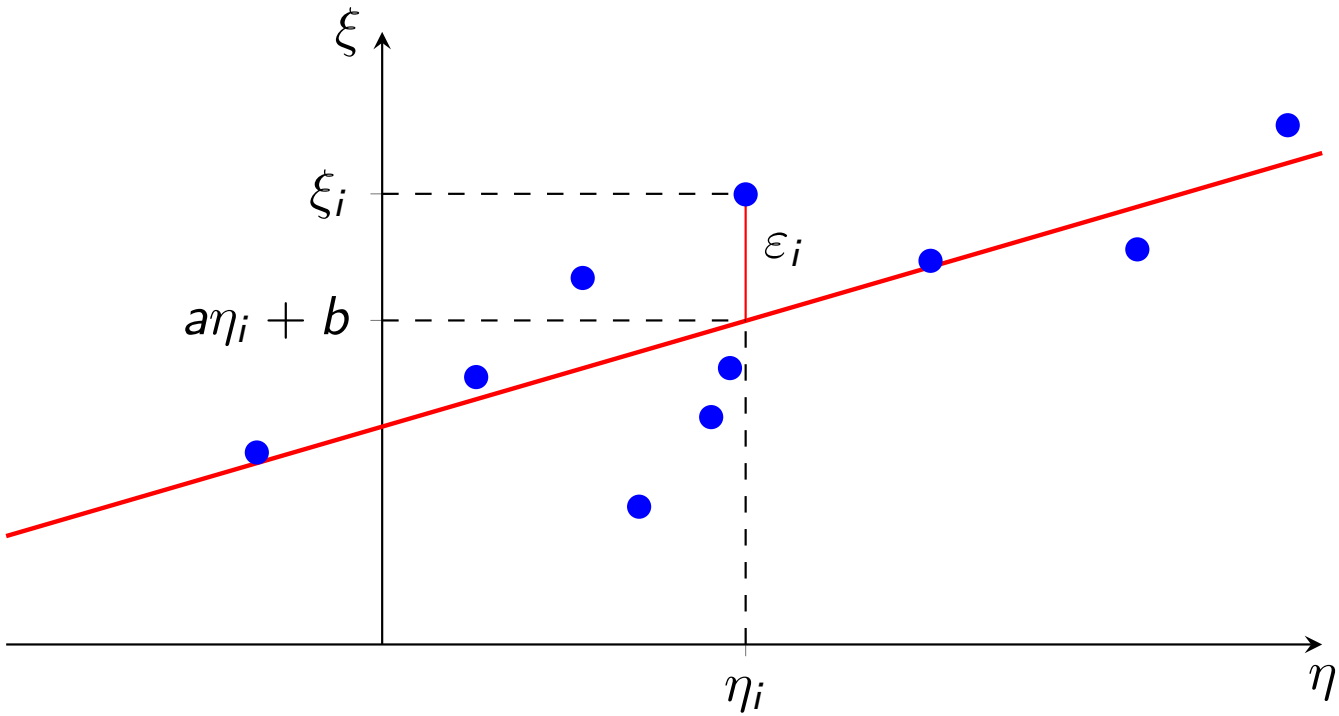
Lineáris regresszió: $\xi \approx a\eta + b$

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

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

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

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):**

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

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

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

- Az egyenletrendszer megoldása:

$$a = \frac{\mathrm{r}_n(\xi,\eta)\,\mathrm{D}_n^*(\xi)}{\mathrm{D}_n^*(\eta)}, \qquad b = \overline{\xi} - a\,\overline{\eta}.$$

Itt $\mathrm{r}_n(\xi,\eta)$ a Pearson-féle korrelációs együttható. (Lásd később.)

Hogyan kapjuk meg az optimális paraméterbeállítást?

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

A b paraméter szerinti deriváltt:

$$0 = \frac{dS}{db} = -2\sum_{i=1}^{n} \big(\xi_i - a\eta_i - b\big) = -2n\big(\overline{\xi} - a\overline{\eta} - b\big)$$

$$\Longrightarrow \qquad 0 = \overline{\xi} - a\overline{\eta} - b \qquad \Longrightarrow \qquad b = \overline{\xi} - a\overline{\eta}$$

Az a paraméter szerinti deriváltt:

$$0 = \frac{dS}{da} = -2\sum_{i=1}^{n} \eta_i\big(\xi_i - a\eta_i - b\big) = -2n\big(\overline{\xi\eta} - a\overline{\eta^2} - b\overline{\eta}\big)$$

$$0 = \overline{\xi\eta} - a\overline{\eta^2} - b\overline{\eta} = \overline{\xi\eta} - a\overline{\eta^2} - (\overline{\xi} - a\overline{\eta})\overline{\eta} = \big(\overline{\xi\eta} - \overline{\xi}\,\overline{\eta}\big) - a\big(\overline{\eta^2} - \overline{\eta}^2\big)$$

$$a = \frac{\overline{\xi\eta} - \overline{\xi}\,\overline{\eta}}{\overline{\eta^2} - \overline{\eta}^2} = \ldots = \frac{\mathrm{r}_n(\xi,\eta)\,\mathrm{D}_n^*(\xi)}{\mathrm{D}_n^*(\eta)}$$

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: $\xi \approx \overline{\xi}$

  A becslés teljes hibája a mintán:

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

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):

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

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

Tulajdonságai:

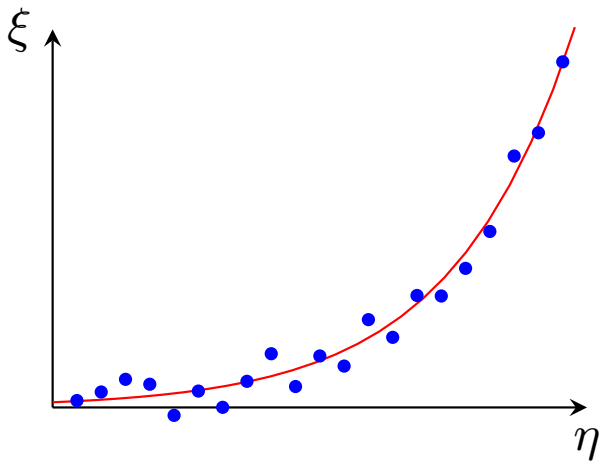
- $0 \leq R^2 \leq 1$.
- Minél nagyobb az R^2 értéke, annál pontosabb a lineáris becslés.
- Ha $R^2 \leq 0.5$, akkor a lineáris becslés nagyon pontatlan.

Miért R^2 a mennyiség neve? Lineáris regressziós esetén: $R^2 = \mathrm{r}_n^2(\xi, \eta)$.

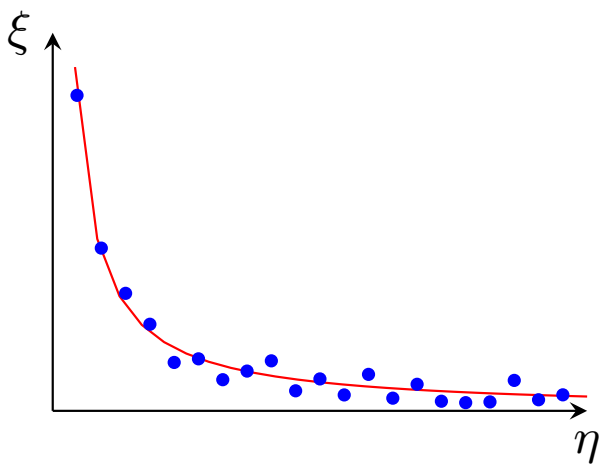
Ez az egyenlőség nemlineáris regresszió esetén már nem teljesül!

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

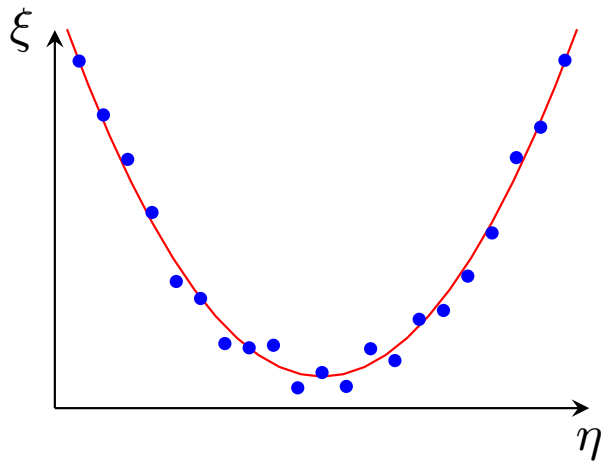
- Exponenciális regresszió: $\xi \approx e^{a\eta+b}$,
- Reciprokos regresszió: $\xi \approx a/\eta + b$,
- Másodfokú regresszió: $\xi \approx a(\eta - b)^2$.



$f_{a,b}(x) = e^{ax+b}$



$f_{a,b}(x) = a/x + b$



$f_{a,b}(x) = a(x - b)^2$

A regresszió általános modellje:

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

A formulában:

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

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

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

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: $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_n$

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

Empirikus kovariancia és **Pearson-féle korrelációs együttható**:

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

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: $\xi \approx a\eta + b$.

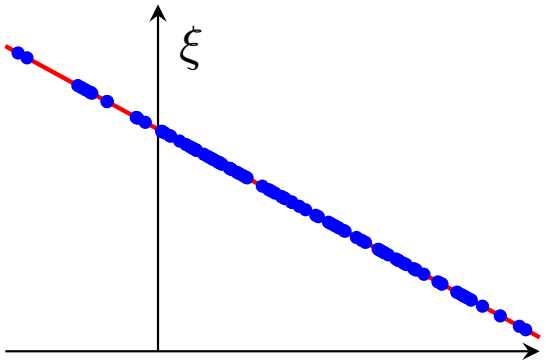
A Pearson-féle korrelációs együttható jellemzi a kapcsolat irányát:

- 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.

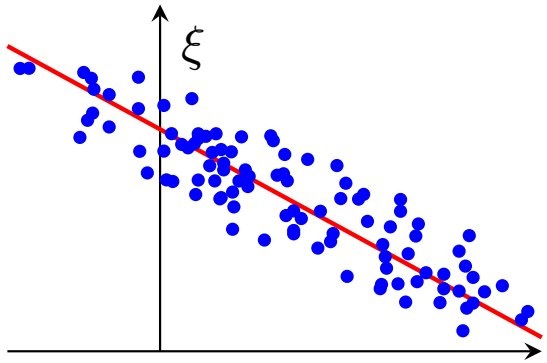
A Pearson-féle korrelációs együttható jellemzi a kapcsolat erősségét:

- 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.

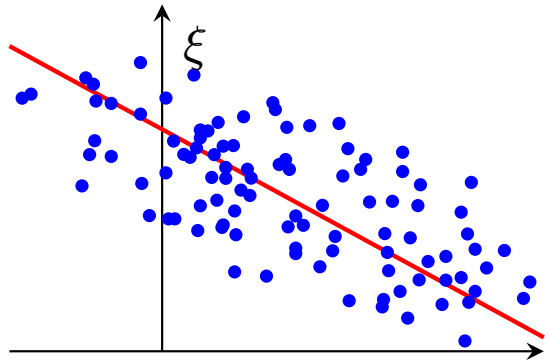
A Pearson-féle korrelációs együttható hátránya: csak a lineáris kapcsolatot képes mérni. Ha a változók között van kapcsolat, de nem lineáris jellegű, akkor ezt nem mindig detektálja.



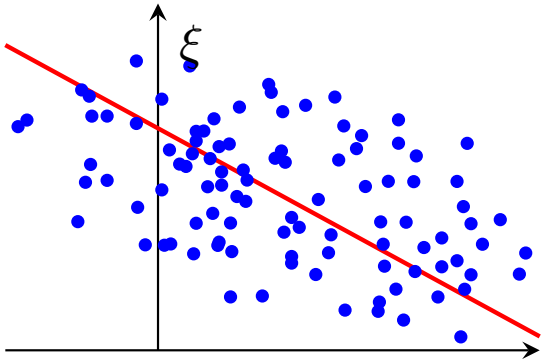
$r_n(\xi, \eta) = -1$



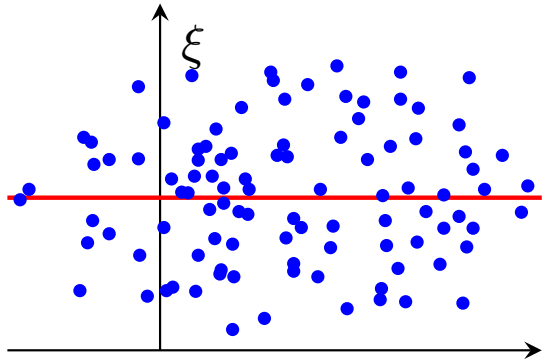
$r_n(\xi, \eta) = -0.9$



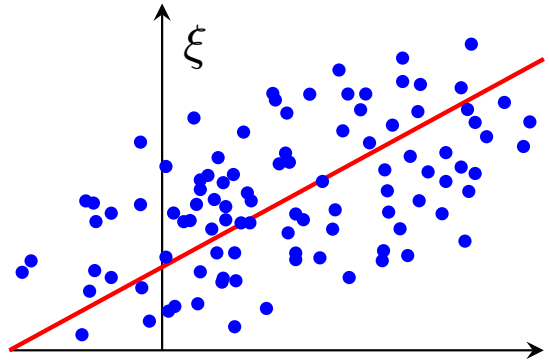
$r_n(\xi, \eta) = -0.7$



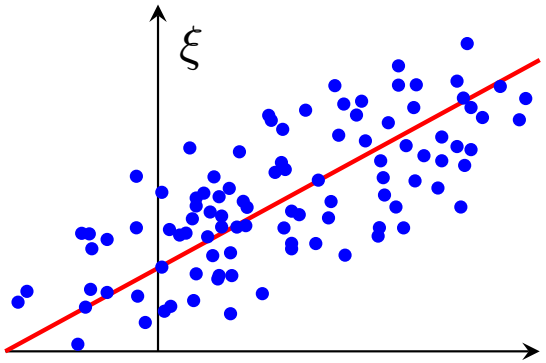
$r_n(\xi, \eta) = -0.5$



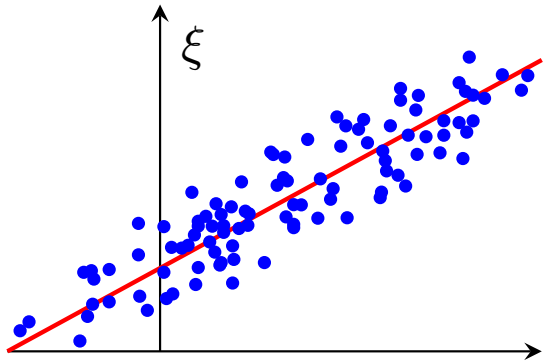
$r_n(\xi, \eta) = 0$



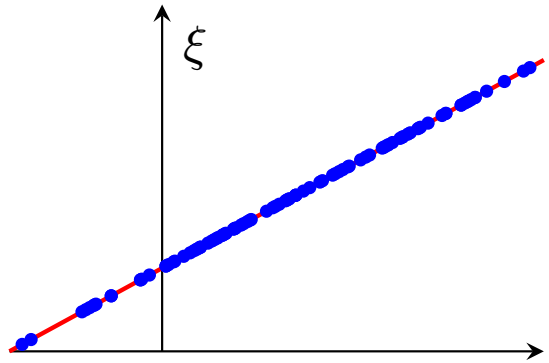
$r_n(\xi, \eta) = 0.5$



$r_n(\xi, \eta) = 0.7$



$r_n(\xi, \eta) = 0.9$



$r_n(\xi, \eta) = 1$

Spearman-féle korrelációs együttható: $\rho_n(\xi, \eta)$.

Fontosabb tulajdonságai:

- Szimmetria: $\rho_n(\xi, \eta) = \rho_n(\eta, \xi)$
- Lehetséges értékek: $-1 \leq \rho_n(\xi, \eta) \leq +1$
- A két változó közötti **rendezési kapcsolatot** jellemzi.

Hogyan jellemzi a rendezési kapcsolatot?

- Ha $\rho_n(\xi, \eta) \approx +1$, akkor pozitív irányú rendezési kapcsolatot látunk: ha a mintában $\eta_i \leq \eta_j$, akkor jellemzően $\xi_i \leq \xi_j$.
- Ha $\rho_n(\xi, \eta) \approx -1$, akkor negatív irányú rendezési kapcsolatot látunk: ha a mintában $\eta_i \leq \eta_j$, akkor jellemzően $\xi_i \geq \xi_j$.
- Ha $\rho_n(\xi, \eta) \approx 0$, akkor nincs rendezési kapcsolat.

Bal oldali ábra: pozitív irányú kapcsolat a változók között.

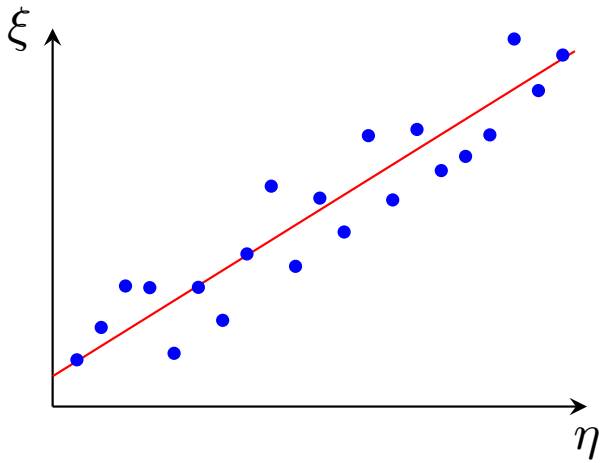
- $r_n(\xi, \eta) = +0.9$: erős lineáris kapcsolat.
- $\rho_n(\xi, \eta) = +0.9$: erős rendezési kapcsolat.

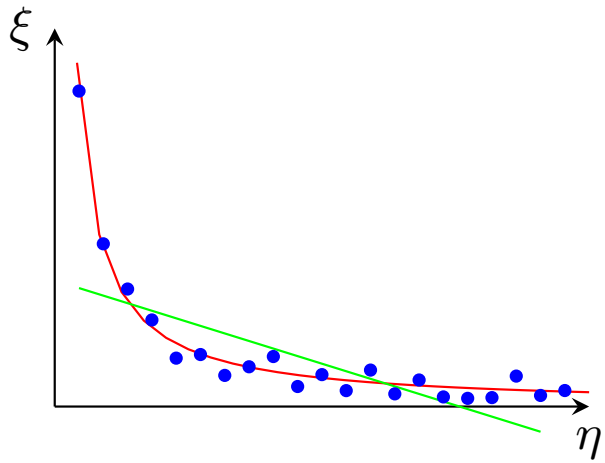
Középső ábra: negatív irányú kapcsolat a változók között.

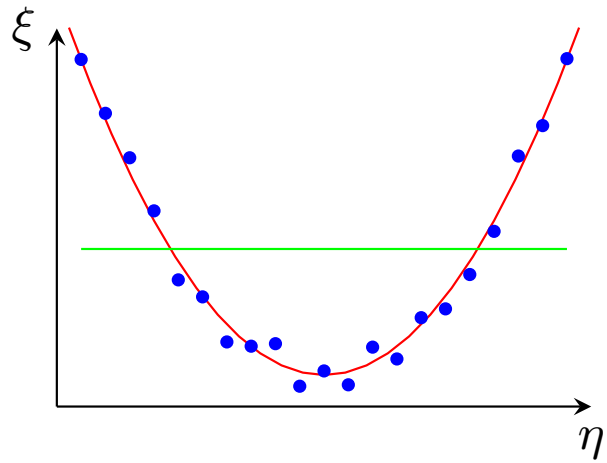
- $r_n(\xi, \eta) = -0.7$: közepesen erős lineáris kapcsolat.
- $\rho_n(\xi, \eta) = -0.9$: erős rendezési kapcsolat.

Jobb oldali ábra: változó irányú kapcsolat.

- $r_n(\xi, \eta) = 0$: nincs lineáris kapcsolat.
- $\rho_n(\xi, \eta) = 0$: nincs rendezési kapcsolat.







Korrelációs teszt

Cél a függetlenség tesztelése összetartozó minták alapján.

Feltevések:

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

Nullhipotézis: H_0 : ξ és η függetlenek.

Megjegyzések:

- A teszt robusztus a normalitásfeltételre, közel szimmetrikus folytonos eloszlású változók esetén lehet alkalmazni.
- A teszt egy kiválasztott korrelációs együttható alapján hoz döntést.
- 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.

  Spearman-féle korreláció: csak a rendezési kapcsolatot érzékeli.

Valószínűségek becslése és tesztelése

Tekintsünk egy tetszőleges tulajdonságot a sokaságban! Véletlenszerűen kiválasztunk egy egyedet, és legyen:

$$A = \text{a kiválasztott egyed rendelkezik a vizsgált tulajdonsággal}$$

Ekkor: $\mathrm{P}(A) = \text{a vizsgált tulajdonság aránya a teljes sokaságban}$

Kérdés: Hogyan becsülhető ez az arány egy statisztikai minta alapján?

Gyakoriság, tapasztalati gyakoriság (frequency):

$$k_n(A) = \text{a mintaelemek közül ennyi rendelkezik a vizsgált tulajdonsággal}$$

Relatív gyakoriság (relative frequency): $r_n(A) = k_n(A)/n$.

A mintaelemek ekkora hányada rendelkezik a vizsgált tulajdonsággal.

Valószínűség (arány) becslése: $\mathrm{P}(A) \approx r_n(A)$.

A relatív gyakoriság erősen konzisztens becslés: $r_n(A) \to \mathrm{P}(A), \quad n \to \infty$.

Feladat: Megvizsgáltunk 200 japán nemzetiségű embert. Közülük rendre 62, 84, 36 illetve 18 esett a 0, az A, a B és az AB vércsoportba. Adjunk becslést a vércsoportok arányára a teljes sokaságon belül!

Véletlenszerűen kiválasztunk egy japán nemzetiségű embert.

$\xi =$ a kiválasztott ember vércsoportja, $\quad R_\xi = \{0, \mathrm{A}, \mathrm{B}, \mathrm{AB}\}$

Értékek	x_i	0	A	B	AB	össz.
Tapasztalati gyak.	$k_n(x_i)$	62	84	36	18	200
Relatív gyak.	$r_n(x_i)$	0.31	0.42	0.18	0.09	1

Tegyük fel, hogy a japán emberek körében a vércsoportok aránya rendre 30%, 40%, 20% illetve 10%! Várhatóan hány megfigyelést kapunk az egyes vércsoportokra a 200 elemű mintán belül?

Értékek	x_i	0	A	B	AB	össz.
Hipotetikus arány	p_i	0.3	0.4	0.2	0.1	1
Várt gyakoriság	np_i	60	80	40	20	200

Khi-négyzet (χ^2) próba valószínűségek tesztelésére

Feltevések:

- ξ egy diszkrét változó, $R_\xi = \{x_1, \ldots, x_K\}$
- hipotetikus valószínűségek: $p_1, \ldots, p_K > 0$, $p_1 + \ldots + p_K = 1$
- nagy mintaméret: $n \geq 5/\min(p_1, \ldots, p_K)$

Nullhipotézis: $H_0 : P(\xi = x_i) = p_i$ minden $i = 1, \ldots, K$ esetén

Próbastatisztika:

$$\chi^2 = \sum_{i=1}^{K} \frac{\left[k_n(x_i) - np_i\right]^2}{np_i} = \sum_{i=1}^{K} \frac{\left[\text{tapasztalati gyak.} - \text{várt gyak.}\right]^2}{\text{várt gyak.}}$$

Kritikus érték: $c_\alpha = F^{-1}_{\chi^2, K-1}(1-\alpha)$

Döntés: akkor fogadjuk el a nullhipotézist, ha $|\chi^2| \leq c_\alpha$.

Feladat: A megadott minta alapján teszteljük le 10%-os szignifikancia szinten azt a nullhipotézist, hogy a japán emberek körében a vércsoportok aránya rendre 30%, 40%, 20% illetve 10%!

Értékek	x_i	0	A	B	AB	össz.
Tapasztalati gyak.	$k_n(x_i)$	62	84	36	18	200
Relatív gyak.	$r_n(x_i)$	0.31	0.42	0.18	0.09	1
Hipotetikus arány	p_i	0.3	0.4	0.2	0.1	1
Várt gyakoriság	np_i	60	80	40	20	200

Próbastatisztika:

$$\chi^2 = \frac{(62-60)^2}{60} + \frac{(84-80)^2}{80} + \frac{(36-40)^2}{40} + \frac{(18-20)^2}{20} = 0.87$$

A kritikus érték: $c_\alpha = F^{-1}_{\chi^2,3}(0.9) = 6.251$

Most $|\chi^2| \leq c_\alpha$, ezért a nullhipotézist elfogadjuk.

Feladat: Azon embereknek a vérében található meg az A típusú antigén, akik az A vagy az AB vércsoportba esnek. Teszteljük le azt, hogy a japán népességen belül az A típusú antigén 60% arányban jelenik meg!

Értékek	x_i	A jelen van	A nincs jelen	össz.
Tapasztalati gyak.	$k_n(x_i)$	102	98	200
Relatív gyak.	$r_n(x_i)$	0.51	0.49	1
Hipotetikus arány	p_i	0.6	0.4	1
Várt gyakoriság	np_i	120	80	200

Próbastatisztika:

$$\chi^2 = \frac{(102-120)^2}{120} + \frac{(98-80)^2}{80} = 6.75$$

A kritikus érték: $c_\alpha = F^{-1}_{\chi^2,1}(0.9) = 2.706$

Most $\chi^2 > c_\alpha$, ezért a nullhipotézist elvetjük.