Kiegyenlítő számítások: Matematikai statisztika

Barsi Árpád

Mi a statisztika?

- A Wikipedia szerint: „a valóság számszerű információinak megfigyelésére, összegzésére, elemzésére és modellezésére irányuló gyakorlati tevékenység és tudomány”
- („számhasonlítás” – Kosztolányi Dezső: Nyelvművelés - Válasz Schöpflin Aladárnak. Nyugat · 1933)
- Etimológia: az olasz *statista* magyarul „államférfi”, politikus
- Ismert vicc szerint?

A **statiszta** a filmforgatásokon, a filmekben, stb. – olyan mellékszereplő, akinek általában elmondandó szövege nincs. Többnyire csak egy-két jelenetben fordul elő. A feladata többnyire csak az, hogy növelje a szereplő személyek számát olyan jelenetekben, ahol sok emberre, „tömegre” van szükség (innen eredeztethető az elnevezés is)
[Wikipedia]

Részterületei

- Leíró (deskriptív, empirikus) statisztika
- Következtető (induktív, inferencia) statisztika
- Exploratív (analítikus) statisztika, adatbányászat (data mining)

Történelem helyett

- Még két brit kutató
- Karl Pearson (1857 – 1936)
  - Matematikus, fizikus, történész, germanista…
  - Híres könyve (The grammar of science) → Einstein 1902
  - Érdemei: korrelációs együttható, momentumok, χ^2-eloszlás, hisztogram, PCA, hipotézis-vizsgálat
- Ronald Fisher (1890 – 1962)
  - Matematikus, genetikus, evolúciókutató…
  - Érdemei: variancia-analízis, maximum likelihood, lin. Diszkrimináns (Iris data set), F-eloszlás, kísérlettervezés

A minta

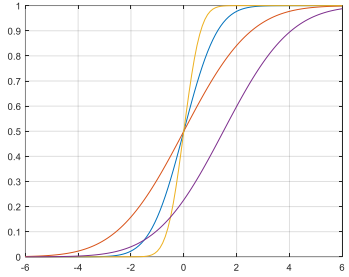
- Definíciószerűleg: valamely valószínűségi változóra vonatkozó véges számú független kísérlet vagy megfigyelés
- A minta nagysága, elemszáma: n
- A minta elemei:
  - ξ esetén: $\mathbf{L} = [L_1, L_2, \ldots, L_n]$
  - $\boldsymbol{\xi} = [\xi_1, \xi_2, \ldots, \xi_k]$ esetén:

$$\underset{(n,k)}{\mathbf{L}} = \left[\begin{array}{c|cccc} & \xi_1 & \xi_2 & \cdots & \xi_k \\ \hline 1 & L_{11} & L_{12} & \cdots & L_{1k} \\ 2 & L_{21} & L_{22} & \cdots & L_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ n & L_{n1} & L_{n2} & \cdots & L_{nk} \end{array}\right]$$

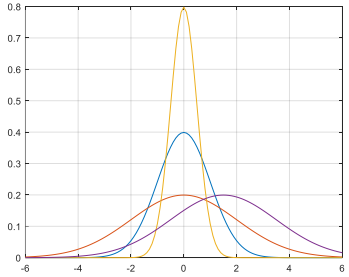
- A mintavételezés folyamata

A minta jellemzése

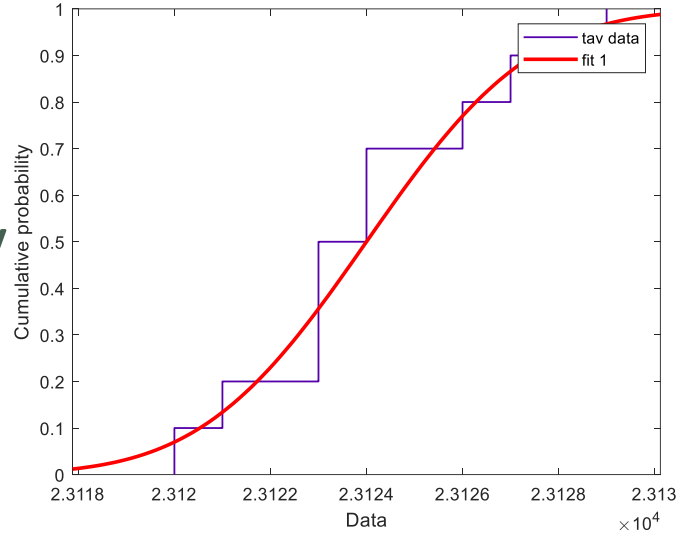
- Tapasztalati eloszlásfüggvény

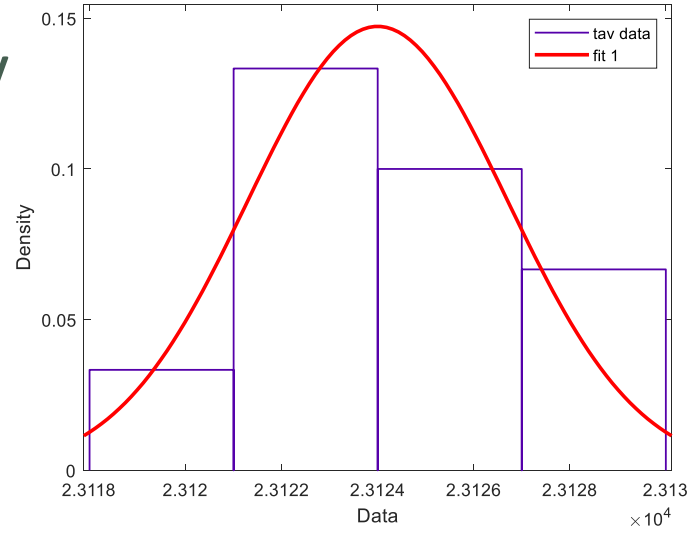


- Tapasztalati sűrűségfüggvény (hisztogram)

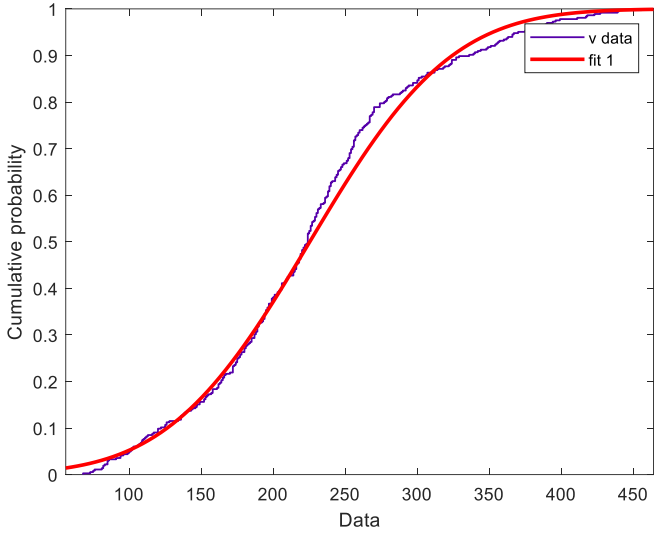


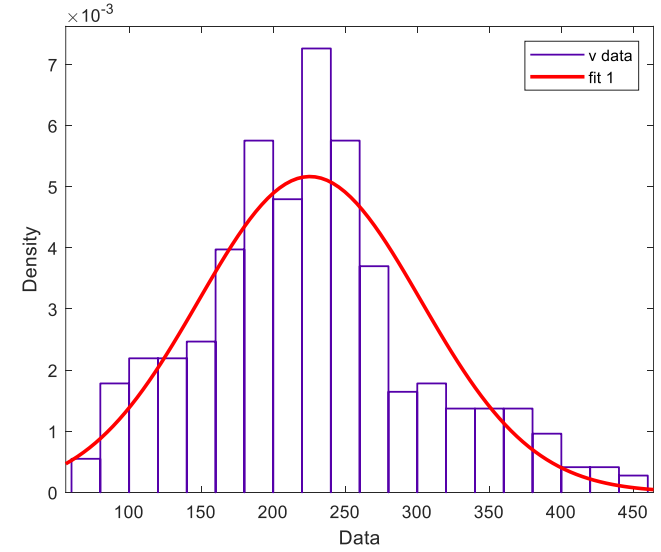
Távmérés (n=10)





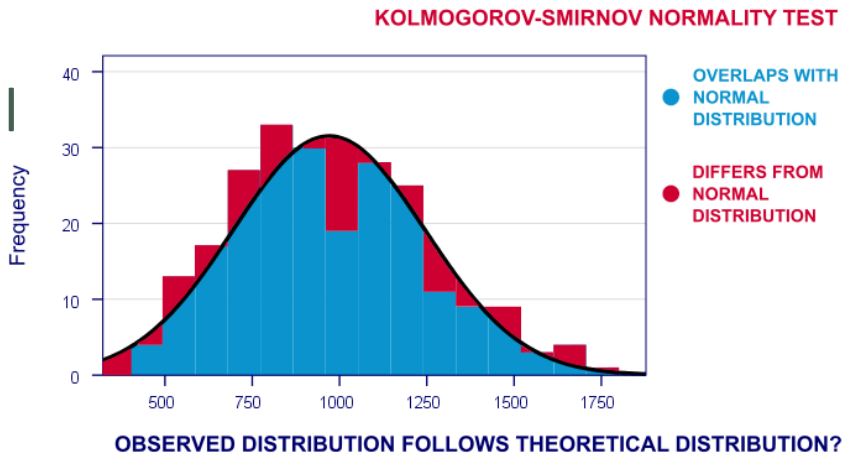
Vízállás (n=365)





Normalitásvizsgálat

- Cél: igazolni, hogy a valószínűségi változó normális eloszlású-e
- Matematikai háttér: (normális) eloszlásfüggvény illeszkedésvizsgálata
- Megoldásai:
  - Vizuálisan
    - Pl. hisztogram segítségével
  - Momentumok kiszámításával
    - Pl. ferdeségi és lapultsági együttható számításával
  - Próba alkalmazásával
    - Pl. Kolmogorov-Szmirnov-teszt
    - Pl. Lilliefors-teszt



Tapasztalati jellemzők

- Pontbecslés és intervallumbecslés
- Várható érték ⟹ mintaközép (átlag)

$$\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i \qquad \mathbf{L} = [L_1, L_2, \ldots, L_n]$$

- Szórásnégyzet

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(L_i - \bar{L}\right)^2 \qquad \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(L_i - \bar{L}\right)^2$$

- Kovariancia

$$\hat{c}_{ij} = \frac{1}{n-1}\sum_{s=1}^{n}\left(L_{si} - \bar{L}_i\right)\cdot\left(L_{sj} - \bar{L}_j\right)$$

$$\mathbf{L}_{(n,k)}$$

- Medián
  - Rendezett minta középső eleme vagy két középső átlaga

Egy kis emlékeztető a múlt óráról

- A valószínűség értékének kifejezése elméletileg

$$p = P(a \le \xi < b) = \underbrace{F(b) - F(a)}_{CDF} = \underbrace{\int_a^b f(t)dt}_{PDF}$$

- Gyakorlati számításokban
  - Tapasztalati eloszlás vagy sűrűségfüggvény

Becslés normális eloszlás esetén

- Standardizált normális eloszlás képleteivel

$$P(a \le \xi < b) = F(b) - F(a)$$

$$p = P(a \le \xi \le b) = P\left(\frac{a-m}{\sigma} \le \frac{\xi - m}{\sigma} \le \frac{b-m}{\sigma}\right) = \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right)$$

- Ahol $N(m,\sigma) \rightarrow N(0,1)$

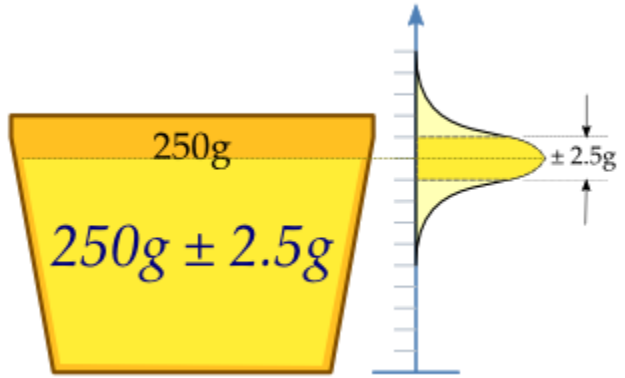
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}} dt \rightarrow \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Konfidencia intervallum definíciója

- Ismeretlen *q* paraméterre
- $[L_1, L_2, \ldots, L_n]$ minta birtokában
- *a* és *b* intervallumra
- *p* valószínűségi szinten

$$p = P(a \le q < b)$$

- *p* elnevezése: konfidencia szint
- *a* és *b* elnevezése: konfidencia intervallum



Konfidencia intervallum

- Várható értékre – szimmetrikus intervallum a mintaközéphez képest

$$p = P(a \le q \le b)$$

- Szórásokkal általában

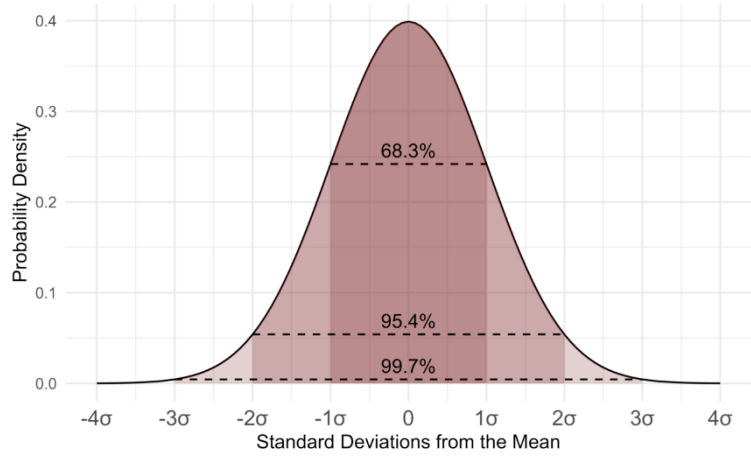
$$p = P(m - u\sigma \le q \le m + u\sigma) = P\left(\frac{m - u\sigma - m}{\sigma} \le q' \le \frac{m + u\sigma - m}{\sigma}\right)$$

$$p = P(-u \le q' \le +u) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1$$

$$P(-\sigma \le \xi - m \le \sigma) = 0.6827$$

$$P(-2\sigma \le \xi - m \le 2\sigma) = 0.9545$$

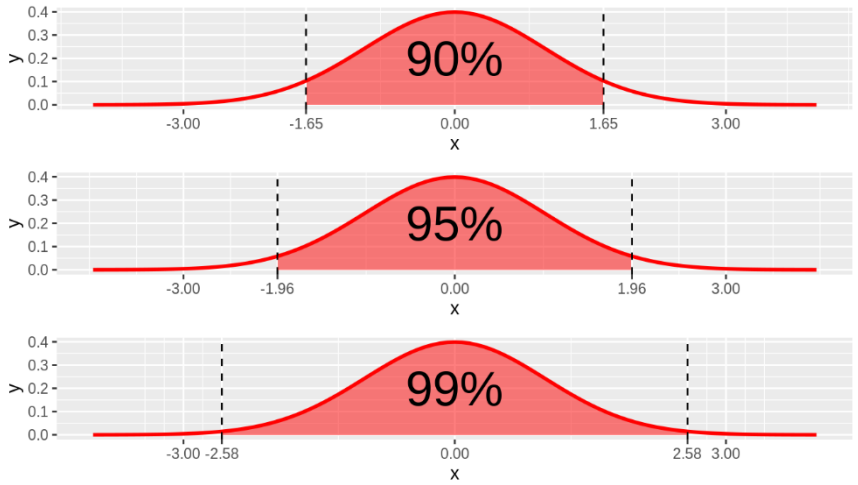
$$P(-3\sigma \le \xi - m \le 3\sigma) = 0.9973$$



Konfidencia intervallum – várható értékre

- Inverz feladat: szimmetrikus intervallum legyen adott *p* szinten, standardizált normális eloszlás összefüggései segítségével

$$P\left(-u_p \le \frac{m-\bar{L}}{\frac{\hat{\sigma}}{\sqrt{n}}} \le +u_p\right) = p = 1-\alpha$$



- Ebből pl. *p*=0.95 vagy másként 95%-os szinten

$$P\left(\bar{L}-u_p\frac{\hat{\sigma}}{\sqrt{n}} \le m \le \bar{L}+u_p\frac{\hat{\sigma}}{\sqrt{n}}\right) = 0.95 \qquad p = 2\Phi\left(u_p\right)-1 \qquad u_p = 1.96$$

Konfidencia intervallum – szórásra

- A számítás elve hasonló, azonban az előzőektől eltérően NEM a standardizált normális eloszlást, hanem a χ^2-eloszlás összefüggései kellenek.

- Gyakorlati példa következik…

Hipotézisvizsgálat

- Definíció: egy vagy több valószínűségi változó eloszlására vagy annak paramétereire vonatkozó feltevések vizsgálata
- Munkamódszer:
  - Nullhipotézis: H_0
  - Alternatív hipotézis: H_1
  - Döntés: statisztikai próba
  - Lehetséges kimenet:

	H_0 hipotézist	
	elfogadjuk	elutasítjuk
H_0 fennáll	helyes döntés (True Positive)	1. fajú hiba (False Negative)
H_0 nem áll fenn	2. fajú hiba (False Positive)	helyes döntés (True Negative)

Statisztikai próba egyetlen várható értékre

- H_0: $m = m_0$
- Ha ismert a szórás: egymintás **u-próba** *p* konfidencia szinten

$$u = \frac{m_0 - \bar{L}}{\frac{\sigma}{\sqrt{n}}} \qquad |u| \le u_p \quad elfogadás \qquad |u| > u_p \quad elvetés$$

- Ha nem ismert a szórás: egymintás **t-próba** *p* konf. szinten, *f* szab. fokon

$$t = \frac{m_0 - \bar{L}}{\frac{\hat{\sigma}}{\sqrt{n}}} \qquad |t| \le t_{p,f} \quad elfogadás \qquad |t| > t_{p,f} \quad elvetés \qquad f = n - 1$$

Statisztikai próba két várható értékre

- H_0: $m_1 - m_2 = \delta$ vagy $m_1 = m_2$
- Ha ismert a szórás: kétmintás **u-próba** *p* konfidencia szinten

$$u = \frac{\overline{L}_1 - \overline{L}_2 - \delta}{\sqrt{n_2\sigma_1^2 + n_1\sigma_2^2}}\sqrt{n_1 n_2}$$

$|u| \le u_p$ *elfogadás*

$|u| > u_p$ *elvetés*

- Ha nem ismert a szórás: kétmintás **t-próba** *p* konf. szinten, *f* szab. fokon

$$t = \frac{\overline{L}_1 - \overline{L}_2 - \delta}{\sqrt{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}}\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

$|t| \le t_{p,f}$ *elfogadás*

$|t| > t_{p,f}$ *elvetés*

$f = n_1 + n_2 - 2$

Statisztikai próba egyetlen szórásra

- H_0: $\sigma^2 = \sigma_0^2$
- **χ^2-próba** *p* konfidencia szinten

$$\hat{\chi}_f^2 = f\,\frac{\hat{\sigma}^2}{\sigma_0^2}$$

$$\hat{\chi}_f^2 \le \chi_{\alpha,f}^2 \quad \textit{elfogadás}$$

$$\hat{\chi}_f^2 > \chi_{\alpha,f}^2 \quad \textit{elvetés}$$

$$f = n-1 \qquad \alpha = 1-p$$

Statisztikai próba két szórásra

- H_0: $\sigma_1^2 = \sigma_2^2$
- **F-próba** *p* konfidencia szinten

$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

$F \le F_{\alpha, f_1, f_2}$ *elfogadás*

$F > F_{\alpha, f_1, f_2}$ *elvetés*

$f_1 = n_1 - 1$

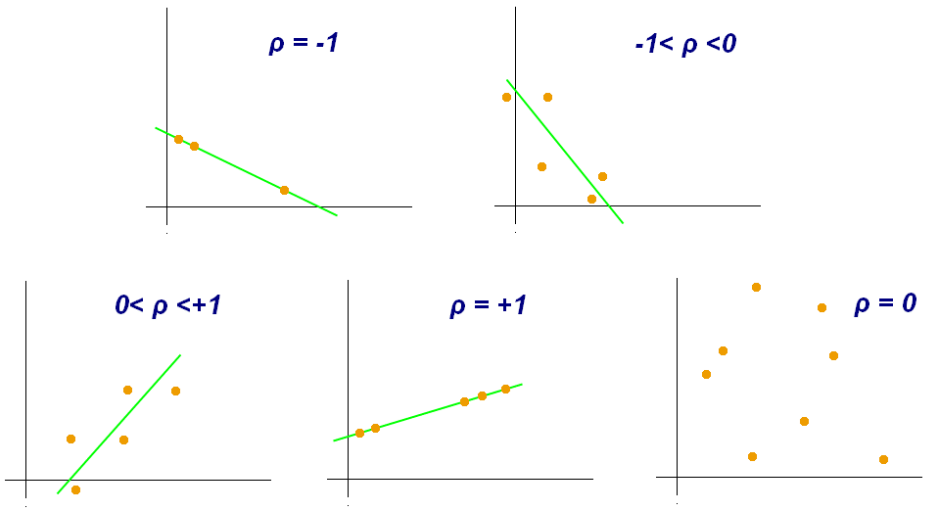
$f_2 = n_2 - 1$

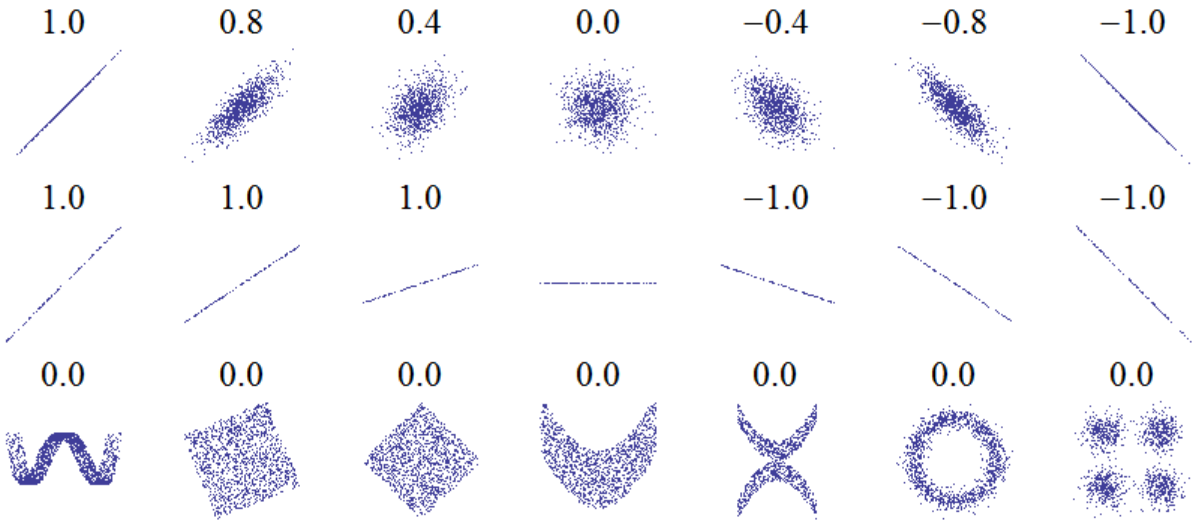
$\alpha = 1 - p$

Korreláció és regresszió

- Meghatározása tapasztalati jellemzőként

$$\hat{r}_{ij} = \frac{\hat{c}_{ij}}{\hat{\sigma}_i \cdot \hat{\sigma}_j}$$





Köszönöm a figyelmet!