Biostatisztika

Kunosné Nedényi Fanni, Szűcs Gábor

Szegedi Tudományegyetem, Bolyai Intézet

2018/19 őszi félév

Mi is az a biostatisztika?

Ide majd még jön valami, valamikor...

Események valószínűsége

A valószínűségszámítás a matematika egyik ága, melynek célja a véletlen jelenségekhez kapcsolódó valószínűségek meghatározása. Alapfogalmak:

- **Véletlen kísérlet:** Egy véletlen jelenség megfigyelése.
- **Kimenetelek:** A véletlen kísérlet lehetséges eredményei.
- **Esemény:** A kísérlet aktuális kimenetelével kapcsolatos állítás. Egy esemény akkor **következik be**, ha a véletlen kísérlet olyan kimenetelt ad, melyre az állítás igaz.
- **Valószínűség:** Annak az esélye, hogy az esemény bekövetkezik.

Példa:

- Véletlen kísérlet: feldobunk egy szabályos dobókockát.
- Kimenetelek: 1, 2, 3, 4, 5, 6.
- Egy esemény: $A =$ páros számot dobunk. Ez akkor következik be, ha a 2, 4, 6 értékek valamelyikét dobjuk, egyébként nem következik be.
- Az A esemény valószínűsége: $P(A) = 3/6 = 50\%$.

Ezen a kurzuson jellemzően az lesz majd a kísérlet, hogy véletlenszerűen kiválasztunk egy vagy több egyedet egy ember/állat/növény populációból. A „véletlenszerűen" szó itt azt jelenti, hogy mindegyik egyedet ugyanakkora eséllyel választjuk ki.

Feladat: Magyarországon az emberek 52 illetve 24 százalékának a vérében található meg az A illetve a B típusú antigén. Mindkét antigén az emberek 8 százalékánál található meg. Véletlenszerűen kiválasztunk egy magyar embert, és leteszteljük az antigénekre. Tekintsük a következő eseményeket:

$A =$ a kiválasztott ember rendelkezik az A típusú antigénnel

$B =$ a kiválasztott ember rendelkezik a B típusú antigénnel

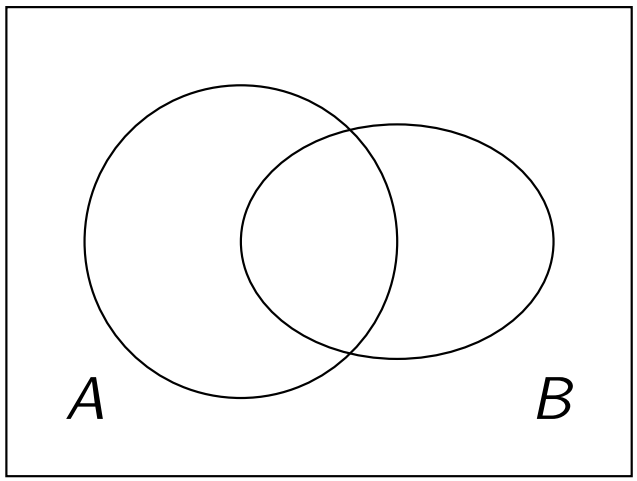
Most minden embert azonos eséllyel választunk ki, ezért a tulajdonságok bekövetkezési valószínűsége azonos lesz a tulajdonságok teljes populáción belül mért részarányával:

$P(A) =$ az A antigén aránya a teljes populáción belül $= 52\% = 0{,}52$

$P(B) =$ a B antigén aránya a teljes populáción belül $= 24\% = 0{,}24$

$P(A \text{ és } B) =$ a két antigén együttes megjelenésének aránya $= 8\% = 0{,}08$

Az alábbi ábrán a magyar emberek populációját ábrázoljuk a két antigén szempontjából. A mellette lévő táblázat a vércsoportokat foglalja össze.



	van A	nincs A
van B	AB	B
nincs B	A	0

Amit tudunk: $P(A) = 52\%$, $P(B) = 24\%$, $P(A \text{ és } B) = 8\%$.

Feladat: Határozzuk meg a vércsoportok részarányát!

$P(\text{a kiválasztott ember az AB vércsoportba esik}) = P(A \text{ és } B) = 8\%$

$P(\text{A vércsoport}) = P(A \text{ igen, de } B \text{ nem}) = P(A) - P(A \text{ és } B) = 44\%$

$P(\text{B vércsoport}) = P(B \text{ igen, de } A \text{ nem}) = P(B) - P(A \text{ és } B) = 16\%$

$P(\text{0 vércsoport}) = 100\% - \text{az előző három összege} = 32\%$

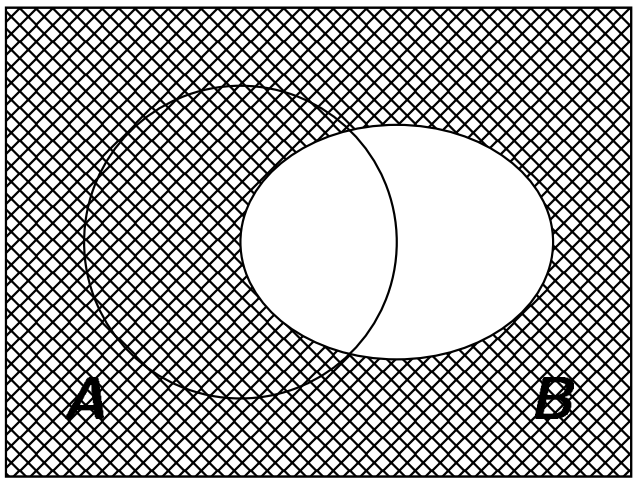
A kurzuson a valószínűség a teljes populáción belüli arányt jelenti. Időnként szükségünk lesz arra, hogy az arányokat egy részpopuláción belül vizsgáljuk. Az A eseménynek a B eseményre vett **feltételes valószínűsége**:

$$P(A|B) = \frac{P(A \text{ és } B)}{P(B)}.$$

A feltételes valószínűség jelentése:

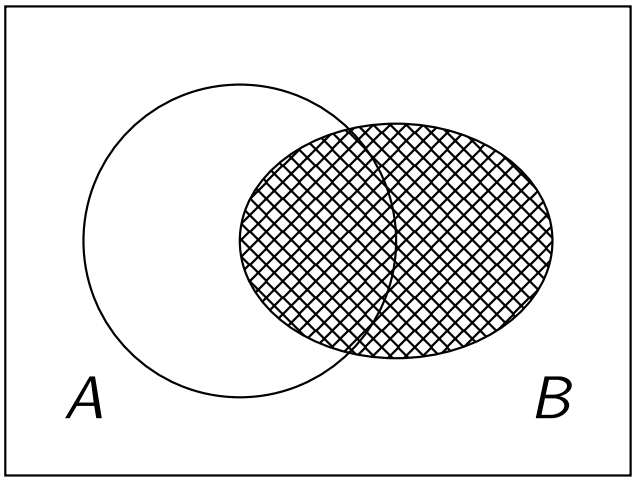
$$\begin{aligned} P(A|B) &= \text{az } A \text{ tulajdonság aránya a } B \text{ részpopuláción belül} \\ &= \text{az } A \text{ esemény valószínűsége, ha tudjuk, hogy } B \text{ bekövetkezik} \end{aligned}$$

Feladat: Mennyi $P(A|B)$ az előző feladatban?



$$P(A|B) = \frac{P(A \text{ és } B)}{P(B)} = \frac{8\%}{24\%} = \frac{1}{3} = 33\%$$

Feladat: Mekkora az A típusú antigénnel rendelkező emberek aránya azon emberek között, akik nem rendelkeznek a B antigénnel?



$$P(A \mid \text{nem } B) = \frac{P(A \text{ és nem } B)}{P(\text{nem } B)} = \frac{44\%}{76\%} = 58\%,$$

$$P(\text{nem } B) = 100\% - P(B) = 76\%,$$

$$P(A \text{ és nem } B) = P(\text{A vércsoport}) = 44\%.$$

Értelmezzük, hogy mit kaptunk:

- Ha véletlenszerűen kiválasztunk egy embert a teljes populációból, akkor 52% valószínűséggel található meg nála az A típusú antigén.
- Ha tudjuk, hogy a kiválasztott ember rendelkezik a B antigénnel, akkor 33% az esélye, hogy az A is antigén megtalálható nála.
- Ha viszont azt tudjuk, hogy nem rendelkezik a B típusú antigénnel, akkor 58% az esélye, hogy az A antigén megtalálható nála.
- Tehát a B antigén jelenléte csökkenti az A antigén megjelenési esélyét:

$$P(A|B) = 33\% < 58\% = P(A \mid \text{nem B})$$

Legyenek A és B tetszőleges események. Bebizonyítható, hogy ekkor az alábbi három egyenlőség ekvivalens, tehát következnek egymásból:

1. $P(A \text{ és } B) = P(A)P(B)$
2. $P(A|B) = P(A)$
3. $P(B|A) = P(B)$

Amennyiben ezen egyenlőségek közül bármelyik (és ezáltal mindegyik) teljesül, akkor azt mondjuk, hogy A és B **független események**.

A függetlenség szemléletesen azt jelenti, hogy a két esemény nem hat egymásra, nem akadályozzák, és nem is segítik elő egymás bekövetkezését.

Lássuk, hogyan következik az első egyenlőségből a második:

$$P(A|B) = \frac{P(A \text{ és } B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A).$$

Feladat: A vércsoportos feladatban A és B független események?

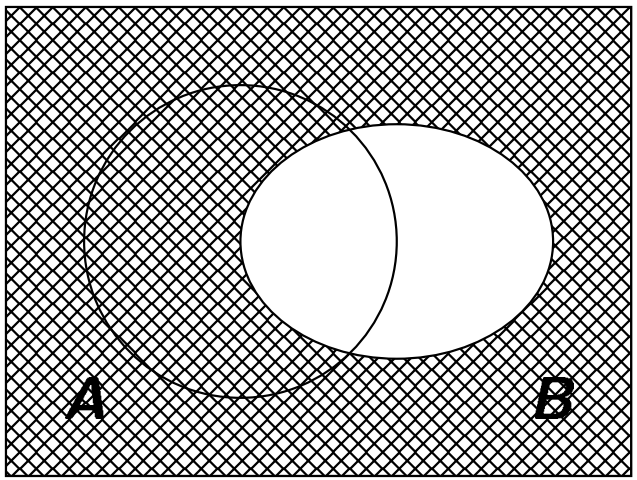
Nem, ugyanis $P(A|B) = 33\% < 52\% = P(A)$.

Feladat: A feladatban a két antigén aránya: $P(A) = 52\%$ és $P(B) = 24\%$. Mikor lenne a két antigén megjelenése független egymástól?

A két antigén megjelenése akkor független, ha

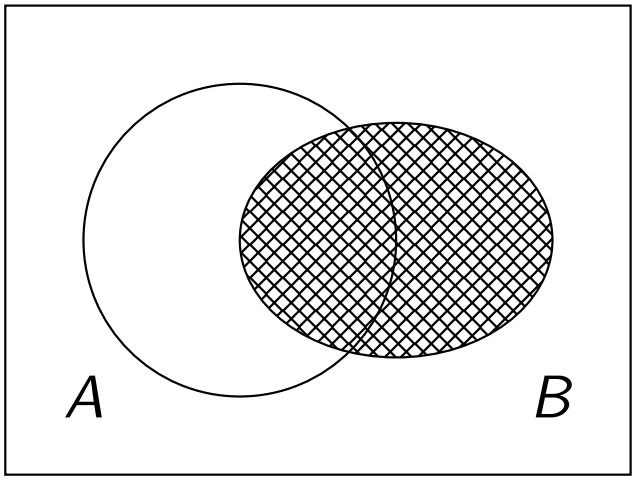
$$P(A \text{ és } B) = P(A)P(B) = 0{,}52 \cdot 0{,}24 = 0{,}125 = 12{,}5\%$$

Feladat: Mennyi lenne a feltételes valószínűségek értéke ebben az esetben?



Az A típusú antigénnel rendelkező emberek aránya a B csoporton belül:

$$P(A|B) = \frac{P(A \text{ és } B)}{P(B)} = \frac{12{,}5\%}{24\%} = 52\%$$



Az A típusú antigénnel rendelkező emberek aránya a B csoporton kívül:

$$P(A \,|\, \text{nem } B) = \frac{P(A \text{ és nem } B)}{P(\text{nem } B)} = \frac{39{,}5\%}{76\%} = 52\%,$$

$$P(A \text{ és nem } B) = P(A) - P(A \text{ és } B) = 39{,}5\%.$$

Tehát a független esetben az A antigénnel rendelkező emberek aránya (=kiválasztási valószínűsége) azonos az alábbi három populáción belül:

- a teljes populáción belül: $P(A) = 52\%$,
- a B típusú antigénnel rendelkező emberek részpopulációján belül: $P(A|B) = 52\%$,
- a B típusú antigénnel nem rendelkező emberek részpopulációján belül: $P(A \,|\, \text{nem } B) = 52\%$.

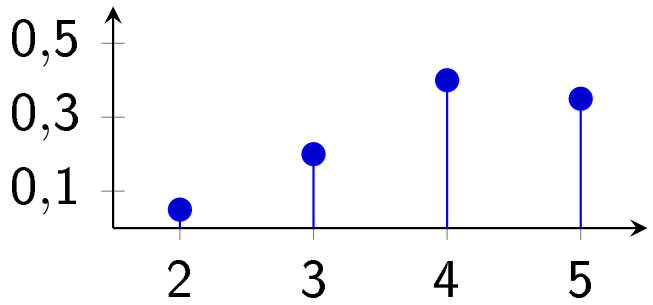
Diszkrét valószínűségi változók

A biológiai vizsgálatok során gyakran felmerül az a kérdés, hogy mi az eloszlása egy mennyiségnek (életkor, testtömeg, utódok száma, stb.) egy populáción belül. Válasszunk ki véletlenszerűen egy egyedet a populációból, és legyen ξ a vizsgált mennyiség értéke a kiválasztott egyed esetében. Mivel az egyedet véletlenszerűen választjuk, a ξ érték egy véletlen szám lesz.

- **Valószínűségi változó:** Egy véletlen kísérletből származó véletlen szám (véletlen mennyiség). Jele: ξ (kszi), η (éta), stb.
- **Értékkészlet:** A változó lehetséges értékeinek a halmaza. Jele: R_ξ, R_η
- **Diszkrét valószínűségi változó:** A változó értékkészlete egy véges vagy végtelen sorozat. Mi tipikusan két esettel fogunk találkozni:
  - az értékkészlet véges halmaz;
  - VAGY minden lehetséges érték egész szám.
- **Folytonos valószínűségi változó:** A változó értékkészlete egy véges vagy végleten intervallum.

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	5%	20%	40%	35%



A ξ egy valószínűségi változó, értékkészlete $R_\xi = \{2, 3, 4, 5\}$. Ez egy véges halmaz, tehát a ξ diszkrét változó. A fészket véletlenszerűen választottuk, ezért a ξ pontosan akkora valószínűséggel veszi fel az egyes értékeket, amennyi ezen értékek aránya a teljes (fészek-) populáción belül:

$$P(\xi = 2) = 0{,}05, \quad P(\xi = 3) = 0{,}2, \quad P(\xi = 4) = 0{,}4, \quad P(\xi = 5) = 0{,}35.$$

Legyen ξ diszkrét valószínűségi változó. A $p_x = P(\xi = x)$ valószínűségeket a változó **valószínűségeloszlásának** nevezzük. Véletlenszerű kiválasztás esetén a valószínűségeloszlás azonos a populáción belül mért arányokkal.

Feladat: Mennyi a ξ változó lehetséges értékeinek összvalószínűsége?

$$P(\xi = 2) + P(\xi = 3) + P(\xi = 4) + P(\xi = 5) = 0{,}05 + 0{,}2 + 0{,}4 + 0{,}35 = 1$$

Feladat: A fészkek mekkora hányadában található legfeljebb 3 tojás?

$$P(\text{legfeljebb 3 tojás}) = P(\xi \leq 3) = P(\xi = 2) + P(\xi = 3) = 0{,}25.$$

Feladat: Melyik tojásszám a leggyakoribb a populációban?

A 4-es érték a leggyakoribb, a fészkek 40%-ában ennyi tojás található.

Feladat: Átlagosan hány tojás található a fészkekben?

A tojások átlagos száma: $\mathrm{E}(\xi) = 2 \cdot 0{,}05 + 3 \cdot 0{,}2 + 4 \cdot 0{,}4 + 5 \cdot 0{,}35 = 4{,}05$.

Legyen ξ diszkrét valószínűségi változó.

- **Módusz:** A ξ változó legnagyobb valószínűségű értéke.
  Jelentése: a ξ változó leggyakoribb értéke a teljes populáción belül.
- **Várható érték:** $\mathrm{E}(\xi) = \sum_{x \in R_\xi} xP(\xi = x)$.
  Jelentése: a ξ változó átlagos értéke a teljes populáción belül.

Milyen módon számszerűsíthetjük egy ξ diszkrét változó szóródását?

- **Várható értéktől való átlagos eltérés:** $\sum_{x \in R_\xi} \left| x - \mathrm{E}(\xi) \right| P(\xi = x)$
- **Variancia:**

$$\mathrm{Var}(\xi) = \sum_{x \in R_\xi} \left(x - \mathrm{E}(\xi) \right)^2 P(\xi = x)$$

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

A szóródás mérésére a várható értéktől való átlagos eltérés egy egyszerű mutatószám lenne, de sajnos ennek rosszak a matematikai tulajdonságai. Emiatt inkább a szórást szoktuk alkalmazni a szóródás mérésére. A két érték jellemzően közel van egymáshoz:

$$\text{szórás} \approx \text{várható értéktől való átlagos eltérés}$$

Emiatt az alkalmazásokban a szórást magát is úgy értelmezzük, mint az átlagtól való átlagos eltérés. A varianciára csak azért van szükségünk, mert abból számoljuk ki a szórást.

Feladat: Mennyi a várható értéktől való átlagos eltérés és a szórás a jelen feladatban? (A várható érték $\mathrm{E}(\xi) = 4{,}05$.)

x	2	3	4	5
$\lvert x - \mathrm{E}(\xi)\rvert$	2,05	1,05	0,05	0,95
$(x - \mathrm{E}(\xi))^2$	$2{,}05^2$	$1{,}05^2$	$0{,}05^2$	$0{,}95^2$
$P(\xi = x)$	0,05	0,2	0,4	0,35

Várható értéktől való átlagos eltérés:

$$2{,}05 \cdot 0{,}05 + 1{,}05 \cdot 0{,}2 + 0{,}05 \cdot 0{,}4 + 0,95 \cdot 0{,}35 = 0{,}665$$

Variancia:

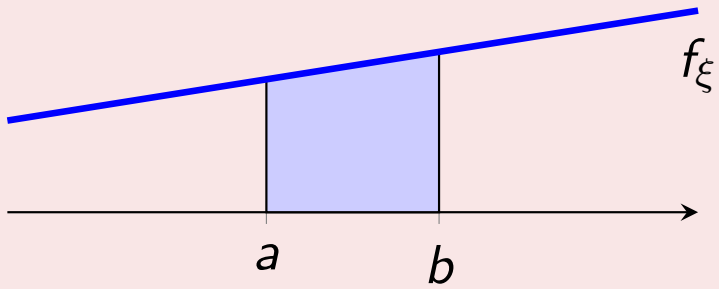
$$\mathrm{Var}(\xi) = 2,05^2 \cdot 0{,}05 + 1{,}05^2 \cdot 0{,}2 + 0{,}05^2 \cdot 0{,}4 + 0,95^2 \cdot 0{,}35 \approx 0{,}75$$

Szórás: $\mathrm{D}(\xi) = \sqrt{0{,}75} \approx 0{,}87$.

Folytonos valószínűségi változók

Egy valószínűségi változó **folytonos**, ha értékkészlete egy véges vagy végleten intervallum. A ξ folytonos változó **sűrűségfüggvénye** egy olyan $f_\xi : \mathbb{R} \to \mathbb{R}$ függvény, melyre tetszőleges a és b számok esetén:

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



Tekintünk egy mennyiséget (például a testtömeget) egy populáción belül. Véletlenszerűen kiválasztunk egy egyedet, és legyen ξ a mennyiség értéke ezen egyed esetében. Ekkor:

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

A folytonos változók és a sűrűségfüggvények néhány tulajdonsága:

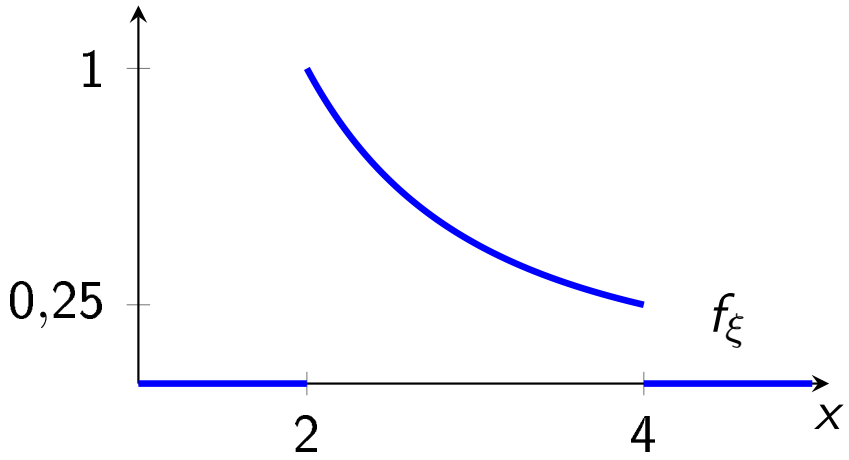
1. $\int_{-\infty}^{\infty} f_\xi(x)dx = 1$.
2. $f_\xi(x) \geq 0$ minden x valós szám esetén.
3. A ξ változó értékkészlete azon x számok halmaza, melyekre $f_\xi(x) > 0$.
4. Tetszőleges a szám esetén $P(\xi = a) = 0$.

Rövid indoklás a fenti állításokhoz:

1. $\int_{-\infty}^{\infty} f_\xi(x)dx = P(-\infty \leq \xi \leq \infty) = 1$.
2. Tegyük fel, hogy az f_ξ függvény negatív egy $[a, b]$ intervallumon. Ekkor $\int_a^b f_\xi(x)dx < 0$, tehát $\int_a^b f_\xi(x)dx \neq P(a \leq \xi \leq b)$, ami ellentmondás.
3. Ha $f_\xi = 0$ az $[a, b]$ intervallumon, akkor $P(a \leq \xi \leq b) = \int_a^b 0dx = 0$.
   Ha $f_\xi > 0$ az $[a, b]$ halmazon, akkor $P(a \leq \xi \leq b) = \int_a^b f_\xi(x)dx > 0$.
   Tehát a ξ változó oda eshet, ahol $f_\xi > 0$.
4. $P(\xi = a) = P(a \leq \xi \leq a) = \int_a^a f_\xi(x)dx = 0$.

Feladat: Egy állatpopulációban legyen ξ egy véletlenszerűen kiválasztott egyed tömege. A változó az alábbi sűrűségfüggvénnyel írható le.

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



Feladat: A teljes görbe alatti terület valóban 1?

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

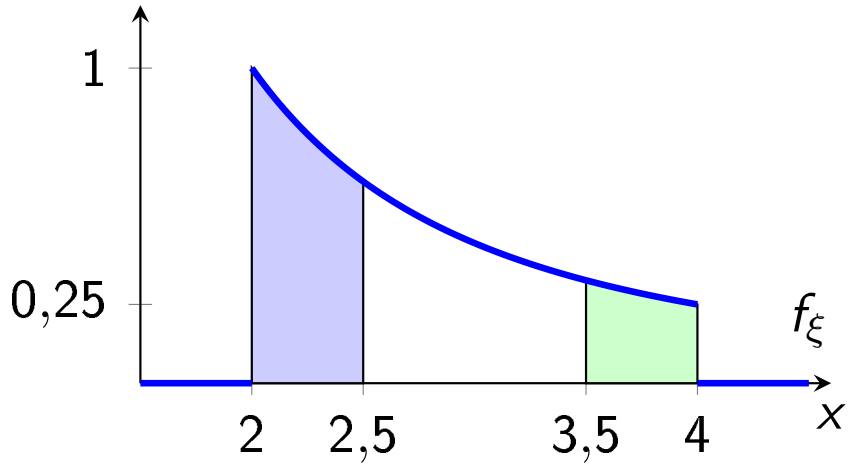
$$= 4\left[\frac{x^{-1}}{-1}\right]_2^4 = 4\left[-\frac{1}{x}\right]_2^4 = 4\left[\left(-\frac{1}{4}\right) - \left(-\frac{1}{2}\right)\right] = 4 \cdot 0{,}25 = 1$$

Feladat: Milyen értékeket vehet fel a ξ változó?

A változó értékkészlete: $R_\xi = [2, 4]$.

Feladat: Mennyi az esélye annak, hogy a ξ változó 2,5-nél kisebb értéket vesz fel? Mennyi a valószínűsége annak, hogy a ξ nagyobb, mint 3,5?

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



$$P(\xi < 2{,}5) = P(2 \le \xi \le 2{,}5) = 4 \int_2^{2{,}5} x^{-2} dx = 4 \left[-\frac{1}{x} \right]_2^{2{,}5}$$

$$= 4 \left[\left(-\frac{1}{2{,}5} \right) - \left(-\frac{1}{2} \right) \right] = 4 \cdot 0{,}1 = 0{,}4,$$

$$P(\xi > 3{,}5) = P(3{,}5 \le \xi \le 4) = 4 \int_{3{,}5}^4 x^{-2} dx = \cdots \approx 0{,}14.$$

Legyen ξ folytonos valószínűségi változó!

- **Móduszok:** Az f_ξ függvény lokális maximumhelyei.
- **Várható érték:** $\mathrm{E}(\xi) = \int_{-\infty}^{\infty} x f_\xi(x) dx$.
  Jelentése: ξ átlagos értéke a teljes populációban.
- **Variancia:** $\mathrm{Var}(\xi) = \int_{-\infty}^{\infty} (x - \mathrm{E}(\xi))^2 f_\xi(x) dx$.
- **Szórás:** $\mathrm{D}(\xi) = \sqrt{\mathrm{Var}(\xi)}$.
  Jelentése: a várható értéktől való átlagos eltérés a populációban.

A varianciára adható egy könnyebben számolható formula is:

$$\mathrm{Var}(\xi) = \int_{-\infty}^{\infty} \big(x - \mathrm{E}(\xi)\big)^2 f_\xi(x) dx$$

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

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

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

Feladat: Mennyi a ξ változó módusza, várható értéke és szórása a jelen feladatban?

A függvénynek csak egy maximumhelye van, az $x = 2$ helyen, ez a módusz.

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

$$= 0 + 4\int_{2}^{4} x^{-1} dx + 0 = 4\Big[\ln x\Big]_2^4 = 4\Big[\ln 4 - \ln 2\Big] = 2{,}77,$$

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

$$\mathrm{Var}(\xi) = \int_{-\infty}^{\infty} x^2 f_\xi(x)dx - \big(\mathrm{E}(\xi)\big)^2 = 8 - (2{,}77)^2 \approx 0{,}33,$$

$$\mathrm{D}(\xi) = \sqrt{\mathrm{Var}(\xi)} = \sqrt{0{,}33} = 0{,}57.$$

Tehát a ξ változó a 2 érték (módusz) közelébe esik a legnagyobb eséllyel. A változó átlagos értéke 2,77, a várható értéktől való átlagos eltérés 0,57.

Egy tetszőleges ξ valószínűségi változó **eloszlásfüggvénye** a következő módon van definiálva: $F_\xi : \mathbb{R} \to [0,1]$, $F_\xi(t) = P(\xi < t)$.

Ha a ξ értéket úgy kapjuk, hogy véletlenszerűen kiválasztunk egy egyedet egy populációból, és megmérünk egy kérdéses mennyiséget, akkor

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

Tetszőleges a és b valós számok esetén teljesülnek az alábbi egyenlőségek:

1. $P(\xi < a) = F_\xi(a)$,
2. $P(\xi \geq a) = 1 - F_\xi(a)$,
3. $P(a \leq \xi < b) = F_\xi(b) - F_\xi(a)$.

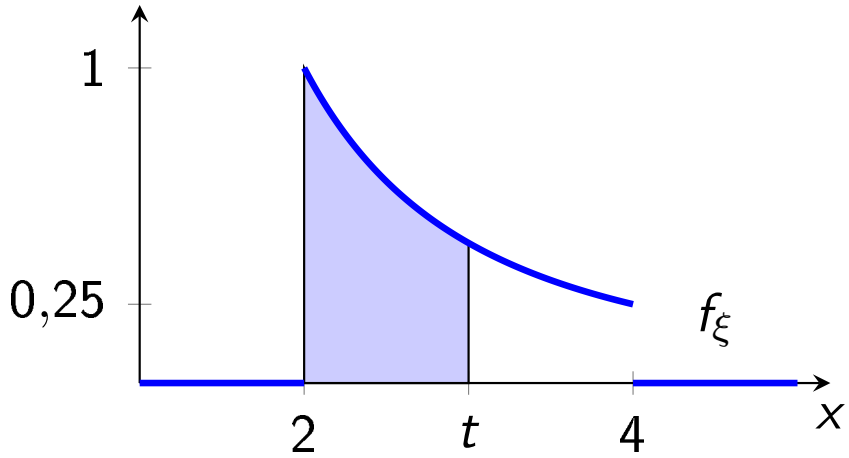
Hogyan kapjuk meg ezeket az azonosságokat?

1. Ez csak az eloszlásfüggvény definíciója.
2. $P(\xi \geq a) = 100\% - P(\xi < a) = 1 - F_\xi(a)$,
3. $P(a \leq \xi < b) = P(\xi < b) - P(\xi < a) = F_\xi(b) - F_\xi(a)$.

Feladat: Hogyan írható fel az eloszlásfüggvény a jelen feladatban?

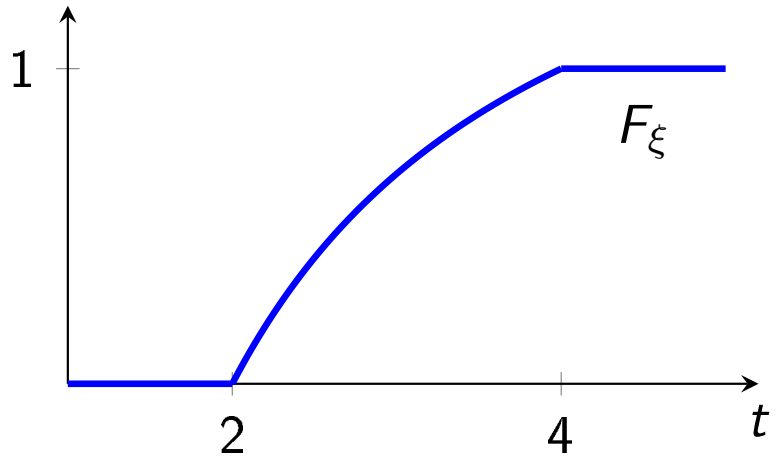
Sűrűségfüggvény:

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



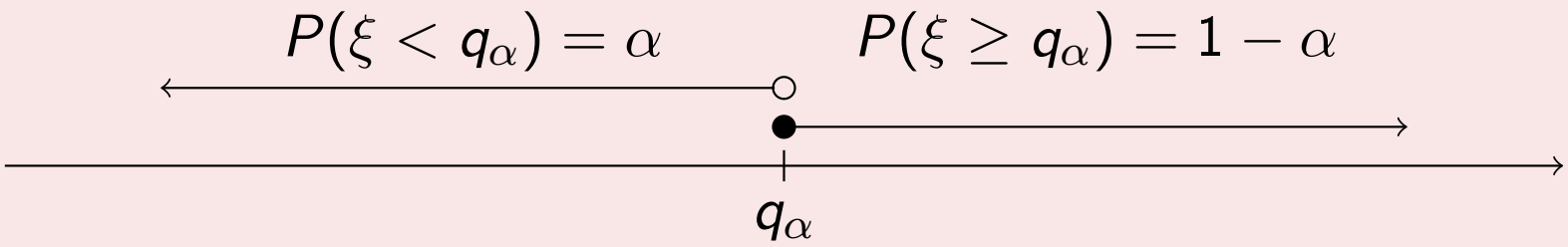
Eloszlásfüggvény: $F_\xi(t) = P(\xi < t)$

- Ha $t < 2$: $F_\xi(t) = 0$.
- Ha $t > 4$: $F_\xi(t) = 1$.
- Ha $2 \le t \le 4$:



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

Legyen ξ tetszőleges valószínűségi változó, és legyen $\alpha \in (0,1)$. A ξ változó **α-kvantilise** egy olyan q_α valós szám, melyre $P(\xi < q_\alpha) = \alpha$.



A kvantilis jelentése: a vizsgált ξ mennyiség a teljes populáción belül

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

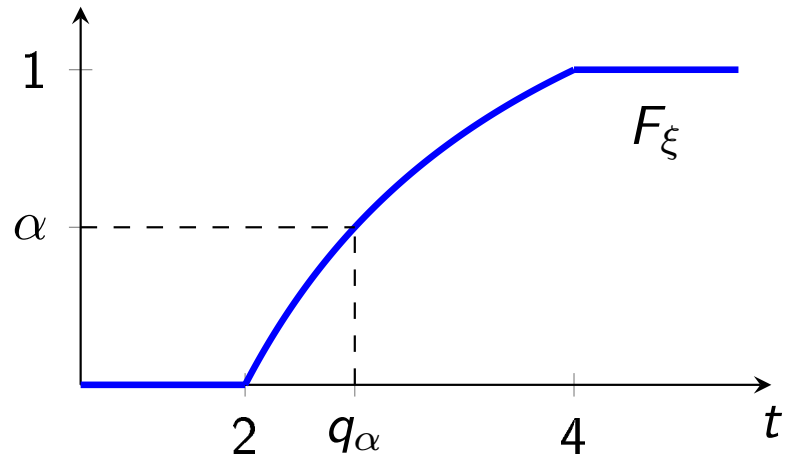
Megjegyzés: Az α-kvantilis nem mindig létezik, és ha létezik, akkor nem feltétlenül egyértelmű.

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 a mediánt valamint az alsó és a felső kvartilist a jelen feladatban.

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

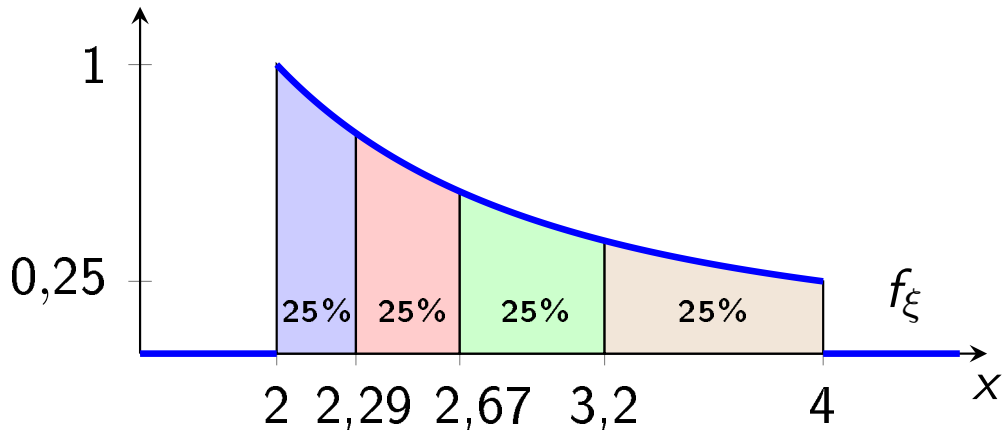


Tetszőleges $0 < \alpha < 1$ szám esetén

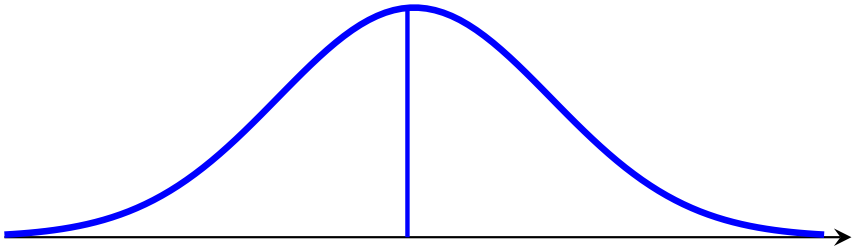
$$\alpha = P(\xi < q_\alpha) = F_\xi(q_\alpha) = 2 - 4/q_\alpha.$$

Ebből következik, hogy $q_\alpha = 4/(2-\alpha)$. A kérdéses értékeket az alábbi táblázat tartalmazza. A medián és a két kvartilis négy részre bontja fel a változó értékkészletét, és a változó mindegyik részbe 25% eséllyel esik bele.

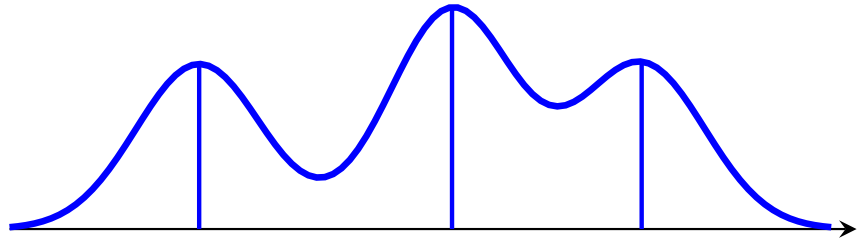
α	25%	50%	75%
q_α	2,29	2,67	3,2



A módoszok száma alapján kétfajta sűrűségfüggvényt különböztetünk meg: **egymóduszú** és **többmóduszú** sűrűségfüggvényt.



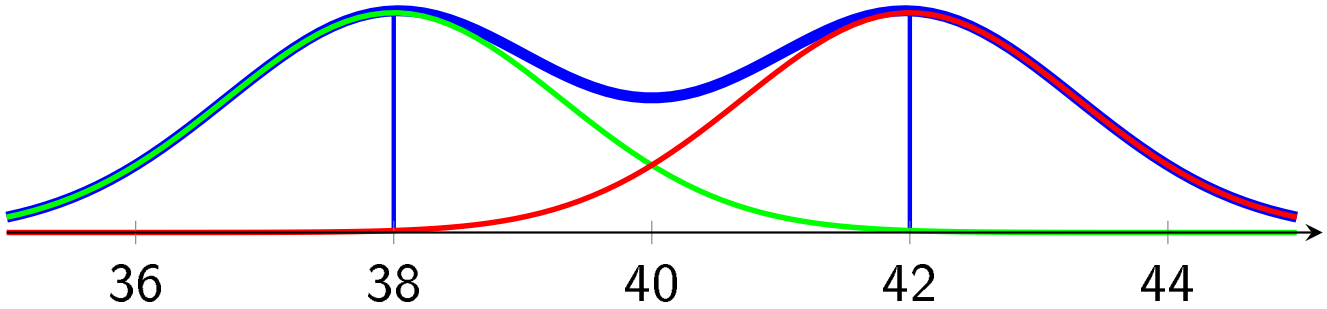
egymóduszú



többmóduszú

A több módusz gyakran arra utal, hogy a populációt több részpopulációra lehet felbontani, melyeken belül a vizsgált ξ mennyiség már egymóduszú.
Példa: lábméret eloszlása a felnőtt népességen belül.

- 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.

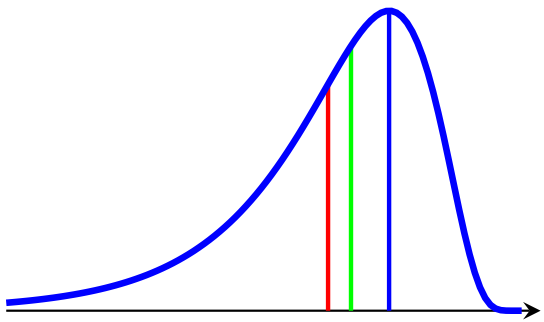


Tegyük fel, hogy a sűrűségfüggvénynek csak egyetlen módusza van. A módusz, a medián és a várható érték jelentése:

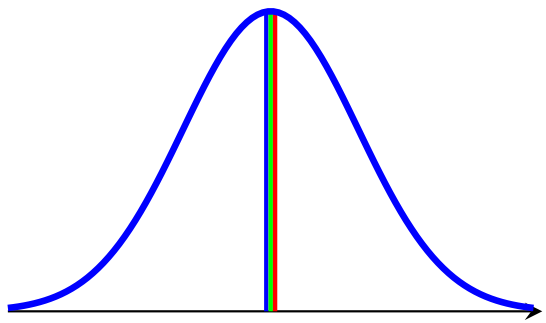
- Módusz: A változó ezen érték közelébe esik a legnagyobb eséllyel.
- Medián: A változó „középső” értéke.
- Várható érték: A változó átlagos értéke.

Ha a sűrűsgfüggvény szimmetrikus, akkor a három mennyiség megegyezik.
Ha a sűrűségfüggvény nem szimmetrikus, akkor jellemzően(!):

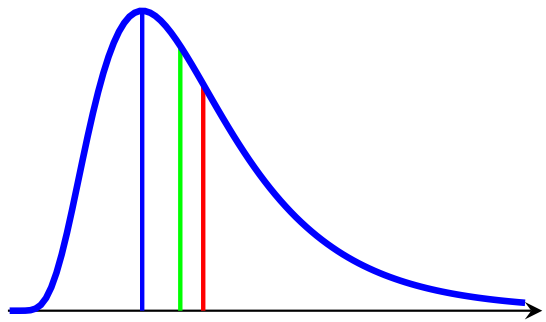
- Balra ferde sűrűségfüggvény esetén: várható érték $<$ medián $<$ módusz
- Jobbra ferde sűrűségfüggvény esetén: módusz$<$medián$<$várható érték



balra ferde függvény



szimmetrikus függvény

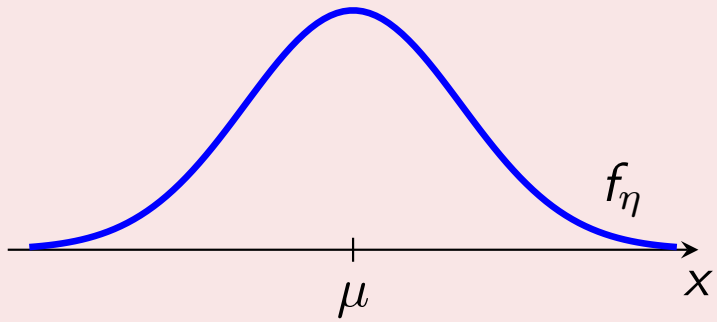


jobbra ferde függvény

A normális eloszlás

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

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



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

A normális eloszlás fontosabb alkalmazásai:

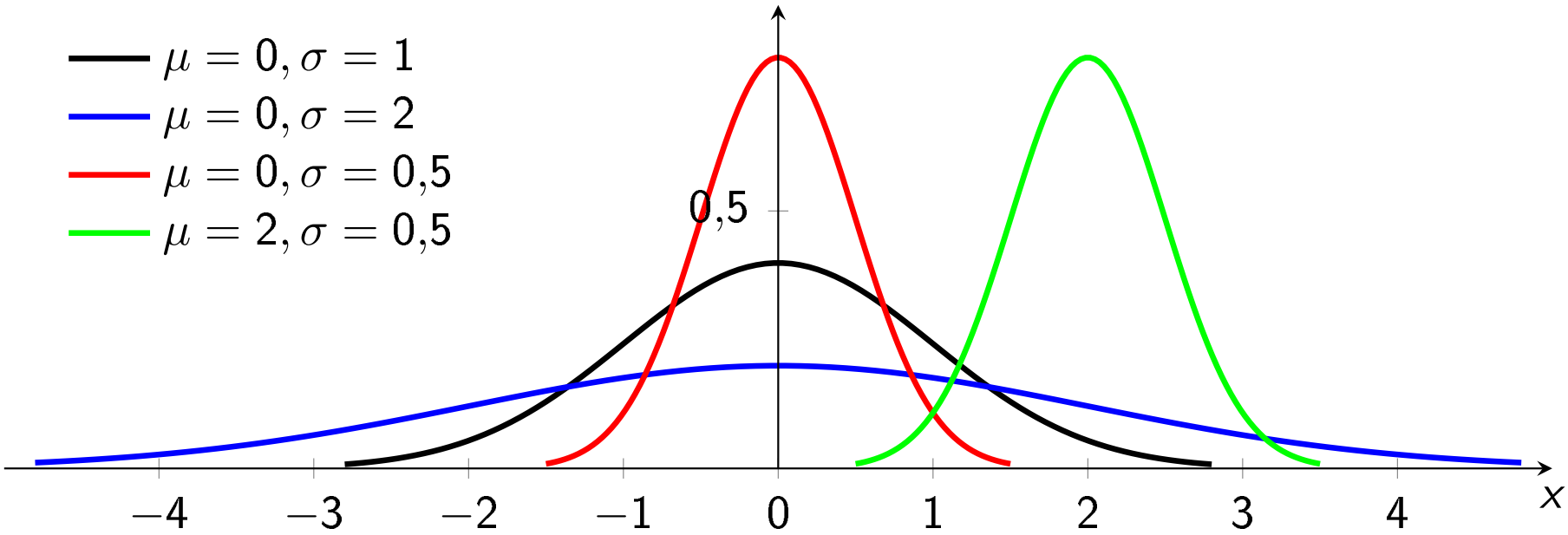
- Mérési hibák modellezése: mért érték = igazi érték + mérési hiba, ahol a mérési hiba normális eloszlást követ.
- Élettudományok: számos mennyiség (testmagasság, vérnyomás, IQ) normális vagy a normálisból származtatott eloszlást követ.

A normális eloszlás tulajdonságai:

- $f_\eta(x) > 0$ minden x valós számra, ezért $R_\eta = \mathbb{R}$.
- $\mathsf{E}(\eta) = \mu$ és $\mathsf{D}(\eta) = \sigma$.
- A sűrűségfüggvény szimmetrikus, ezért módusz = medián = $\mathsf{E}(\xi) = \mu$.

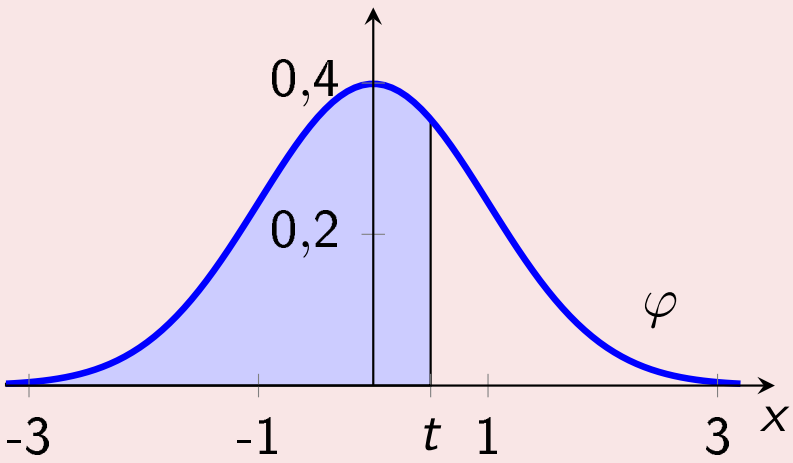
Hogyan hat a két paraméter a sűrűségfüggvényre:

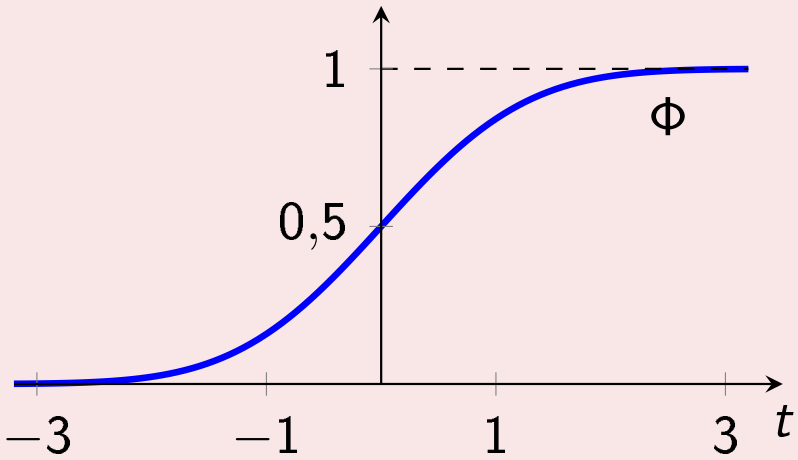
- σ: a sűrűségfüggvény alakját határozza meg,
- μ: eltolás, a sűrűségfüggvény szimmetriatengelye.



A $\mu = 0$ és $\sigma = 1$ paraméteres normális eloszlást **standard normális** eloszlásnak nevezzük. Jelölésben: $\eta_{0,1}$. Sűrűség- és eloszlásfüggvénye:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad \Phi(t) = P(\eta_{0,1} < t) = \int_{-\infty}^{t} \varphi(x)dx.$$





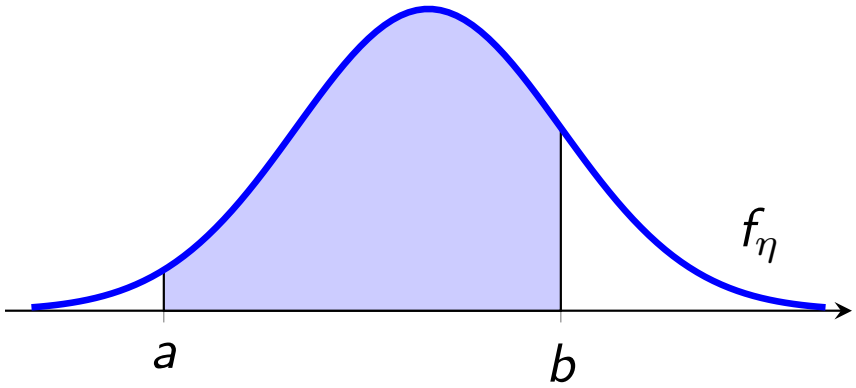
A Φ függvény tulajdonságai:

$$\Phi(t) \begin{cases} < 0{,}5, & \text{ha } t < 0, \\ = 0{,}5, & \text{ha } t = 0, \\ > 0{,}5, & \text{ha } t > 0, \end{cases} \qquad \Phi(-t) = 1 - \Phi(t).$$

Ha η normális eloszlású, akkor tetszőleges a és b valós számokra:

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

Probléma: ezt az integrált nem tudjuk papíron kiszámolni.



Legyen η normális eloszlású μ várható értékkel és σ szórással. Ekkor az $(\eta - \mu)/\sigma$ valószínűségi változót η **standardizáltjának** nevezzük. Megmutatható, hogy ez az új változó standard normális eloszlás követ.

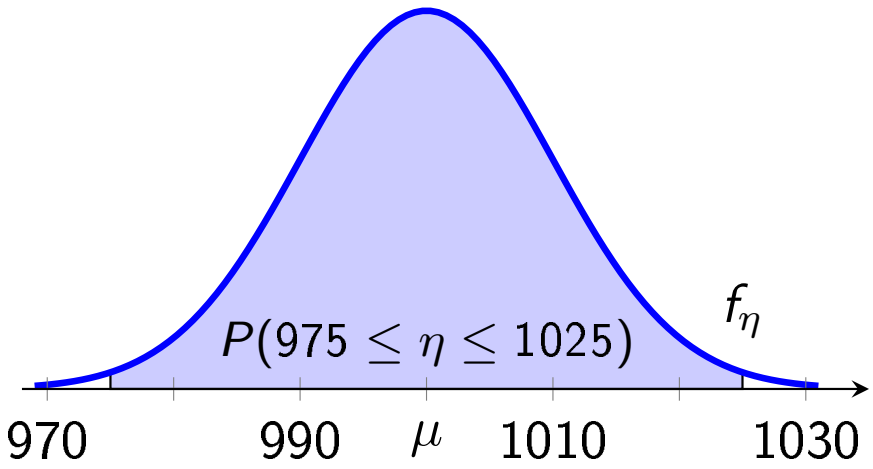
Ha η normális eloszlású változó, akkor standardizálással:

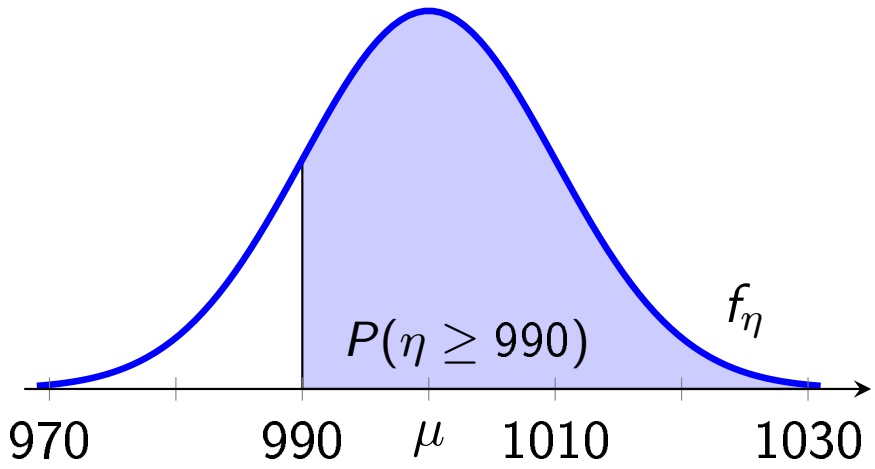
$$P(a \leq \eta < b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{\eta-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Feladat. Egy tejgyárban az 1 literes dobozos tej csomagolását automata töltőberendezés végzi, és a dobozokba töltött mennyiség egy normális eloszlású valószínűségi változó, melynek várható értéke a névleges tartalom és szórása $\sigma = 10$ ml. Véletlenszerűen kiválasztunk egy dobozt.

- Mennyi annak a valószínűsége, hogy a doboz legfeljebb 2,5%-kal tér el a névleges tartalomtól?
- Mennyi annak az esélye, hogy a doboz legalább 990 ml tejet tartamaz?

Legyen η a kiválasztott dobozban található mennyiség. Az η változó normális eloszlású $\mu = 1000$ ml várható értékkel és $\sigma = 10$ ml szórással. A következő valószínűségekre (=területekre) vagyunk kíváncsiak, de ezek most nem számolhatóak ki integrálással:





Az első valószínűség standardizálással határozható meg:

$$P\big(975 \le \eta \le 1025\big) = P\left(\frac{975-1000}{10} \le \frac{\eta-\mu}{\sigma} \le \frac{1025-1000}{10}\right)$$
$$= P\big(-2{,}5 \le \eta_{0,1} \le 2{,}5\big) = \Phi(2{,}5) - \Phi(-2{,}5) = 0{,}9938 - 0{,}0062 = 0{,}9876,$$

Ez azt jelenti, hogy a tejesdobozok 98,76%-a tartalmaz 975 ml és 1025 ml közötti tejet. Itt felhasználtuk azt, hogy

$$\Phi(-2{,}5) = 1 - \Phi(2{,}5) = 1 - 0{,}9938 = 0{,}0062.$$

A második valószínűség az első mintájára:

$$P(\eta \ge 990) = P\left(\frac{\eta-\mu}{\sigma} \ge \frac{990-1000}{10}\right) = P\big(\eta_{0,1} \ge -1\big)$$
$$= 1 - P\big(\eta_{0,1} < -1\big) = 1 - \Phi(-1) = 1 - \big[1 - \Phi(1)\big]$$
$$= 1 - \big[1 - 0{,}84\big] = 0{,}84.$$

Feladat. Adjunk meg egy olyan $[a, b]$ intervallumot, amire teljesül, hogy a tejesdobozok 95%-a ebbe az intervallumba esik: $P(a \le \eta \le b) = 0{,}95$.

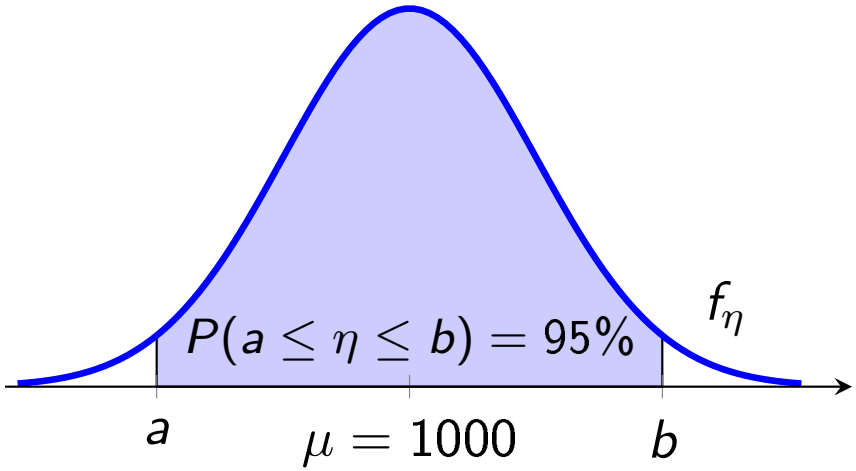
Az intervallumot $[\mu - c\sigma, \mu + c\sigma]$ alakban fogjuk keresni. Ismét csak standardizálással:

$$0{,}95 = P\big(\mu - c\sigma \le \eta \le \mu + c\sigma\big) = P\left(-c \le \frac{\eta - \mu}{\sigma} \le c \right)$$

$$= P\big(-c \le \eta_{0,1} \le c\big) = \Phi(c) - \Phi(-c) = \Phi(c) - \big[1 - \Phi(c)\big] = 2\Phi(c) - 1\,.$$

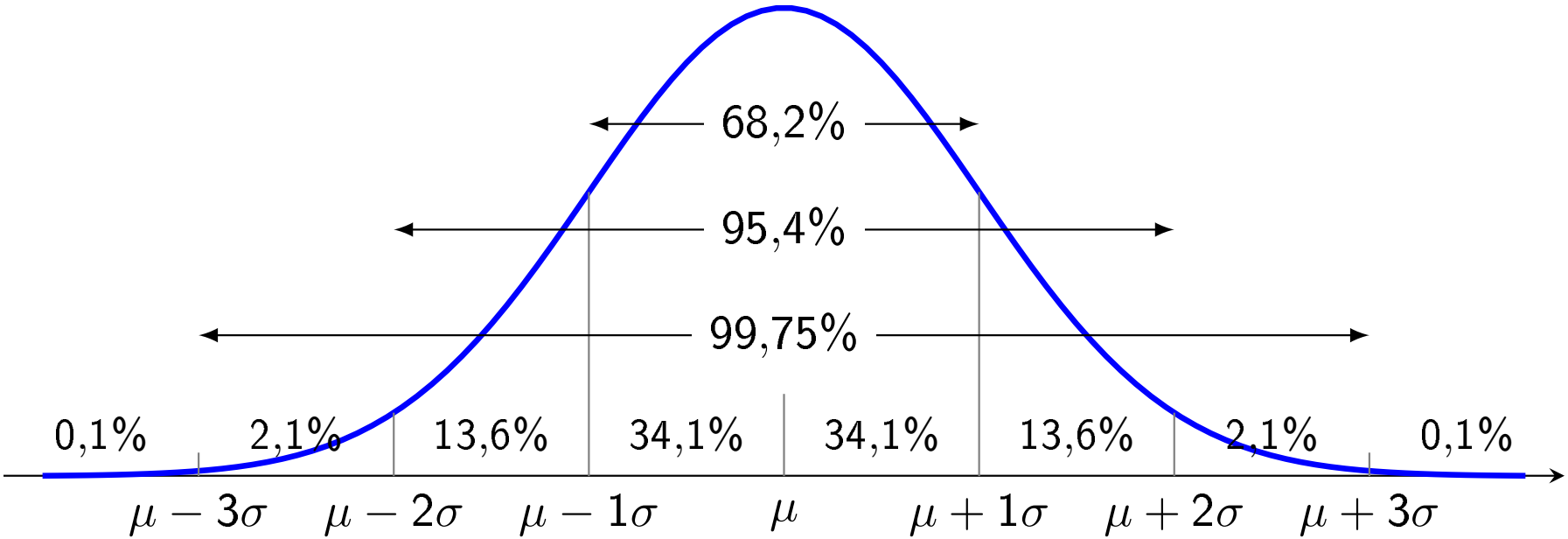
Ebből azt kapjuk, hogy $\Phi(c) = 0{,}975 = \Phi(1{,}96)$, tehát $c = 1{,}96$.

Tehát a kérdéses intervallum: $[\mu - 1{,}96\sigma, \mu + 1{,}96\sigma] = [980{,}4, 1019{,}6]$.



Közelítő intervallum a 2σ-szabállyal: $[\mu - 2\sigma, \mu + 2\sigma] = [980, 1020]$.

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



Legyen η normális eloszlású változó. Ekkor:

- 1σ-szabály: $P\big(\mu - \sigma \leq \eta \leq \mu + \sigma\big) \approx 68\%$,
- 2σ-szabály: $P\big(\mu - 2\sigma \leq \eta \leq \mu + 2\sigma\big) \approx 95\%$,
- 3σ-szabály: $P\big(\mu - 3\sigma \leq \eta \leq \mu + 3\sigma\big) \approx 99{,}75\%$.

Statisztikai alapfogalmak

Legyen adva egy populáció, és tekintsünk egy mennyiséget az egyedeken (életkor, testtömeg, utódok száma, stb.). Véletlenszerűen kiválasztunk egy egyedet, és ξ jelöli a vizsgált mennyiséget a kiválasztott egyed esetében.

Valószínűségszámítás: Ha ismerjük a ξ változó valószínűségeloszlását vagy sűrűségfüggvényét, akkor ki tudjuk számolni a következő értékeket:

- $\mathrm{E}(\xi) =$ a vizsgált mennyiség átlagos értéke a populáción belül,
- $\mathrm{D}(\xi) =$ a vizsgált mennyiség szórása a populáción belül,
- $P(a \leq \xi \leq b) =$ arány a teljes populáción belül.

Matematikai statisztika: Nem ismerjük a ξ változó valószínűségeloszlását vagy sűrűségfüggvényét, ezért nem tudjuk kiszámolni ezeket az értékeket. Ehelyett megfigyeléseket végzünk a ξ változóra, és a kapott minta alapján vonunk le következtetéseket. Célok:

- Becsléselmélet: Adjunk becslést a várható értékre, szórásra, stb.
- Hipotézisvizsgálat: Adott egy állítás a ξ mennyiséggel kapcsolatban. (Pl: $\mathrm{E}(\xi) = 2$.) Döntsük el, hogy ez az állítás igaz vagy hamis.

Statisztikai alapfogalmak:

- **Háttérváltozó:** Az a ξ valószínűségi változó, melyet vizsgálunk.
- **Statisztikai minta (statistical sample):** $\xi_1, \ldots, \xi_n$ valószínűségi változók, független megfigyelések a ξ változóra. Jellemzően: véletlenszerűen kiválasztunk n egyedet a teljes populációból.
- **Mintarealizáció (realization, observations):** a $\xi_1, \ldots, \xi_n$ változók megfigyelés során kapott konkrét értékei.
- **Mintaméret (sample size):** a megfigyelések száma (n).

Hogyan is történik ez a gyakorlatban:

- Kíváncsiak vagyunk egy ξ mennyiség eloszlására egy populációban.
- Megtervezzük a mintavételezést és a statisztikai kiértékelést. Ezen a ponton a mintaelemek valószínűségi változók: még nem tudjuk, hogy mik lesznek a megfigyelt értékek.
- Elvégezzük a mintavételezést, ezzel megkapjuk a realizációt, tehát a mintaelemek konkrét értékeit.
- Elvégezzük a statisztikai elemzést a realizáción. (Mi a továbbiakban nagyrészt ezzel a lépéssel foglalkozunk.)

Leíró statisztikák (descriptive statistics)

Egy ξ háttérváltozó várható értékét, varianciáját és szórását a következő módon becsülhetjük meg egy $\xi_1, \ldots, \xi_n$ minta alapján:

- **Empirikus várható érték, mintaátlag (sample mean)**:

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

- **Empirikus variancia (sample variance)**:

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

- **Empirikus szórás (standard deviation)**: $\mathsf{D}_n(\xi) = \sqrt{\mathsf{Var}_n(\xi)} \approx \mathsf{D}(\xi)$

Miért így van definiálva az empirikus variancia?

$$\mathsf{Var}(\xi) = \mathsf{E}\Big(\big[\xi - \mathsf{E}(\xi)\big]^2\Big) \approx \frac{\big[\xi_1 - \mathsf{E}(\xi)\big]^2 + \cdots + \big[\xi_n - \mathsf{E}(\xi)\big]^2}{n} \approx \mathsf{Var}_n(\xi)$$

Az előző oldalon felsorolt becslések **erősen konzisztensek**, tehát

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

Ez azt jelenti, hogy 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: kicsit megnöveljük ezeket az értékeket.

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

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

Nagy mintaméret esetén a korrigálás csak kis mértékben változtat a becsléseken. Kis mintaméret esetén viszont jelentős a növekedés.

A korrigálás során kapott becslések kis n esetén pontosabban, mint az eredeti becslések, de az erős konzisztencia is megmarad:

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

Feladat: A kar férfi hallgatóinak testmagasságát vizsgáljuk, jelölje ξ egy véletlenszerűen kiválasztott férfi hallgató magasságát. Megfigyeléseket végzünk a változóra, a következő realizációt kapjuk: 180, 175, 188, 168, 173, 183. Adjunk becslést a testmagasság átlagára és szórására.

$$\overline{\xi} = \mathsf{E}_6(\xi) = \frac{180 + 175 + 188 + 168 + 173 + 183}{6} = 177{,}8 \approx \mathsf{E}(\xi),$$

$$\mathsf{Var}_6(\xi) = \frac{(180 - 177{,}8)^2 + \cdots + (183 - 177{,}8)^2}{6} = 43{,}81 \approx \mathsf{Var}(\xi),$$

$$\mathsf{D}_6(\xi) = \sqrt{43{,}81} = 6{,}62 \approx \mathsf{D}(\xi).$$

A kis mintaméret miatt ($n = 6$) a szórást jobb a korrigált szórással becsülni:

$$\mathsf{Var}_6^*(\xi) = \frac{6}{5}\,43{,}81 = 52{,}57, \qquad \mathsf{D}_6^*(\xi) = \sqrt{52{,}57} = 7{,}25 \approx \mathsf{D}(\xi).$$

Foglaljuk össze, hogy mit kaptunk:

- átlagos testmagasság a populációban $= \mathsf{E}(\xi) \approx 177{,}8$,
- a testmagasság szórása a populációban $= \mathsf{D}(\xi) \approx 7{,}25$.

Ezt a két értéket publikációkban így szokták közölni: $177{,}8 \pm 7{,}25$ cm.

Ha van egy mintarealizációnk, akkor a mintaátlag egy becslés az ismeretlen várható értékre. Ha egy másik mintavételből származó másik realizációval dolgozunk, akkor egy másik becslést kapunk ugyanarra a várható értékre. A mintaátlag egy valószínűségi változó, ami a realizációtól függ.

Tétel. A mintaátlag várható értéke és szórása:

$$\mathrm{E}(\bar{\xi}) = \mathrm{E}(\xi) \qquad \text{és} \qquad \mathrm{D}(\bar{\xi}) = \mathrm{D}(\xi)/\sqrt{n}.$$

Értelmezzük a kapott eredményeket:

- Ha minden lehetséges realizációból kiszámolnánk a mintaátlagot, akkor átlagban a várható értéket kapnánk. Ez egy jó tulajdonság, amit **tozítatlanságnak** nevezünk.
- Ha minden lehetséges realizációból kiszámolnánk a mintaátlagot, akkor ezek az értékek átlagosan $\mathrm{D}(\xi)/\sqrt{n}$ mértékben térnek el a becsülni kívánt $\mathrm{E}(\xi)$ várható értéktől. Tehát átlagosan ennyit tévedünk a becslés során.
  Vegyük észre: $\mathrm{D}(\xi)/\sqrt{n} \to 0$, amint $n \to \infty$. Ez azt jelenti, hogy egyre nagyobb minta alapján egyre kisebb hibával tudunk becsülni.

Standard hiba (standard error of the mean, s.e.m.): $\mathrm{SE} = \mathrm{D}_n^*(\xi)/\sqrt{n}$.
Jelentése: a $\mathrm{D}(\overline{\xi})$ szórás becslése a minta alapján.

- Ha a standard hiba kicsi, akkor a mintaátlag minden realizáció esetén pontos becslése lesz a várható értéknek.
- Ha a standard hiba nagy, akkor vannak olyan realizációk, melyekre a mintaátlag pontatlan becslést ad a várható értékre.

Feladat: Határozzuk meg a standard hibát a jelen feladatban.

Amit tudunk: $n = 6$, $\mathrm{E}_6(\xi) = 177{,}8$, $\mathrm{D}_6^*(\xi) = 7{,}25$.

Ekkor: $\mathrm{SE} = 7{,}25/\sqrt{6} = 2{,}96$.

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, várhatóan ennyi a becslés hibája.
- Ezt a két értéket így szokták közölni: $177{,}8 \pm 2{,}96$ (SE).

A ξ valószínűségi változó α-kvantilise egy olyan q_α valós szám, melyre $P(\xi < q_\alpha) = \alpha$. Jelentése: a populáción belül a vizsgált ξ mennyiség az egyedek α hányadánál kisebb, mint q_α.

Az α-kvantilis becslésére egy $\xi_1, \ldots, \xi_n$ statisztika minta alapján több módszer is létezik. Mi most nem adunk precíz matematikai formulát a becslésre, csak a becslés alapötletét ismertetjük.

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

Például: **empirikus medián:**

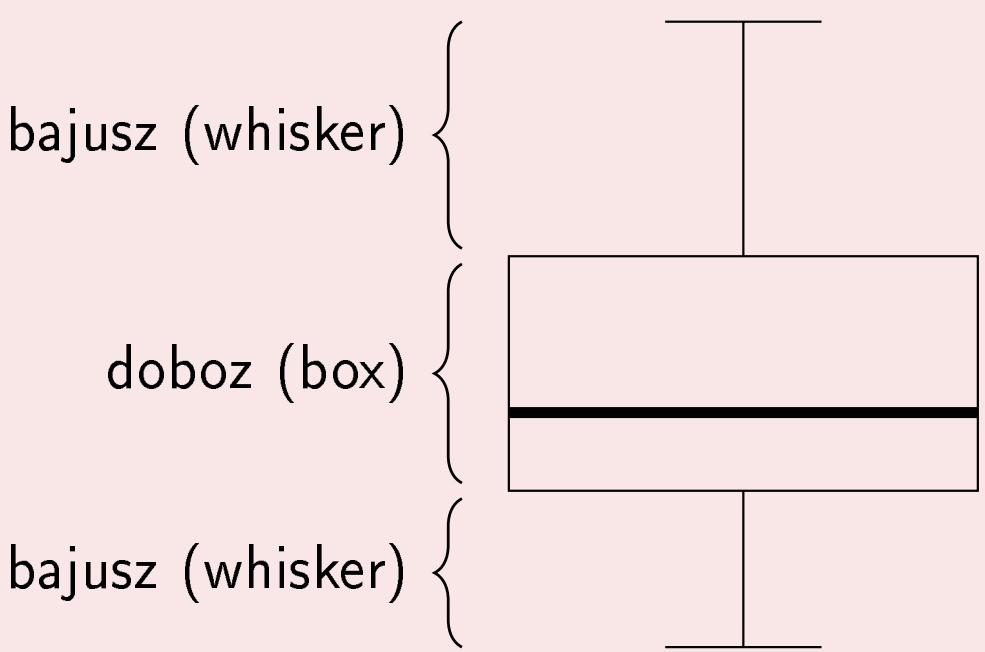
$$\hat{q}_{50\%} = \begin{cases} \text{a középső mintaelem}, & \text{ha } n \text{ páratlan}, \\ \text{a két középső átlaga}, & \text{ha } n \text{ páros}. \end{cases}$$

Feladat: Adjunk becslést a testmagasság elméleti mediánjára a kar férfi hallgatóinak populációjában.

A rendezett minta: 168, 173, 175, 180, 183, 188. A becslés:

$$q_{50\%} \approx \hat{q}_{50\%} = \text{két középső mintaelem átlaga} = 177{,}5.$$

A **boxplot** egy olyan grafikon, mely az alábbi statisztikai mutatószámokat ábrázolja egyszerű formában:



max = legnagyobb mintaelem

$\hat{q}_{75\%}$ = felső kvartilis

$\hat{q}_{50\%}$ = medián

$\hat{q}_{25\%}$ = alsó kvartilis

min = legkisebb mintaelem

További mutatószámok:

- **Terjedelem (range)** = max − min = a boxplot magassága,
- **Interkvartilis távolság (interquartile range):**

  IQR = felső kvartilis − alsó kvartilis = a doboz magassága.

Konfidencia intervallumok (confidence intervals)

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

- **Pontbecslés:** Az ismeretlen mennyiséget egyetlen számmal becsüljük meg, és reménykedünk benne, hogy nem tévedünk nagyot.
- **Intervallumbecslés:** Egy intervallumot adunk meg, mely nagy megbízhatósággal tartalmazza a kérdéses mennyiséget.

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

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

- A megbízhatóság általában 90%, 95% vagy 99% szokott lenni, a biostatisztikában tipikusan a 95%-ot használják.
- A konfidencia intervallum hasonló módon definiálható tetszőleges más mutatószámra is (szórás, variancia, medián, stb.)

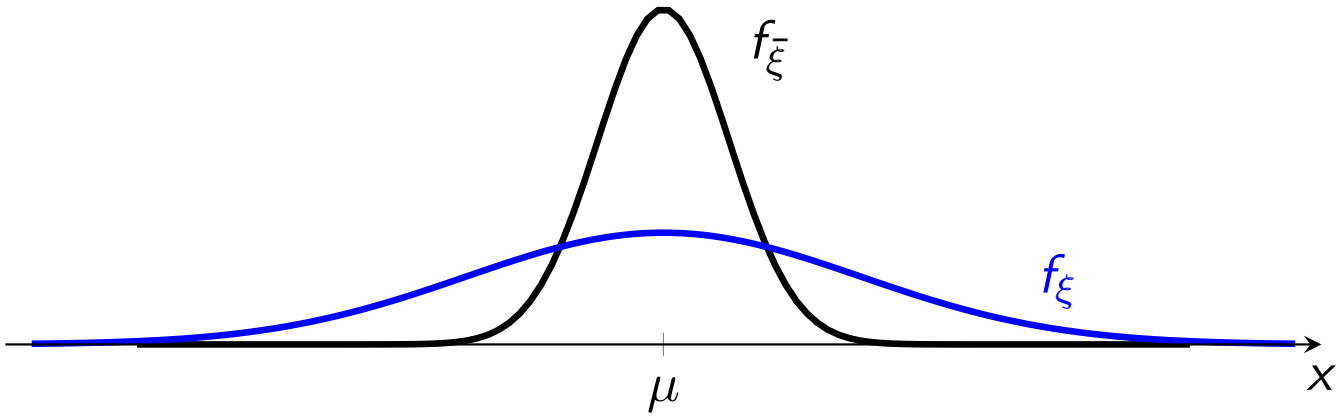
Feladat: Legyen ξ normális eloszlású valószínűségi változó ismeretlen μ várható értékkel és ismert σ szórással. Egy $\xi_1, \ldots, \xi_n$ statisztikai minta alapján adjunk konfidencia intervallumot a várható értékre.

Tétel. Ha a ξ háttérváltozó normális eloszlású, akkor a $\xi_1 + \cdots + \xi_n$ összeg és a $\overline{\xi} = (\xi_1 + \cdots + \xi_n)/n$ mintaátlag is normális eloszlású változó.

Jelölje $\mu_{\overline{\xi}}$ és $\sigma_{\overline{\xi}}$ a mintaátlag várható értékét és szórását. Ekkor

- $\mu_{\overline{\xi}} = \mathsf{E}(\overline{\xi}) = \mathsf{E}(\xi) = \mu$,
- $\sigma_{\overline{\xi}} = \mathsf{D}(\overline{\xi}) = \mathsf{D}(\xi)/\sqrt{n} = \sigma/\sqrt{n}$.

Az alábbi ábrán a ξ háttérváltozó és a $\overline{\xi}$ mintaátlag sűrűségfüggvénye látható:



Először megadunk egy olyan intervallumot, mely $1-\alpha$ valószínűséggel tartalmazza a $\overline{\xi}$ változót. Az intervallumot most is $[\mu_{\overline{\xi}} - c\sigma_{\overline{\xi}}, \mu_{\overline{\xi}} + c\sigma_{\overline{\xi}}]$ alakban keressük. Standardizálással:

$$1-\alpha = P\big(\mu_{\overline{\xi}} - c\sigma_{\overline{\xi}} \le \overline{\xi} \le \mu_{\overline{\xi}} + c\sigma_{\overline{\xi}}\big) = P\left(-c \le \frac{\overline{\xi} - \mu_{\overline{\xi}}}{\sigma_{\overline{\xi}}} \le c\right)$$

$$= P\big(-c \le \eta_{0,1} \le c\big) = \Phi(c) - \Phi(-c) = \Phi(c) - \big[1 - \Phi(c)\big] = 2\Phi(c) - 1$$

Tehát $\Phi(c) = 1 - \alpha/2$, amiből $c = \Phi^{-1}(1-\alpha/2)$. Ezt az értéket ki tudjuk keresni a táblázatból tetszőleges $\alpha \in (0,1)$ esetén.

A fenti nagy formulát a következő módon tudjuk továbbalakítani:

$$1-\alpha = P\big(\mu_{\overline{\xi}} - c\sigma_{\overline{\xi}} \le \overline{\xi} \le \mu_{\overline{\xi}} + c\sigma_{\overline{\xi}}\big) = P\big(-\overline{\xi} - c\sigma_{\overline{\xi}} \le -\mu_{\overline{\xi}} \le -\overline{\xi} + c\sigma_{\overline{\xi}}\big)$$

$$= P\big(\overline{\xi} + c\sigma_{\overline{\xi}} \ge \mu_{\overline{\xi}} \ge \overline{\xi} - c\sigma_{\overline{\xi}}\big) = P\left(\overline{\xi} + c\frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{\xi} - c\frac{\sigma}{\sqrt{n}}\right)$$

De hát ez éppen egy konfidencia intervallum az $\mathrm{E}(\xi) = \mu$ ismeretlen várható értékre:

$$1-\alpha = P\left(\mathrm{E}(\xi) \in \left[\overline{\xi} - c\frac{\sigma}{\sqrt{n}}\,,\, \overline{\xi} + c\frac{\sigma}{\sqrt{n}}\right]\right)$$

Legyen ξ normális eloszlású változó ismert σ szórással. Ekkor a változó várható értékére a következő formában adható $1-\alpha$ megbízhatóságú konfidencia intervallum:

$$\left[\overline{\xi} - c\frac{\sigma}{\sqrt{n}}\,,\, \overline{\xi} + c\frac{\sigma}{\sqrt{n}}\right], \qquad c = \Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

Feladat: Tegyük fel, hogy a kar férfi hallgatóinak testmagassága normális eloszlású $\sigma = 7$ cm szórással. Adjunk 95% megbízhatóságú konfidencia intervallumot a testmagasság várható értékére (az átlagos testmagasságra).

A minta: 180, 175, 188, 168, 173, 183.

A mintaméret és a mintaátlag: $n = 6$, $\overline{\xi} = 177{,}8$.

Most $\alpha = 5\% = 0{,}05$, tehát $c = \Phi^{-1}(0{,}975) = 1{,}96$.

Az intervallum:

$$\left[177{,}8 - 1{,}96\frac{7}{\sqrt{6}}\,,\, 177{,}8 + 1{,}96\frac{7}{\sqrt{6}}\right] = \left[172{,}2\,,\, 183{,}4\right].$$

De mi ennek az intervallumnak a jelentése?

Probléma: a ξ háttérváltozó igazi szórását sosem tudjuk.

Megoldás: helyettesítsük a szórást a becslésével: $\sigma \approx \mathrm{D}_n^*(\xi)$. Ennek az az ára, hogy a c értéket a **Student-eloszlás** táblázatából kell kikeresni.

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

$$\left[\overline{\xi} - c\frac{\mathrm{D}_n^*(\xi)}{\sqrt{n}}\,,\, \overline{\xi} + c\frac{\mathrm{D}_n^*(\xi)}{\sqrt{n}}\right] = \left[\overline{\xi} - c\,\mathrm{SE}\,,\, \overline{\xi} + c\,\mathrm{SE}\right], \quad c = \Phi_{n-1}^{-1}\left(1-\frac{\alpha}{2}\right).$$

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

Feladat: Adjunk 95% megbízhatóságú konfidencia intervallumot a kar férfi hallgatóinak átlagos testmagasságra ismeretlen szórás esetén!

Most: $n = 6$, $\overline{\xi} = 177{,}8$, $\mathrm{D}_6^*(\xi) = 7{,}25$, $c = \Phi_5^{-1}(0{,}975) = 2{,}57$.

Az intervallum:

$$\left[177{,}8 - 2{,}57\frac{7{,}25}{\sqrt{6}}\,,\, 177{,}8 + 2{,}57\frac{7{,}25}{\sqrt{6}}\right] = \left[170{,}2\,,\, 185{,}4\right].$$

Kérdés: Hogyan értelmezhető a kapott eredmény?

A mintavételezés során a véletlen sok különböző mintarealizációt sorsolhat ki nekünk. Ezek két csoportba sorolhatóak:

- „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: Ismeretlen szórás esetén miért kaptunk bővebb intervallumot?

Nem volt ismert a szórás, ami további bizonytalanságot jelentett. Emiatt egy kis „ráhagyással" kellett számolnunk: nagyobb lett a c érték, ami bővebb intervallumot eredményezett.

Kérdés: Hogyan értelmezhető az intervallum: $\left[\overline{\xi} - c\,\mathrm{SE}\,,\, \overline{\xi} + c\,\mathrm{SE}\right]$?

A konfidencia intervallum felírásakor a $\overline{\xi}$ mintaátlagból indulunk ki, ugyanis ez egy jó becslése a várható értéknek. Erre a becslésre mérjük fel a $c\,\mathrm{SE}$ szorzatot két oldalra. Ebben a szorzatban két dolog jelenik meg:

- A standard hiba számszerűsíti, hogy mennyire jól becsli a mintaátlag a várható értéket, mekkora „ráhagyással" kell számolni a konfidencia intervallum felírásakor.
- A c értékben a megbízhatóság jelenik meg:

  nagyobb megbízhatóság $\Rightarrow$ magasabb c érték $\Rightarrow$ bővebb intervallum.

Kérdés: Miért nem számolunk 99,99%-os megbízhatósággal?

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

A 95%-os választás jó egyensúlyt jelent a két cél (magas megbízhatóság és szűk konfidencia intervallum) között. A megbízhatóság további növelése drasztikusan szélesebb intervallumot eredményez. Csak akkor dolgozunk magasabb megbízhatósággal, ha a standard hiba alacsony.

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

A levezetésnek a következő tétel volt az alapja: ha a ξ háttérváltozó normális eloszlású, akkor a $\overline{\xi}$ mintaátlag is normális eloszlású változó.

Tétel. Ha a minta nem normális eloszlásból jön, de a mintaméret elég nagy, akkor a $\overline{\xi}$ mintaátlag közel normális eloszlású.

A tételnek az a következménye, hogy a kapott intervallum egy közelítő konfidencia intervallum a várható értékre tetszőleges ξ háttérváltozó esetén:

$$P\Big(\mathrm{E}(\xi) \in \big[\overline{\xi} - c\,\mathrm{SE}\,,\, \overline{\xi} + c\,\mathrm{SE}\big]\Big) \approx 1 - \alpha\,.$$

Kérdés: Mit jelent ebben az esetben az „elég 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.
- **Null-hipotézis (H_0, null hypothesis):** Egy állítás a ξ változóra.
- **Alternatív hipotézis (H_A, alternative hypothesis):** Egy másik állítás a ξ változóra.
- A hipotézisvizsgálat célja: A két hipotézis közül valamelyik igaz. Döntsük el a statisztikai minta alapján, hogy H_0 vagy H_A igaz.

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

A továbbiakban a kurzuson az alternatív hipotézis mindig a nullhipotézis tagadása lesz. Azt kell eldönteni, hogy H_0 igaz vagy nem. 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 statisztikai minta alapján kiszámoljuk a **próbastatisztika (test statistic)** értékét: s_n.
- Meghatározzuk a **kritikus értéket (critival value)**: c.
- Ha $|s_n| \leq c$, akkor **elfogadjuk (accept)** a nullhipotézist.
  Ha $|s_n| > c$, akkor **elvetjük (reject)** a nullhipotézist.

Az egész olyan, mint egy bírósági tárgyalás:

- A nullhipotézis a vádlott szava („ártatlan vagyok").
- A statisztikai minta a bizonyítékok halmaza.
- A próbastatisztika (s_n) azt fejezi ki, hogy a vádlott szava mennyire van ellentmondásban a bizonyítékokkal.
- A c kritikus érték egy küszöbérték. Ha $|s_n| \leq c$, akkor a bíró hisz a vádlottnak, és felmenti. Ha $|s_n| > c$, akkor nem hisz neki, és elítéli.

Feladat: A kar férfi hallgatóinak testmagasságát vizsgáljuk, jelölje ξ egy véletlenszerűen kiválasztott férfi hallgató magasságát. Mit állíthatunk ξ várható értékéről, az átlagos testmagasságról a teljes populáción belül?

- Megfigyelt értékek: 180, 175, 188, 168, 173, 183.
- Becslések: $\mathrm{E}(\xi) \approx \overline{\xi} = 177{,}8$, $\mathrm{D}(\xi) \approx \mathrm{D}_6^*(\xi) = 7{,}25$.

Teszteljük a következő nullhipotézist: $H_0 : \mathrm{E}(\xi) = 175$.

Látni fogjuk, hogy a várható értéket a t-próba segítségével lehet tesztelni:

- Próbastatisztika:

$$s_n = \frac{\overline{\xi} - 175}{\mathrm{D}_n^*(\xi)/\sqrt{n}} = \frac{177{,}8 - 175}{7{,}25/\sqrt{6}} = 0{,}946,$$

- A kritikus érték: $c = 2{,}571$. (Miért ennyi? Majd később kiderül.)
- Döntés: $|s_n| \leq c$, tehát a nullhipotézist elfogadjuk. A megfigyelt értékek nincsenek ellentmondásban a nullhipotézis állításával.

Kérdés: Biztosan jól döntöttünk? Biztos, hogy a nullhipotézis igaz?

Sajnos nem: ha félrevezető a minta, amivel dolgozunk, akkor helytelen következtetést vonhatunk le, és hibás döntést hozunk?

Milyen hibákat véthetünk a hipotézisvizsgálat során:

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

$$\alpha = P(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ igaz}).$$

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

$$\beta = P(\text{elfogadjuk } H_0\text{-t} \mid H_0 \text{ hamis}).$$

Még egy fogalom:

$$\textbf{erő (power)} = P(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ hamis}) = 1 - \beta.$$

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?

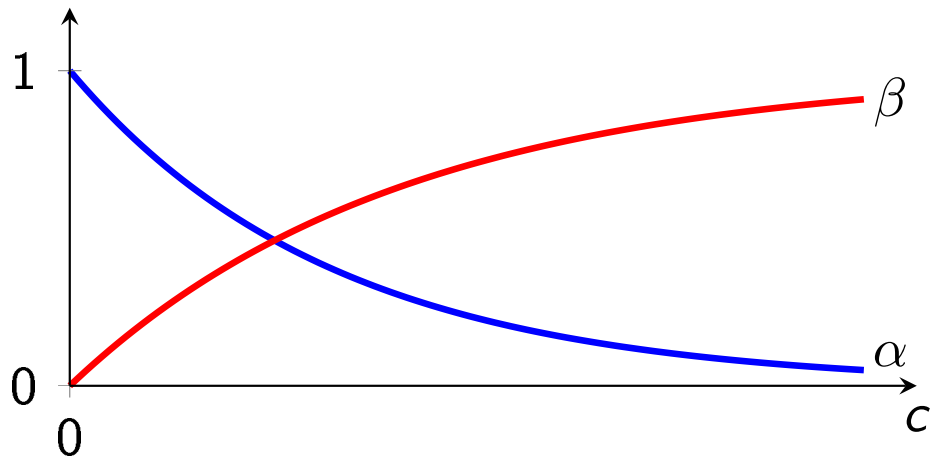
- Akkor vetjük el a nullhipotézist, ha $|s_n| > c$.
- A nullhipotézis, a tesztelési módszer és a statisztikai minta adott: az s_n próbastatisztika értékét nem tudjuk befolyásolni.
- A c kritikus értéket (=mennyire szigorú a bíró) mi választjuk.

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

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

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

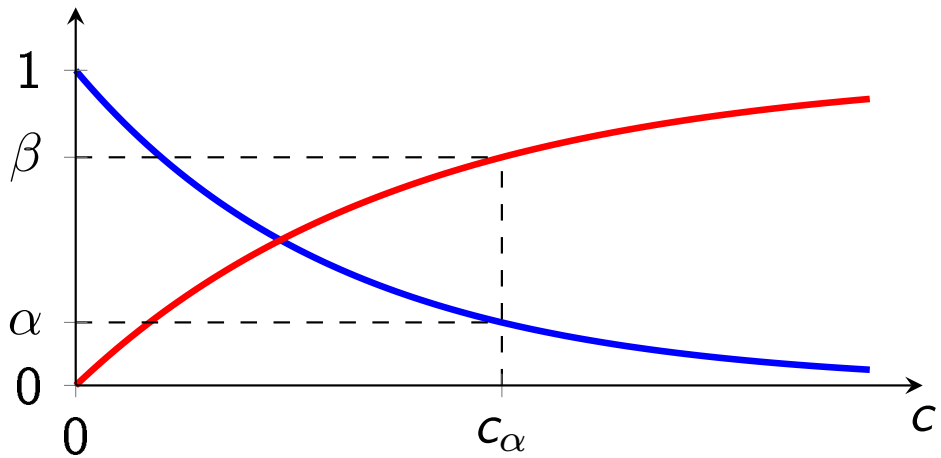
Adott n mintaméret esetén a kétfajta hiba nagysága egymással ellentétesen változik, ha módosítjuk a kritikus értéket:



A hipotézisvizsgálat során az α elsőfajú hibát (**szignifikancia szintet**) előre meg szoktuk adni, és a kritikus értéket ennek megfelelően választjuk. A szignifikancia szint kicsi (tipikusan 1%, 5% vagy 10%) szokott lenni (ártatlanok védelme). A β másodfajú hibára nincsen ráhatásunk.

A kritikus érték meghatározása:

- A feladat megadja az α szignifikancia szintet (=elsőfajú hiba).
- Meghatározzuk a hozzá tartozó kritikus értéket (c_α) és tesztelünk.
- A β másodfajú hiba lehet kicsi vagy nagy is, erre nincs ráhatásunk.



A hipotézisvizsgálat során megjelenő valószínűségek:

	elfogadjuk	elvetjük
H_0 igaz	$1-\alpha$ (nagy)	α (kicsi)
H_0 hamis	β (nem ismert)	$1-\beta$ (nem ismert)

Hogyan lehet értelmezni a hipotézisvizsgálat eredményét?

- Ha elfogadjuk a nullhipotézist, az nem jelent semmit sem:
  - lehetséges, hogy a nullhipotézis igaz, tehát jól döntöttünk,
  - lehetséges, hogy hamis, és másodfajú hibát vétettünk.
- Ha elvetjük a nullhipotézist, az már jelent valamit:
  - lehetséges ugyan, hogy a nullhipotézis igaz, és elsőfajú hibát vétettünk, de ennek kicsi az esélye, ez ritkán történik meg,
  - a nullhipotézis elvetése tipikusan azt jelenti, hogy a nullhipotézis hamis.

Az általunk tanult tesztelési módszerek esetében $\beta \to 0$, ha $n \to \infty$.
Tehát ha növeljük a mintaméretet, akkor a másodfajú hiba is alacsony lesz.
Ez azt jelenti, hogy ezeknél módszereknél nagy mintaméret esetén a nullhipotézis elfogadása már tényleg arra utal, hogy a nullhipotézis igaz.

Az egymintás t-próba

Egymintás t-próba (One sample t test)

Cél a ξ valószínűségi változó várható értékének tesztelése egy $\xi_1, \ldots, \xi_n$ statisztikai minta alapján.

- Feltevések:
  - ξ normális eloszlású változó ismeretlen μ várható értékkel,
  - μ_0 egy tetszőleges hipotetikus érték.
- Nullhipotézis: $H_0 : \mu = \mu_0$.
- Próbastatisztika: (t-próba esetén hagyományosan t_n a jele)

$$t_n = \frac{\overline{\xi} - \mu_0}{\mathrm{D}^*_n(\xi)/\sqrt{n}} = \frac{\overline{\xi} - \mu_0}{\mathrm{SE}}.$$

- Kritikus érték: $c_\alpha = \Phi^{-1}_{n-1}(1 - \alpha/2)$.
- Döntés: akkor fogadjuk el a nullhipotézist, ha $|t_n| \leq c_\alpha$.

Feladat: A kar férfi hallgatóinak testmagasságát vizsgáljuk, jelölje ξ egy véletlenszerűen kiválasztott férfi hallgató magasságát. Mit állíthatunk ξ várható értékéről, az átlagos testmagasságról a teljes populáción belül?

- Megfigyelt értékek: 180, 175, 188, 168, 173, 183.
- Becslések: $\mathrm{E}(\xi) \approx \overline{\xi} = 177{,}8$, $\mathrm{D}(\xi) \approx \mathrm{D}_6^*(\xi) = 7{,}25$.

Teszteljük 5%-os szignifikancia szinten azt, hogy $H_0 : \mathrm{E}(\xi) = 175$.

Tegyük fel, hogy a testmagasság normális eloszlást követ a populáción belül. Ekkor a t-próba alkalmazható.

- Hipotetikus érték, szignifikancia szint: $\mu_0 = 175$, $\alpha = 0{,}05$.
- Próbastatisztika:

$$t_n = \frac{\overline{\xi} - \mu_0}{\mathrm{D}_n^*(\xi)/\sqrt{n}} = \frac{177{,}8 - 175}{7{,}25/\sqrt{6}} = 0{,}946,$$

- A kritikus érték: $c_\alpha = \Phi_{n-1}^{-1}(1 - \alpha/2) = \Phi_5^{-1}(0{,}975) = 2{,}571$.
- Döntés: $|t_n| \leq c$, tehát a nullhipotézist elfogadjuk. A várható érték nem különbözik **szignifikáns (=statisztikailag kimutatható) mértékben** a 175-ös értéktől.

Mi a gondolat a t-próba mögött? A mintaátlag jó becslése a μ igazi várható értéknek, tehát

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

A $H_0 : \mu = \mu_0$ nullhipotézist teszteljük.

- Ha a nullhipotézis igaz, akkor

$$t_n \approx \frac{\mu - \mu_0}{\mathrm{SE}} = 0.$$

- Ha a nullhipotézis nem igaz, akkor

$$t_n \approx \frac{\mu - \mu_0}{\mathrm{SE}} \neq 0.$$

A nullhipotézist akkor fogadjuk el, ha $|t_n| \leq c_\alpha$, tehát ha t_n nullához közeli szám. Ez 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.

Fejtsük ki egy kicsit jobban az előző oldalt! Mikor fogadjuk el H_0-t?

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

Amit kaptunk, az az $1-\alpha$ megbízhatóságú konfidencia intervallum a normális eloszlás várható értékére. Ekkor

$$P\big(\text{elfogadjuk } H_0\text{-t} \mid H_0 \text{ igaz}\big) = P\Big(\mu_0 \in \left[\overline{\xi} - c_\alpha \,\mathrm{SE}\,,\, \overline{\xi} + c_\alpha \,\mathrm{SE}\right] \mid \mu = \mu_0\Big)$$
$$= P\Big(\mu \in \left[\overline{\xi} - c_\alpha \,\mathrm{SE}\,,\, \overline{\xi} + c_\alpha \,\mathrm{SE}\right]\Big) = 1 - \alpha.$$

Ebből következik, hogy

$$P\big(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ igaz}\big) = 1 - P\big(\text{elfogadjuk } H_0\text{-t} \mid H_0 \text{ igaz}\big) = \alpha.$$

Az előző oldalon levezetett számolásnak több fontos következménye van:

- 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.
- Ha a minta normális eloszlásból jön, akkor a t-próba pontosan betartja az előírt elsőfajú hibát:

$$P\big(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ igaz}\big) = \text{megadott szignifikancia szint.}$$

- Ha a minta nem normális eloszlásból származik, de a mintaméret elég nagy, akkor a t-próba használható a várható érték tesztelésére. Ebben az esetben a próba csak közelítőleg tartja be az előírt elsőfajú hibát:

$$P\big(\text{elvetjük } H_0\text{-t} \mid H_0 \text{ igaz}\big) \approx \text{megadott szignifikancia szint.}$$

Lefutattam a t-próbát 5%-os szignifikancia szinten a testmagasságokra az R programmal, az alábbi eredményt kaptam:

```
        One Sample t-test
data:  magassag
t = 0.95723, df = 5, p-value = 0.3824
alternative hypothesis:  true mean is not equal to 175
95 percent confidence interval:  170.2246 185.4420
sample estimates:  mean of x 177.8333
```

Értelmezzük, hogy milyen információ van az outputban:

- Egymintás t-próba a „magassag" nevű adatsoron.
- Próbastatisztika: $t = 0.95723$, szabadsági fok (degrees of freedom): $\mathrm{df} = 5$.
- Nullhipotézis és alternatív hipotézis: $H_0 : \mu = 175$, $H_A : \mu \neq 175$.
- 95%-os konfidencia intervallum: $[170.2246, 185.4420]$.
- Mintaátlag: 177.8333

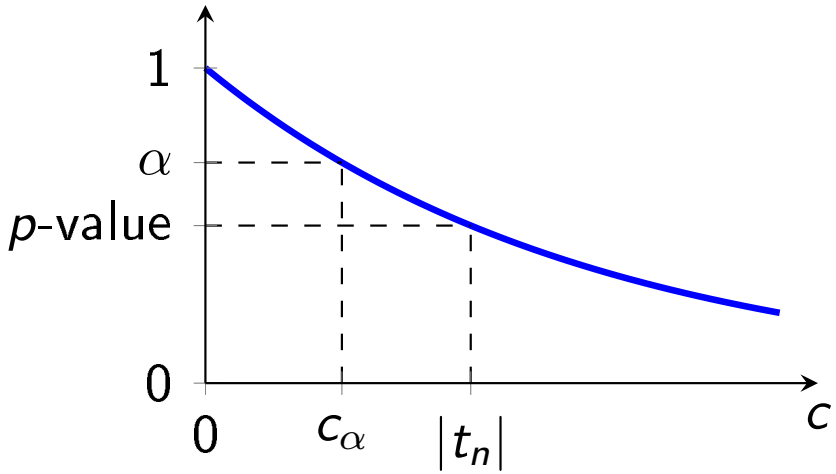
A program által adott értékek kissé eltérnek attól, amit mi kaptunk: nálunk sok volt a kerekítési hiba. Felmerülő kérdések:

- Hol a kritikus érték és a döntés? És mi az a „p-value"?

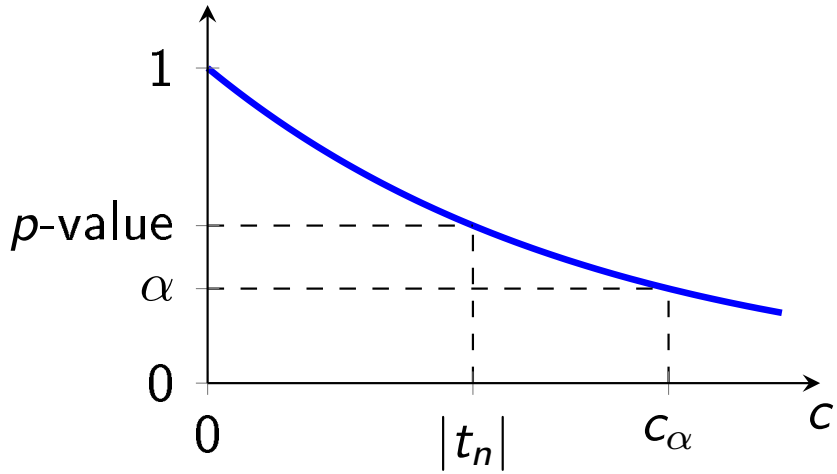
A **p-érték (p-value)** az a határ szignifikancia szint, amikor még éppen elfogadjuk a nullhipotézist, tehát $c_{p\text{-value}} = t_n$. Ekkor

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

A statisztikai programok tesztelés során gyakran nem a kritikus értéket, hanema p-értéket adják meg. A p-érték 0 és 1 közé esik, és értelmezhető olyan módon, hogy mennyire „hihető” a nullhipotézis az adott statisztikai minta mellett. A nullhipotézist akkor vetjük el, ha a p-érték alacsony.



elvetjük H_0-t



elfogadjuk H_0-t

Az előző feladatban: $p\text{-value} = 0.3824 > \alpha = 0.05$, tehát H_0-t elfogadjuk.

Kétmintás hipotézisvizsgálat

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 segítségével becsléseket végezhetünk:

- $\overline{\xi} = \mathsf{E}_n(\xi) \approx \mathsf{E}(\xi)$, $\mathsf{D}^*_n(\xi) \approx \mathsf{D}(\xi)$,
- $\overline{\eta} = \mathsf{E}_m(\eta) \approx \mathsf{E}(\eta)$, $\mathsf{D}^*_m(\eta) \approx \mathsf{D}(\eta)$.

A minták tipikusan kétfajta kapcsolatban állhat egymással:

- **Független minták (independent samples):** A minták között nincs kapcsolat, mert a két minta független mintavételezésből származik. Például: egymástól függetlenül veszünk mintát két részpopulációból.
- **Összetartozó minták (paired samples, related samples):** A ξ_i és az η_i megfigyelés minden i esetén a populáció ugyanazon egyedére vonatkozik, ezért ezek az értékek nem függetlenek egymástól. Ebben az esetben mindig $n = m$.

Feladat: Döntsük el, hogy az alábbi példákban független vagy összetartozó mintákról van szó.

1.
   - $\xi_1, \ldots, \xi_n$: n véletlenszerűen kiválasztott férfi hallgató testmagassága,
   - $\eta_1, \ldots, \eta_m$: m véletlenszerűen kiválasztott női hallgató testmagassága.

   Független minták: a minták független megfigyelésekből jönnek.
2.
   - $\xi_1, \ldots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben,
   - $\eta_1, \ldots, \eta_n$: ugyanezen hallgatók édesanyjának testmagassága.

   Összetartozó minták: a megfigyelések azonos egyedekre vonatkoznak.
3.
   - $\xi_1, \ldots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben,
   - $\eta_1, \ldots, \eta_n$: n férfi hallgató testmagassága egy 5 évvel ezelőtti független felmérésben.

   Független minták: a minták független megfigyelésekből jönnek.
4.
   - $\xi_1, \ldots, \xi_n$: n férfi hallgató testmagassága egy mai felmérésben,
   - $\eta_1, \ldots, \eta_n$: ugyanezen hallgatók testmagassága egy 5 évvel ezelőtti felmérésben.

   Összetartozó minták: a megfigyelések azonos egyedekre vonatkoznak.

A ξ és az η valószínűségi változó **együttesen normális eloszlást követ**, ha tetszőleges a és b valós számok esetén $a\xi + b\eta$ normális eloszlású. Ez egy kicsivel több annál, hogy ξ és η normális eloszlású.

Tegyük fel, hogy

- ξ és η együttesen normális eloszlásúak,
- a várható értékek (μ_ξ és μ_η) ismeretlenek,
- $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_n$ összetartozó minták,

Célunk a következő nullhipotézist tesztelni: $H_0 : \mu_\xi = \mu_\eta$.

Gondolatmenet:

- $(+1)\xi + (-1)\eta = \xi - \eta$ normális eloszlású változó.
- $\xi_1 - \eta_1, \ldots, \xi_n - \eta_n$ statisztikai minta a $\xi - \eta$ változóra.
- $\mathrm{E}(\xi - \eta) = \mathrm{E}(\xi) - \mathrm{E}(\eta) = \mu_\xi - \mu_\eta$, ezért $H_0 \Leftrightarrow \mathrm{E}(\xi - \eta) = 0$.
- Teszteltjük a $H_0 : \mathrm{E}(\xi - \eta) = 0$ nullhipotézist egymintás t-próbával.

Foglaljuk össze, hogyan lehet összetartozó minták várható értékét tesztelni:

Páros t-próba (paired samples t test)

Cél a várható értékek tesztelése összetartozó minták esetén.

- Feltevések:
  - ξ és η együttesen normális eloszlásúak,
  - a várható értékek (μ_ξ és μ_η) ismeretlenek,
  - $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_n$ összetartozó minták.
- Nullhipotézis: $H_0 : \mu_\xi = \mu_\eta$.
- Próbastatisztika:

$$t_n = \frac{\overline{\xi - \eta} - 0}{\mathrm{D}^*_n(\xi - \eta)/\sqrt{n}}.$$

- Kritikus érték: $c_\alpha = \Phi^{-1}_{n-1}(1 - \alpha/2)$.
- Döntés: akkor fogadjuk el a nullhipotézist, ha $|t_n| \leq c_\alpha$.

Feladat: Teszteljük azt, hogy a kar jelenlegi férfi hallgatóinak átlagos testmagassága nem változott ez elmúlt 5 év folyamán. ($\alpha = 5\%$)

Legyen

- ξ = véletlenszerűen kiválasztott férfi hallgató testmagassága ma,
- η = ugyanezen hallgató testmagassága 5 évvel ezelőtt,
- $\xi - \eta$ = testmagasság változása 5 év alatt.

Várható értékek:

- $\mu_\xi = \mathsf{E}(\xi)$ = férfi hallgatók átlagos testmagassága ma,
- $\mu_\eta = \mathsf{E}(\eta)$ = ugyanezen hallgatók átlagos magassága 5 évvel ezelőtt,
- $\mu_\xi - \mu_\eta = \mathsf{E}(\xi - \eta)$ = átlagos magasságváltozás 5 év alatt.

Nullhipotézis: $H_0 : \mu_\xi = \mu_\eta$. Helyette: $H_0 : \mathsf{E}(\xi - \eta) = 0$.

Statisztikai minták: $(n = 6)$

- Kiválasztott hallgatók magassága ma: 180, 175, 188, 168, 173, 183.
- Ugyanezen hallgatók magassága 1 éve: 175, 172, 184, 167, 170, 178.
- Minta a $\xi - \eta$ változóra: 5, 3, 4, 1, 3, 5.

Minta a $\xi - \eta$ változóra: 5, 3, 4, 1, 3, 5.

Becslések:

- Mintaátlag: $\overline{\xi - \eta} = 3{,}5 \approx \mathrm{E}(\xi - \eta)$.
- Korrigált empirikus szórás: $\mathrm{D}_6^*(\xi - \eta) = 1{,}52 \approx \mathrm{D}(\xi - \eta)$.

Nullhipotézis: $H_0 : \mathrm{E}(\xi - \eta) = 0$.

Egymintás t-próba:

- Próbastatisztika:

$$t_n = \frac{\overline{\xi - \eta}}{\mathrm{D}_n^*(\xi - \eta)/\sqrt{n}} = \frac{3{,}5 - 0}{1{,}52/\sqrt{6}} = 5{,}64.$$

- A kritikus érték: $c_\alpha = \Phi_{n-1}^{-1}(1 - \alpha/2) = \Phi_5^{-1}(0{,}975) = 2{,}571$.
- Döntés: $|t_n| > c_\alpha$, tehát a nullhipotézist elvetjük. A populációban az átlagos testmagasság szignifikáns módon változott az elmúlt 5 év folyamán.

A következő módszerrel független minták várható értékét tesztelhetjük:

Kétmintás t-próba (independent samples t test)

Cél a várható értékek tesztelése független minták esetén.

- Feltevések:
  - ξ és η normális eloszlásúak,
  - a várható értékekek (μ_ξ és μ_η) és a szórások (σ_ξ és σ_η) ismeretlenek,
  - a szórások megegyeznek: $\sigma_\mu = \sigma_\eta$,
  - $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_m$ független minták.
- Nullhipotézis: $H_0 : \mu_\xi = \mu_\eta$.
- Próbastatisztika: $t_{n,m} = (\overline{\xi} - \overline{\eta}) / \mathrm{D}_{n,m}$, ahol

$$\mathrm{D}_{n,m} = \sqrt{\left[(n-1)\,\mathrm{Var}^*_n(\xi) + (m-1)\,\mathrm{Var}^*_m(\eta)\right] \frac{n+m}{nm(n+m-2)}}$$

- Kritikus érték: $c_\alpha = \Phi^{-1}_{n+m-2}(1-\alpha/2)$.
- Döntés: akkor fogadjuk el a nullhipotézist, ha $|t_{n,m}| \leq c_\alpha$.

Mi a kétmintás t-próba alapötlete? A mintaátlag jó becslés a várható értékre, ezért

$$t_{n,m} = \frac{\overline{\xi} - \overline{\eta}}{\mathrm{D}_{n,m}} \approx \frac{\mu_\xi - \mu_\eta}{\mathrm{D}_{n,m}}.$$

A $H_0 : \mu_\xi = \mu_\eta$ nullhipotézist teszteljük.

- Ha a nullhipotézis igaz, akkor

$$t_{n,m} \approx \frac{\mu_\xi - \mu_\eta}{\mathrm{D}_{n,m}} = 0.$$

- Ha a nullhipotézis nem igaz, akkor

$$t_{n,m} \approx \frac{\mu_\xi - \mu_\eta}{\mathrm{D}_{n,m}} \neq 0.$$

A nullhipotézist akkor fogadjuk el, ha $t_{n,m}$ nullához közeli szám. Ez logikus ötlet, hiszen

- ha $t_{n,m} \approx 0$, akkor az arra utal, hogy H_0 igaz,
- ha $t_{n,m} \not\approx 0$, akkor az arra utal, hogy H_0 nem igaz.

Feladat: Teszteljük azt a nullhipotézist, hogy a kar férfi hallgatóinak átlagos testmagassága nem változott ez elmúlt 5 év folyamán. ($\alpha = 5\%$)

Legyen

- $\xi =$ véletlenszerűen kiválasztott férfi hallgató testmagassága ma,
- $\eta =$ véletlenszerű hallgató magassága az 5 évvel ezelőtti populációban.

Várható értékek:

- $\mu_\xi = \mathsf{E}(\xi) =$ férfi hallgatók átlagos testmagassága ma,
- $\mu_\eta = \mathsf{E}(\eta) =$ férfi hallgatók átlagos testmagassága 5 évvel ezelőtt.

Statisztikai minták két független felmérésből:

- Kiválasztott hallgatók magassága ma: 180, 175, 188, 168, 173, 183.
- Kiválasztott hallgatók magassága 5 éve: 171, 178, 183, 168, 175.

Becslések:

- $\mu_\xi \approx \overline{\xi} = 177{,}8, \quad \sigma_\xi \approx \mathrm{D}^*_6(\xi) = 7{,}25.$
- $\mu_\eta \approx \overline{\eta} = 175, \quad \sigma_\eta \approx \mathrm{D}^*_5(\eta) = 5{,}87.$

Nullhipotézis: $H_0 : \mu_\xi = \mu_\eta.$

Feltehető, hogy a minták normális eloszlásból származnak és az elméleti szórások azonosak $(\sigma_\xi = \sigma_\eta)$.

Kétmintás t-próba:

- Próbastatisztika: $t_{n,m} = (177{,}8 - 175)/3{,}845 = 0{,}73,$ ugyanis

$$\mathrm{D}_{n,m} = \sqrt{\left[(6-1)7{,}25^2 + (5-1)5{,}87^2\right] \frac{6+5}{6 \cdot 5 \cdot (6+5-2)}} = 3{,}845.$$

- Kritikus érték: $c_\alpha = \Phi_9^{-1}(0{,}975) = 2{,}262.$
- Döntés: $|t_{n,m}| \leq c_\alpha,$ ezért a nullhipotézist elfogadjuk. A kar férfi hallgatóinak átlagos testmagassága az elmúlt 5 évben nem változott szignifikáns mértékben.

F-próba (F test)

Cél a szórások tesztelése független minták esetén.

- Feltevések:
  - ξ és η normális eloszlásúak, a szórások (σ_ξ és σ_η) ismeretlenek,
  - $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_m$ független minták.
- Nullhipotézis: $H_0 : \sigma_\xi = \sigma_\eta$.
- Próbastatisztika, kritikus érték: nem tanuljuk.

Welch-próba (Welch test)

Ugyanaz, mint a kétmintás t-próba, de nem kell a szórások egyenlősége.

- Feltevések:
  - ξ és η normális eloszlásúak,
  - a várható értékekek (μ_ξ és μ_η) ismeretlenek,
  - $\xi_1, \ldots, \xi_n$ és $\eta_1, \ldots, \eta_m$ független minták.
- Nullhipotézis: $H_0 : \mu_\xi = \mu_\eta$.
- Próbastatisztika, kritikus érték: nem tanuljuk.

A tanult statisztikai módszerek (becslések, tesztek) mögött matematikai tételek állnak. Ezek a tételek garantálják, hogy megfelelő feltételek mellett ezek a módszerek jól működnek. (Például a becslések erősen konzisztensek, a tesztek betartják az előírt elsőfajú hibát.) De mennyire hetékonyak ezek a módszerek akkor, ha a szükséges feltételek nem teljesülnek?

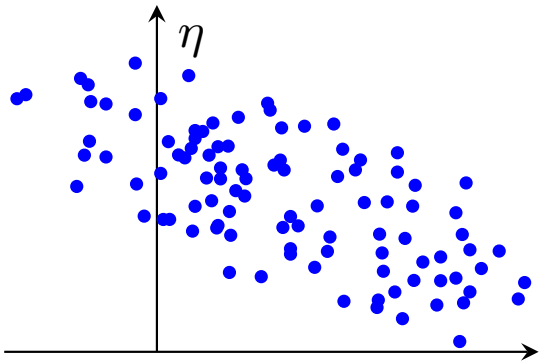
Egy statisztikai módszer **robusztus (robust)** egy feltételre nézve, ha a módszer a feltétel elhagyásával is jól alkalmazható. Például:

- A tanult t-próbák és a Welch-próba robusztus a normalitási feltételre nézve, ezek a próbák nem normális eloszlásra is alkalmazhatóak, ha a mintaméret elég nagy. (50 mintaelem elég szokott lenni.)
- A kétmintás t-próba nem robusztus a szórásfeltételre nézve. Ha a szórások nem azonosak, akkor a Welch-próbát kell alkalmazni.

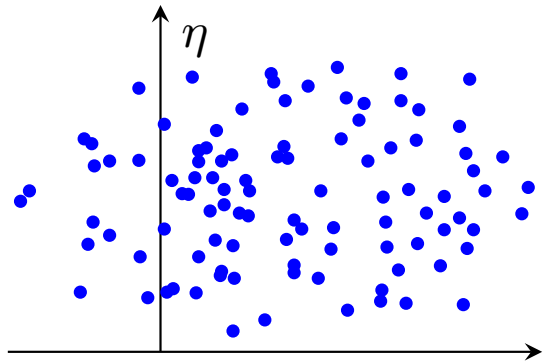
Kovariancia és korreláció

Eddig azt vizsgáltuk, hogy egy ξ mennyiségnek milyen az eloszlása egy populáción belül. A továbbiakban két mennyiség (ξ és η) együttes viselkedésével foglalkozunk. Főleg az a kérdés, hogy milyen irányú és milyen erősségű kapcsolat van a két változó között. A fontosabb esetek:

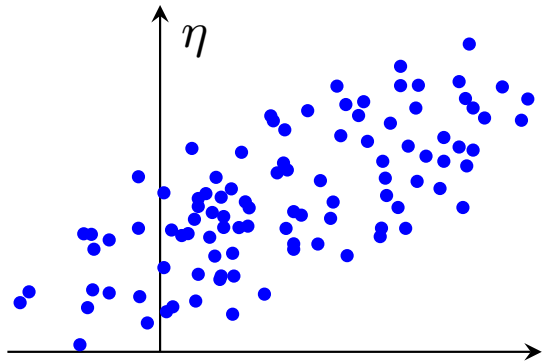
- Pozitív irányú kapcsolat: a ξ és az η (jellemzően) azonos irányba mozdul el. (Például: testmagasság és testsúly.)
- Negatív irányú kapcsolat: ξ és η (jellemzően) egymással ellentétes irányba mozog.
- Független változók: nincs kapcsolat az értékek között.



Negatív kapcsolat



Független változók



Pozitív kapcsolat

Legyen ξ és η valószínűségi változó. Ekkor a két változó **kovarianciája (covariance)** illetve **korrelációs együtthatója (correlation coefficient)**:

$$\mathrm{Cov}(\xi,\eta) = \mathrm{E}\Big(\big[\xi - \mathrm{E}(\xi)\big]\big[\eta - \mathrm{E}(\eta)\big]\Big), \quad \mathrm{corr}(\xi,\eta) = \frac{\mathrm{Cov}(\xi,\eta)}{\mathrm{D}(\xi)\,\mathrm{D}(\eta)}.$$

Ha a kovariancia (és ezáltal a korrelációs együttható) értéke nulla, akkor azt mondjuk, hogy a két változó **korrelálatlan (uncorrelated)**.

A kovariancia és a korrelációs együttható fontosabb tulajdonságai:

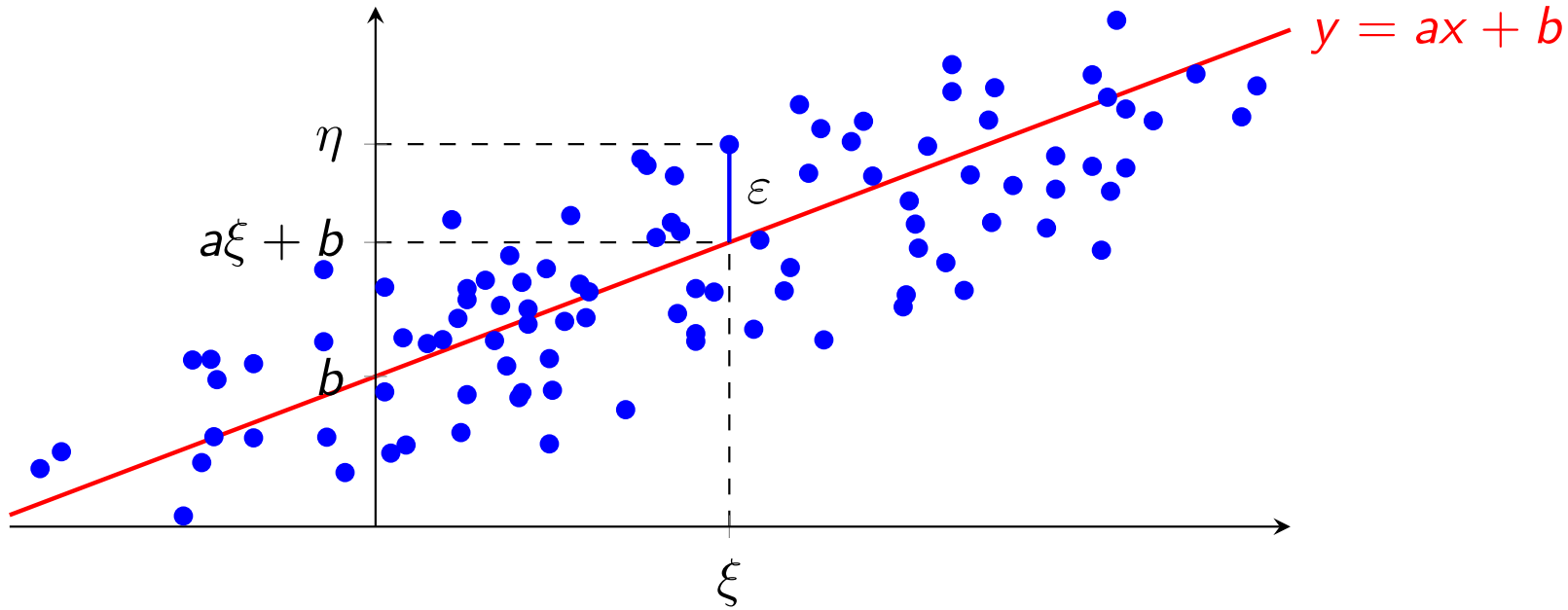
- A kovariancia és a korrelációs együttható a teljes populációt jellemzi valamilyen (de vajon milyen?) szempontból.
- Lehetséges értékek: $\mathrm{Cov}(\xi,\eta) \in \mathbb{R}$, $\mathrm{corr}(\xi,\eta) \in [-1,+1]$.
- Szimmetria: $\mathrm{Cov}(\xi,\eta) = \mathrm{Cov}(\eta,\xi)$, $\mathrm{corr}(\xi,\eta) = \mathrm{corr}(\eta,\xi)$.
- Ha ξ és η függetlenek, akkor korrelálatlanok is. Viszont a korrelálatlanságból nem következik a függetlenség.
- Ha a két változó együttesen normális eloszlású, akkor a függetlenség és a korrelálatlanság ekvivalens fogalmak.

Lineáris regresszió a teljes populáción: szeretnénk megérteni, hogy az η változó értéke milyen módon alakul ki. Vegyük a következő reprezentációt:

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

Elnevezések és modellfeltevések:

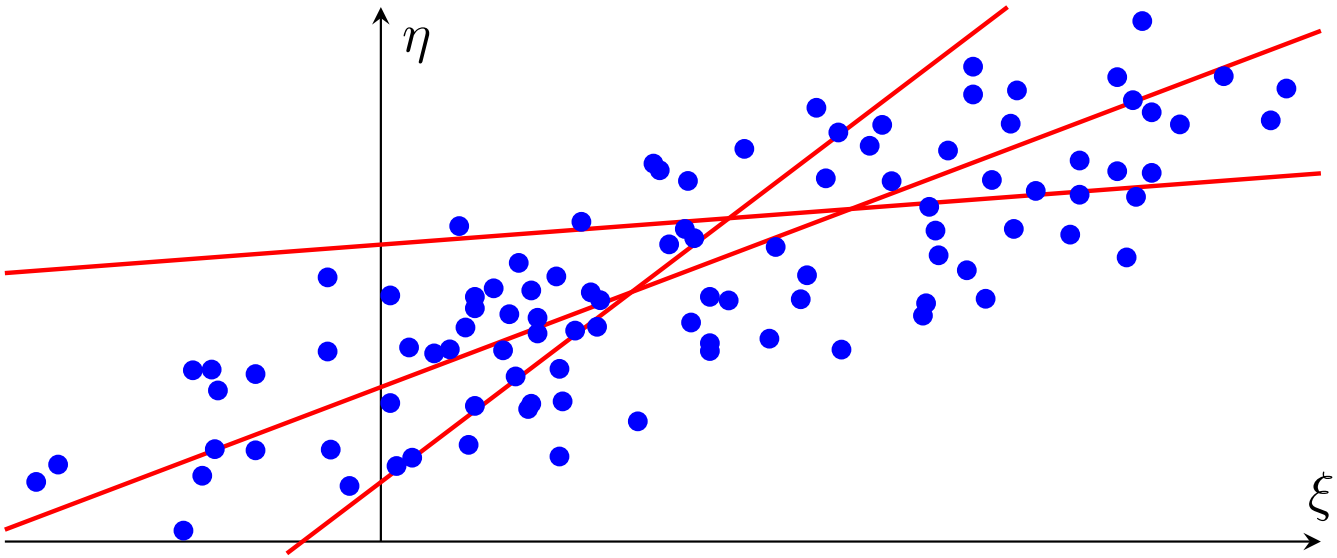
- ξ a **magyarázó változó**, η a **függő változó**,
- ε az egyedre jellemző **hibatag**, ami független a ξ változótól,
- $\mathrm{E}(\varepsilon) =$ átlagos hiba $= 0$.



A regressziós modell lehetséges alkalmazásai:

- Becslés az η változóra: ha a hibatag kicsi, akkor $\eta \approx a\xi + b$.
- Megérteni, hogy milyen hatások szerint alakul ki az η változó.

Probléma: több olyan egyenes is létezhet, ami teljesíti a modellfeltevéseket.



Mi a $\mathrm{D}(\varepsilon)$ szórás jelentése a populációra nézve?

$$\mathrm{D}(\varepsilon) = \varepsilon \text{ átlagos eltérése a 0 várható értéktől}$$
$$= \text{az abszolút hiba átlagos értéke}$$

Cél: azt az egyenest keressük, melynél $\mathrm{D}(\varepsilon)$ minimális, ugyanis ez az egyenes illeszkedik a legjobban a ponthalmazra.

Minimalizáljuk a hibatag szórását az a és b változókban!

$$\mathrm{D}^2(\varepsilon) = \mathrm{D}^2(\eta - a\xi - b) = \mathrm{D}^2(\eta) + a^2\mathrm{D}^2(\xi) - 2a\mathrm{D}(\xi)\mathrm{D}(\eta)\,\mathrm{corr}(\xi,\eta)$$

Deriválással:

$$0 = \frac{\partial \mathrm{D}^2(\varepsilon)}{\partial a} = 2a\mathrm{D}^2(\xi) - 2\mathrm{D}(\xi)\mathrm{D}(\eta)\,\mathrm{corr}(\xi,\eta),$$

amiből következik, hogy

$$a = \frac{\mathrm{corr}(\xi,\eta)\,\mathrm{D}(\eta)}{\mathrm{D}(\xi)}.$$

Azt is tudjuk, hogy $\mathrm{E}(\varepsilon) = 0$, tehát

$$0 = \mathrm{E}(\varepsilon) = \mathrm{E}(\eta - a\xi - b) = \mathrm{E}(\eta) - a\mathrm{E}(\xi) - b,$$

amiből

$$b = \mathrm{E}(\eta) - a\mathrm{E}(\xi) = \mathrm{E}(\eta) - \frac{\mathrm{corr}(\xi,\eta)\,\mathrm{D}(\eta)}{\mathrm{D}(\xi)}\mathrm{E}(\xi).$$

Feladat: A Tisza és a Maros vízhozama számunkra ismeretlen eloszlást követ. Azt tudjuk, hogy a Maros torkolata felett a Tisza vízhozamának a várható értéke 660 m^3/s, szórása 160 m^3/s, míg a Maros vízhozamának a várható értéke 200 m^3/s, szórása 50 m^3/s. A korrelációs együttható 0,8. Egy adott napok a Tisza vízhozama 800 m^3/s. Milyen becslést adhatunk a Maros vízhozamára?

Legyen ξ és η a Tisza illetve a Maros vízhozama torkolat felett. Ekkor

$$\mathrm{E}(\xi) = 660, \quad \mathrm{D}(\xi) = 160, \quad \mathrm{E}(\eta) = 200, \quad \mathrm{D}(\eta) = 50, \quad \mathrm{corr}(\xi, \eta) = 0{,}8.$$

Az előző oldalakon kapott formulák alkalmazásával kapjuk, hogy

$$a = \frac{\mathrm{corr}(\xi, \eta)\mathrm{D}(\eta)}{\mathrm{D}(\xi)} = 0{,}25, \qquad b = \mathrm{E}(\eta) - a\mathrm{E}(\xi) = 35.$$

Tehát a két vízhozamra az alábbi regressziós modell írható fel:

$$\eta = a\xi + b + \text{hibatag} = 0{,}25\xi + 35 + \text{hibatag} \approx 0{,}25\xi + 35.$$

A mai napon $\xi = 800$, tehát $\eta \approx 0{,}25 \cdot 800 + 35 = 235$.

Viszgáljuk meg a varianciákat! A tagok függetlensége miatt

$$\mathrm{D}^2(\eta) = \mathrm{D}^2(a\xi + b + \varepsilon) = \mathrm{D}^2(a\xi + b) + \mathrm{D}^2(\varepsilon).$$

Egy kis számolás után (amit nem részletezünk) a predikciós tag varianciája

$$\mathrm{D}^2(a\xi + b) = \mathrm{corr}^2(\xi, \eta)\mathrm{D}^2(\eta).$$

Ebből következik, hogy a hibatag varianciája

$$\mathrm{D}^2(\varepsilon) = \mathrm{D}^2(\eta) - \mathrm{D}^2(a\xi + b) = \left[1 - \mathrm{corr}^2(\xi, \eta)\right]\mathrm{D}^2(\eta).$$

Foglaljuk össze, hogy mit kaptunk:

- A pedikciós tag varianciája a teljes varianciának $\mathrm{corr}^2(\xi, \eta)$ hányada. „Ekkora mértékben magyarázza a ξ változó az η értékét."
- A hibatag varianciája a teljes varianciának $1 - \mathrm{corr}^2(\xi, \eta)$ hányada. „Ekkora mértékben magyarázza az ε változó az η értékét."

Feladat: Mekkora a predikciós tag illetve a hibatag várható értéke illetve szórása a vízhozamos feladatban? Milyen arányban magyarázza a Tisza vízhozama a Maros vízhozamát?

A modell: $\eta = (0{,}25\xi + 35) + \varepsilon =$ predikciós tag + hibatag.
A predikciós tag várható értéke:

$$\mathrm{E}(\text{predikciós tag}) = 0{,}25\,\mathrm{E}(\xi) + 35 = 0{,}25 \cdot 600 + 35 = 200.$$

A varianciák:

$$\mathrm{D}^2(\text{predikciós tag}) = \mathrm{D}^2(a\xi + b) = \mathrm{corr}^2(\xi, \eta)\mathrm{D}^2(\eta) = 0{,}8^2 \cdot 50^2 = 1600,$$

$$\mathrm{D}^2(\text{hibatag}) = \mathrm{D}^2(\varepsilon) = \left[1 - \mathrm{corr}^2(\xi, \eta)\right]\mathrm{D}^2(\eta) = \left[1 - 0{,}8^2\right] \cdot 50^2 = 900.$$

Foglaljuk össze ezeket egy táblázatban:

Hatás	E	Var	D	Var%
Tisza $(a\xi + b)$	200	1600	40	64%
hibatag (ε)	0	900	30	36%
Maros (η)	200	2500	50	100%

A Tisza 64%, a hibatag 36% arányban magyarázza a Maros vízhozamát.

Milyen módon jellemzi a korrelációs együttható a ξ és az η változó kapcsolatát? Induljunk ki a következő két összefüggésből:

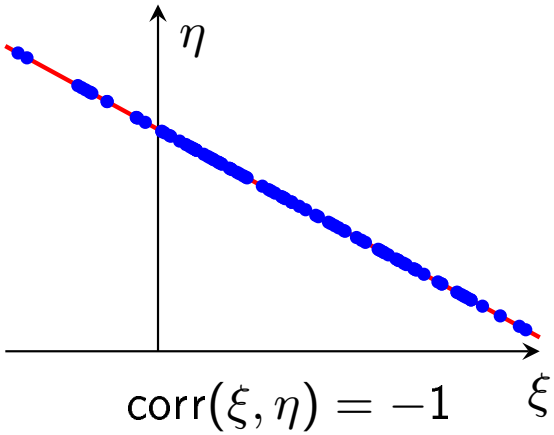
$$a = \frac{\mathrm{corr}(\xi,\eta)\mathrm{D}(\eta)}{\mathrm{D}(\xi)}, \qquad \mathrm{D}^2(\varepsilon) = \left[1 - \mathrm{corr}^2(\xi,\eta)\right]\mathrm{D}^2(\eta).$$

A regressziós egyenes meredeksége (a) alapján:

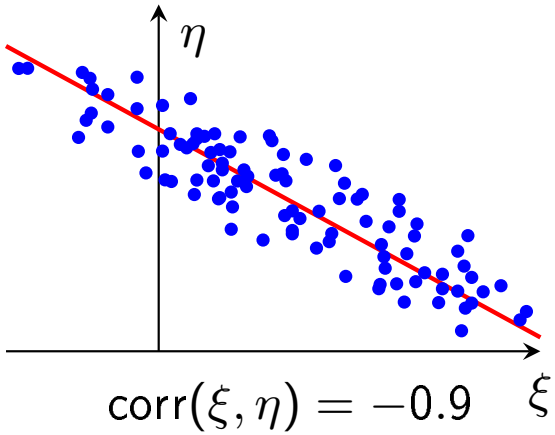
- Ha $\mathrm{corr}(\xi,\eta) > 0$, akkor $a > 0$, tehát a változók között pozitív irányú kapcsolat van.
- Ha $\mathrm{corr}(\xi,\eta) < 0$, akkor $a < 0$, tehát a változók között negatív irányú kapcsolat van.

A hibatag varianciája alapján:

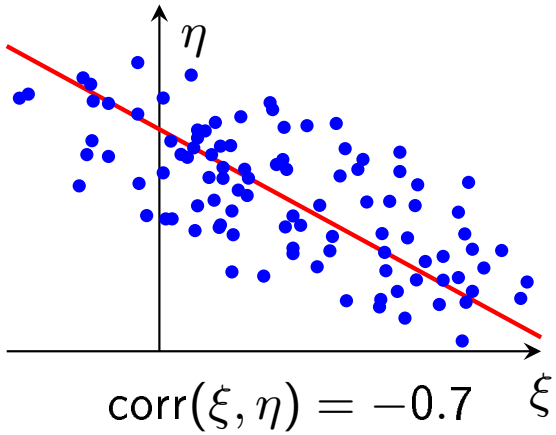
- Ha $\mathrm{corr}(\xi,\eta) \approx \pm 1$, akkor $\mathrm{D}(\varepsilon) \approx 0$. Ebben az esetben jó az illeszkedés a regressziós egyeneshez, a változók közötti kapcsolat erős.
- Ha $\mathrm{corr}(\xi,\eta) \approx 0$, akkor a $\mathrm{D}(\varepsilon)$ szórás nagy. Ebben az esetben nem jó az illeszkedés a regressziós egyeneshez, a változók között gyenge a kapcsolat (vagy akár függetlenek is).



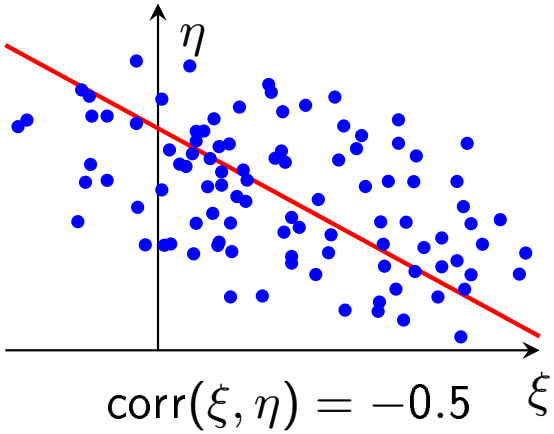
$\mathrm{corr}(\xi, \eta) = -1$



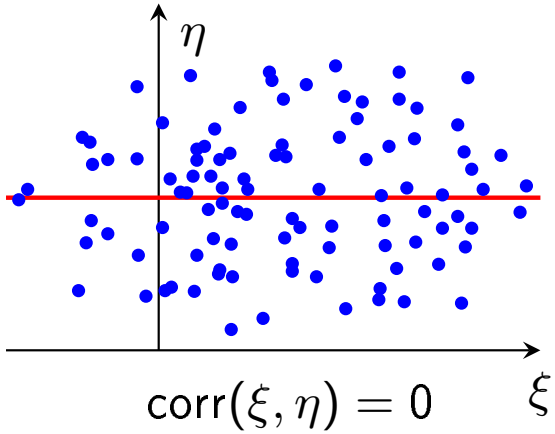
$\mathrm{corr}(\xi, \eta) = -0.9$



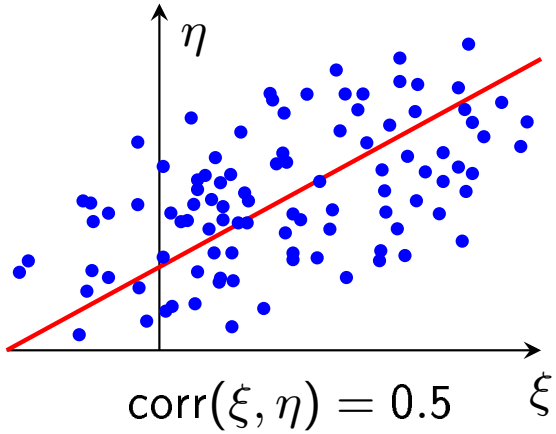
$\mathrm{corr}(\xi, \eta) = -0.7$



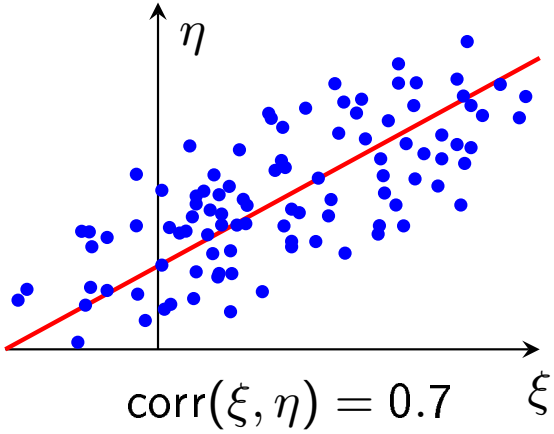
$\mathrm{corr}(\xi, \eta) = -0.5$



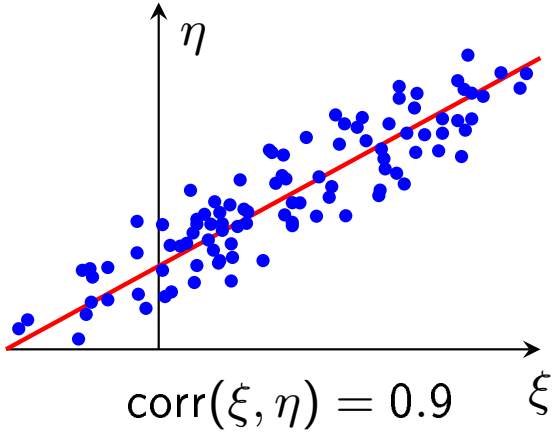
$\mathrm{corr}(\xi, \eta) = 0$



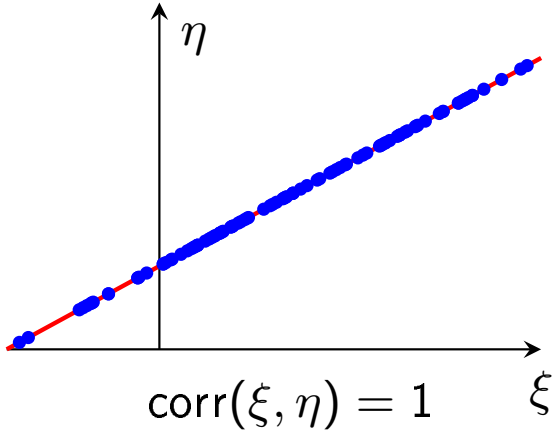
$\mathrm{corr}(\xi, \eta) = 0.5$



$\mathrm{corr}(\xi, \eta) = 0.7$



$\mathrm{corr}(\xi, \eta) = 0.9$



$\mathrm{corr}(\xi, \eta) = 1$

Statisztikai lineáris regresszió

Az előző részben két mennyiség (ξ és η) kapcsolatát vizsgáltuk a teljes populáción belül. Probléma: nem ismerjük a változók elméleti várható értékét, szórását és korrelációs együtthatóját. Megoldás: egy statisztikai minta alapján mindent becsülni fogunk.

Tekintsünk összetartozó mintákat a ξ és η változókra:

- $\xi_1, \ldots, \xi_n$ független megfigyelések ξ-re,
- $\eta_1, \ldots, \eta_n$ az η mennyiség értékei ugyanezen egyedeknél.

Empirikus kovariancia (sample covariance) és **empirikus korrelációs együttható (sample correlation coefficient)**:

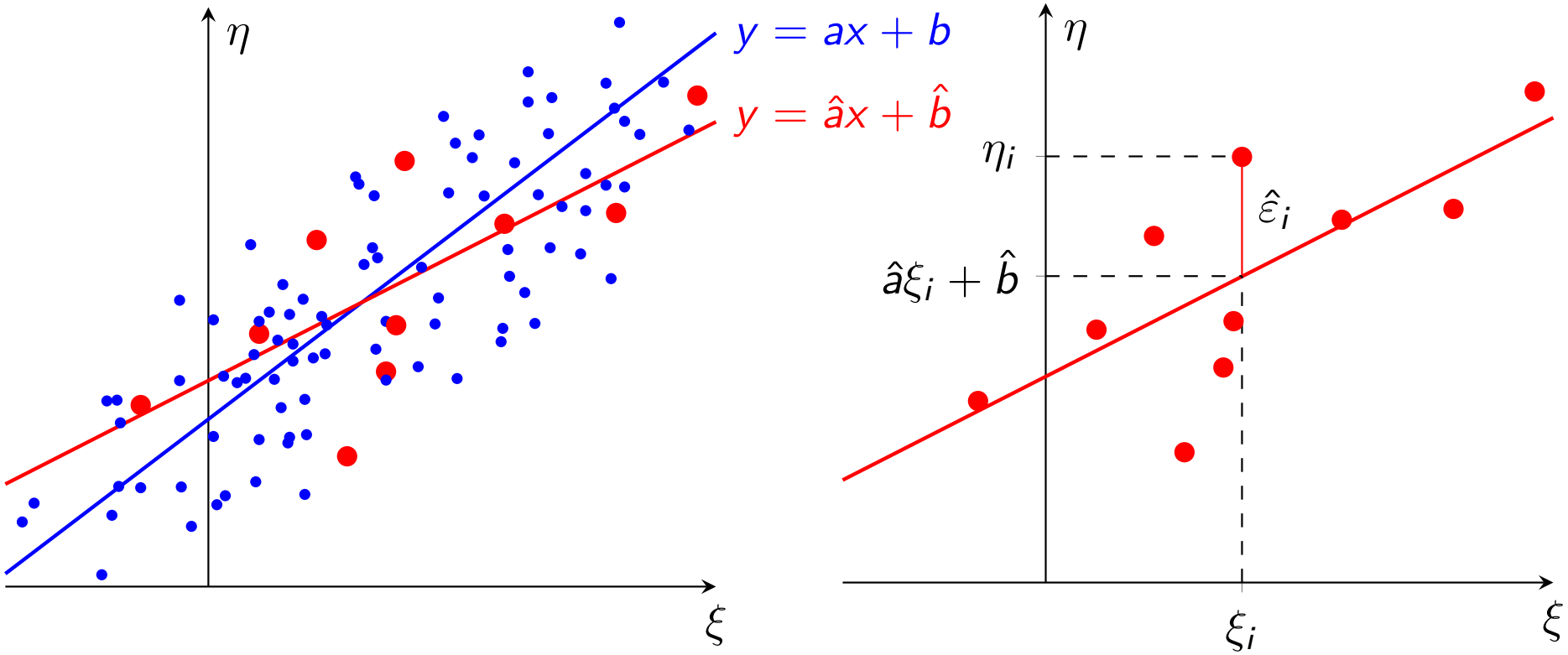
$$\mathrm{Cov}_n(\xi,\eta) = \frac{1}{n-1}\sum_{i=1}^{n}(\xi_i-\overline{\xi})(\eta_i-\overline{\eta}), \qquad \mathrm{corr}_n(\xi,\eta) = \frac{\mathrm{Cov}_n(\xi,\eta)}{\mathrm{D}^*_n(\xi)\,\mathrm{D}^*_n(\eta)}.$$

Lineáris regresszió a teljes populáción:

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

Lineáris regresszió a mintaelemeken:

$$\eta_i = (\hat{a}\xi_i + \hat{b}) + \hat{\varepsilon}_i = \text{predikciós tag} + \text{reziduális}, \qquad i = 1, \ldots, n.$$



Az egyenest a **legkisebb négyzetes becslés (least squares estimation)** alkalmazásával kapjuk meg: keressük azon $\hat{a}$ és $\hat{b}$ számokat, melyekre

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

Az összeget $\hat{a}$ és $\hat{b}$ szerint deriválva:

$$0 = \frac{\partial S}{\partial \hat{a}} = -2 \sum_{i=1}^{n} (\eta_i - \hat{a}\xi_i - \hat{b})\xi_i,$$

$$0 = \frac{\partial S}{\partial \hat{b}} = -2 \sum_{i=1}^{n} (\eta_i - \hat{a}\xi_i - \hat{b}).$$

Az egyenletrendszer megoldása:

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

Vegyük észre: lényegében elemenként becsültünk mindent a és b formulájában.

A kapott becslések erősen konzisztensek, tehát $n \to \infty$ esetén

$$\mathrm{Cov}_n(\xi,\eta) \to \mathrm{Cov}(\xi,\eta), \quad \mathrm{corr}_n(\xi,\eta) \to \mathrm{corr}(\xi,\eta), \quad \hat{a} \to a, \quad \hat{b} \to b.$$

Mennyire jó az illeszkedés a regressziós egyeneshez? A teljes populáción az illeszkedés „jóságát" a $\mathrm{corr}^2(\xi,\eta)$ mennyiség számszerűsíti. Becslése a minta alapján: $\text{R-squared} = \mathrm{corr}_n^2(\xi,\eta)$. Tulajdonságai:

- $0 \leq \text{R-squared} \leq 1$,
- Annál jobb az illeszkedés, minél nagyobb az R-squared értéke,
- Ha $\text{R-squared} \leq 0{,}5$, akkor nagyon rossz az illeszkedés a regressziós egyeneshez, nagyok a hibatagok, ezért a modell alapján nem érdemes becsléseket végezni.

Korrelációs teszt

Cél a ξ és az η valószínűségi változó függetlenségének tesztelése összetartozó minták alapján.

- Feltevés: ξ és η együttesen normális eloszlású.
- Nullhipotézis: H_0 : ξ és η függetlenek.
- Próbastatisztika: egy ronda formula, amiben szerepel az empirikus korreláció.

Megjegyzések:

- Ez a teszt nem robusztus a normalitásfeltételre nézve: nagy minta esetén szimmetrikus eloszlásra még alkalmazható, de dőlt eloszlásra ne használjuk!
  Mit alkalmazzunk dőlt eloszlás esetén? Spearman-féle korreláció.
- Mi a jelentősége a függetlenségvizsgálatnak a regressziós modellben? Ha a próba nem veti el ξ és η függetlenségét, akkor nem érdemes regressziót végezni.

Varianciaanalízis

Korábban kétmintás t-próbával teszteltük a várható értékek azonosságát részpopuláción belül. Most több részcsoportot fogunk vizsgálni egyszerre. Véletlenszerűen kiválasztva egy egyedet tekintsünk két változót:

- $\xi =$ az egyed melyik részcsoportba esik (diszkrét változó),
- $\eta =$ egy vizsgált mennyiség (diszkrét vagy folytonos változó).

Modellezzük az η mennyiséget a következő módon: ha az egyed a j. csoportba esik (tehát $\xi = j$), akkor legyen

$$\eta = a_j + \varepsilon = \text{csoporthatás} + \text{egyedi hatás},$$

ahol

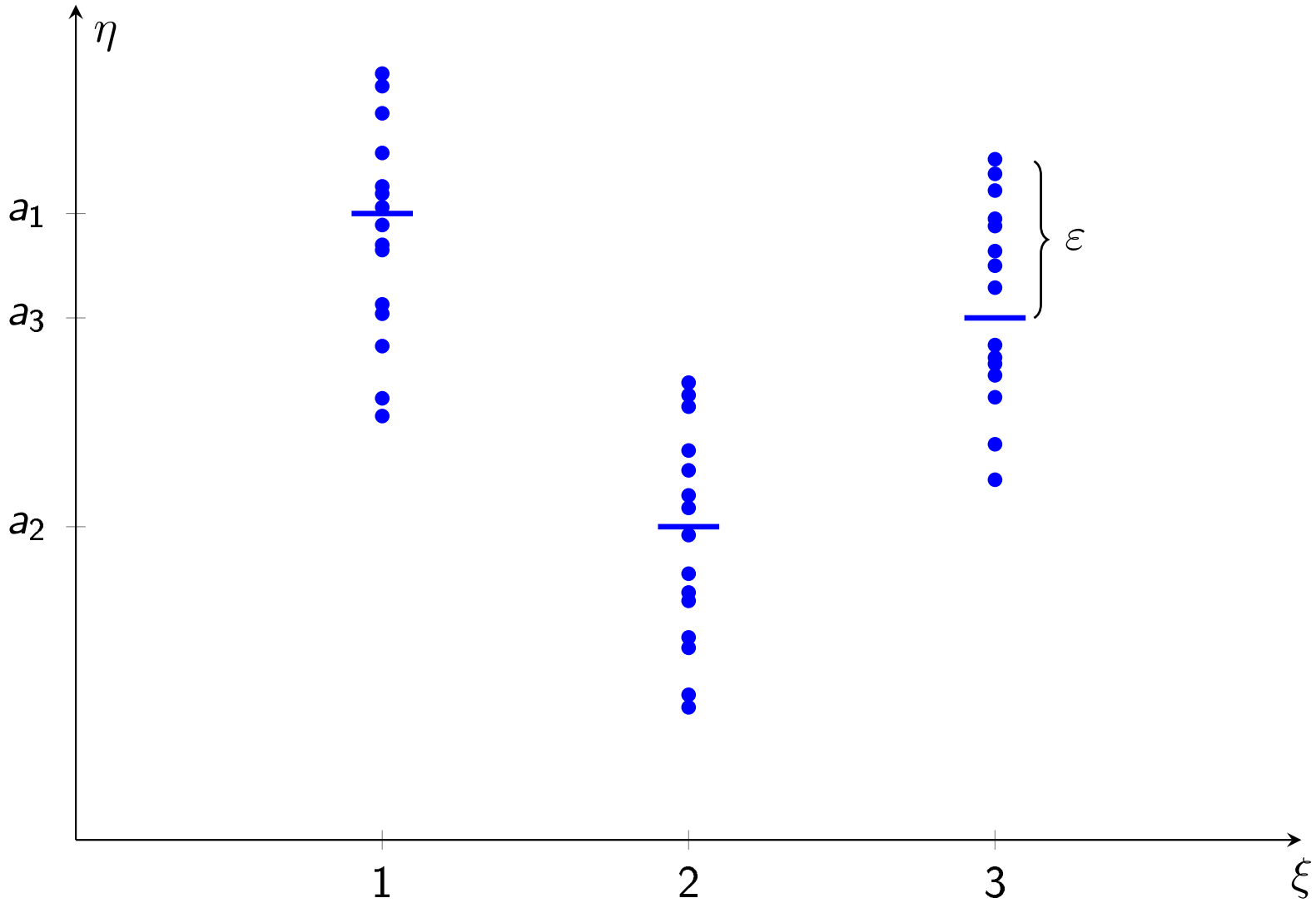
- a_j rögzített valós szám, a j. csoport hatása,
- $\mathrm{E}(\varepsilon) = 0$, és ε független a ξ változótól.

Az η változó átlagos értéke a j. csoportban (feltételes várható érték):

$$\mathrm{E}(\eta \mid \xi = j) = a_j + \mathrm{E}(\varepsilon) = a_j.$$

Tehát a modell:

$$\eta = a_j + \varepsilon = \text{csoporthatás} + \text{egyedi hatás}$$

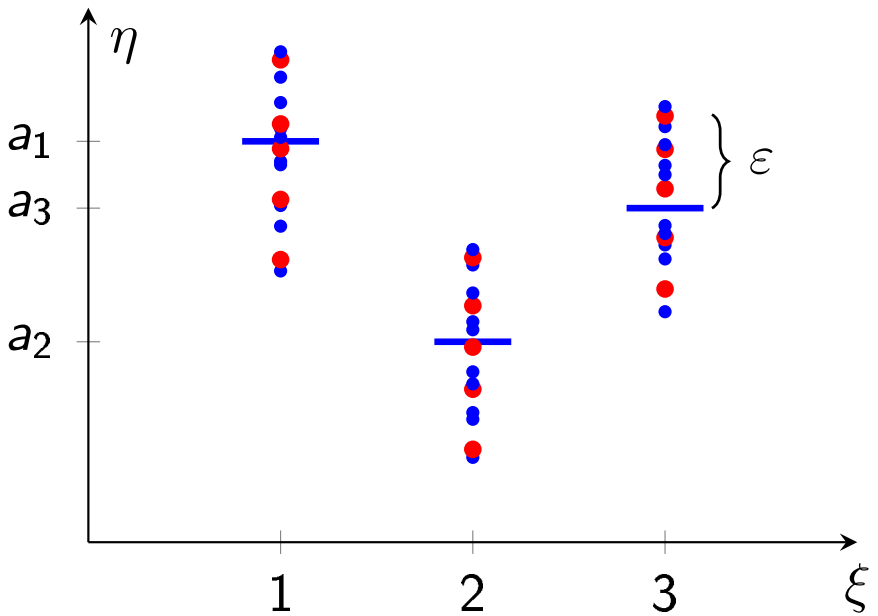


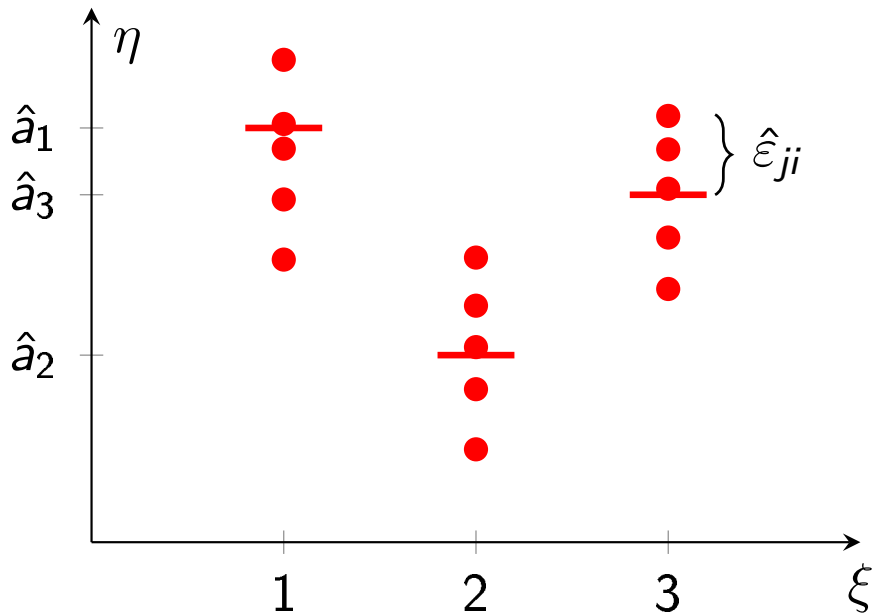
Adjunk becslést a populációátlagra és a csoporthatásokra. A teljes minta:

- minta az 1. részcsoportra: $\eta_{11}, \eta_{12}, \ldots, \eta_{1n_1}$,
- minta a 2. részcsoportra: $\eta_{21}, \eta_{22}, \ldots, \eta_{2n_2}$,
- $\ldots$
- minta az utolsó (r.) részcsoportra: $\eta_{r1}, \eta_{r2}, \ldots, \eta_{rn_r}$.

A teljes minta elemszáma: $n = n_1 + n_2 + \cdots + n_k$. Mintaátlag: $\overline{\eta}$.

Mintaátlag a j. csoportban: $\overline{\eta}_j = (\eta_{j1} + \eta_{j2} + \cdots + \eta_{jn_j})/n_j$.





Legyenek $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_k$ a becslések. A j. csoport egy tetszőleges egyedére vonatkozó **reziduális**:

$$\hat{\varepsilon}_{ji} = \eta_{ji} - \hat{a}_j \approx \eta_{ji} - a_j = \varepsilon_{ji}.$$

A paramétereket a legkisebb négyzetek módszerével fogjuk becsülni:

$$\text{SSW} = \text{a mintaelemek reziduálisainak négyzetösszege}$$
$$= \sum_{j=1}^{r} \sum_{i=1}^{n_j} \hat{\varepsilon}_{ji}^2 = \sum_{j=1}^{r} \sum_{i=1}^{n_j} \left[\eta_{ji} - \hat{a}_j\right]^2 \longrightarrow \min$$

Egy kis számolás után a következő becsléseket kapjuk:

Mennyiség	Elméleti érték	Becslés
j. csoport átlaga	a_j	$\hat{a}_j = \overline{\eta}_j$
Egyedi hatás (hibatag)	$\varepsilon_{ji} = \eta_{ji} - a_j$	$\hat{\varepsilon}_{ji} = \eta_{ji} - \overline{\eta}_j$

Az előző oldalról:

$$\text{SST} = \textbf{sum of squares (total)} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(\eta_{ji} - \overline{\eta})^2$$

$$= \text{milyen mértékben szóródnak az adatok a mintaátlag körül}$$

Ez a szóródás két forrásból származik:

$$\text{SSW} = \textbf{sum of squares (within groups)} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(\eta_{ji} - \overline{\eta}_j)^2$$

$$= \text{milyen mértékben szóródnak az adatok a csoportátlagok körül}$$

$$\text{SSB} = \textbf{sum of squares (between groups)} = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(\overline{\eta}_i - \overline{\eta})^2$$

$$= \text{milyen mértékben szóródnak a csoportátlagok a mintaátlag körül}$$

Megmutatható, hogy $\text{SST} = \text{SSW} + \text{SSB}$.

Varianciaanalízis (Analysis of Variances, ANOVA)

A cél azt tesztelni, hogy minden csoportnak azonos a hatása, tehát a teljes populáción belül azonosak a csoportátlagok.

- Feltevések: az η változó minden csoporton belül normális eloszlást követ, és minden csoporton belül azonos a szórása.
- Nullhipotézis: $H_0 : a_1 = a_2 = \ldots = a_r$.
- Próbastatisztika: $F = \frac{\mathrm{SSB}}{\mathrm{SSW}} \frac{n-r}{r-1}$.

A kapott értékeket az **ANOVA táblázatban** szoktuk összefoglalni:

Effect	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ξ	$r-1$	SSB	$\frac{\mathrm{SSB}}{r-1}$	$\frac{\mathrm{SSB}}{\mathrm{SSW}} \frac{n-r}{r-1}$	p-value
Residuals	$n-r$	SSW	$\frac{\mathrm{SSW}}{n-r}$		
Total	$n-1$	SST			

Ha a csoportonkénti szórások azonosak, akkor $\mathrm{D}(\varepsilon) \approx \mathrm{SSW}/(n-r)$.

Megjegyzések:

- Alapötlet: ha H_0 igaz, akkor $\mathrm{SSB} \approx 0$, tehát $F \approx 0$. Akkor fogadjuk el a nullhipotézist, ha $F \leq c$, ahol c a kritikus érték.
- A teszt robusztus a normalitásfeltételre nézve, de nem robusztus a szórásfeltételre nézve. Ha a szórások nem azonosak, akkor használjuk inkább a Welch-féle ANOVA tesztet.
- Két csoport ($r = 2$) esetén: ANOVA = kétmintás t-próba.
- Azt szeretjük, ha közel ugyanannyi megfigyelés esik minden csoportba.

Milyen módszerrel ellenőrizhető a csoportonkénti szórások egyenlősége?

- Levene-teszt: formálisan kell hozzá a csoportonkénti normalitás, de robusztus erre a feltételre nézve. (Opcióknál: median!)
- Bartlett-teszt: formálisan ehhez is kell a csoportonkénti normalitás, és nem robusztus erre a feltételre nézve. (Ezért kevésbé ajánlott.)
- F-próba: csak $r = 2$ csoport esetén működik, és ez a legérzékenyebb a normalitásra.

Várható érték és szórás tesztelése független minták alapján, illetve a robusztusság a normalitásfeltételre nézve:

Csoportok száma	Várható érték teszt (azonos szórás)	Várható érték teszt (tetszőleges szórás)	Szórás teszt
$r = 2$	Kétmintás t-próba (robusztus)	Welch-próba (robusztus)	F-próba (nem robusztus!!!)
$r \geq 2$	ANOVA (robusztus)	Welch-féle ANOVA (robusztus)	Levene-teszt (robusztus)

Lineáris modellek

A lineáris regresszió és az ANOVA azonos alapötletre épül: egy magyarázó változó megfelelő függvényével modellezzük egy függő változó értékét. Ezt több magyarázó változóval is meg lehet tenni:

- Két (vagy több) magyarázó változó: ξ, ζ.
- Függő változó: η.
- Az egyedre jellemző hibatag: ε, független a magyarázó változóktól.

Lássunk néhány ilyen modellt!

Többszörös lineáris regresszió (multivariate linear regression): Tipikusan akkor alkalmazzuk, ha a magyarázó változók folytonosak, de diszkrét változókra is lehet. A modell:

$$\eta = (a\xi + b\zeta + c) + \varepsilon = \text{predikciós tag} + \text{hibatag}.$$

Például: felnőtt embereket vizsgálunk,

$$\text{testsúly} = a \cdot \text{testmagasság} + b \cdot \text{derékbőség} + c + \varepsilon.$$

Többszempontos varianciaanalízis (multi-factor ANOVA):
A magyarázó változók **faktorok (factors)**, tehát olyan diszkrét változók, melyek részpopulációkat definiálnak. Ha az egyedre $\xi = i$ és $\zeta = j$, akkor

$$\eta = (m + a_i + b_j) + \varepsilon = \text{predikciós tag} + \text{hibatag}.$$

Jelölések:

- m = populációátlag,
- a_i = az első faktor szerint az i-edik részpopuláció hatása,
- b_i = a második faktor szerint a j-edik részpopuláció hatása.

Például: felnőtt embereket vizsgálunk,

- ξ: nem (1=férfi, 2=nő), csoporthatások: a_1, a_2,
- ζ: életmód, szokott-e sportolni (1=soha, 2=időnként, 3=naponta), csoporthatások: b_1, b_2, b_3.

$$\text{testsúly} = \text{populációátlag} + \text{nem hatása} + \text{életmód hatása} + \varepsilon.$$

Kevert modell (mixed model): Akkor alkalmazzuk, ha a magyarázó változók között van faktor és folytonos változó is. Legyen ξ a folytonos változó és ζ a faktor. A modellben ha $\zeta = j$, akkor

$$\eta = (a\xi + b_j + c) + \varepsilon = \text{predikciós tag} + \text{hibatag}.$$

Például: felnőtt embereket vizsgálunk,

$$\text{testsúly} = a \cdot \text{testmagasság} + \text{életmód hatása} + c + \varepsilon.$$

Lineáris modell (linear model): A magyarázó változók tetszőlegesek (folytonosak vagy diszkrétek) lehetnek. A modellben rögzítettek a g és h függvények, melyek segítségével:

$$\eta = a \cdot g(\xi) + b \cdot h(\zeta) + c + \varepsilon.$$

A bemutatott modellek (többszörös lineáris regresszió, többszempontos ANOVA, kevert modell) mind felírhatóak ilyen alakban, tehát ezek speciális esetei a lineáris modellnek.

Valószínűségek becslése és tesztelése

Feladat: Becsüljük meg egy statisztikai minta alapján azon egyedek arányát egy populációban, melyek rendelkeznek egy adott tulajdonsággal.

Válasszunk ki véletlenszerűen egy egyedet a populációból, és legyen:

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

Ekkor: $P(A) = \text{a vizsgált tulajdonság aránya a populációban (ismeretlen)}$

Vezessük be a következő valószínűségi változót:

$$\xi = \begin{cases} 1, & \text{ha a kiválasztott egyed rendelkezik a vizsgált tulajdonsággal,} \\ 0, & \text{ha nem rendelkezik.} \end{cases}$$

Ekkor: $P(\xi = 1) = P(A)$ és $P(\xi = 0) = 1 - P(A)$, tehát

$$\begin{aligned} \mathrm{E}(\xi) &= 0 \cdot P(\xi = 0) + 1 \cdot P(\xi = 1) = 0 + 1 \cdot P(A) = P(A) \\ &= \text{a vizsgált tulajdonság aránya a populációban.} \end{aligned}$$

Tehát igazából egy várható értéket kell megbecsülnünk!

Statisztikai minta: kiválasztunk n egyedet a populációból, és jelölje $\xi_1, \ldots, \xi_n$ a ξ változó értékét a kiválasztott egyedek esetében. Ekkor:

$$\text{a tulajdonság aránya a populációban} = P(A) = \mathsf{E}(\xi) \approx \frac{\xi_1 + \cdots + \xi_n}{n}.$$

Sőt, ez egy erősen konzisztens becslés, tehát $\overline{\xi} \to P(A)$, amint $n \to \infty$.

Gyakoriság, tapasztalati gyakoriság (frequency): $k_A = \xi_1 + \cdots + \xi_n$.
A mintaelemek közül ennyi rendelkezik a vizsgált tulajdonsággal.

Relatív gyakoriság (relative frequency): k_A/n.
A mintaelemek ekkora hányada rendelkezik a vizsgált tulajdonsággal.

A relatív gyakoriság erősen konzisztens becslés: ha $n \to \infty$, akkor

$$k_A/n \to P(A) = \text{a tulajdonság aránya a teljes populációban}$$

Feladat: Megvizsgáltunk 200 japán nemzetiségű embert, közülük 84 esett az A vércsoportba. Adjunk becslést az A vércsoport arányára Japánban!

Az A vércsoport tapasztalati gyakorisága illetve relatív gyakorisága:

$$k_A = 84, \qquad k_A/n = 84/200 = 42\% \approx \text{arány a populációban}.$$

Tudjuk: a tulajdonság aránya a populációban $= P(A) = \mathrm{E}(\xi)$.

A várható értékre vonatkozó módszerek alkalmazásával lehetőség van:

- Konfidencia intervallumot adni a $P(A)$ valószínűségre.
- Tesztelni a $P(A)$ valószínűség értékét. Legyen $p \in [0, 1]$ tetszőleges hipotetikus valószínűség, és tekintsük az alábbi nullhipotézist:

$$H_0 : P(A) = p \qquad \text{tehát} \qquad H_0 : \mathrm{E}(\xi) = p.$$

  Ez a nullhipotézis tesztelhető t-próbával.

FONTOS: Most a ξ háttérváltozó nem normális eloszlást követ, emiatt ezek a módszerek csak nagy mintaméretre működnek. Tipikusan legyen $n \geq 20$, de inkább $n \geq 50$.

Mit tegyünk, ha csak kevés mintaelemünk van?

- Ne használjunk t-próbát!
- Alkalmazzuk a **binomiális próbát**, ugyanis ez tetszőleges n esetén alkalmazható. (Ezt a próbát nem tanuljuk.)

Feladat: Megvizsgáltunk 200 japán nemzetiségű embert, közülük 84 esett az A vércsoportba. Teszteljük azt a nullhipotézist, hogy a japán emberek körében az A vércsoport aránya 40%! Adjunk 95% megbízhatóságú konfidencia intervallumot erre az arányra!

Vezessük be a következő valószínűségi változót:

$$\xi = \begin{cases} 1, & \text{ha a kiválasztott ember az A vércsoportba esik,} \\ 0, & \text{ha nem oda esik.} \end{cases}$$

Nullhipotézis: $H_0 : \mathrm{E}(\xi) = 0{,}4$.

- Statisztikai minta $(\xi_1, \ldots, \xi_{200})$: 84 db 1-es és 116 db 0-ás érték.
- Mintaátlag: $\overline{\xi} = 0{,}42$.
- Korrigált empirikus szórás: $\mathrm{D}^*_n(\xi) = 0{,}495$.
- Standard hiba: $\mathrm{SE} = \mathrm{D}^*_n(\xi)/\sqrt{n} = 0{,}035$.
- Próbastatisztika: $t = (\overline{\xi} - 0{,}4)/\,\mathrm{SE} = 0{,}57$.
- Kritikus érték: $c = \Phi^{-1}_{199}(0{,}975) = 1{,}97$.
- Döntés: $|t| \leq c$, ezért a nullhipotézist elfogadjuk.
- Konfidencia intervallum: $[\overline{\xi} - c\,\mathrm{SE}, \overline{\xi} + c\,\mathrm{SE}] = [0{,}35, 0{,}49]$.

Mit tesztelhetünk még hasonló módszerrel?

- Két vagy több részpopuláción belüli arányok összehasonlítása független minták alapján. Például:
  - $\xi_1, \ldots, \xi_{200}$: 200 japán ember közül ki esik az A vércsoportba (0/1)
  - $\eta_1, \ldots, \eta_{100}$: 100 magyar ember közül ki esik az A vércsoportba (0/1)

  H_0 : a japánoknál és a magyaroknál azonos az A vércsoport aránya

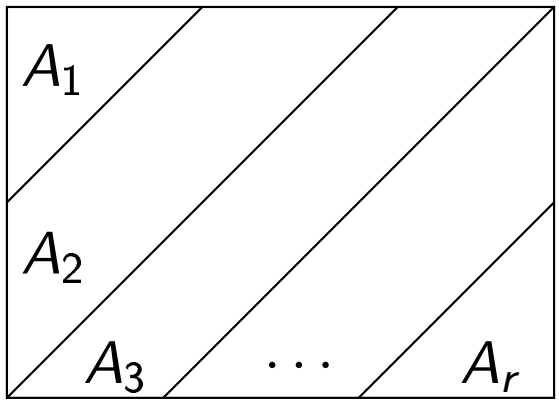
  Tesz: Welch-próba vagy ANOVA
- Egy populáción belüli két arány összehasonlítása összetartozó minták alapján. Például:
  - $\xi_1, \ldots, \xi_{200}$: 200 japán ember közül ki esik az A vércsoportba (0/1)
  - $\eta_1, \ldots, \eta_{200}$: ugyanezen emberek közül ki esik a B vércsoportba (0/1)

  H_0 : a japánoknál azonos az A és a B vércsoport aránya

  Tesz: páros t-próba.

FONTOS: kell a nagy minta! Legalább 50 megfigyelés kell minden egyes változóra. Kis mintaelemszám esetén olyan teszteket kell keresni, melyek speciálisan arányokra vannak kitalálva.

Tegyük fel, hogy a populáció valamely szempont szerint több részcsoportra bontható fel. Ezek a részcsoportok együttesen lefedik az összes egyedet, és minden csoportban van egyed. Feladat: teszteljük a részcsoportok arányát.



Legyen $r \geq 2$ a részcsoportok száma. Véletlenszerűen kiválasztunk egy egyedet, és legyen

$$A_i = \text{a kiválasztott egyed az } i\text{-edik részcsoportba esik}, \qquad i = 1, \ldots, r.$$

Válasszunk ki véletlenszerűen n egyedet a populációból. Ekkor:

- $k_i =$ az i-edik csoport gyakorisága a mintában,
- $k_i/n =$ az i-edik csoport relatív gyakorisága a mintában.

Becslés: $k_i/n \approx P(A_i) =$ az i-edik részcsoport aránya.

Tekintsünk hipotetikus valószínűségeket (részarányokat): $p_1, \ldots, p_m > 0$, ahol $p_1 + \cdots + p_m = 1$. A célunk a következő nullhipotézist tesztelni:

$$H_0 : P(A_i) = p_i \quad \text{minden } i \text{ részcsoport esetén.}$$

Ötlet: ha igaz a nullhipotézis, akkor az n elemű mintában körülbelül np_i olyan egyednek kell lennie, mely az i-edik csoportba tartozik.

χ^2-próba (khinégyzet-próba) valószínűségek tesztelésére

A fenti nullhipotézis tesztelhető a következő próbastatisztikával:

$$\chi^2 = \frac{(k_1 - np_1)^2}{np_1} + \cdots + \frac{(k_r - np_r)^2}{np_r}$$

A kritikus érték a χ^2-eloszlás táblázatából kereshető ki.
Feltétel: nagy minta, $n \geq 5/\min(p_1, \ldots, p_r)$.

Vegyük észre, a fenti összegben minden tag a következő módon áll elő:

$$\frac{(\text{tapasztalati gyakoriság} - \text{várt gyakoriság})^2}{\text{várt gyakoriság}}$$

Feladat: Megvizsgáltunk 200 japán nemzetiségű embert. Közülük rendre 62, 84, 38 illetve 24 esett a 0, az A, a B és az AB vércsoportba. Adjunk becslést a vércsoportok arányára a teljes populáción belül! Teszteljük 90%-os szignifikancia szinten azt a nullhipotézist, hogy a japán emberek körében a vércsoportok aránya rendre 30%, 40%, 20% illetve 10%!

	0	A	B	AB	össz.
Tapasztalati gyakoriság (k_i)	62	84	38	24	200
Relatív gyakoriság (k_i/n)	0,31	0,42	0,19	0,12	1
Hipotetikus valószínűség (p_i)	0,3	0,4	0,2	0,1	1
Várt gyakoriság (np_i)	60	80	40	20	200

Próbastatisztika:

$$\chi^2 = \frac{(62-60)^2}{60} + \frac{(84-80)^2}{80} + \frac{(38-40)^2}{40} + \frac{(24-20)^2}{20} = 1{,}17\,.$$

A kritikus érték a χ^2-eloszlás táblázatából: $c = 6{,}25$.
Most $|\chi^2| \leq c$, tehát a nullhipotézist elfogadjuk.

Függetlenségvizsgálat

Kérdés: milyen módon tesztelhető egy ξ és egy η mennyiség függetlensége?
Válasz: attól függ, hogy milyen változókról van szó...

- Korrelációs teszt a Pearson-féle korrelációs együtthatóval:
  - Csak normális eloszlású változókra.
  - A korrelációs együttható segítségével tényleg a függetlenséget teszteli.
- Korrelációs teszt a Spearman-féle korrelációs együtthatóval:
  - Folytonos eloszlású változókra, nem csak normálisra.
  - A korrelációs együttható segítségével tényleg a függetlenséget teszteli.
- ANOVA, kétmintás t-próba, Welch-próba:
  - Egyik változó diszkrét (csoportokat definiál), a másik folytonos.
  - Azt teszteli, hogy nincs csoporthatás, tehát az η változó várható értékére nem hat a ξ változó. Ez jóval kevesebb, mint a függetlenség!
- χ^2-próba (khinégyzet-próba) függetlenségre:
  - Nem azonos a valószínűségeknél tanult χ^2-próbával!
  - Mindkét változó diszkrét, a függetlenséget teszteli.
  - Kell a nagy mintaméret!

Adott két mennyiség a populációban, ξ és η. Összetartozó minták:

- $\xi_1, \ldots, \xi_n$: a ξ mennyiség értéke n kiválasztott egyednél,
- $\eta_1, \ldots, \eta_n$: az η mennyiség értéke ugyanezen egyedeknél.

H_0: a két mennyiség független egymástól.
Tesztelési módszer a változók típusa szerint:

	η diszkrét	η normális	η folytonos
ξ diszkrét	χ^2-próba (nagy minta!)	ANOVA	ANOVA (nagy minta!)
ξ normális	ANOVA	Pearson-korreláció	Spearman-korreláció (nagy minta!)
ξ folytonos	ANOVA (nagy minta!)	Spearman-korreláció (nagy minta!)	Spearman-korreláció (nagy minta!)

Illeszkedésvizsgálat

Feladat: becsüljük meg és teszteljük le egy ξ mennyiség eloszlását a teljes populációban a $\xi_1, \ldots, \xi_n$ megfigyelések alapján!

Legyen ξ **diszkrét**, és legyen $R_\xi = \{x_1, \ldots, x_k\}$ az értékkészlete.
Feladat: becsüljük meg és teszteljük a $P(\xi = x_i)$ valószínűségeket.
Becslés:

$$P(\xi = x_i) \approx \text{relatív gyakoriság} = \text{az } x_i \text{ érték aránya a mintában}$$

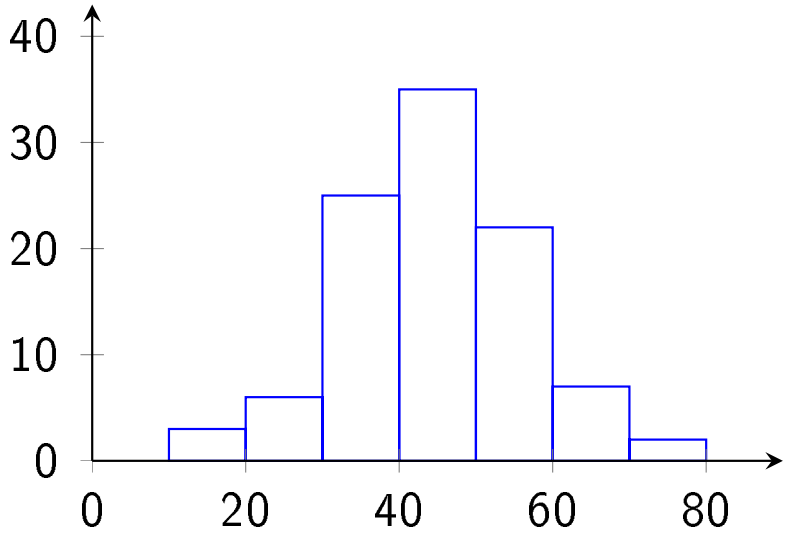
Tesztelés: adottak $p_1, \ldots, p_k > 0$ hipotetikus arányok, $p_1 + \cdots + p_k = 1$.

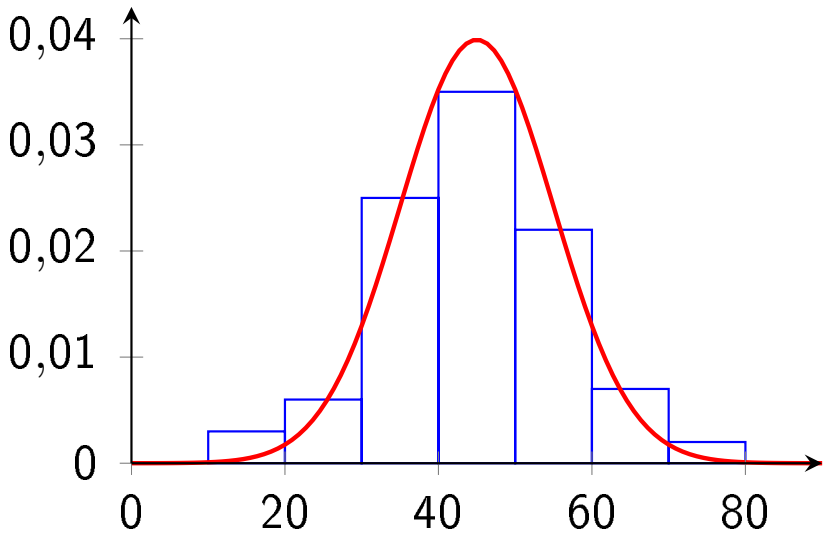
$$H_0 : P(\xi = x_i) = p_i \quad \text{minden } i\text{-re}$$

Tesztelési módszer: χ^2-próba a valószínűségekre. Probléma megoldva.

A továbbiakban csak azzal az esettel foglalkozunk, amikor ξ folytonos eloszlású egy ismeretlen f_ξ sűrűségfüggvénnyel.

Sűrűségfüggvény becslése **hisztogrammal**: bontsuk fel a számegyenest azonos (mondjuk $h > 0$) hosszúságú intervallumokra. Minden intervallumra állítsunk egy olyan magas oszlopot, ahány elem esik az adott intervallumba. Az így kapott hisztogrammnak még nincs sok köze a sűrűségfüggvényhez.





Osszuk minden oszlop magasságát nh-val! Az így kapott új hisztogramm összterülete pontosan 1 lesz.

Tétel. Ha $n \to \infty$ és $h \to 0$, akkor az átskálázott hisztogramm konvergál az ismeretlen sűrűségfüggvényhez. Ilyen módon grafikus becslést adhatunk a sűrűségfüggvényre.

A legtöbb statisztikai program nem olyan módon ábrázolja a boxplotot, ahogyan azt korábban tanultuk. Általában felmérnek a doboz aljára és tetejére $1.5 \cdot \text{IQR}$ távolságot, és a bajúszt csak eddig ábrázolják. Az ezen kívül eső megfigyeléseket **outlier** értékeknek nevezzük, ezek egyesével vannak ábrázolva a boxploton.

Normalitásvizsgálat **boxplot** segítségével: amennyiben a minta normális eloszlásból jön, akkor a boxplotnak két speciális tulajdonsága van:

- A boxplot körülbelül szimmetrikus a mediánra.
- Az outlier értékek a teljes minta legfeljebb 1%-át teszik ki. Az ennél több outlier arra utal, hogy a minta nem normális eloszlásból jön. Ezt a tulajdonságot jól szemlélteti ez a Wikipedia ábra.

Mi lehet még az oka a sok outlier értéknek? Mérési hibák, jegyzőkönyvezési hibák, stb. Az adatelemzés során az outlier értékeket külön-külön meg szokták vizgálni, nem hibából származnak-e.

Rendezzük növekvő sorrendbe a megfigyelt értékeket: $\xi_1^* \leq \ldots \leq \xi_n^*$.
Ezt nevezzük **rendezett mintának**. Legyen q_α a ξ változó α-kvantilise.

Tétel. Nagy mintaelemszám esetén $\xi_i^* \approx q_{i/(n+1)}$ minden i-re.

Q-Q plot: Koordináta-rendszerben ábrázoljuk a $(q_{i/(n+1)}, \xi_i^*)$ pontokat.

Eloszlásvizsgálat Q-Q plot segítségével: kíváncsiak vagyunk arra, hogy a minta egy adott (például normális) eloszlásból származik-e.

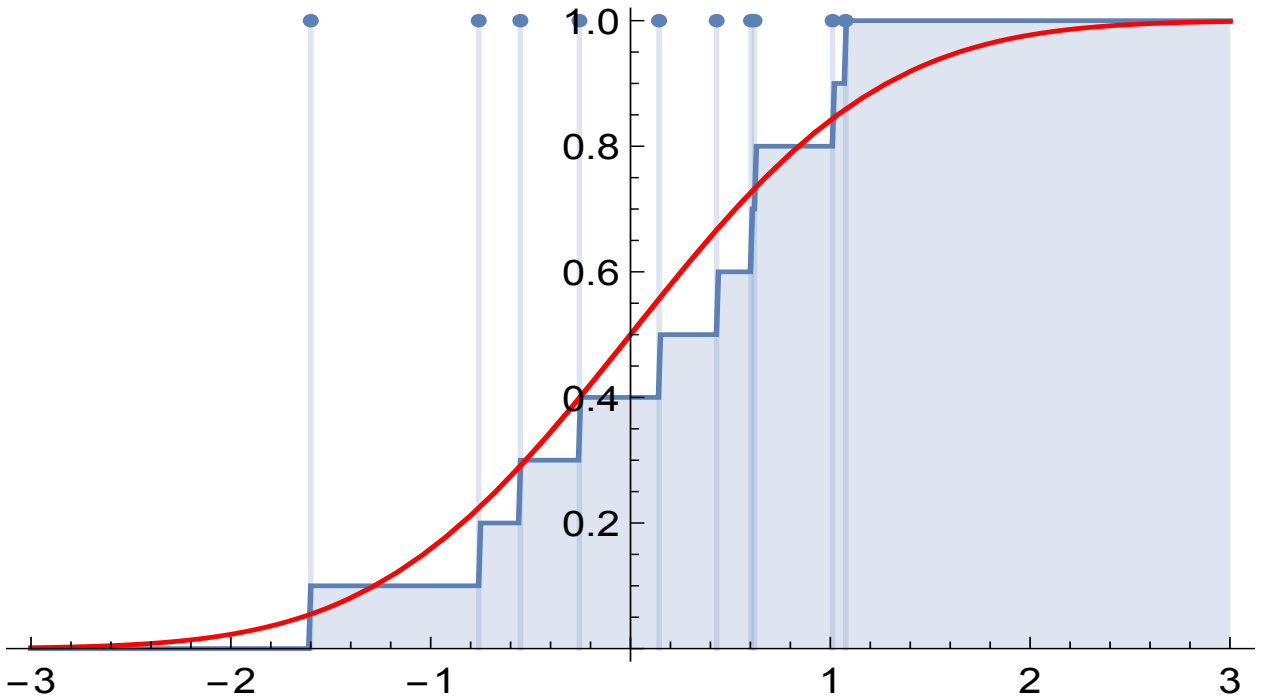
- Kiszámoljuk a kérdéses eloszlás kvantiliseit: $q_{1/(n+1)}, \ldots, q_{n/(n+1)}$.
- Ábrázoljuk a Q-Q plotot.
- Ha a minta a kérdéses eloszlásból származik, akkor $\xi_i^* \approx q_{i/(n+1)}$ minden i-re, tehát minden pont az $x = y$ egyenes közelébe esik.
- Ha a minta nem a kérdéses eloszlásból jön, akkor $\xi_i^* \not\approx q_{i/(n+1)}$ bizonyos megfigyelésekre, tehát egyes pontok nem illeszkednek az $x = y$ egyeneshez.
- Hátrány: nem egzakt módszer, hanem szubjektív döntés.
- Előny: azt is le lehet olvasni az ábráról, hogy a minta mely érték-tartományban milyen mértékben illeszkedik a kérdéses eloszláshoz.

Empirikus eloszlásfüggvény (sample distribution function):

$F_n(x) =$ az x értéknél kisebb elemek relatív gyakorisága a mintában

Tulajdonságok:

- Minden x valós számra $F_n(x) \in [0, 1]$.
- Az F_n függvény egy lépcsős függvény. A mintaelemeknél van ugrása, és minden mintaelemnél $1/n$ az ugrás nagysága.
- Nagy mintaméret esetén $F(x) = P(\xi < x) \approx F_n(x)$.



A matematikai statisztika alaptétele: Jelölje F_ξ a minta valódi eloszlásfüggvényét. Ekkor

$$\max_{x\in\mathbb{R}} \left|F_n(x) - F_\xi(x)\right| \to 0, \qquad n \to \infty.$$

A tétel következményei:

- A matematikai statisztikának van értelme: a mintában van elég információ ahhoz, hogy mindent meg tudjunk becsülni.
- Tesztelhetjük az eloszlásfüggvényt.

Kolmogorov–Szmirnov-próba

Cél: teszteljük egy ξ változó eloszlásfüggvényét egy minta alapján. Legyen F_ξ az igazi eloszlásfüggvény (ismeretlen) és F_0 egy tetszőleges hipotetikus eloszlásfüggvény. Nullhipotézis: $H_0 : F_\xi = F_0$. Próbastatisztika:

$$D_n = \max_{x\in\mathbb{R}} \left|F_n(x) - F_0(x)\right|.$$

Akkor fogadjuk el a nullhipotézist, ha $D_n \approx 0$.

A Kolmogorov–Szmirnov-próba tulajdonsága:

- Csak konkrét eloszlást lehet vele tesztelni. Például:
  H_0 : a ξ változó normális eloszlású 0 várható értékkel és 1 szórással.
- Tetszőleges eloszlás esetén alkalmazható, de főleg folytonos változókra szokták alkalmazni.
- Hátránya: kis mintaelemszám esetén alacsony az ereje, a hamis nullhipotéziseket is elfogadja.

Cél a normalitás tesztelése:

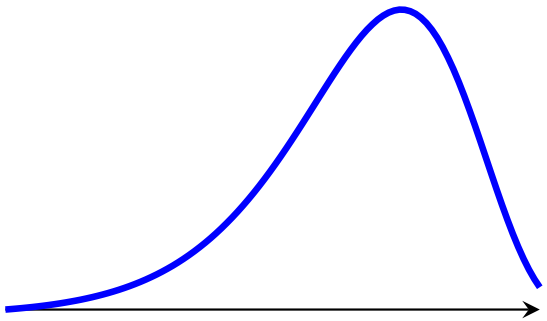
H_0 : a ξ változó normális eloszlást követ valamilyen paraméterekkel

Fontosabb tesztelési módszerek:

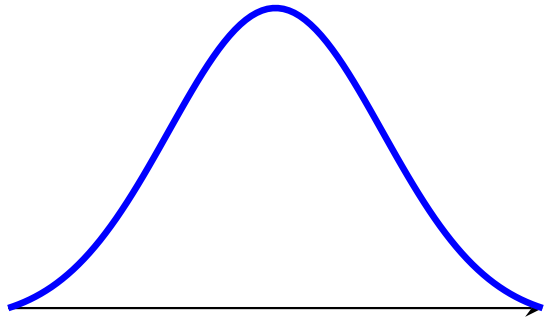
- **Lilliefors-teszt:** Becslést végez a várható értékre és a szórásra a minta alapján, majd azt teszteli a K–Sz-próbával, hogy a minta normális eloszlást követ $\mu = \mathrm{E}_n(\xi)$ és $\sigma = \mathrm{D}^*_n(\xi)$ paraméterekkel.
- **Shapiro–Wilk-teszt:** Sokak szerint ez a legjobb normalitásteszt, kis mintaelemszám esetén is magas az ereje.

Skewness (ferdeség): egy olyan statisztikai mutatószám, mely a minta alapján jellemzi a sűrűségfüggvény szimmetriáját. Tulajdonságai:

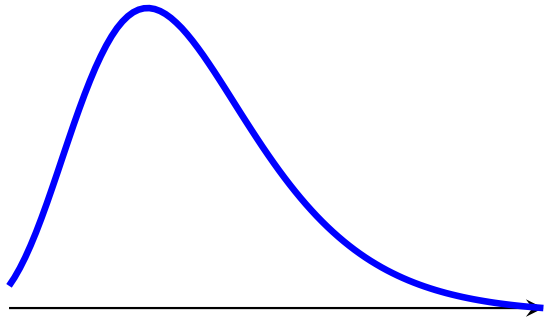
- skewness ≈ 0: (közel) szimmetrikus sűrűségfüggvény
- skewness > 0: jobbra ferde sűrűségfüggvény
- skewness < 0: balra ferde sűrűségfüggvény



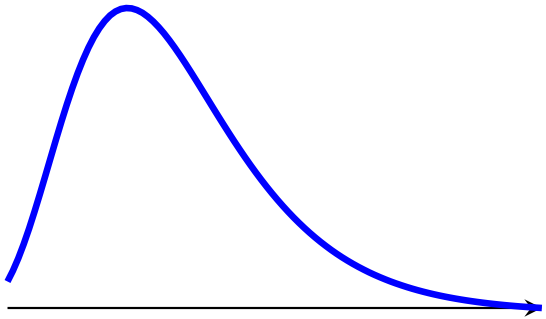
skewness $= -1$



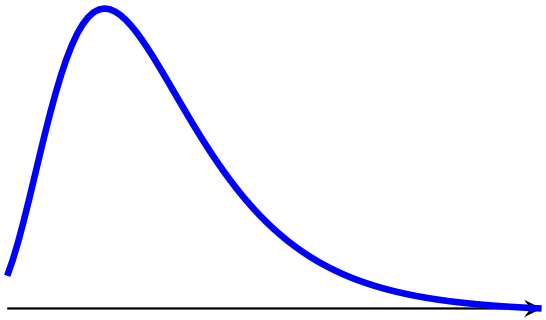
skewness $= 0$



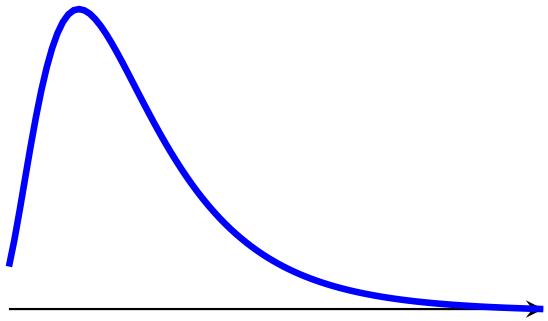
skewness $= 1$



skewness $= 1.2$



skewness $= 1.5$



skewness $= 2$

Néhány további gondolat

Az adatelemzés akkor hatékony, ha a statisztiai minta jól reprezentálja a teljes populációt. Hogyan kaphatunk **reprezentatív mintát**?

Véletlenszerű mintavételezés: véletlenszerűen kiválasztunk egyedeket a teljes populációból.

- Előny: egyszerűen és olcsón megvalósítható (biztos?)
- Hátrány: a reprezentativitáshoz nagy mintaméretre lehet szükség.

Irányítottan összeállított minta: a teljes populáción belüli arányokat figyelembe véve magunk állítunk össze egy mintát.

- Előny: kisebb mintaméter, mint véletlenszerű mintavételezésnél.
- Hátrány: előzetes ismeretekre van szükség a populációról; bonyolult és gyakran drága.
- Gyakran súlyozást alkalmaznak a populációarányok reprezentálásához.

Egyes statisztika programok (SPSS, R Commander) típusokba sorolják a változókat, és csak azokat az eljárásokat engedik futtatni, amik megfelelnek az adott típusnak. Milyen típusokról tanultunk eddig:

- Folytonos változó: mindig valós szám az értéke, az értékkészlete egy intervallumon.
- Diszkrét változó: valós szám vagy szöveg is lehet az értéke, de az értékkészlete véges.

Még egy kifejezés:

- Faktor: csoportokat definiáló változó. Mindig diszkrét.

Milyen típusokba sorolján egyes programok a változókat:

- Skálaváltozó: értelmezhetőek a matematikai műveletek (összeadás, átlagolás). Például: testmagasság, utódok száma, vizsgajegy(?).
- Ordinális változó: nem értelmezhetőek a matematikai műveletek, de van rendezés az értékek között. Például: rendfokozatok, ordinális skálák.
- Nominális változó: nem értelmezhetőek a matematikai műveletek és rendezés sincs az értékek között. Például: nem, nemzetiség.