

Normális eloszlás

Grafikonok az R-ben

Normális eloszlás

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

Parametrikus tesztek előfeltétele gyakran normális eloszlás.

- ▶ Folytonos változók,
- ▶ 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 közelít).

Hol fordul elő a normális eloszlás?

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

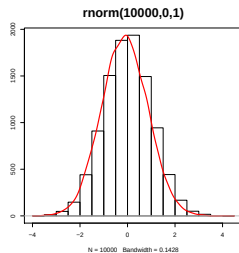
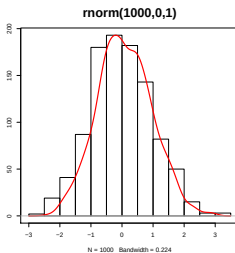
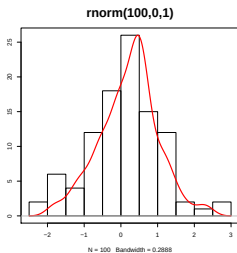
- ▶ IQ – az átlag önkényesen rögzített érték! (Mentális életkor/valós életkor*100.)
- ▶ Emberek magassága és súlya nemenként.
- ▶ Egy adott tóból kifogott halak súlya és hossza.
- ▶ Háztartások éves bevétele egy adott országban.
- ▶ Egy vizsgán szerzett osztályzatok, ha a minta elég nagy.
- ▶ Klíma, pl. minimum és maximum hőmérsékletek adott időn át.

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).

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.

N paramétere

μ : populáció feltételezett átlaga.

σ : populáció feltételezett szórása.

$\bar{x}$: minta átlaga.

s : 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 : populáció feltételezett átlaga.

σ : populáció feltételezett szórása.

$\bar{x}$: minta átlaga.

s : 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:

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

N paraméterei

μ : populáció feltételezett átlaga.

σ : populáció feltételezett szórása.

$\bar{x}$: minta átlaga.

s : 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:

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

minta szórása:

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

z-transzformáció

Szórás függ az elemszámtól és az átlagtól $\Rightarrow$ eltérő átlagok eloszlása nem összehasonlítható.

Megoldás: standardizálás z-értékre. Minden egyes elemre:

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

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

Ez az eljárás a z-transzformáció.

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: $\sigma = \mu + \sigma$

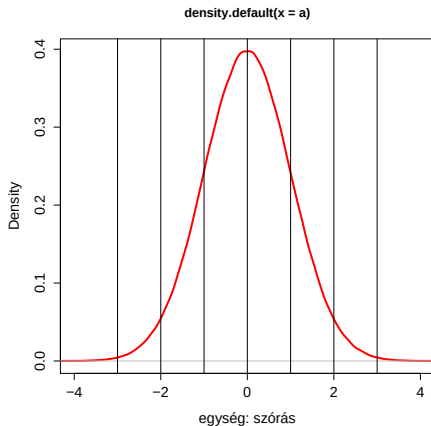
$$z_i = \frac{(\mu + \sigma) - \mu}{\sigma} = 1$$

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

Standard normális eloszlás

sűrűségfüggvény:

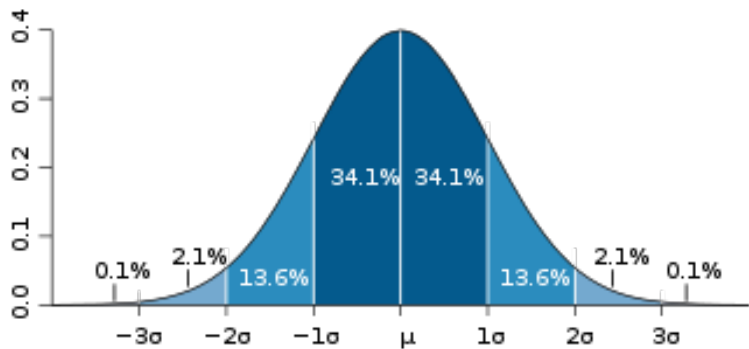
x-tengely: egységnyi szórás, tartomány: $\sigma = -\infty \cdots + \infty$.



Függvény 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



Mintaméret és standard hiba

n (minta elemszáma):

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:

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

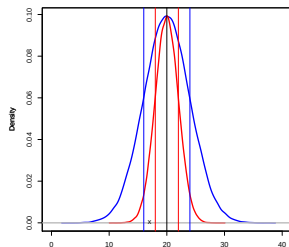
Ha n alacsony, a standard hiba nagyobb, mert a nevezőben $\sqrt{n}$ szerepel $\Rightarrow$ nagyobb standard hiba esetén a kritikus érték nagyobb lesz, ezért H_0 elutasításának lehetősége kisebb.

Példa

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

1. populáció alsó kritikus értéke : $20 - 1 \cdot 1.96 = 18.04$, 2. populációé: $20 - 2 \cdot 1.96 = 16.08$.

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.

R

R keresési útvonalai

Ha egy csomagot egy adott könyvtárba akarunk telepíteni:

```
install.packages("package-to-install","target.library")
```

Beállított keresési útvonalak lekérdezése:

```
.libPaths()
```

Házi feladat

languageR könyvtár ratings objektuma.

`names(ratings)`: változók (oszlopok) neve.

Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Házi feladat

languageR könyvtár ratings objektuma.
`names(ratings)`: változók (oszlopok) neve.

Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Ábrázolás:

```
plot(ratings$Length,ratings$Frequency)
```

első érték: x-tengely, második érték: y-tengely.

Házi feladat

languageR könyvtár ratings objektuma.

`names(ratings)`: változók (oszlopok) neve.

Hipotézis: A rövidebb állat- és növénynevek gyakoribbak.

Ábrázolás:

```
plot(ratings$Length,ratings$Frequency)
```

első érték: x-tengely, második érték: y-tengely.

Korrelációs együtthatók:

```
cor.test(ratings$Length,ratings$Frequency,method="pearson"
```

```
alternatív ...method="spearman", method="kendall"
```

$$r = -0.4281462$$

$$\rho = -0.4311981$$

$$\tau = -0.316297$$

Grafikus paraméterek

Rengeteg paraméteren lehet állítani. Hogyan lehet ezekről tudni?

- ▶ grafikus parancs opcionális argumentumai. Lekérdezés: `?boxplot`, `?plot`, `?barplot` stb.
- ▶ parancsok súgója gyakran utal további hasznos parancsokra, pl. `line()`, `title()`, `abline()` stb.
- ▶ `par()`: rengeteg paraméter, pl. tengelyek feliratozása (felirat mérete, elhelyezése, egységek mérete), tengelyek aránya stb.

Első lépés súgó. Felépítés: (1) kötelező és opcionális argumentumok listája, (2) argumentumok rövid magyarázata, (3) részletek: többnyire innen derül ki a releváns infó, ha még nem ismerjük a parancsot, gyakran hivatkozások is, (4) lásd még - hasznos, esetenként hasznosabb, további parancsok, (5) példák - ezek általában túl bonyolultak, ezért nem túl hasznosak.

Néhány hasznos paraméter

Általában alkalmazhatók grafikus parancs argumentumaként:

```
plot(x,y,xlim=c(0,length(x)),ylim=c(0,1),main="első  
ábrám 2012-ben",xlab="ez az x-tengely",ylab="ez az  
y-tengely",col=2,cex.main=1.7,cex.axis=1.3,cex.lab=1.3)
```

- ▶ **xlab=, ylab=:** "x-tengely felirata", "y-tengely felirata".
- ▶ **main=** "Ábra címe".
- ▶ **xlim, ylim:** Ábrázolt értékek tól-ig. Főleg y-tengelynél fontos, ha összehasonlítható ábrákat akarunk. Pl százalékos ábrázolásnál `ylim = c(0,100)`, azaz 0–100%-ig.
Egyenlőségjel előtti szóköz opcionális.
- ▶ **col:** színek, vagy névvel, vagy számmal. Pl. `col=2` és `col="red"` azonosak.
- ▶ **cex:** `cex.main`, `cex.axis`, `cex.names`, `cex.lab` stb.
Default: `cex=1`, ehhez képest cím, mérőszámok, címkék betűmérete nagyobb (1.3, 1.7) vagy kisebb (0.7).

Ábra mentése

Mentés pdf-ként:

```
plot(blablabla)
dev.print("célfájl",device=pdf)
```

Mentés ábraként, pl. png (jpg és tif nem jó, mert képtömörítéssel készül!):

```
png(file = "fajlnev", bg = "white")
plot(blablabla)
dev.off()
```

Alternatíva Windows-os R-ben: jobb egérgombbal, kattintgatva

Ha nem adunk meg elérési útvonalat: mentés aktuális könyvtárba (getwd() paranccsal megtudható).

Boxplotok

Ábrázolás módja: y-tengely: függő változó interkvartilis eloszlása,
x-tengely: csoportok, esetleg további tagolással.

```
boxplot(függőváltozó~függetlenváltozó)
```

Ha további csoportosítás:

```
boxplot(függőváltozó~csoport*függetlenváltozó)
```

Például:

```
boxplot(ratings$Frequency~ratings$Class*ratings$Complex)
```

Egyszerű és összetett állat- és növénynevek gyakorisága.

hasznos paraméterek boxplot() argumentumaként:

```
col=c("red","blue")
```

```
names=c("állat","növény")
```

```
ha ismétlődnie kell: names=rep(c("állat","növény"),2)
```

```
rep() ismétlés(mit,hányszor).
```


Feladatok

1. feladat: ratings fájlban található adatok alapján tetszőleges pontdiagram készítése és mentése.

ratings fájlban szereplő adatok alapján tetszőleges pontdiagram készítése és mentése.

2. feladat: ratings fájlban található adatok alapján tetszőleges boxplot készítése.

Célábra: cím, angol vagy magyar nyelvű tengelyfelirat.