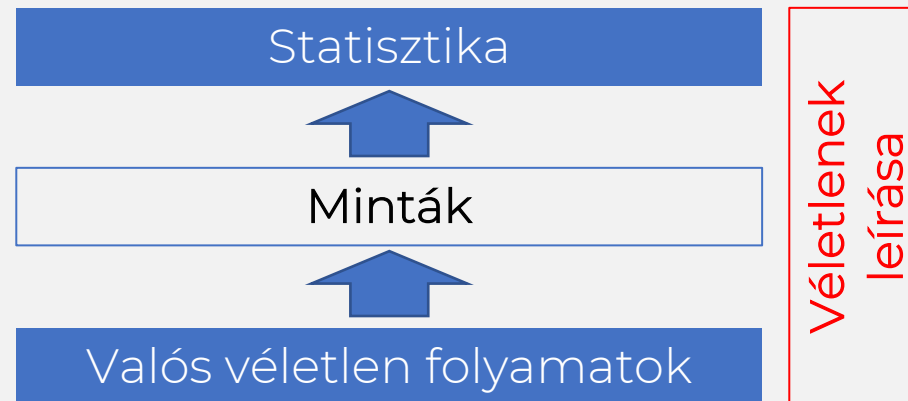


ADATOK MEGISMERÉSE

(2) Exploratory Data Analysis

STATISZTIKA ÉS VALÓSZÍNŰSÉGSZÁMÍTÁS

- A statisztika a valószínűségszámítás elméletét használja a véletlen folyamatok megfigyelésének leírására, de
 - A valószínűségszámítás magát a **véletlen folyamatot** írja le
 - A statisztika a **folyamat által generált mintákat** figyeli meg

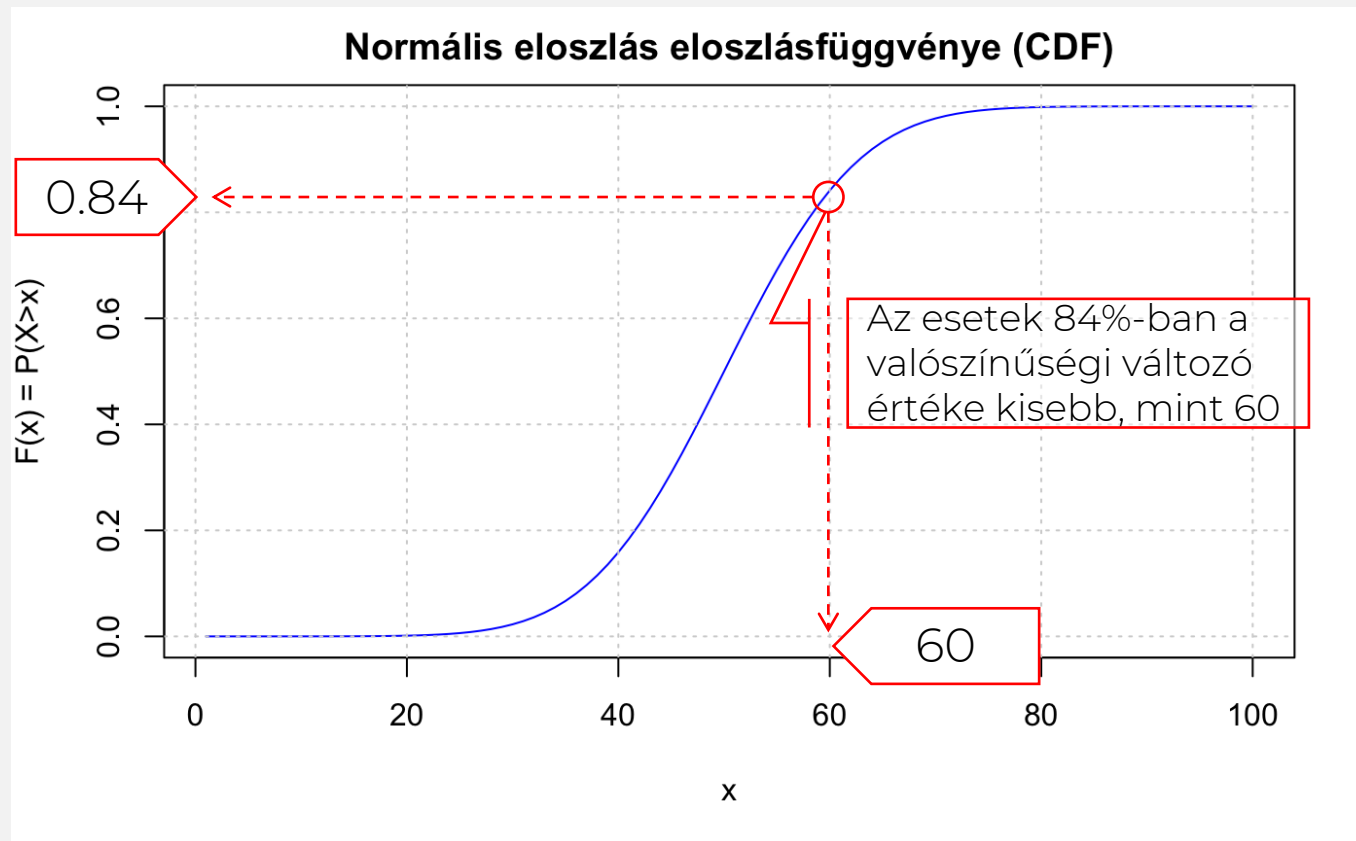


ELOSZLÁSOK LEÍRÁSA

VALÓSZÍNŰSÉGI ELOSZLÁSFÜGGVÉNY

- CDF, Cumulative Distribution Function

$$F(x) = P(X < x)$$

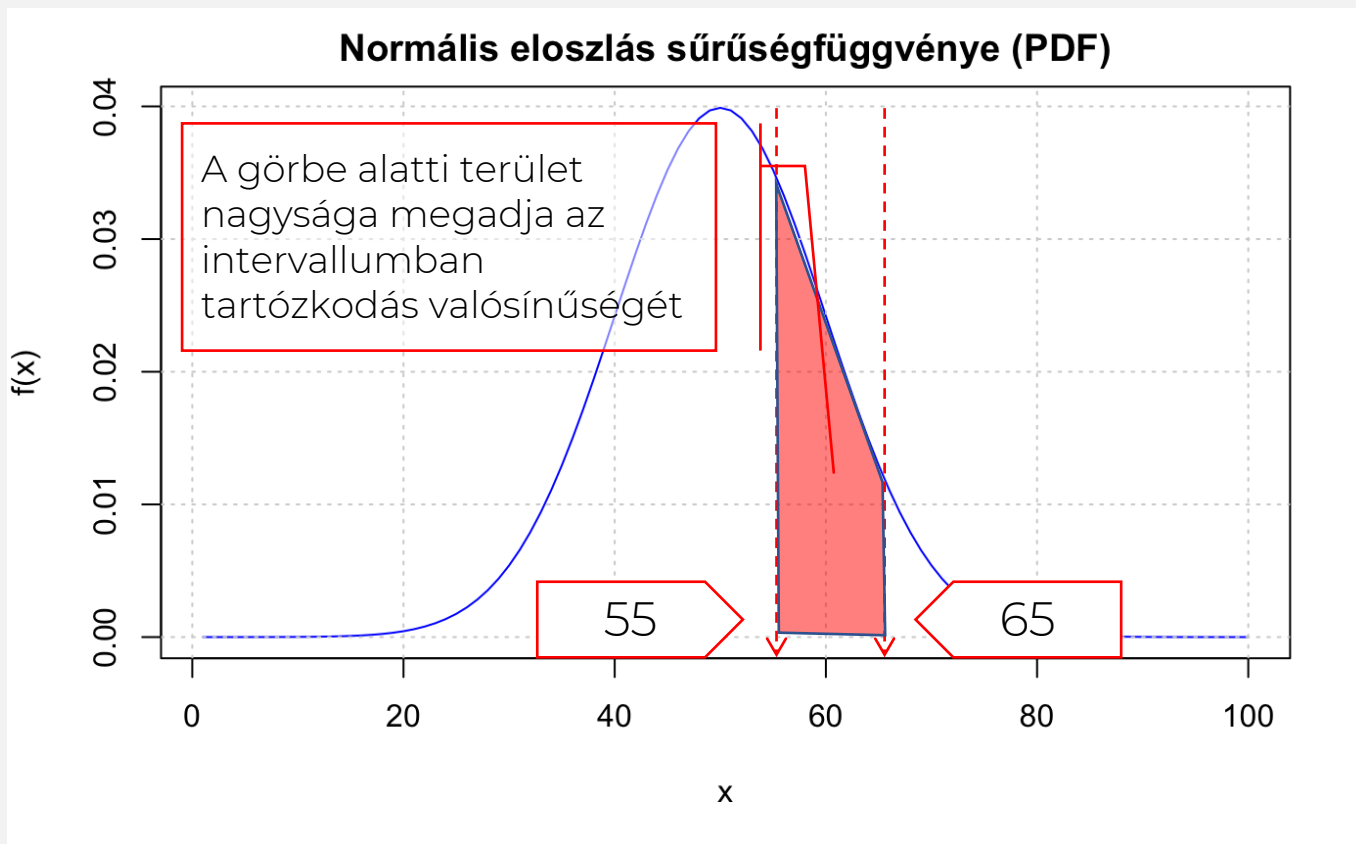


ELOSZLÁSOK LEÍRÁSA

VALÓSZÍNŰSÉGI SŰRŰSÉGFÜGGVÉNY

- PDF, Probability Density Function

$$P(x \in [a, b]) = \int_a^b f(x) dx$$



ELOSZLÁSOK LEÍRÁSA

VÁRHATÓ ÉRTÉK, SZÓRÁSNÉGYZET

■ Várható érték

- Amire tippelve a hibánk várható értékben a legkisebb lesz
- Folytonos értékkészletű
- $E\{X\} = \mu = \int_{-\infty}^{\infty} x \cdot f(x) dx$
- Diszkrét értékkészletű
- $E\{X\} = \mu = \sum_i x_i P_i$

■ Szórásnégyzet

- A tippelési hibánk négyzetének várható értéke
- Folytonos értékkészletű
- $D^2\{X\} = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$
- Diszkrét értékkészletű
- $D^2\{X\} = \sigma^2 = \sum_i (x_i - \mu)^2 P_i = \sum_i x_i^2 P_i - \mu^2$

ELOSZLÁSOK LEÍRÁSA

CENTRÁLIS MOMENTUMOK

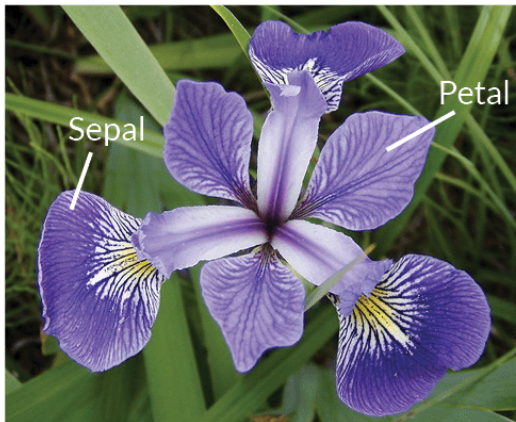
- Egy valószínűségi változó k -dik centrális momentuma a következőképpen számolható
 - Folytonos értékkészletű esetben
 - $\mu_k = \int_{-\infty}^{\infty} (x - \mu)^k f(x) dx$
 - Diszkrét értékkészletű esetben
 - $\mu_k = \sum_i (x_i - \mu)^k P_i$
- Tulajdonságok
 - A nulladik centrális momentum értéke $\mu_0 = 1$
 - Az első centrális momentum értéke $\mu_1 = 0$
 - A második centrális momentum értéke $\mu_2 = \sigma^2$

AZ IRIS ADATHALMAZ

ADATOK ÉS FELÉPÍTÉSE

- Három írisz fajta 50-50 egyedének levél méretei

Csészelevél hossz		Csészelevél szélesség		Sziromlevél hossz		Sziromlevél szélesség		Írisz fajta	
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species					
5.1	3.5	1.4	0.2	setosa	Objektum 1				
5.0	3.2	1.3	0.2	versicolor	Objektum 2				



Iris Versicolor



Iris Setosa



Iris Virginica

EGYVÁLTOZÓS LEÍRÁS

ADATOK TÁROLÁSÁNAK VIZSGÁLATA

- Az adathalmaz attribútumai
 - **attributes()**
 - **names()**, **colnames()**, **rownames()**
- Tárolási típus
 - **typeof()**
- Értelmezési osztály
 - **class()**
- Adathalmaz mérete
 - **dim()**, **nrow()**, **ncol()**
- Felépítés
 - **str()**
- Adatok összegzése
 - **summary()**
- Példa adatok
 - **head()**, **tail()**
- Faktor adatok vizsgálata
 - **levels()**

VIZUÁLIS VIZSGÁLATOK

- Ponthalmazok ábrázolása
 - **plot()**
 - Ponthalmaz kirajzolása x, y koordináta rendszerben
 - **table()**
 - Gyakorisági (kontingencia) táblázat
 - **hist()**
 - Hisztogram
 - **density()**
 - Tapasztalati sűrűségfüggvény, ePDF
 - **ecdf()**
 - Tapasztalati eloszlásfüggvény, eCDF

JELLEMZŐ ÉRTÉKEK

CENTRUMRA JELLEMZŐ ÉRTÉKEK

- Átlag, tapasztalati várható érték

$$\bar{\mu} = \bar{x} = \frac{1}{N} \sum_{i=1..N} x_i$$

- `mean(x, na.rm = TRUE)`

- Medián

- Középérték, melytől ugyanannyi kisebb, mint nagyobb értéket vesz fel az attribútum
 - `median(x, na.rm = TRUE)`

- Módusz

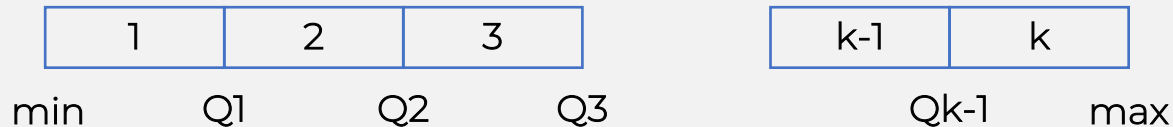
- A leggyakrabban felvett érték
 - `tabulate()`: megszámolja, hogy az egyes értékek(1-től) hányszor fordulnak elő
 - `which.max()`: visszaadja a legnagyobb elem indexét
 - `match()`: egy vektorban egy másik vektor elemei melyik pozícióban helyezkednek el

SZÓRTSÁGRA JELLEMZŐ ÉRTÉKEK

■ Értéktartomány

- **min()**: minimális érték
- **max()**: maximális érték
- **range()**: terjedelelem

■ Kvantilis értékek



- A medián (Q1, k=2)
- Kvartilis (Q1, Q2, Q3, k=4)
- Decilis (k=10)
- Percentilis (k=100)
- **quantile(x, na.rm = TRUE, probs = seq(0, 1, 0.1))**
- **fivenum()**: kvartilisek, a min és max értékek
- **boxplot()**: kvartilisek, a min és max értékek ábrázolása
- **boxplot.stats()**: nincs ábrázolás

SZÓRTSÁGRA JELLEMZŐ ÉRTÉKEK 2.

- **Korrigált tapasztalati szórásnégyzet**

- Átlagtól való négyzetes eltérés átlaga, **variancia**

$$\bar{\sigma}^2 = \frac{1}{N-1} \sum_{i=1..N} (x_i - \bar{x})^2 = \bar{\mu}_2 = E\{X^2\} - E^2\{X\}$$

- **var()**: korrigált tapasztalati szórásnégyzet

- **Korrigált tapasztalati szórás**

- A variáció négyzetgyöke (**standard deviáció**)

$$\bar{\sigma} = \frac{1}{\sqrt{N-1}} \sqrt{\sum_i (x_i - \bar{x})^2}$$

- **sd()**: korrigált szórás

SZÓRTSÁGRA JELLEMZŐ ÉRTÉKEK 3.

- Standard hiba

- A szórás és az elemszám négyzetgyökének hányadosa

$$\bar{\sigma}_{std} = \frac{\bar{\sigma}}{\sqrt{N}}$$

- Relatív szórás

- A szórás és a várható érték hányadosa

$$\bar{\sigma}_{rel} = \frac{\bar{\sigma}}{\bar{\mu}}$$

- Abszolút közepes eltérés

- A várható értéktől való abszolút eltérés átlaga

$$d_{abs} = \frac{1}{N} \sum_i \|x_i - \bar{x}\|$$

SZÓRTSÁGRA JELLEMZŐ ÉRTÉKEK 4.

■ Ferdeség (skewness)

- Az adatok sűrűségfüggvénye mennyire szimmetrikus

$$\bar{\gamma}_1 = \frac{1}{N-1} \frac{\sum (x_i - \bar{x})^3}{\bar{\sigma}^3} = \frac{\bar{\mu}_3}{\bar{\sigma}^3}$$

- $\bar{\gamma}_1 = 0$: A tapasztalati eloszlás szimmetrikus
- $\bar{\gamma}_1 > 0$: Az eloszlás jobbra nyúlik el
- $\bar{\gamma}_1 < 0$: Az eloszlás balra nyúlik el
- **PerformanceAnalytics::skewness()**

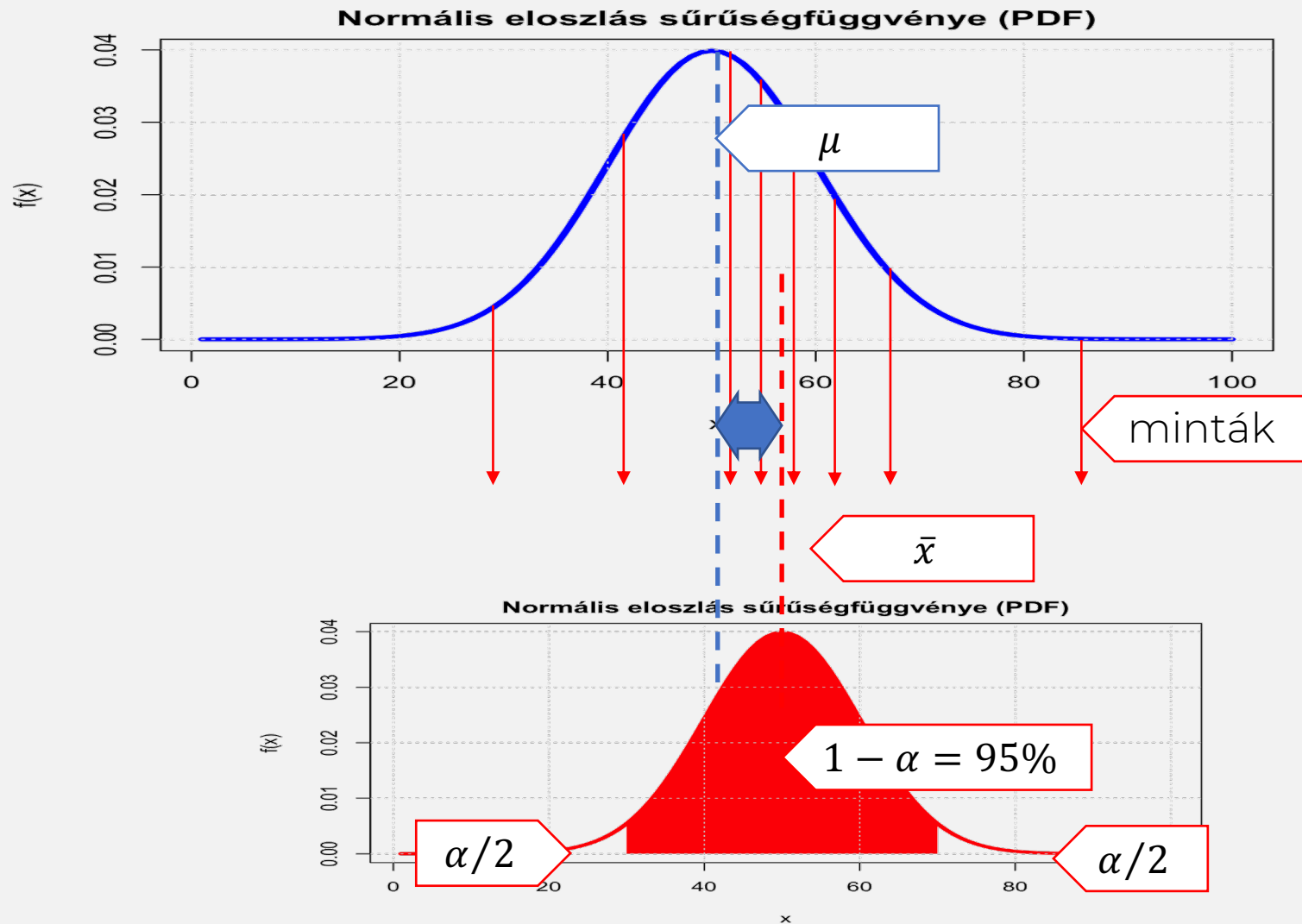
■ Csúcsosság (kurtosis)

- A sűrűségfüggvény mennyire csúcsos, lapult a normális eloszláshoz képest

$$\bar{\gamma}_2 = \frac{1}{N-1} \frac{\sum (x_i - \bar{x})^4}{\bar{\sigma}^4} - 3 = \frac{\bar{\mu}_4}{\bar{\sigma}^4} - 3$$

- $\bar{\gamma}_2 = 0$: normális eloszlás esetében
- $\bar{\gamma}_2 > 0$: csúcsosabb
- $\bar{\gamma}_2 < 0$: laposabb
- **PerformanceAnalytics::kurtosis()**

KONFIDENCIA INTERVALLUM



KONFIDENCIA INTERVALLUM 2.

- Intervallumbecslés eszköze
 - Egy becsült paraméter tartózkodása egy intervallumban
 - Adott szignifikanciaszintez: $(1 - \alpha)$

- Normális eloszlás esetében:

$$P(\bar{x} - \theta < \mu < \bar{x} + \theta) = 1 - \alpha$$

, ahol

$$\theta = z \cdot \frac{\sigma_w}{\sqrt{N}}$$

- z : a normális eloszlás $(1 - \alpha/2)$ rendű kvantilise
- $\frac{\sigma_w}{\sqrt{N}}$: standard hiba
- A konfidencia intervallum meghatározására gyakorlatban a Studen-féle t-eloszlást szokás alkalmazni

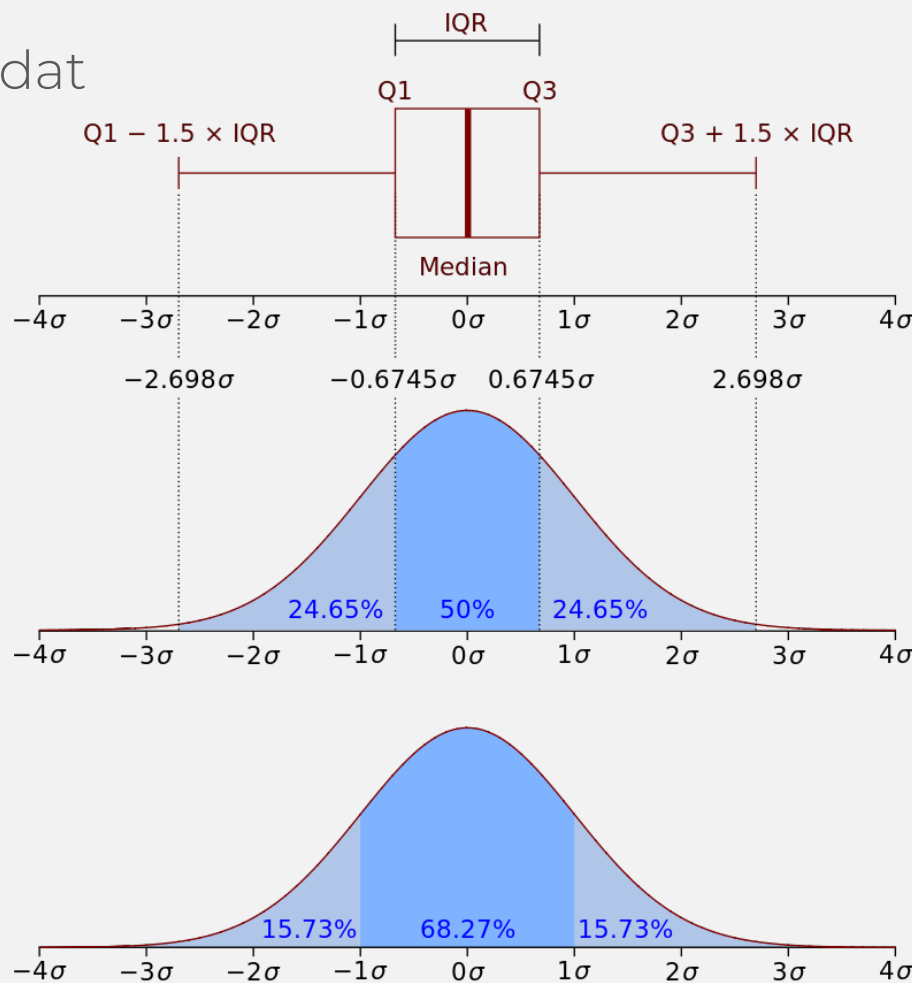
OUTLIER DETEKCÍÓ

- Outlier
 - Modellen kívüli, kilógó adat

- Meghatározása

- $x_i < Q_1 - 1.5IQR$
- $x_i > Q_3 + 1.5IQR$
- , ahol
- $IQR = Q_3 - Q_1$

- `boxplot()`
- `boxplot.stat()`



NORMALITÁS VIZSGÁLAT VIZUÁLIS

- Standardizálás

- Adatok konvertálása az összehasonlíthatóság érdekében
 - Várható érték 0, a szórás 1 legyen

- Sűrűségfüggvény alapján

- A standardizált adatok sűrűségfüggvényének összehasonlítása $N(0,1)$ sűrűségfüggvénnyel
- **dnorm()**: a normális eloszlás sűrűségfüggvénye

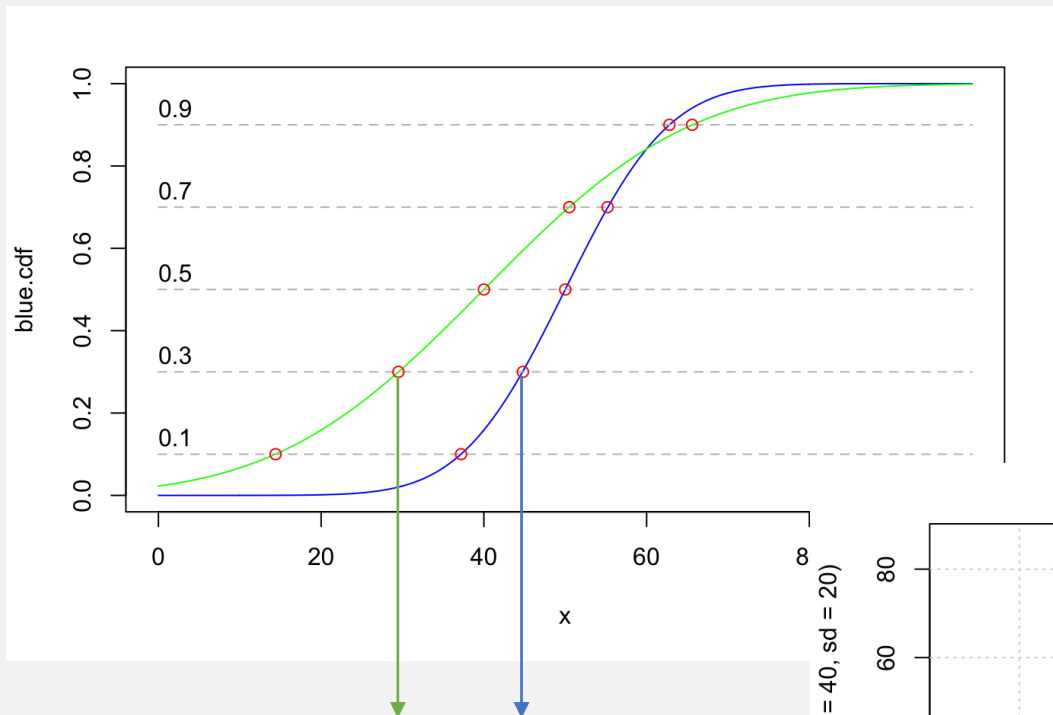
- Eloszlásfüggvény alapján (P-P plot)

- A standardizált adatok eloszlásfüggvényének összehasonlítása $N(0,1)$ eloszlásfüggvénnyel
- **pnorm()**: a normális eloszlás eloszlásfüggvénye

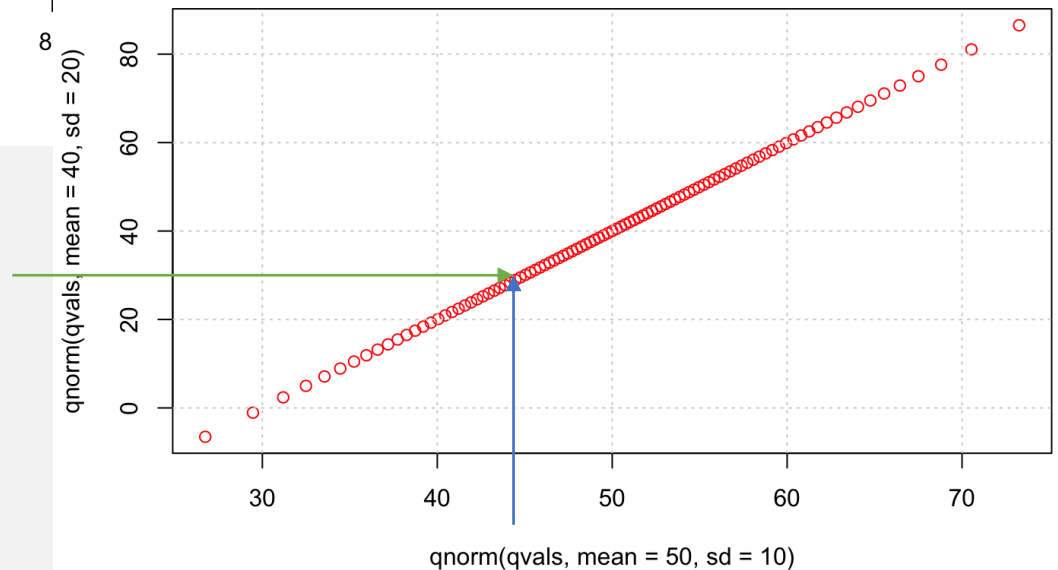
- Kvantilisek alapján (Q-Q plot)

- Kevés adat esetében
- Összetartozó kvantilis értékek ábrázolása x-y koordináta rendszerben
- A Q-Q pontok milyen közel vannak az 1 meredekségű egyeneshez
- **qqnorm()**, **qqline()**
- **qqplot()**
- **car::qqPlot()**

Q-Q PLOT



Ugyanazon
értékhez tartozó
kvantilisek



NORMALITÁS VIZSGÁLAT

STATISZTIKAI PARAMÉTEREK ALAPJÁN

- **Átlag, medián, módusz**
 - Normális eloszlás esetében a mind a három érték megegyezik egymással
 - A különbségeket vizsgáljuk
- **Fedesség**
 - Normális eloszlás esetében 0
- **Csúcsosság**
 - Normális eloszlás esetében 0

NORMALITÁS VIZSGÁLAT

SZIGNIFIKANCIA TESZT

- Nullhipotézis (H_0)
 - Az adathalmazunk normális eloszlást követ
- Szignifikanciaszint
 - A hipotézisünk valamilyen valószínűséggel igaz
 - A valószínűség, mely esetében H_0 -t **elvetjük** a **szignifikanciaszint** (jellemzően $p < 0.05$)
- Normalitás tesztek
 - Kolmogorov-Smirnov teszt
 - **Shapiro-Wilk teszt**
 - **`shapiro.test()`**
 - Pearson's chi-squared teszt

TÖBBVÁLTOZÓS ELEMZÉS

VIZUÁLIS VIZSGÁLATOK

- 2D
 - `plot(x, y)`
- 3D
 - `scatterplot3d()`
- Páronként
 - `plot(<matrix>)`
- Eloszlások
 - Sűrűségfüggvény
 - Eloszlásfüggvény (P-P plot)
 - Kvantilis értékek (Q-Q plot)
 - `qqplot()`
 - `qqline()`

KOVARIANCIA

- Szórásnégyzet általánosítása
- **Tapasztalati kovariancia** n db x_i és y_i mintára a következőképpen számítható

$$\bar{\sigma}_{x,y} = \frac{1}{N} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- **cov()**
- Kovariancia mátrix
 - Kettő, vagy több változó közötti kovariancia értékét.

$$Cov(\mathbf{X}) = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,N} \\ \sigma_{2,1} & \sigma_2^2 & \dots & \sigma_{2,N} \\ \dots & \dots & \dots & \dots \\ \sigma_{N,1} & \sigma_{N,2} & \dots & \sigma_N^2 \end{bmatrix}$$

KORELLÁCIÓ

- Kovariancia normalizált értéke

$$\bar{r}_{x,y} = \frac{\bar{\sigma}_{x,y}}{\bar{\sigma}_x \bar{\sigma}_y}$$

- $[-1, 1]$ között vehet fel értéket:
 - Pozitív lineáris korreláció: $r = 1$
 - Nincs korreláció, az értékek korrelálatlanok: $r = 0$
 - Negatív lineáris korreláció: $r = -1$
 - Erős korrelációs kapcsolat: $|r| \geq 0.7$
- `cor()`
- Kovariancia mátrix

HASONLÓSÁG, KÜLÖNBÖZŐSÉG, TÁVOLSÁG, SZOMSZÉDOSSÁG

DEFINÍCIÓK

- Adott két objektum
 - $x \rightarrow (x_1, x_2, \dots, x_n)$
 - $y \rightarrow (y_1, y_2, \dots, y_n)$
- Hasonlóság (similarity)
 - A két objektum mennyire hasonlít egymásra
 - $s(x, y) \in [0, 1]$
- Különbözőség
 - A hasonlóság komplementere
 - $d(x, y) = 1 - s(x, y)$
- Távolság
 - Hasonlatos a különbözőséghez
 - $d(x, y) \in [0, \infty)$
- Az egyes metrikák egymásba transzformálhatók

BINÁRIS ADATOK

- Kontingencia táblázat

- Elempárok előfordulásának száma

- **table()**

x_i/y_j	1	0	Σ
1	a	b	$a + b$
0	c	d	$c + d$
Σ	$a + c$	$b + d$	n

- Simple matching

- $s(x, y) = \frac{a+c}{a+b+c+d}$

- Jaccard index

- $s_J(x, y) = \frac{a}{a+b+c}$

- Dice hasonlóság

- $s_D(x, y) = \frac{2|x \cap y|}{|x|+|y|} = \frac{2a}{a+b+a+c} = \frac{2a}{2a+b+c}$

NOMINÁLIS, ORDINÁLIS ADATOK

- Nominális adatok
 - Kategorikus, felsorolás, diszkrét típusú adatok
 - $s(x, y) = \frac{\text{egyező attribútumok száma}}{\text{összes attribútum}}$
- Sorrend típusú adatok
 - Intervallum típusú adatokká konvertáljuk

INTERVALLUM ÉS ARÁNYSKÁLA TÍPUSÚ ADATOK

- Minkowski metrika

$$d(x, y) = \left[\sum_i |x_i - y_i|^p \right]^{\frac{1}{p}}$$

- Manhattan metrika, $p = 1$

$$d(x, y) = \sum_i |x_i - y_i|$$

- Euklideszi metrika, $p = 2$

$$d(x, y) = \left[\sum_i (x_i - y_i)^2 \right]^{\frac{1}{2}}$$

- Csebisev metrika, $p = \infty$

$$d(x, y) = \lim_{p \rightarrow \infty} \left[\sum_i |x_i - y_i|^p \right]^{\frac{1}{p}} = \max_i |x_i - y_i|$$

- Mahanalobis távolság

$$d(x, y) = \sqrt{(\mathbf{x} - \mathbf{y})^T \mathbf{C} (\mathbf{x} - \mathbf{y})}$$

SZOMSZÉDOSSÁG

- Objektumok esetében
 - Egy adott távolságnál közelebb van
- Gráfok esetében
 - Szomszédossági mátrix alapján

$$A_G = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n} \\ \dots & \dots & \dots & \dots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n} \end{bmatrix}$$