

Normális eloszlás

Normális eloszlás

A statisztikai tesztek előfeltétele többnyire unimodális eloszlás (egy módusz van).

Parametrikus tesztek előfeltétele gyakran normális eloszlás.

- ▶ Módusz nagyjából az eloszlás közepén található, megegyezik a mediánnal és az átlaggal,
- ▶ onnan mindkét irányban szimmetrikusan csökken,
- ▶ megközelítőleg harang formájú (Gauß-görbe).
- ▶ aszimptotikus (0-hoz tart, de soha nem éri el),
- ▶ feltétel általában: folytonos változók, tehát nem csak egész, hanem valós számok.



Normal Distribution



Paranormal Distribution

Hol fordul elő a normális eloszlás?

A legtöbb, természete szerint parametrikus adat normális eloszlású, ha a minta elég nagy.

- ▶ Egy adott tóból kifogott halak tömege és hossza.
- ▶ Egy városban naponta megtett utak hossza méterben.
- ▶ Klíma, pl. a május 1-i minimum és maximum hőmérséklet több éven át.
- ▶ A rövid /i/ magánhangzó tartama 100 magyar beszélő spontán beszédében.
- ▶ Az IQ is ilyen, szándékosan normális eloszlásra van kalibrálva. Az átlag önkényesen rögzített érték: mentális életkor/valós életkor*100.

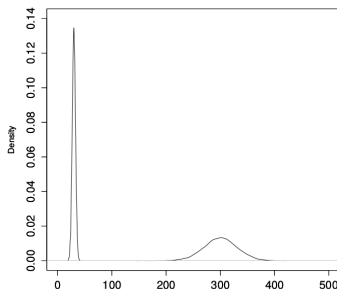
Hol nem találunk normális eloszlást?

Sok humántudományi területen, például:

- ▶ Pénzügyi mutatók és gazdasági adatok, pl. magyar keresőképes lakosság bevétele.
- ▶ Árváltozások, hozamok, tőzsdei értékek, részvények, árfolyamok.
- ▶ Emberek életkora (korfa).
- ▶ Reakcióidő: rövid idő gyakoribb.

Felhasználás

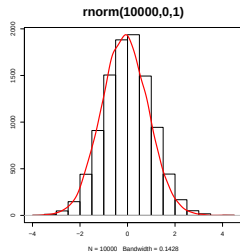
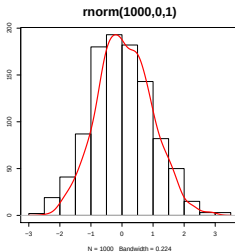
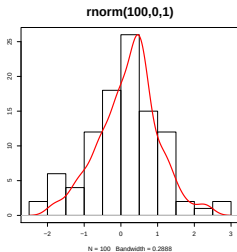
Egy állattelepen kutyákat és sertéseket tenyésztene. Rendszeresen ellenőrzik, hogy vannak-e túl sovány vagy túlsúlyos egyedek, és őket külön vizsgálatnak vetik alá. Az adatokat évtizedek óta gyűjtik, és ezek normális eloszlást mutatnak.



Hogyan lehet megbízható határértékeket találni a két, meglehetősen eltérő súlycsoporthoz tartozó állatnál az átlagostól lényegesen eltérő tömeg megállapítására?

Normális eloszlás (N)

Függ n -től – az elemszámtól, és k -tól – az osztályok/kategóriák számától. Itt: $n = 100, 1000, 10000$ egy-egy normális eloszlású minta mérete.



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

`hist(a)`: hisztogramm. `plot(density(a))`: sűrűségfüggvény.

Újabb ábra egy már meglévő ábrára: a két függvény között `par(new=T)` futtatása. (Ld. 5_statisztika.R.txt fájl.)

N paramétere

μ : a teljes populáció feltételezett átlaga.

σ : a teljes populáció feltételezett szórása.

$\bar{x}$: a rendelkezésre álló minta átlaga.

s : a rendelkezésre álló minta szórása.

minta átlaga:

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

N paraméterei

μ : a teljes populáció feltételezett átlaga.

σ : a teljes populáció feltételezett szórása.

$\bar{x}$: a rendelkezésre álló minta átlaga.

s : a rendelkezésre álló minta szórása.

minta átlaga:

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

minta varianciája:

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

N paraméterei

μ : a teljes populáció feltételezett átlaga.

σ : a teljes populáció feltételezett szórása.

$\bar{x}$: a rendelkezésre álló minta átlaga.

s : a rendelkezésre álló minta szórása.

minta átlaga:

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

minta varianciája:

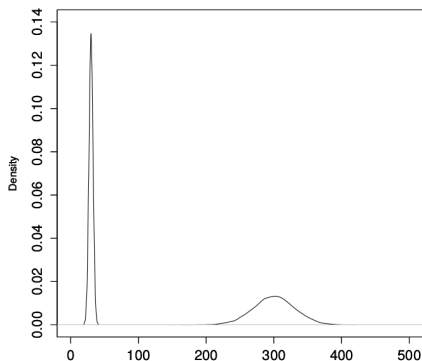
$$v = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

minta szórása:

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

Probléma

Az eloszlás szórása függ az elemszámtól és az átlagtól $\Rightarrow$ eltérő átlagok eloszlása nem összehasonlítható.



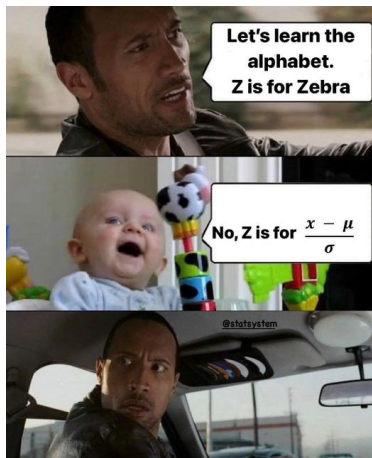
z-transzformáció

Megoldás: standardizálás z-értékre. Minden egyes elemre:

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

azaz minden egyes elem értékből kivonjuk a populáció (itt: minta) átlagát, és elosztjuk a populáció feltételezett szórásával (tehát a nevezőben itt n van, és nem $n - 1$).

Ez az általánosító eljárás a z-transzformáció. A normál eloszlású adatokat így standard normális eloszlásúvá alakítjuk. A függvénye `scale()`.



Standard normális eloszlás

Populáció normális eloszlásának jellemzői: $N(\mu, \sigma)$.

átlag: $\mu = \bar{x}$

z-transzformáció:

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

szórás: $\sigma = \mu + \sigma$

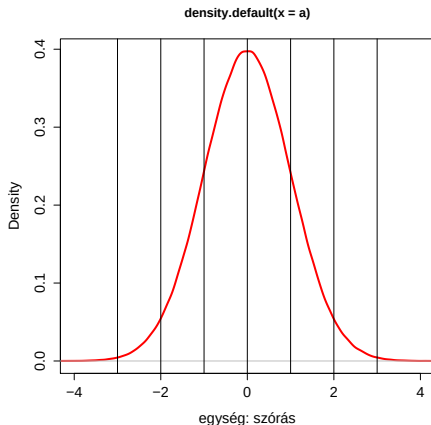
$$z_i = \frac{(\mu + \sigma) - \mu}{\sigma} = 1$$

standard normális eloszlás jellemzése: $N(0, 1)$

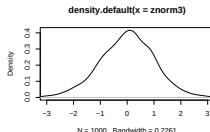
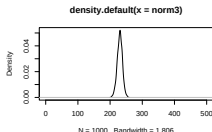
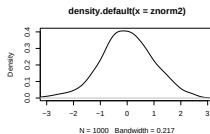
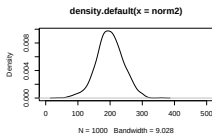
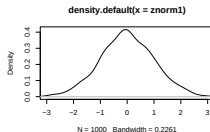
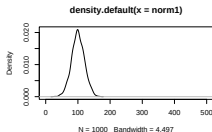
Standard normális eloszlás

Sűrűségfüggvény:

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



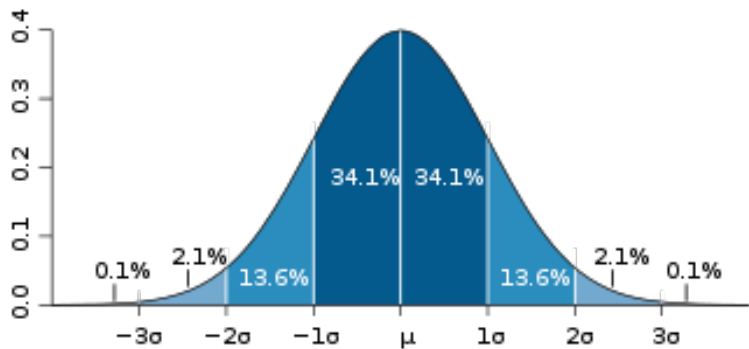
Hozzunk létre normális eloszlású vektorokat eltérő átlagokkal és szórásokkal, majd standardizáljuk őket z-transzformációval.
Szükséges parancssor: 5_statisztika_R.txt



A standardizálással kapott függvények közös jellemzői

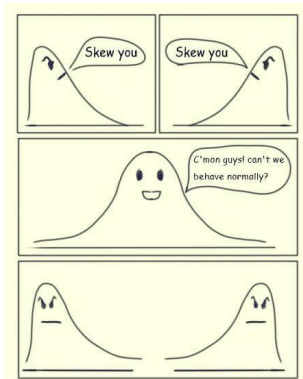
- ▶ az x -tengely és a sűrűségfüggvény által bezárt terület összege $= 1$.
- ▶ Az esetek 50%-a az átlagtól balra helyezkedik el.
- ▶ $\sigma = -1 \cdots +1$ közötti tartomány az esetek **68,27%-át** tartalmazza.
- ▶ $\sigma = -2 \cdots +2$ közötti tartomány az esetek **95,45%-át** tartalmazza.
- ▶ $\sigma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

Standard normális eloszlás



Ferdeség

Ha az eloszlás szimmetrikus, a **ferdeség** (skewness) 0. Pozitív ferdeség esetén a jobb oldal laposabb, a módusz és a medián az átlagnál alacsonyabb, azaz balra esik → jobbra ferde eloszlás.



Negatív ferdeség esetén a bal oldal laposabb, a módusz és a medián az átlagnál magasabb → balra ferde eloszlás.

Transzformációk

Az unimodális, de nem szimmetrikus, azaz jobbra vagy balra ferde eloszlások gyakran átalakíthatóak normális eloszlásúvá.

Szokásos eljárások:

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

R

Programcsomagok telepítése az R-ben

Az R nyílt forráskódú szoftver, bárki fejleszthet hozzá csomagokat. Elérhető csomagok listája az R mirror oldalakon, Packages menüpont alatt.

Telepítés interneten keresztül:

```
install.packages("corrplot")
```

Tükör kiválasztása (minél közelebb, annál kevesebb adatforgalmat generálunk).

Linux: csomagokat érdemes superuser-ként telepíteni, akkor a root könyvtárból bárhonnan elérhetők lesznek. Vagyis az R megnyitása: `sudo R`.

Windowson rendszergazdaként érdemes megnyitni az R-t a csomag telepítése előtt: kattintás jobb egér gombbal. Utána zárjuk be, és indítsuk újra sima userként.

Korrelációs mátrix létrehozása

Letöltjük az óra honlapjáról a ratings.cor.RData fájlt a statisztika könyvtárba. Betöltés az R-be a `load()` függvénnyel. A pontos kód az 5_statisztika_R.txt fájlból kimásolható.

Betöltjük az R-be az előzőleg telepített `corrplot` csomagot a `library(corrplot)` függvénnyel. Idézőjel ekkor már nem kell!

A ratings.cor adatmátrix csupa numerikus oszlopot tartalmaz, amik között korrelációt lehet számolni. Hozzunk létre egy korrelációs mátrixot.

```
m = cor(ratings.cor)
```

Nézzük meg a tartalmát az `m` beírásával.

Korrelációs mátrix ábrázolása

```
corrplot(m)
```

Színes körök mutatják a korreláció erősségét. Színárnyalat: sötétpirostól a sötétkékig -1 és $+1$ között. A felső bal cellánál kezdődő átló a mérőszám önmagával való összehasonlítását jelzi, mindig igaz rá, hogy $r = +1$.

Ha a korreláció irányát jobban szeretnénk vizualizálni, ellipsziseket is rajzolhatunk:

```
corrplot(m, method="ellipse")
```

Ez jól látszik, ha az együttható értékét számmal írjuk ki:

```
corrplot(m, method="number")
```


Adatok szűrése logikai vektorokkal

Ha egy pontdiagramban a növényeket és az állatokat különböző színekkel akarjuk megjeleníteni, külön csoportokra kell osztanunk a 81 sorból álló táblázatot.

Egy logikai vektort hozunk létre:

```
noveny = ratings$Class == "plant"
```

A két egyenlőségjel abban különbözik az egytől, hogy itt nem egy plant nevű új objektumot hozunk létre, hanem egy olyat, amiben azok a sorok vannak jelölve, amikre igaz, hogy a Class oszlopban "plant" kategória szerepel.

A noveny objektum egy 81 elemből álló vektor (adatok egydimenziós sorozata), amiben a plant-ot tartalmazó soroknál TRUE, a többinél FALSE szerepel.

Növény- és állatnevek ábrázolása külön színnel

Az x tengelyen a szóhossz, az y tengelyen a gyakoriság jelenik meg.
Hivatkozás rájuk: `ratings$Length[noveny]` ,
`ratings$Frequency[noveny]`.

```
plot(ratings$Length[noveny], ratings$Frequency[noveny],  
col="green")
```

Az állatokra létrehozhatunk egy külön logikai vektort:
`allat = ratings$Class == "animal"`

Ha ugyanebbe az ábrába akarunk rajzolni, a `par(new=TRUE)`
függvénnyel jelezzük, hogy nem akarunk új ábrát létrehozni. Majd
barnával az állatok:

```
plot(ratings$Length[allat], ratings$Frequency[allat],  
col="brown")
```

Probléma: az R a két osztályhoz tartozó értékek alapján maga határozza meg a tengelyek minimum és maximum értékét. Kézzel állítás az `xlim =`, `ylim=` paraméterek segítségével.

A teljes halmaz minimum és maximum értékeinek lekérdezése:

```
range(ratings$Length)
range(ratings$Frequency)
```

A szóhosszoknál 3–10 betűt érdemes ábrázolni, a gyakoriságoknál 1.5 és 8 közötti értéket.

```
plot(ratings$Length[noveny], ratings$Frequency[noveny],
     col="green", xlim = c(3,10), ylim=c(1.5,8)
     (new=T)
plot(ratings$Length[allat], ratings$Frequency[allat],
     col="brown", xlim = c(3,10), ylim=c(1.5,8))
```