Normális eloszlás

Minta és populáció

Hipotézisek tesztelésénél egy populációról szeretnénk valamit állítani. De legtöbbször csak egy minta áll rendelkezésünkre.

Eljárás: a minta eloszlásából következtetünk a populáció eloszlására.

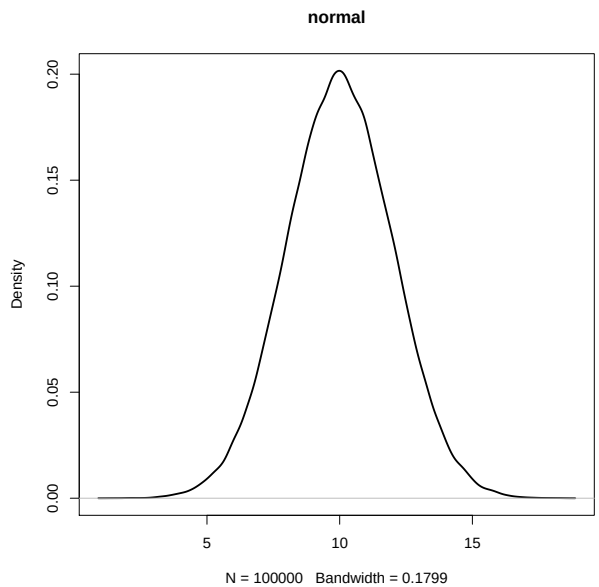
Alapfeltétel: az eloszlásról matematikai ismeretekkel kell rendelkeznünk.

Hipotézis tesztelése valószínűség alapján: mennyire valószínű, hogy egy adott minta illeszkedik a feltételezett eloszláshoz?

Normális eloszlás

A normális eloszlásról pontos ismereteink vannak.

- Egy módusza van, ami nagyjából az eloszlás közepén található, megegyezik a mediánnal és az átlaggal,
- az értékek onnan mindkét irányban szimmetrikus csökkennek,
- megközelítőleg harang formájú (Gauß-görbe).
- aszimptotikus (0-hoz közelít).



Hol fordul elő a normális eloszlás?

A legtöbb folytonos változó normális eloszlású, ha a minta elég nagy.

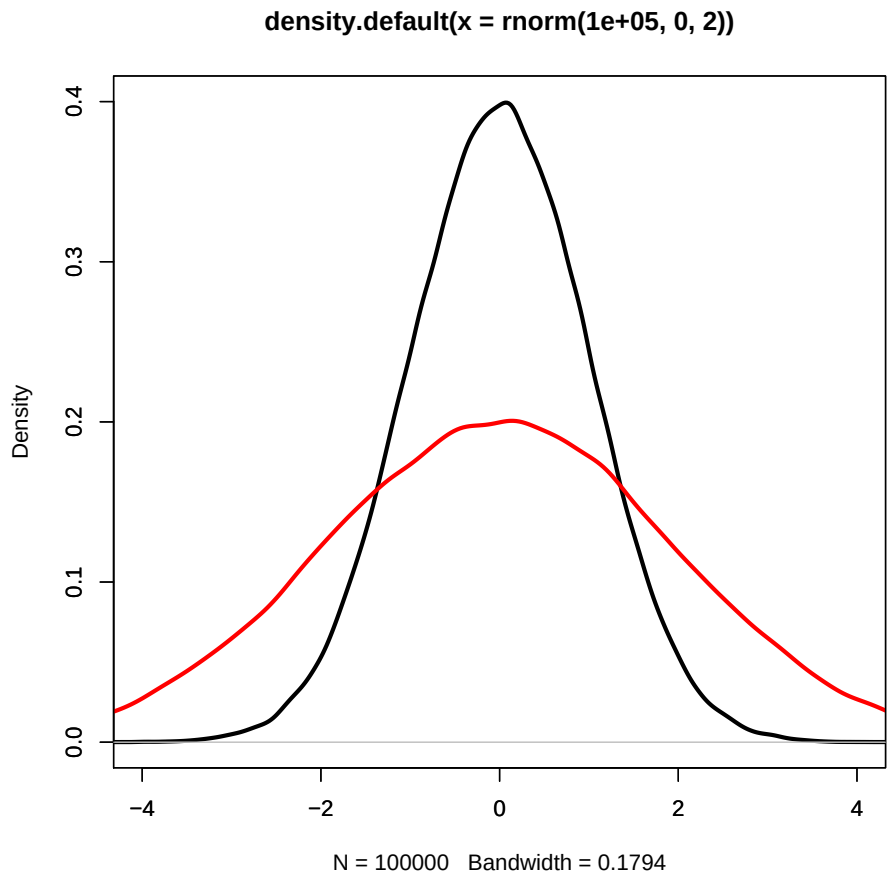
- IQ – az átlag önkényesen rögzített érték! (Mentális életkor/valós életkor*100.)
- Emberek magassága és súlya nemenként.
- Egy adott tóból kifogott halak tömege és hossza.
- Háztartások éves bevétele egy adott országban.
- Klíma, pl. minimum és maximum hőmérsékletek július 15-én 100 éven keresztül.

Hol nem találunk normális eloszlást?

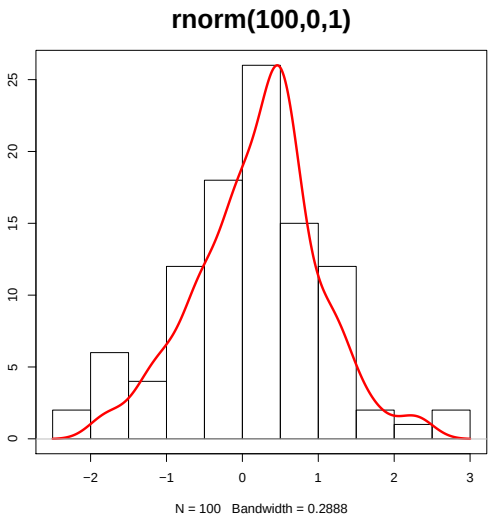
- Pénzügyi mutatók és gazdasági adatok.
- Árváltozások, hozamok, tőzsdei értékek, részvények, árfolyamok.
- Emberek élettartama.
- Műszaki és elektronikus termékek élettartama.
- Várakozási idő sorban állva.
- Biztonsági adatok (pl. autóbalesetek).

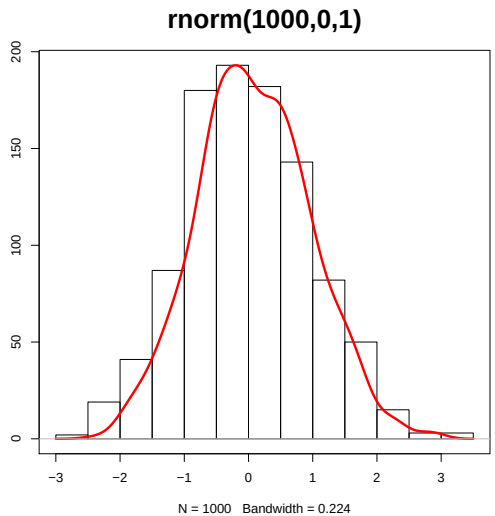
A normális eloszlást az átlaggal és a szórással jellemezzük.

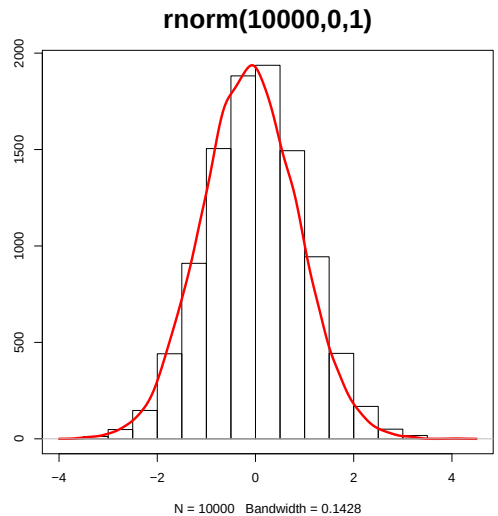
Az átlag a görbe maximuma, a szórás pedig a meredeksége. A nagyobb szórás laposabb görbét eredményez, azaz az értékek nagyobb része esik messze az átlagtól.



Függ n-től – az elemszámtól, és k-tól – a kategóriák számától, azaz hogy milyen pontossággal adjuk meg az értékeket, pl. 1 mm vagy 1000 mm = 1 m távolságot mérünk-e.
Minél nagyobb az elemszám, a minta eloszlása annál jobban közelíti a populáció normális eloszlását. Itt: $n = 100, 1000, 10000$.







A fenti ábrák előállítása:

100 random szám húzása normális eloszlásból, , $\bar{x} = 0$ átlaggal, $s = 1$ szórással.
`a = rnorm(100,0,1)`

Hisztogramm: egy adott érték milyen gyakran fordul elő. A leggyakoribb értékek az átlag körül helyezkednek el. Minél távolabb vannak tőle, annál ritkábbak.
`hist(a)`

Sűrűségfüggvény: folytonossá alakított interpolált érték. Pl. ha méterekben mérünk, a sűrűségfüggvény az egész számok közötti értékekre is értelmezhető.
`plot(density(a))`

N paraméterei

$\bar{x}$: minta átlaga, s: minta szórása.
Minket csak az érdekel, hogy az egyes értékek milyen messze vannak az átlagtól, az nem, hogy melyik irányban. Ezért kiszámoljuk az egyes értékek távolságát az átlagtól, majd a távolságokat egyenként négyzetre emeljük, és összeadjuk. Így kapjuk meg a varianciát. A szórás a variancia gyöke.

minta átlaga:

$$\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

minta varianciája:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}$$

minta szórása:

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}}$$

z-transzformáció

Minket azonban nem a minta érdekel, hanem a populáció. Ennek paraméterei:
μ: populáció feltételezett átlaga.
σ: populáció feltételezett szórása.
Probléma: a konkrét minta szórása függ az elemszámtól és az átlagtól $\Rightarrow$ eltérő átlagok eloszlása nem összehasonlítható.
Trükk: standardizálás z-értékre. Feltételezzük, hogy a mintánk átlaga azonos a populáció átlagával. Minden egyes elemre:

$$z_i = \frac{x_i - \mu}{\sigma}$$

azaz minden egyes elem mérőszámát kivonjuk a populáció (itt: minta) átlagából, és elosztjuk a szórással.

Ez az eljárás a z-transzformáció, az értékek eloszlása standard normális eloszlást mutat.

Standard normális eloszlás

Normális eloszlás jellemzői: $N(\mu, \sigma)$.

átlag: $x_i = \mu = \bar{x}$
z-transzformáció:

$$z_i = \frac{\mu - \mu}{\sigma} = 0$$

szórás: $\sigma = \mu + \sigma$

$$z_i = \frac{(\mu + \sigma) - \mu}{\sigma} = 1$$

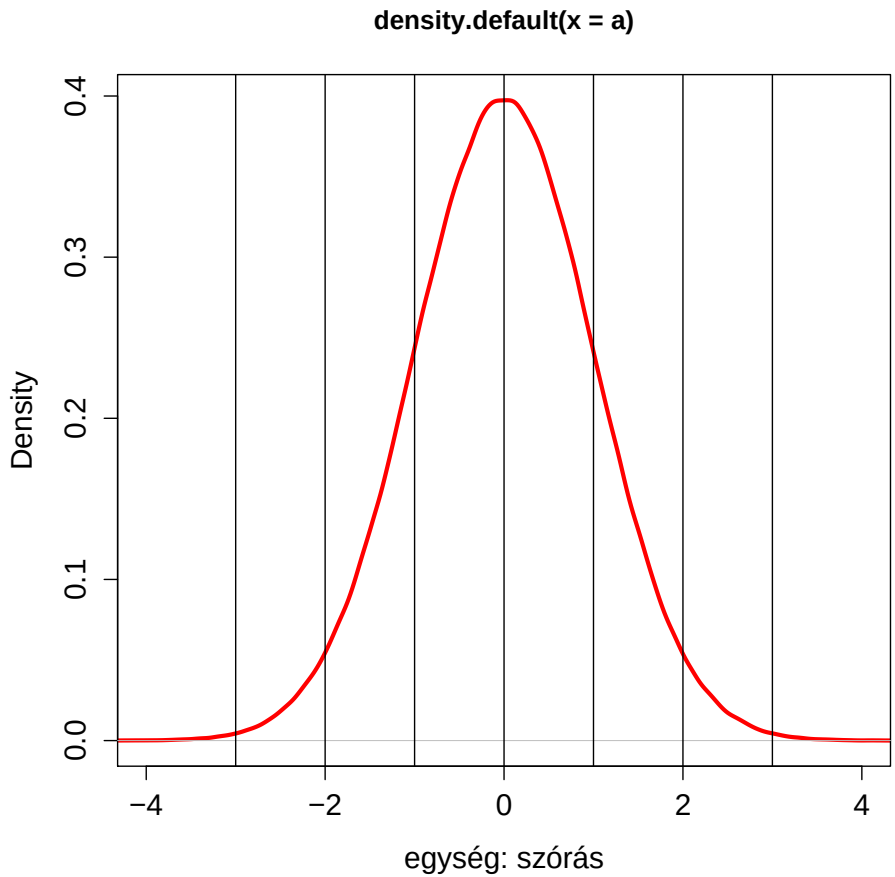
Standard normális eloszlás jellemzése: $N(0, 1)$. Vagyis az átlaga mindig 0, a szórása mindig 1. Előny: bármilyen normál eloszlású minta átalakítható standard normális eloszlásra, aminek az alakját pontosan ismerjük.

Standard normális eloszlás

Sűrűségfüggvény:

x-tengely: egységnyi szórás, tartomány: $\sigma = -\infty \cdots + \infty$.

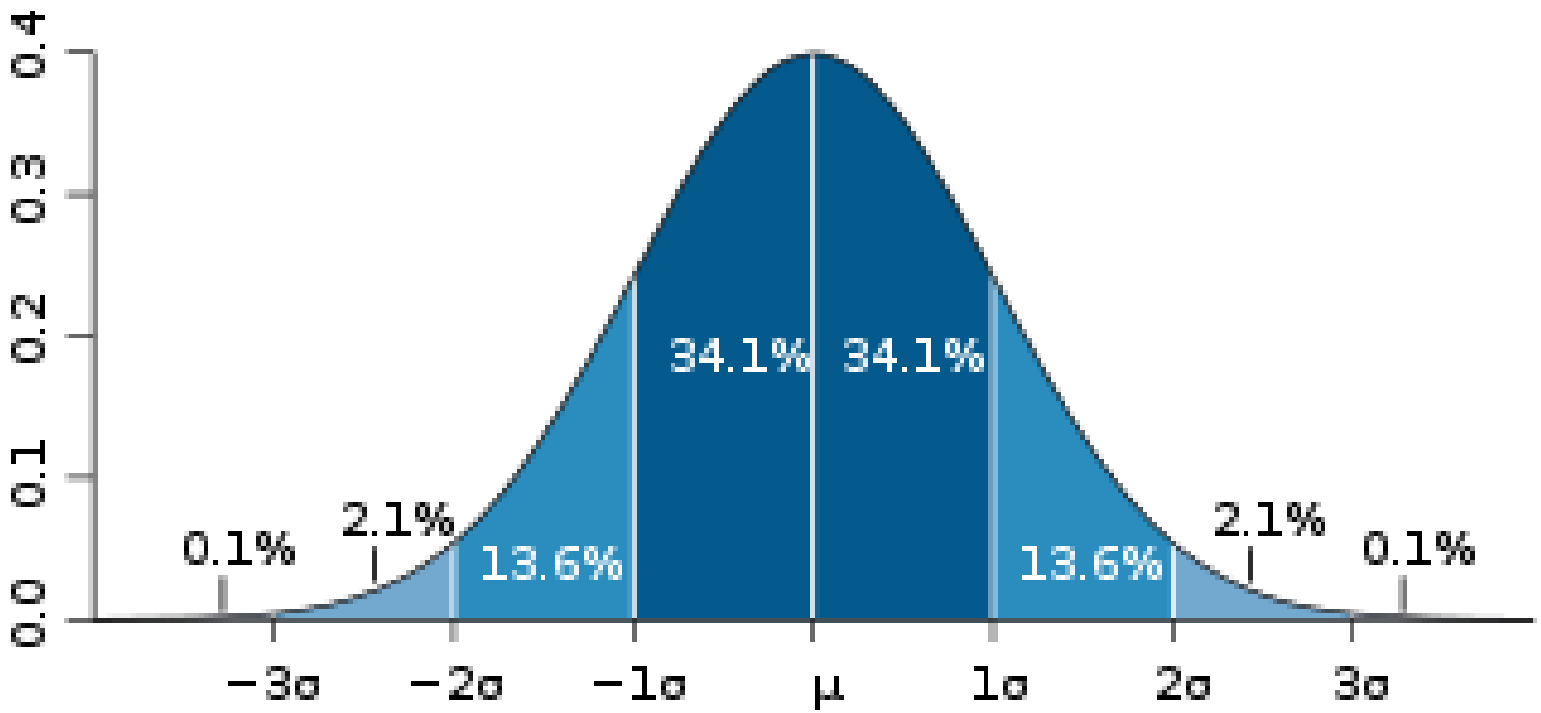
y-tengely: annak a valószínűsége, hogy egy adott x (a transzformáció után z) előfordul a populációban.



A standard normális eloszlás függvényének jellemzői

- az x-tengely és a sűrűségfüggvény által bezárt terület összege $= 1$.
- Az esetek 50%-a az átlagtól balra helyezkedik el.
- $\sigma = -1 \cdots + 1$ közötti tartomány az esetek 68,27%-át tartalmazza.
- $\sigma = -2 \cdots + 2$ közötti tartomány az esetek 95,45%-át tartalmazza.
- $\sigma = -3 \cdots + 3$ közötti tartomány az esetek 99,73%-át tartalmazza.

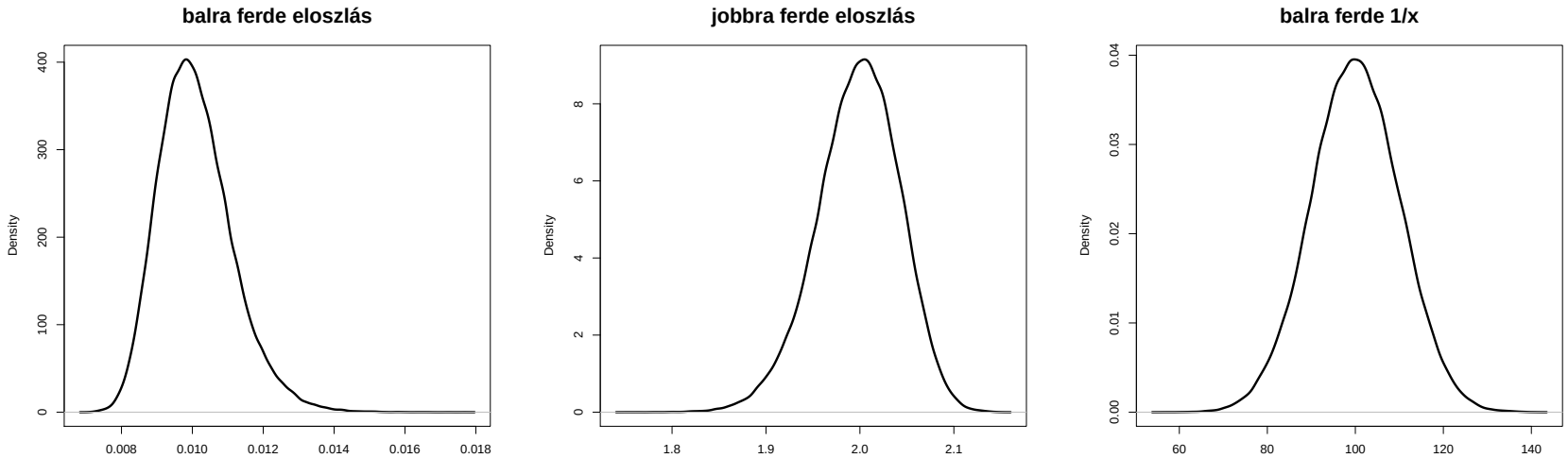
Standard normális eloszlás



Normális eloszlás tesztje

Kolmogorov-Szmirnov vagy Wilk-Shapiro próba.

R-funkció: `shapiro.test(vektor)`



Ha $p > 0,05$, elfogadjuk, hogy a minta normális eloszlású (ld. később).

Transzformációk

Unimodális, jobbra vagy balra ferde eloszlások gyakran átalakíthatóak normális eloszlásúvá.

Szokásos eljárások:

- $x = log(x)$
- $x = 1/x$
- $x = \sqrt{x}$
- ...

Becslés

Inferenciális statisztika: oksági vagy relációs alapú statisztika. A minta értékei alapján következtet a populációra.

DE: a minta alapján a populációra csak **becsléseket** tehetünk.

1. probléma: különböző minták különböző átlagokat eredményeznek, még véletlenszerű kiválasztás esetén is.

2. probléma: véges az időnk, csak véges mintával tudunk dolgozni.

Egy mintából következtetés sok mintára

Feltételezzük, hogy a mintánk átlaga megegyezik a populációéval. Továbbá feltételezzük, hogy ha sok mintánk lenne, akkor az egyes minták átlagai normális eloszlást mutatnának a populáció átlaga körül. Vagyis:

n minta $\bar{x}$ átlagai normális eloszlást mutatnak a populáció μ átlaga körül, ha a populáció szórása σ.

A mintaátlagok μ körüli szórása egyenlő a standard hibával, azaz $se = \frac{s}{\sqrt{n}}$

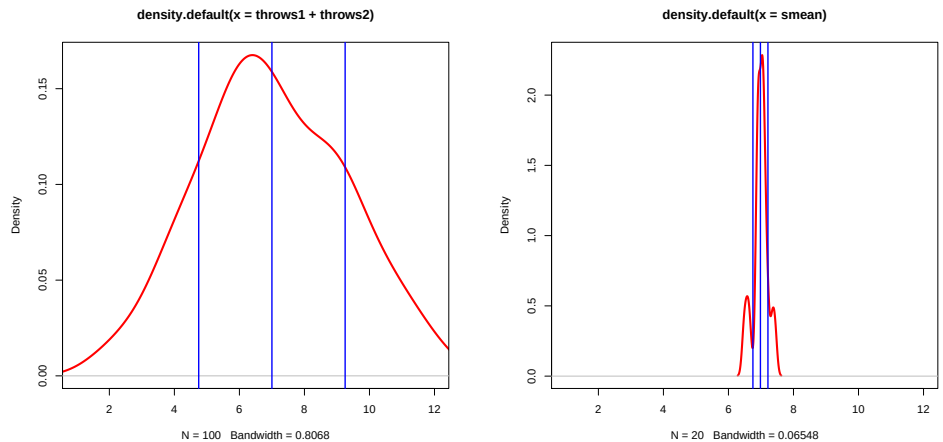
A szórás egyes adatpontok mintaátlagtól való távolságát fejezi ki.

A standard hiba a mintaátlagok populációátlagtól való távolságát fejezi ki. Ez az érték értelemszerűen sokkal kisebb, mint az egyes minták szórása.

Előny: egyetlen minta átlaga és szórása alapján következtethetünk a populáció ismeretlen értékeire.

A minta és a populáció szórása

Két kockával dobunk 100-szor.
Bal: egyetlen, 100 dobásból álló minta összegei. Jobb: 20 minta átlagai, amelyek egyenként 100 dobásból állnak.



Minta átlaga $= 7$, átlagok átlaga $= 6{,}98$.
Minta szórása $= 2{,}25$, átlagok szórása $= 0{,}23$.

Standard hiba EGYETLEN mintából számolva:
$2,25/\sqrt{100} = 0,225$
$\rightarrow$ a 20 mintából számolt szórás jó közelítése 20 minta híján is.

Pontbecslés

Véletlen minta átlaga függ a véletlentől, azaz egy **becsült pont**.

Megmérjük egy véletlenszerűen kiválasztott, 300 fős, férfi egyetemistából álló csoport testmagasságát.
s = 6,3 cm

A minta részmintáiból számolt átlagok szórása függ az elemszámtól: a tíz fős minták szórása a minta átlaga körül $se = 6,3/\sqrt{10} = 1,99$, ötven fős mintáké $se = 6,3/\sqrt{50} = 0,89$, stb.

$\Rightarrow$ minél nagyobb az elemszám, annál kisebb a szórás, azaz az egyes mintaátlagok annál jobban közelítik a populáció átlagát.

Feladat

Fájl letöltése:

http://clara.nytud.hu/~mady/courses/statistics/materials/testmagassag.txt

300 férfi egyetemi hallgató testmagassága.

Testmagasság adatai testmagassag.txt nevű fájlban.
átlag: `mean()`
szórás: `sd()`
gyök: `sqrt()`

Hogyan számoljuk ki az első tíz fő testmagasságának átlagát?
`mean(testmagassag$height[1:10])`

Teljes minta standard hibája?
`sd(testmagassag)/sqrt(300)`
0,36

Konfidenciaintervallum

Kérdés: igaz-e, hogy a véletlen minta átlaga beleesik az ismeretlen populációátlag körül szóródó mintaátlagokba?

Nehézség: μ-t nem ismerjük, csak $\bar{x}$-et.
$\Rightarrow$ döntés nem lehetséges, csak egy adott valószínűségi határon, azaz **konfidenciaintervallumon** belüli valószínűség megállapítása.

Kérdés: igaz-e, hogy $\bar{x}$ 95%-os valószínűséggel beleesik a μ körül standard hibával szóródó mintaátlagok tartományába?

Konfidenciaszint ebben az esetben: $p = 0,95$.

Kiindulás

- Véletlenszerű minták átlagai normális eloszlásúak.
- Átlagok 95%-a $\pm$ 1,96*szórás (s), itt s/$\sqrt{n}$, azaz 1,96*standard hiba (se).
- Keresett μ a populáció eloszlásának középpontja (szimmetria feltételezése miatt).

tehát:

$p(-1,96 * se + \mu < \bar{x} < \mu + 1,96 * se) = 0,95$

Cél: a 95%-os konfidenciaintervallumon belüli határértékek meghatározása negatív és pozitív irányban.

Konfidenciaintervallum $\bar{x}$ alapján

$$p(-1,96 * se + \mu < \bar{x} < \mu + 1,96 * se) = 0,95$$

$$-\mu$$

$$p(-1,96 * se < \bar{x} - \mu < 1,96 * se) = 0,95$$

$$* -1$$

$$p(1,96 * se > \mu - \bar{x} > -1,96 * se) = 0,95$$

$$+\bar{x}$$

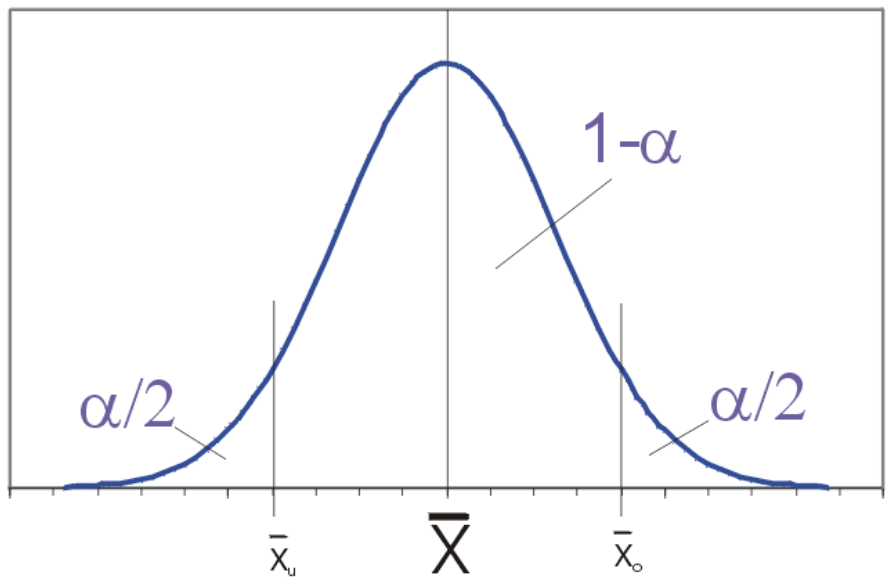
$$p(1,96 * se + \bar{x} > \mu > \bar{x} - 1,96 * se) = 0,95$$

$$p(-1,96 * se + \bar{x} < \mu < \bar{x} + 1,96 * se) = 0,95$$

Konfidenciaszint

Konfidenciaintervallum: értéktartomány, amely a becsülendő paramétert előre rögzített valószínűséggel tartalmazza.

Konfidenciaintervallumon kívüli tartomány: $\alpha = 1 - p$.



Ha $\bar{x}$ nem esik a 95%-os konfidenciaintervallumba, akkor is tartozhat az adott populációhoz! Tévedés valószínűsége 5%, ez az ún. alfa-hiba.

Kiindulási hipotézis tesztelése

Hipotézis állítása falszifikáción keresztül, azaz az állításunk **ellenhipotézisét** teszteljük.

Az empirikus vizsgálatokban általában abban vagyunk érdekeltek, hogy vizsgált érték $1 - p$, azaz α tartományba essen.
$\Rightarrow$ szignifikanciaszintet α értékével szokás megadni, azaz 0,05 vagy 5%.

Ha azt akarjuk bizonyítani, hogy egy adott minta nem tartozik az adott p konfidenciaintervallumba, akkor a mintának negatív és pozitív irányban az $\alpha/2$ tartományba kell tartoznia. Tehát egy szimmetrikus, azaz kétoldalas tesztnél az azonosság elutasítása 2,5%-ra teljesül.

A mintaméret jelentősége a hipotézistesztelés szempontjából

A $p = 0,95$-es konfidenciahatár kritikus értékei:

alsó kritikus érték: $-1.96 * se + \bar{x}$

felső kritikus érték: $+1.96 * se + \bar{x}$

A standard hiba kiszámítása:

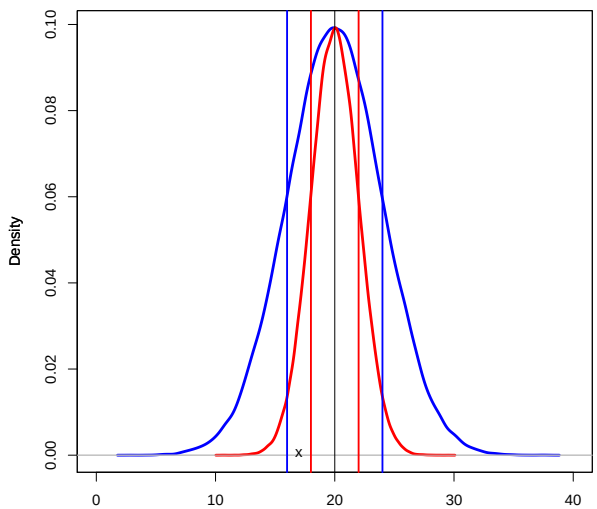
$$se = \frac{s}{\sqrt{n}}$$

Ha n alacsony, a standard hiba nagyobb, mert a nevezőben $\sqrt{n}$ szerepel $\Rightarrow$ nagyobb standard hiba esetén a kritikus érték nagyobb lesz, ezért H_0 elutasításának lehetősége kisebb.

Kisebb és nagyobb standard hiba és hipotézistesztelés

Két populáció, amelyek átlaga $\mu = 20$. 1. populáció szórása $se = 1$, 2. populációé $se = 2$.

1. populáció alsó kritikus értéke : 20–1*1.96 = 18.04, 2. populációé: 20–2*1.96 = 16,08.
Egy olyan minta, amely átlaga $\bar{x} = 17$, ehhez a populációhoz tartozik?



1. populáció: $17 < 18.04 \Rightarrow H_0$-t elutasíthatjuk.
2. populáció: $17 > 16.08 \Rightarrow H_0$-t nem utasíthatjuk el.

Feladat

Számoljuk ki a `testmagassag` R-objektum első tíz elemének átlagát. Beleesik a teljes, 300 elemű minta 95%-os konfidenciaintervallumába?

Első tíz elem átlagának kiszámítása:
`mean(testmagassag$height[1:10])`
175.9037

95%-os konfidenciaintervallum határai?
A régi szép időkben megnéztük az adott $\alpha/2$ tartományra megadott (standardizált!) z-értéket a függvénytáblázatban.

Manapság letöltjük a `gmodels` nevű R-csomagot, és lekérdezzük a határokat a
`ci` függvénnyel.
`ci(testmagassag$height,0.95) Estimate CI lower CI upper Std. Error 178.0349657 177.3166967 178.7532346 0.3649871`