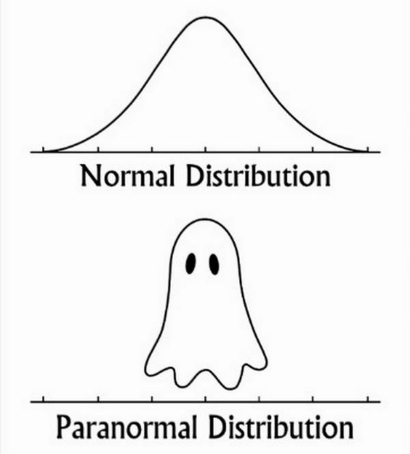
Normális eloszlás


Normal Distribution
Paranormal Distribution

Normális eloszlás

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

Parametrikus tesztek előfeltétele gyakran normális eloszlás.

- Módusz nagyjából az eloszlás közepén található, megegyezik a mediánnal és az átlaggal,
- onnan mindkét irányban szimmetrikus csökkenés,
- 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,

Hol fordul elő a normális eloszlás?

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

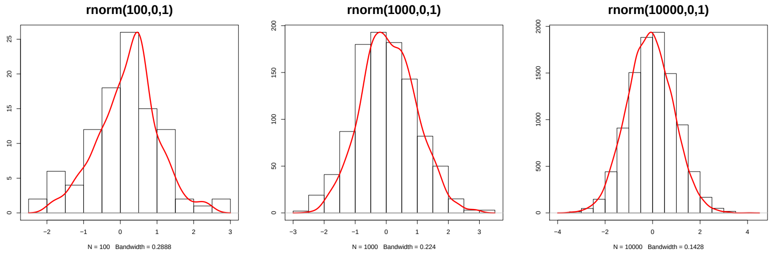
- Egy adott tóból kifogott halak tömege és hossza.
- Egy városban naponta megtett utak hossza méterben.
- Klíma, pl. minimum és maximum hőmérsékletek adott időn át.
- A rövid /i/ magánhangzó tartama 100 magyar beszélő spontán beszédében.
- Az IQ is ilyen, szándékosan normális eloszlásra van kalibrálva.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

Hol nem találunk normális eloszlást?

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

Normális eloszlás (N)

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



`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.

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:

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

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:

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

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:

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

minta varianciája:

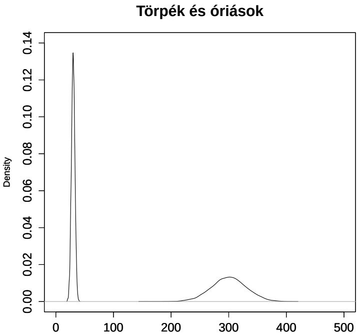
$$v = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

minta szórása:

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

Probléma

A törpék átlagos testmagassága 30 cm, a szórás ennek 1/10-e, azaz 3 cm. Az óriások átlagosan 300 cm magasak, a testmagasság szórása 30 cm.



DE: a szórás 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 értékét kivonjuk a populáció (itt: minta) átlagából, 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()`.


Let's learn the alphabet.
Z is for Zebra
No, Z is for $\frac{x - \mu}{\sigma}$
@statsystem

Standard normális eloszlás

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

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

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

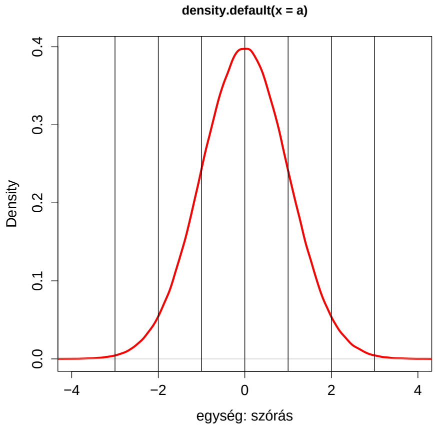
szórás: $\sigma = \mu + \sigma$

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

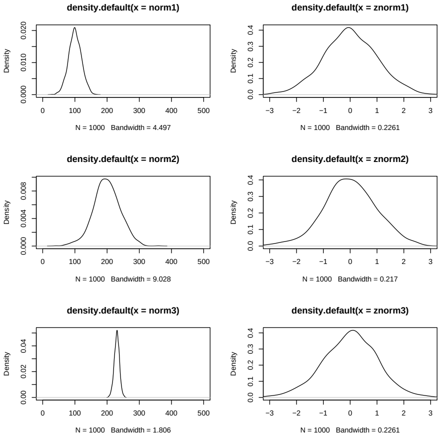
standard normális eloszlás jellemzése: $N(0, 1)$

Standard normális eloszlás

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



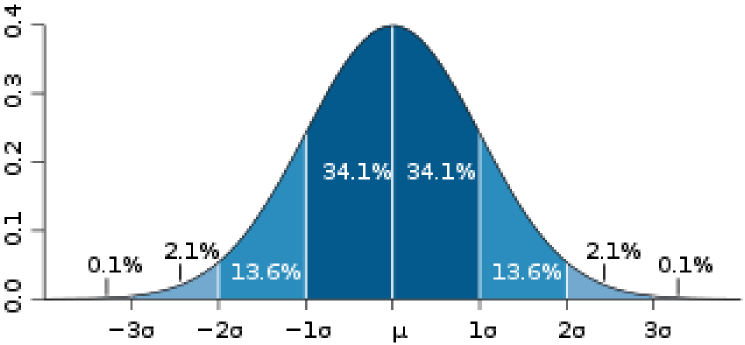
Hozzunk létre normális eloszlású vektorokat eltérő átlagokkal és szórásokkal, majd standardizáljuk őket z-transzformácóval.
Szükséges parancssor: 5_statisztika_R.txt



A standardizálással kapott függvények közös jellemzői

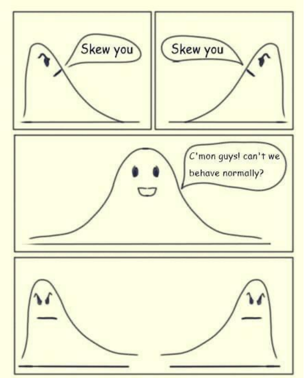
- az x-tengely és a sűrűségfüggvény által bezárt terület összege $= 1$.
- Az esetek 50%-a az átlagtól balra helyezkedik el.
- $\sigma = -1 \cdots + 1$ közötti tartomány az esetek 68,27%-át tartalmazza.
- $\sigma = -2 \cdots + 2$ közötti tartomány az esetek 95,45%-át tartalmazza.
- $\sigma = -3 \cdots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

Standard normális eloszlás



Ferdeség

Ha az eloszlás szimmetrikus, a ferdesége (skewness) 0. Pozitív ferdeség esetén a jobb oldal laposabb, az átlag mediántól jobbra esik, azaz magasabb, mint a medián → jobbra ferde eloszlás.



Negatív ferdeség esetén a bal oldal laposabb, az átlag a mediánnál alacsonyabb → balra ferde eloszlás.

Transzformációk

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

Szokásos eljárások:

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