Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

# Biostatisztika – 1.

Dr. Dinya Elek –Dr. Solymosi Róbert: Biometria a klinikumban
Dr. Dinya Elek: Biostatisztika

Hullám Gábor

MŰEGYETEM 1782

# Változók típusai

A változók ***jellegük szerint lehetnek:***

- Minőségi - kvalitatív leírók
  - Pl.: végzettség, nem, lakhely
- Mennyiségi - kvantitatív leírók
  - Pl.: életkor, vérnyomás

A változók ***értékkészletük szerint lehetnek:***

- Diszkrét
  - Pl.: testvérek száma
- Folytonos
  - Pl.: testmagasság

# Változók értékeinek skálája

- **Nominális:** a változó lehetséges értékei között csak az azonos vagy nem azonos reláció értelmezett (=,≠).
- **Ordinális**: a változó lehetséges értékei között a kisebb v. nagyobb (<,>) reláció (rangsorolás) is értelmezett.
  (+ az előbbi relációk) pl.: termék minőségi osztályok
- **Intervallum**: a változó értékei között értelmezhető a távolság (különbség), pl.: hőmérséklet
  (nincs valódi nullpontja a skálának)
- **Arányskála**: mind a négy alapművelet értelmezett a változó értékein, pl.: beavatkozás költsége

# Teljes populáció vs. minta

**Sokaság (populáció)**

- Mindazon elemek halmaza, amelyre a statisztikai következtetés irányul.
- A jellemző változók (ismérvek) által felvett vagy felvehető értékek halmaza.

- Két típus vizsgálat időtartama szerint:
  - Időpontban: **álló sokaság**
  - Időintervallumban: **mozgó sokaság**

# Teljes populáció vs. minta

**Minta**

- Egy adott véges sokaságból kiválasztott véges számú elemek halmaza
- Megmért, rendelkezésre álló adathalmaz
- A minta vizsgálata alapján vonunk le következtetéseket a sokaságra
- Reprezentativitás: egyes változók értékeinek az aránya a mintában megegyezik a sokaságbeli aránnyal

# Mintavételi típusok

- **Véletlenszerű mintavételezés** (random sampling)
- **Rétegzett mintavételezés** (stratified sampling)
  - alpopulációk *(rétegek, strata) kialakítása*
  - minden rétegből véletlenszerű mintavételezés
- **Csoportos (többlépcsős) mintavételezés** (cluster sampling)
  - A populáció csoportokra bontása egy ismérv mentén, majd ebből véletlenszerű mintavételezés
  - Opcionálisan: mintákból alcsoportok kialakítása, majd újabb mintavételezés
- **Véletlen besorolásos vizsgálatok (random assignment)**
  - Résztvevők kiválasztása előzetesen megállapított kritériumok szerint
  - véletlenszerű eljárással csoportba sorolás
- **Kényelmi mintavételezés**
  - Nem ismert a populáció, paraméterei nem becsülhetőek
- **Kvóta minta**
  - Részben ismert populáció, paraméterei becsültek
  - Akár több ismérv szerint is lehet, pl.: a megkérdezettek 50-50% legye nő illetve férfi, és 20% legyen fővárosi, 80% legyen vidéki.

# Mintavételi hiba

- Mintavételi hiba: a minta ‘eltérése’ a teljes populációhoz képest
  - ‘Csak egy részletet látunk, nem a teljes képet’
  - **Függ: a (1) mintaszámtól és a (2) mintavételi módszertől**
  - Minél nagyobb a minta, annál nagyobb bizonyossággal vonhatunk le következtetéseket a sokaságra vonatkozóan
  - Véletlen mintánál számítható érvényesen

- Nem mintavételi hiba:
  - Adatfelvételnél keletkezhet:
    - Szisztematikus vagy véletlen hiba
    - Hiányzás (MCAR / MAR / NMAR)
  - Adattisztítási és imputálási lépésekkel kezelhető

# Változók – gyakoriság (minta alapján)

- (Abszolút) **Gyakoriság**: egy változó által felvehető lehetséges értékekre jutó megfigyelések száma.
  - Diszkrét esetben: pl.: 4 fej és 6 írás a 10 pénzfeldobásból
  - Folytonos esetben csak intervallumra vonatkoztatva értelmezhető !! Pl.: testmagasság *X<175 cm vagy 180,4 cm < X < 180,8 cm*

- **Relatív gyakoriság:** egy változó által felvehető lehetséges értékekre vonatkoztatott gyakoriság elosztva a teljes minta nagyságával

# Események – véletlen kísérlet (=minta)

- **Elemi esemény:** minden lehetséges kimenetel ($e_1, e_2, \ldots, e_n$), amiről a kísérlet elvégzése után eldönthető, hogy bekövetkezett vagy sem
  - $e_1 \wedge e_2 \wedge \ldots \wedge e_n = \emptyset$
- **Eseménytér:** a kísérlet összes kimenetele, az összes elemi esemény halmaza ($\Omega$).
  - $e_1 U e_2 U \ldots U e_n = \Omega$
- **Véletlen esemény:** az eseménytér egy részhalmaza
- **Biztos esemény:** a kísérlet során biztosan (minden kimenetelnél) bekövetkezik.
- **Ellentett esemény**: akkor és csak akkor következik be, ha az eredeti esemény nem következik be

# Valószínűség

- Egy adott $A$ esemény valószínűsége ($P(A)$) az a számérték, amely körül az esemény relatív gyakorisága ($f_A$) ingadozik, ha egyre több kísérletet végzünk.

$$P(A) = \lim_{n \to \infty} f_A$$

- $P(A)$ tehát azt mutatja meg, hogy az $A$ esemény az összes kísérlet mekkora hányadában (%) következik be.
- Minden $A$ eseményre $0 \leq P(A) \leq 1$
- $P(\setminus A) = 1 - P(A)$ , ahol $\setminus A$ az $A$ esemény *ellentettje*

- Egy teljes eseményrendszer valószínűségeinek sorozatát valószínűségeloszlásnak nevezzük

# Valószínűségi változó

- Valószínűségi változó: olyan változó, melynek értékei egy véletlen kísérlet lehetséges kimenetelei
- Kapcsolódó fogalmak:
  - Kolmogorov-féle valószínűségi mező
    - Eseménytér ($\Omega$)
    - Eseményalgebra ($F$)
    - Valószínűségi mérték ($P$)
- Eloszlásfüggvény: leírja, hogy egy valószínűségi változó milyen valószínűséggel vehet fel egy adott értéket.

# Diszkrét eloszlások

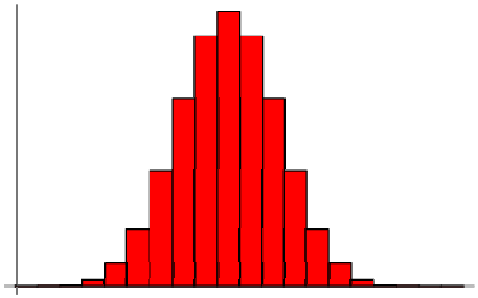
Diszkrét valószínűségi változók esetén

- Binomiális eloszlás
- Poisson-eloszlás
- Bernoulli-eloszlás
- Hipergeometrikus eloszlás
- Diszkrét egyenletes eloszlás

# Diszkrét eloszlások

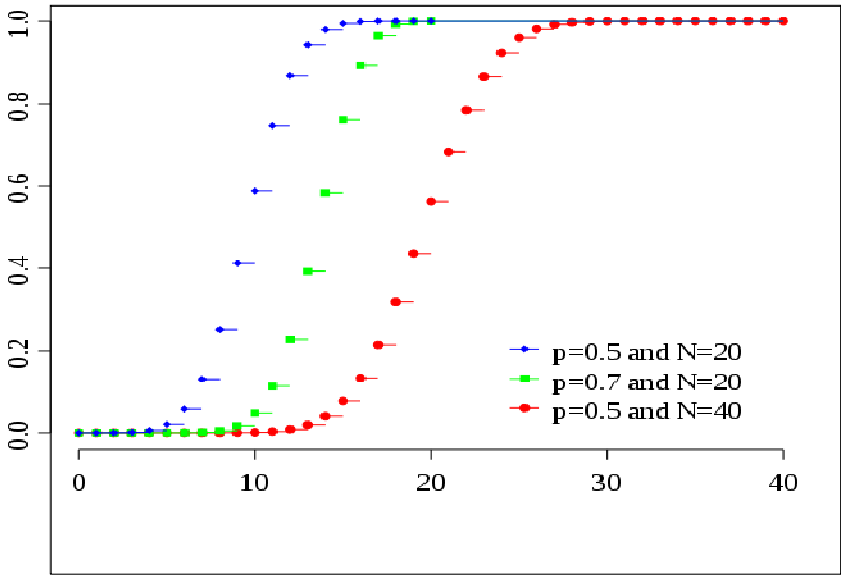
## Binomiális eloszlás

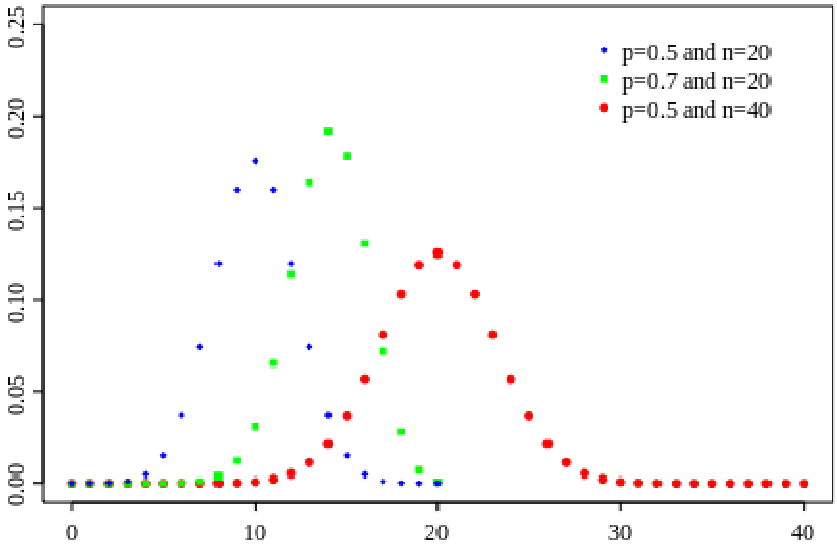
- Valószínűség tömegfüggvény (pmf)
- Kumulatív eloszlásfüggvény (cdf)



$$F(k;n,p) = \Pr(X \le k) = \sum_{i=0}^{\lfloor k \rfloor} \binom{n}{i} p^i (1-p)^{n-i}$$

$$Pr(k;n,p) = \Pr(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



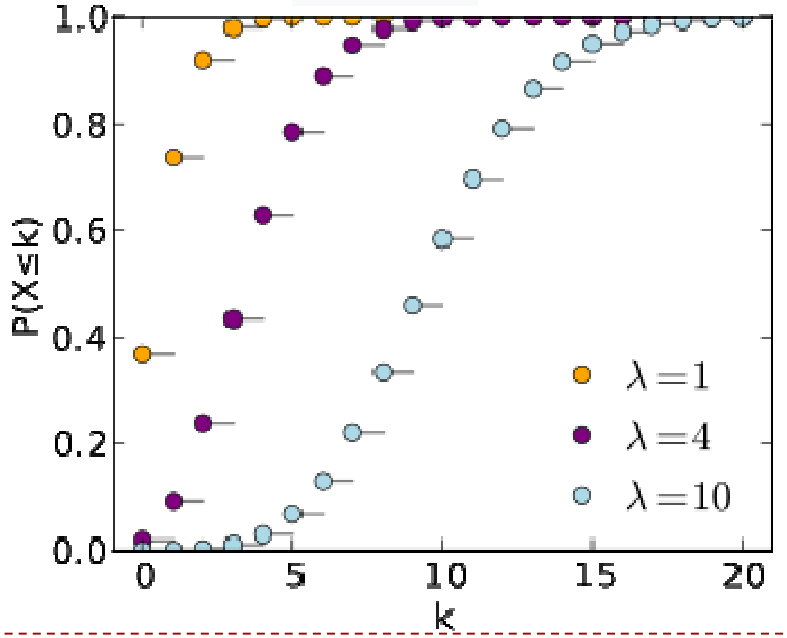



# Diszkrét eloszlások

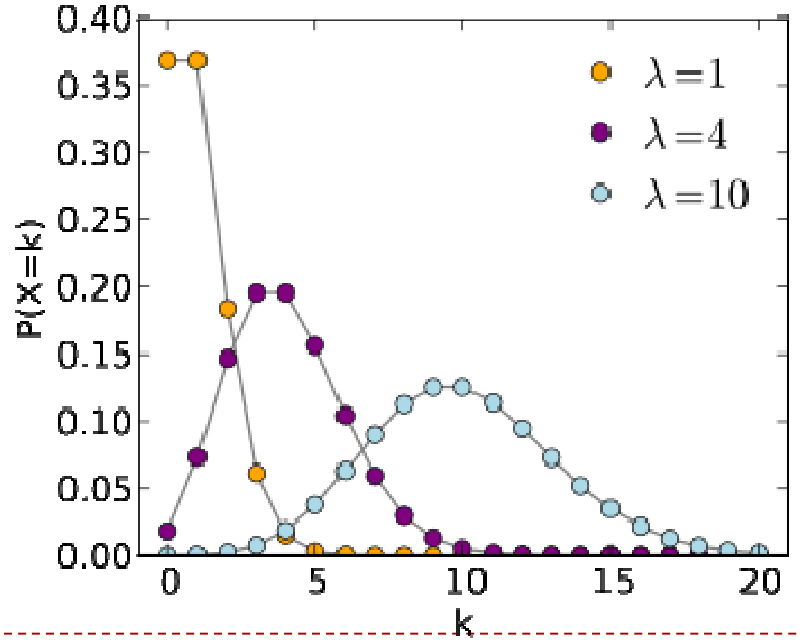
## Poisson eloszlás

- Valószínűség tömegfüggvény (pmf)
- Kumulatív eloszlásfüggvény (cdf)

$$F(k,\lambda) = e^{-\lambda}\sum_{i=0}^{\lfloor k \rfloor}\frac{\lambda^i}{i!}$$




$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

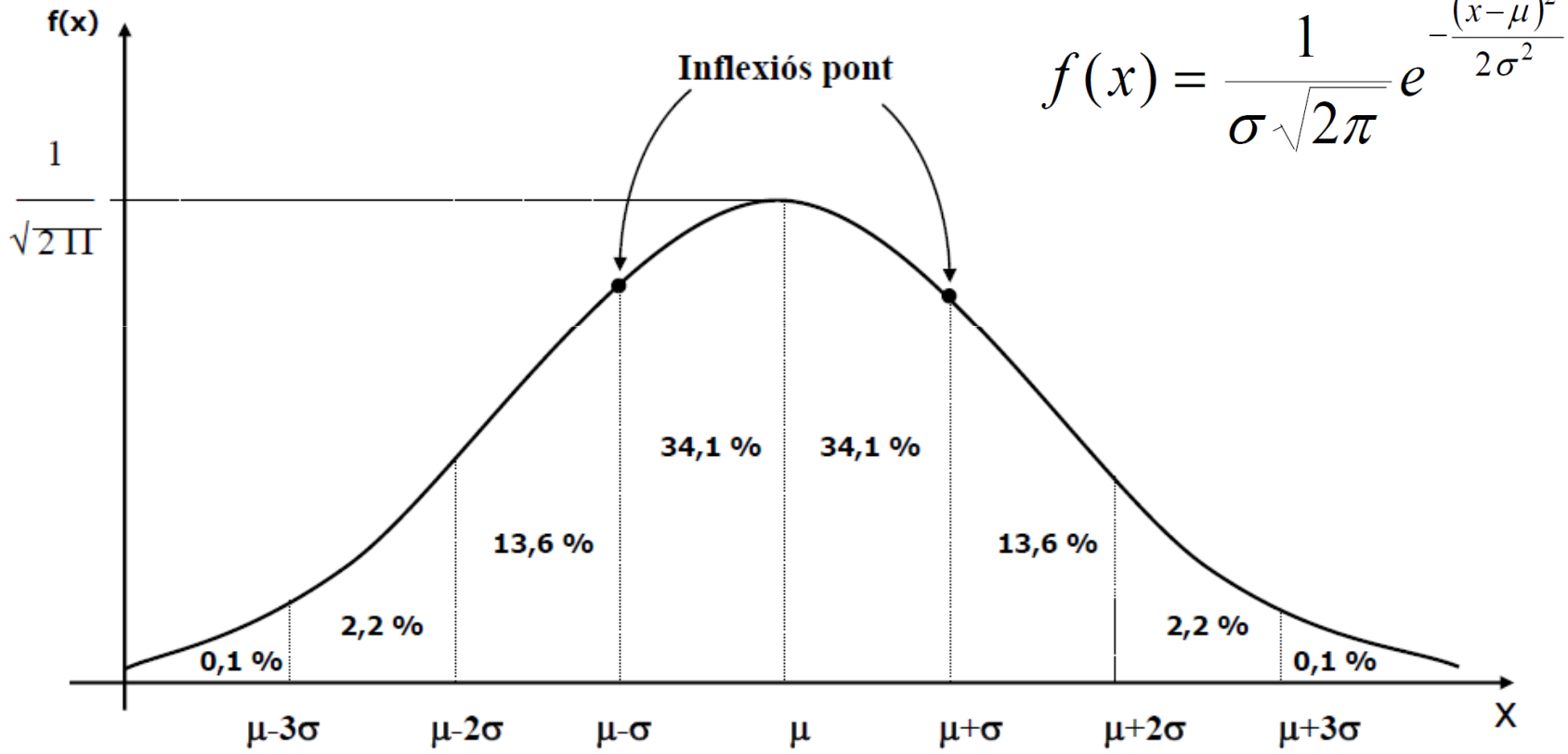


# Folytonos eloszlások

- Normális eloszlás
- Exponenciális eloszlás
- Khí-négyzet eloszlás
- F-eloszlás
- T-eloszlás

# Folytonos eloszlások – normál eloszlás

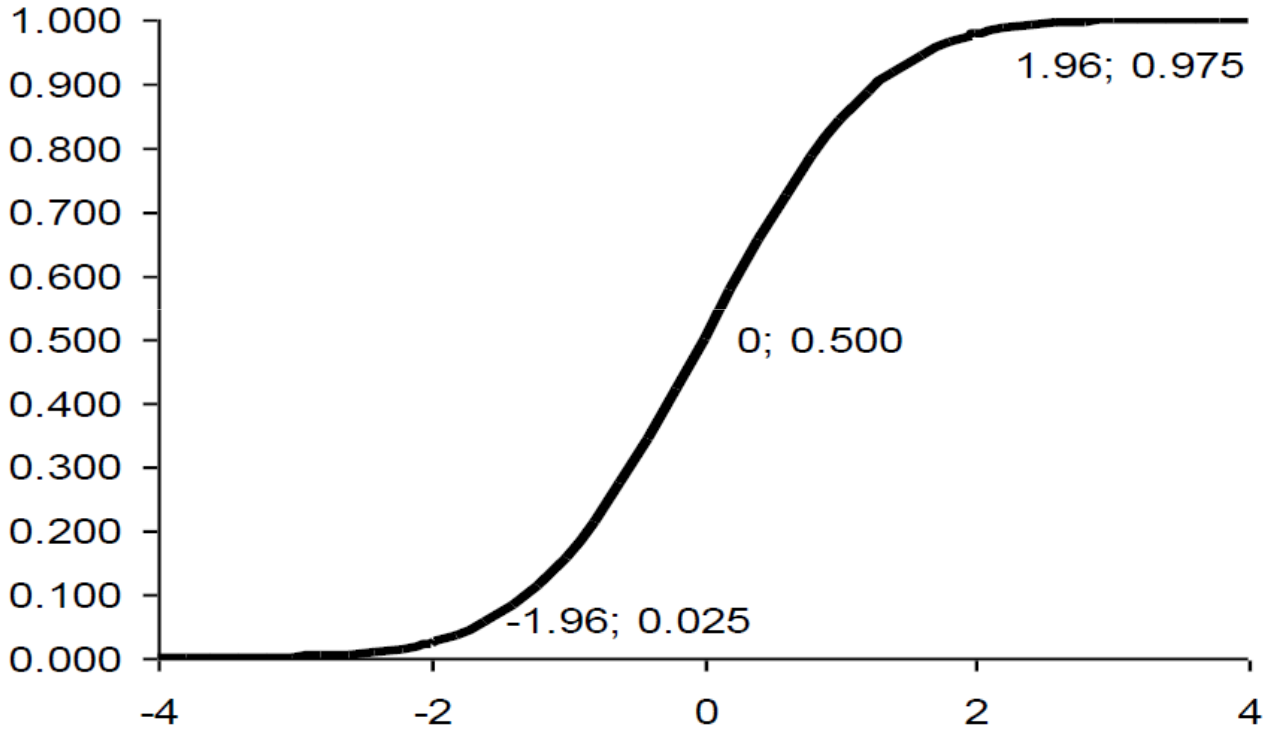
- Valószínűségsűrűség függvény (pdf)




$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# Folytonos eloszlások – normál eloszlás

- Kumulatív eloszlásfüggvény (cdf)



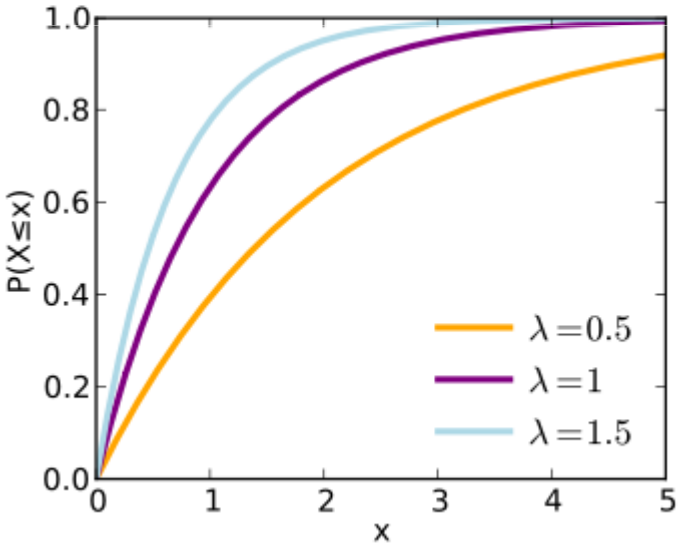

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

# Folytonos eloszlások

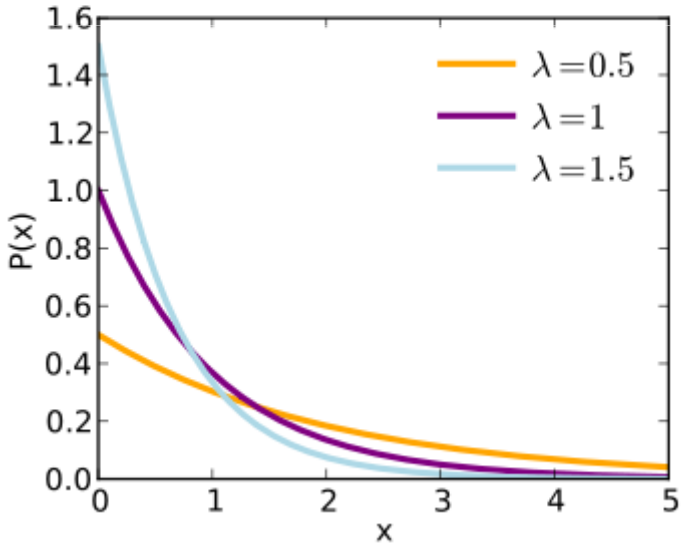
## Exponenciális eloszlás

- Valószínűségsűrűség függvény (pdf)
- Kumulatív eloszlásfüggvény (cdf)

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0 \end{cases}$$




$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$



# Eloszlás jellemzése középértékkel

Középértékek két típusa:

- Számított középérték
- Helyzeti középérték

Követelmények:

- a) Számított középérték: közbenső helyet foglaljon el: $x_{min} \leq$ középérték $\leq x_{max}$
- b) Helyzeti középérték: gyakori érték legyen
- c) Legyen könnyen meghatározható és egyértelmű

# Számított középértékek

Átlag

- Egyszerű
- Súlyozott (figyelembe veszi az azonos értékek gyakoriságát)

Típusai:

- Számtani (Aritmetikai) átlag
- Harmonikus átlag
- Mértani (Geometriai) átlag
- Négyzetes (Kvadratikus) átlag

# Számított középértékek

Átlagok:

Számtani (aritmetikai)

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\bar{x}_a = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}$$

Mértani (geometriai)

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

$$\bar{x}_g = \sqrt[\sum_{i=1}^{n} f_i]{\prod_{i=1}^{n} x_i^{f_i}}$$

Harmonikus

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

$$\bar{x}_h = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

Négyzetes (kvadratikus)

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

$$\bar{x}_q = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 \cdot f_i}{\sum_{i=1}^{n} f_i}}$$

# Helyzeti középértékek

Módusz (*Mo*)

- Diszkrét eset: a legnagyobb gyakorisággal előforduló érték
- Folytonos eset: az az érték, ahol a gyakorisági görbe a maximumot veszi fel

Kvantilisek ($K_j^k$)

- az összes előforduló érték j/k (j=1,2,…,k-1) része kisebb és 1-(j/k) része nagyobb
- Pl.: $K_2^3$ : k=3, j=2 az összes érték 2/3-a kisebb, mint $K_2^3$, 1/3-a nagyobb

Típusai:

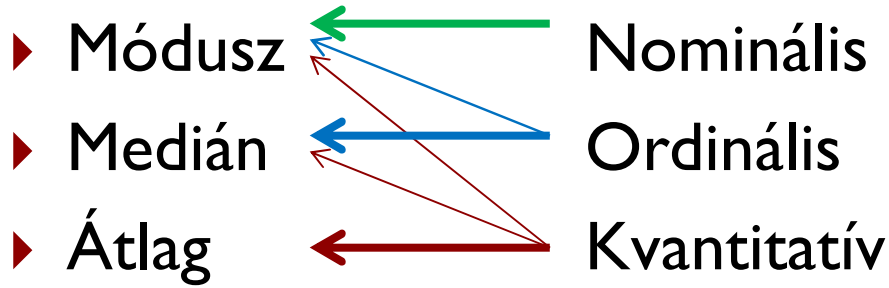
- k=2 Medián (*Me*)
- k=3 Tercilis
- k=4 Qvartilis (*Q*) : *Q1, Q2 = Me, Q3*
- k=10 Decilis
- K=100 Percentilis

# Medián

Medián (*Me*)

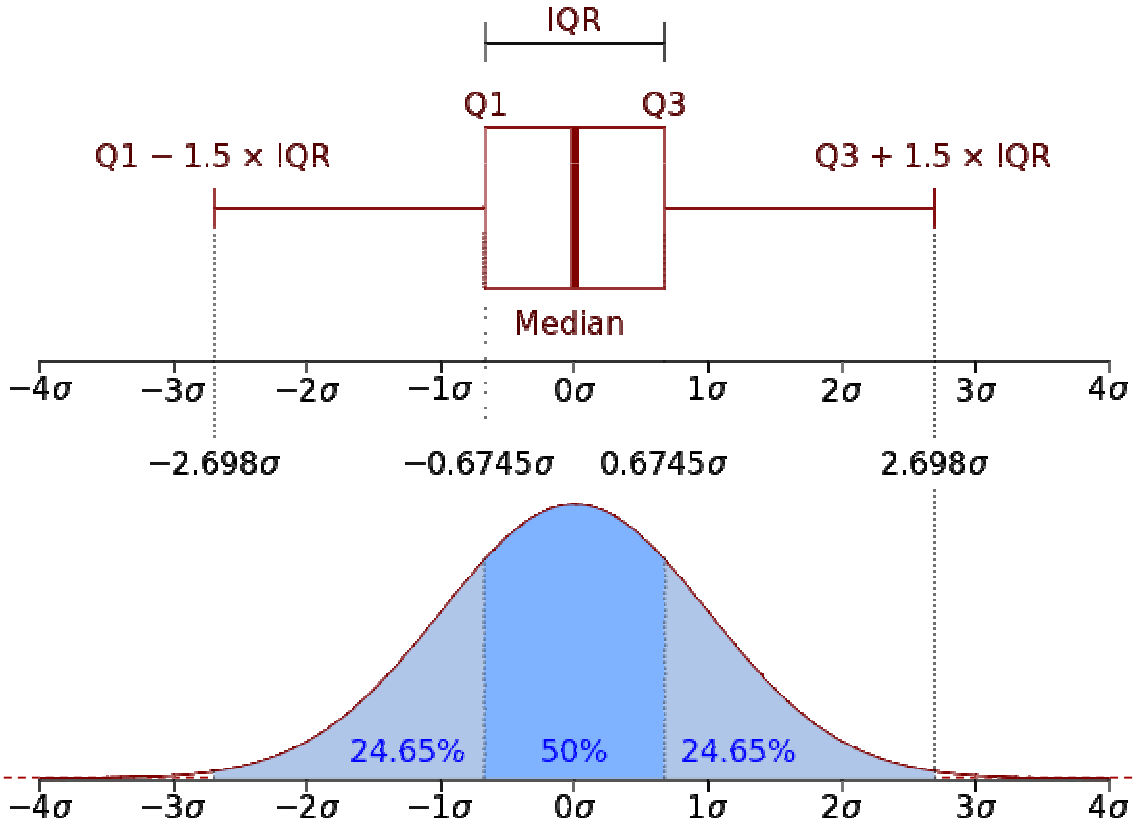
- Nagyságság szerint rendezett értékek közül a középső érték
- A szélső értékek nem befolyásolják

Mikor melyik középértéket célszerű alkalmazni?



# Kvartilisek

- Q1: alsó  kvartilis
- Q3: felső kvartilis



# A szóródás mérése

- **Szóródás:** egy változó értékeinek a *különbözősége*

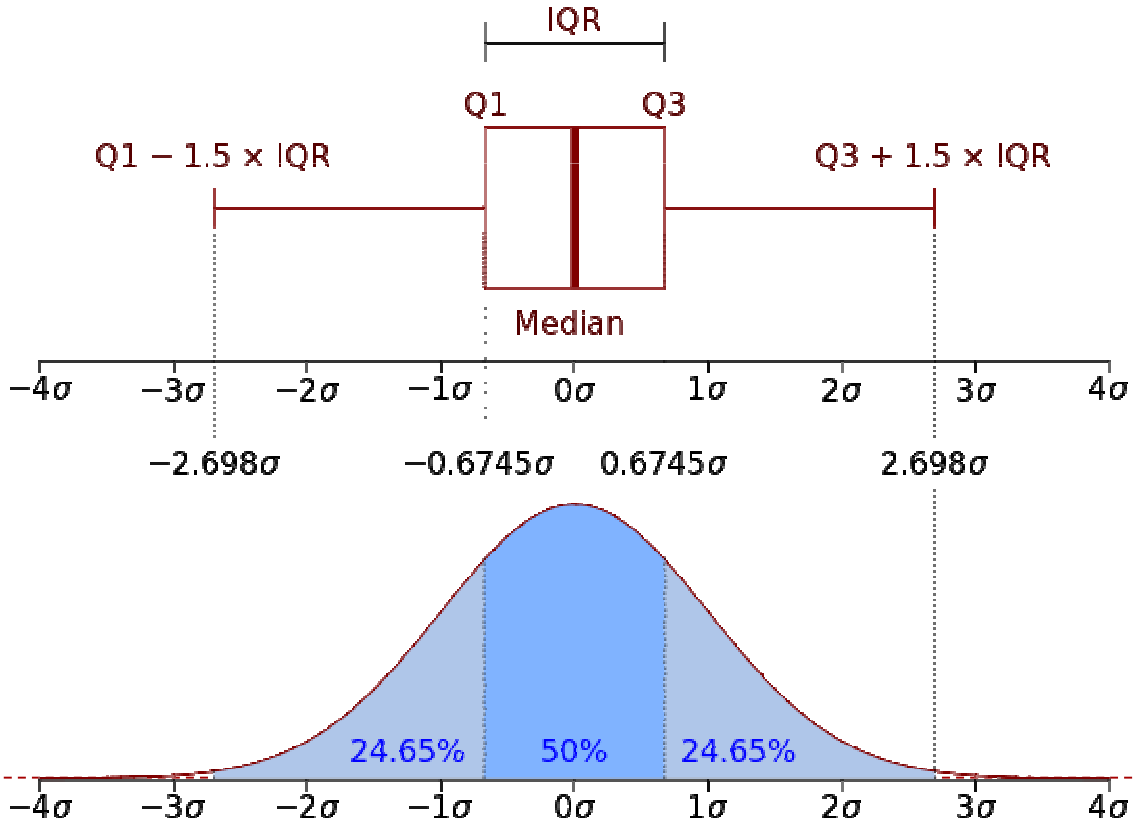
- **Mérése:** egy változó értékeinek egymás közötti vagy a változó valamely középértékétől vett különbségei alapján

**Szóródási mutatók**

- A szóródás terjedelme
- Átlagos abszolút eltérés
- Szórásnégyzet (variancia), szórás, relatív szórás
- Koncentráció

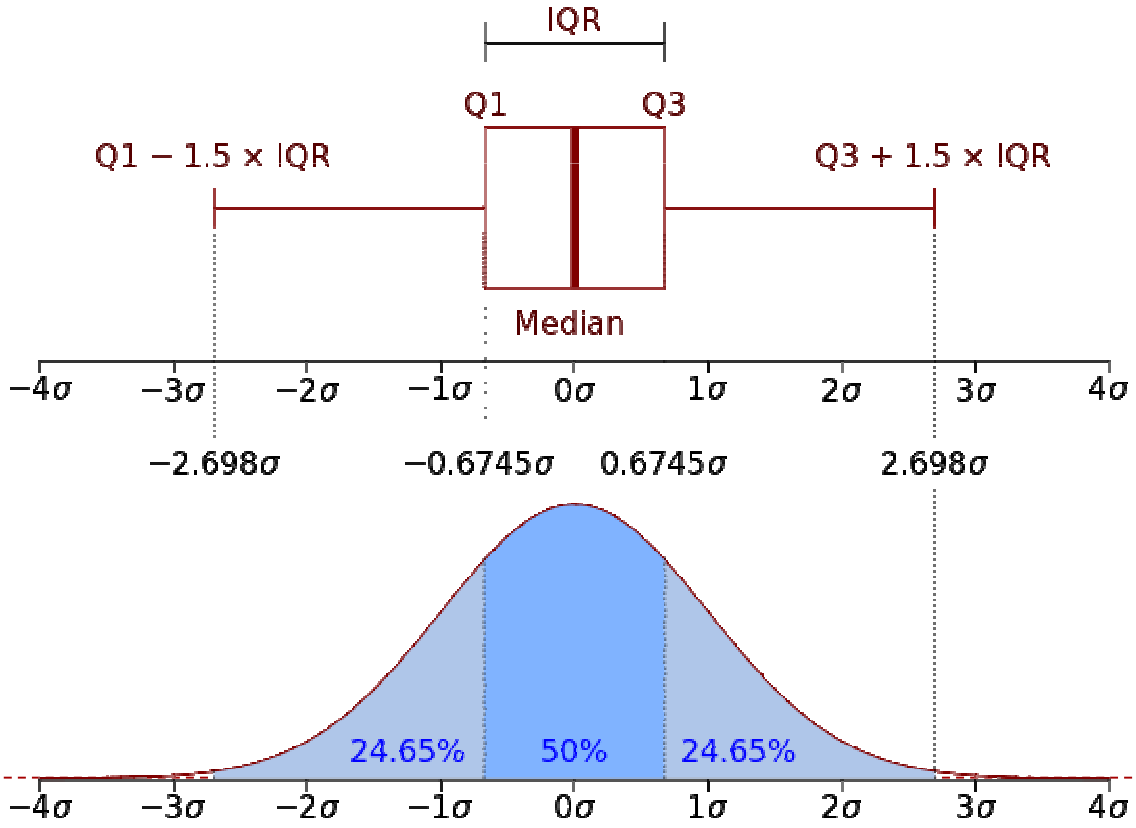
# A szóródás terjedelme - IQR

- Terjedelem (range) : $T = x_{max} - x_{min}$
- IQR (interkvartilis terjedelem): Q3-Q1



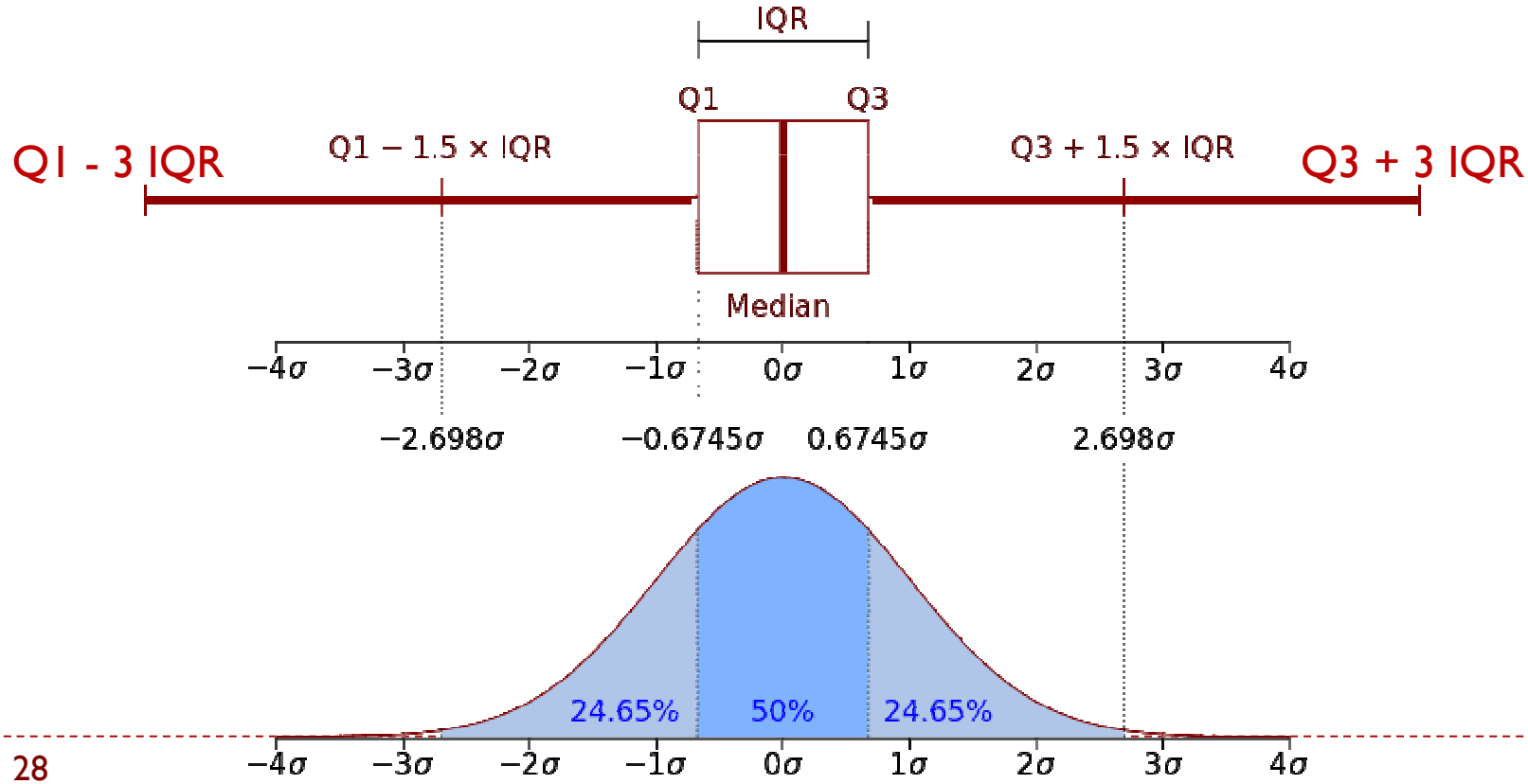
# Terjedelem - outlier

- Q1 – 1,5 IQR : alsó küszöb, ez alatt már (enyhe) outlier
- Q3 + 1,5 IQR : felső küszöb, e felett már (enyhe) outlier



# Terjedelem – extrém outlier

- Q1 – 3 IQR : alsó küszöb, ez alatt már extrém outlier
- Q3 + 3 IQR : felső küszöb, e felett már extrém outlier



# Szórás és variancia

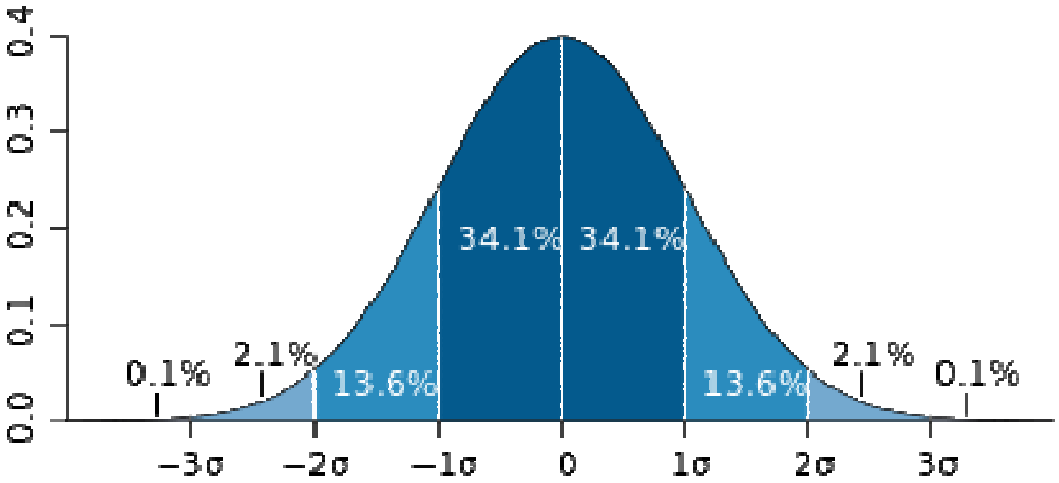
## Variancia (szórásnégyzet) : $\sigma^2$ vagy $s^2$

- A számtani átlagtól való négyzetes eltérés várható értéke

$$\operatorname{Var}(X) = \mathrm{E}\left[(X - \mu)^2\right].$$

$$Var(x) = s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N} \qquad s^{*2} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N\text{-}1}$$

# Szórás




- Szórás (Standard deviation) Korrigált szórás (Sample standard deviation)

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

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

# A szórás tulajdonságai

- Ha minden *y* értékhez ugyanazt a konstans számot hozzáadjuk (*y*+a), a szórás változatlan marad.
- Ha minden *y* értéket ugyanazzal a *k* konstans számmal megszorozzuk, (*k**y), a szórás is k-szorosára változik.
- Az eltérésnégyzet-összeg az átlagtól való eltérésekkel számolva a legkisebb
- A szórásnégyzet felírható a négyzetes átlag és a számtani átlag négyzetének a különbségeként.
- A sokaságot jellemző teljes szórásnégyzet (variancia) megegyezik a részsokaságok külső és belső szórásnégyzetének összegével (ANOVA): $\sigma^2 = \sigma_B^2 + \sigma_K^2$

# További szóráshoz kapcsolódó mutatók

- Variációs együttható – relatív szórás (V)

$$V\% = \frac{s}{\bar{x}} * 100$$

- Átlag szórása – Standard error of mean (SEM)

$$S_{\bar{x}} = \frac{S}{\sqrt{N}}$$

# Eloszlás aszimmetria

Az aszimmetria Pearson-féle **A-mutatószáma:**

- Szimmetrikus eloszlás esetén: A **= 0**
- Jobb oldali aszimmetria esetén: A **> 0**
- Bal oldali aszimmetria esetén: A **< 0**

$$A = \frac{\bar{x} - Mo}{\sigma}$$

Az aszimmetria **F-mutatószáma:**

$$F = \frac{(Q_3 - Me) - (Me - Q_1)}{(Q_3 - Me) + (Me - Q_1)}$$

- Szimmetrikus eloszlás esetén: F **= 0**
- Jobb oldali aszimmetria esetén: F **> 0**
- Bal oldali aszimmetria esetén: F **< 0**

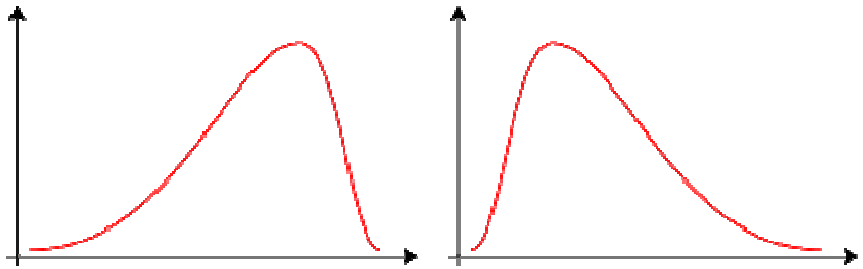
# Eloszlás aszimmetria - ferdeség

## Ferdeség (skewedness)

- A ferdeség az eloszlás középérték körüli aszimmetriájának mértékét jelzi.
- A pozitív ferdeség a pozitív értékek irányába nyúló aszimmetrikus eloszlást jelez, míg a negatív ferdeség a negatív értékek irányában torzított.

$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$$

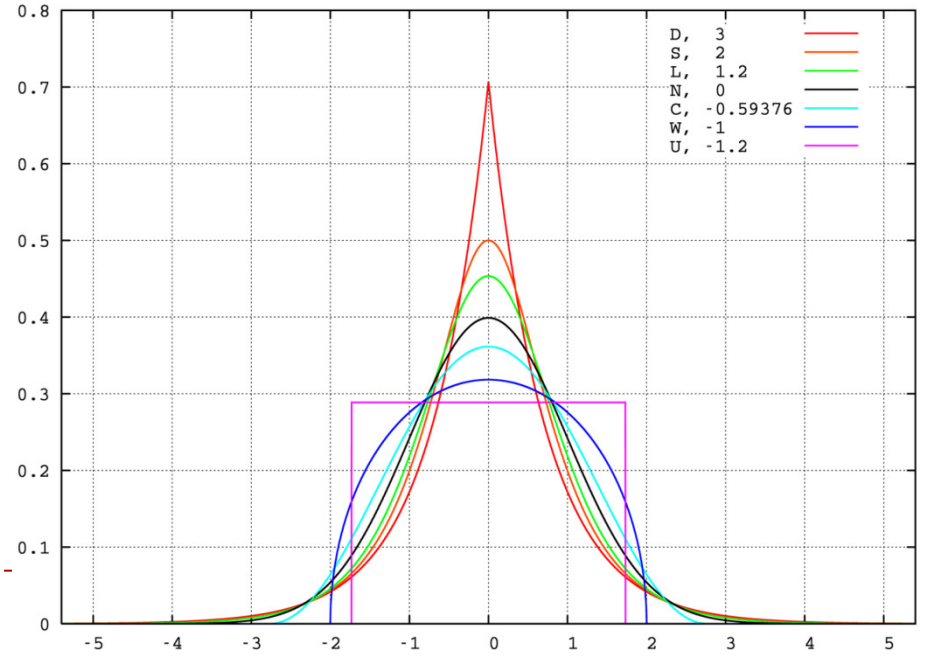


# Eloszlás aszimmetria - csúcsosság

- Egy adathalmaz csúcsosságát számítja ki.
- A függvény a normális eloszláshoz viszonyítva egy eloszlás csúcsosságát vagy laposságát adja meg.
- A pozitív értékek viszonylag csúcsos, a negatív értékek viszonylag lapos eloszlást jelentenek.

$$\operatorname{Kurt}[X] = \operatorname{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4} = \frac{\operatorname{E}[(X-\mu)^4]}{(\operatorname{E}[(X-\mu)^2])^2},$$

$$-\frac{3(n-1)^2}{(n-1)(n-2)}$$



# Intervallum becslés

Célkitűzés:

- Határozzuk meg egy becsült paraméter körül azt az intervallumot amibe egy előre meghatározott valószínűséggel esik a minta alapján.
  - Pl.: populációs átlag (μ) – minta áltag (x^)

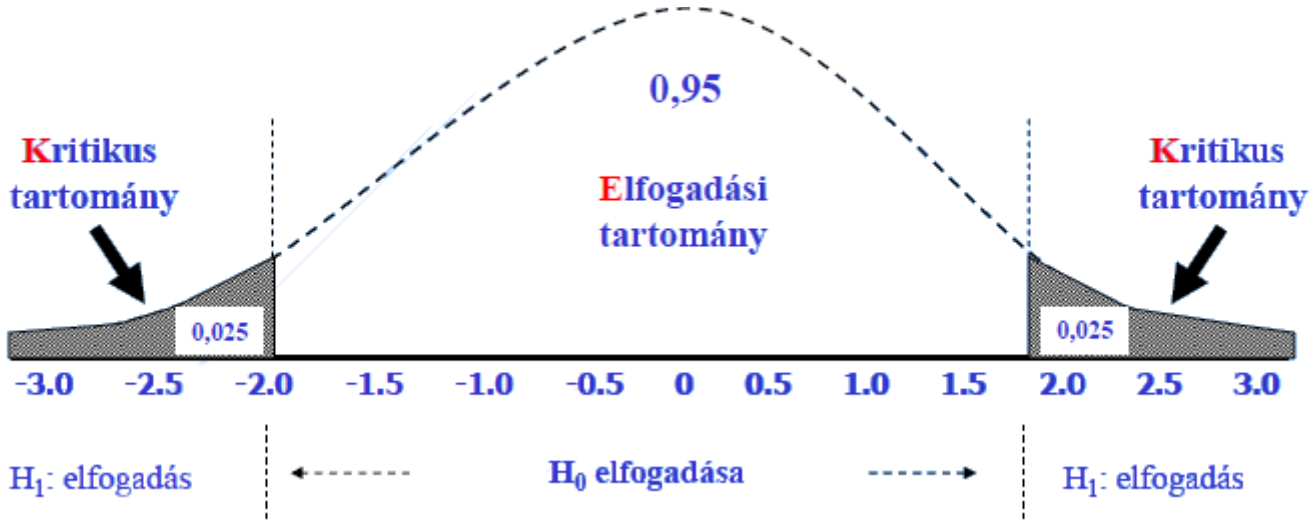
- A becsülendő értéket (μ) pontosan nem tudjuk, de a minta alapján számított érték (x^) körül van:
  - nagy (1-α) valószínűséggel a fenti intervallumban, és
  - kis (α) valószínűséggel esik ezen kívülre.

$$CI_L = \bar{x} - z^* \frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{N}} = CI_U$$

# Konfidenciaintervallum

- Ha $\alpha$ = 0,05 (5%), akkor $\mu$ 95%–ban ebben a konfidenciaintervallumban lesz benne és 5%–ban pedig ezen kívül
- $\alpha$ = 0,05 szignifikancia szint esetén z=1.96

$$P(\bar{x} - 1.96\frac{\sigma}{\sqrt{N}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{N}}) = 1 - \alpha = 0.95$$



# Konfidencia intervallum

- Nem ismert $\sigma$, csak $s$ a minta alapján esetén:

$$CI_L = \bar{x} - t^* \frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{x} + t^* \frac{\sigma}{\sqrt{N}} = CI_U$$

  - *A $t^*$ értékét a Student–eloszlás alapján határozzuk meg $df = N–1$ szabadsági fok ismerete mellett.*
- A t-eloszlás a normális eloszláshoz tart, ha $n \rightarrow \infty$.
- Minél nagyobb mintából becslünk annál jobb lesz az átlag szórásának becslése is.
- Általában, n>30 esetre a konfidencia intervallumhoz a normális eloszlás táblázata megfelelő.

# Hatáserősség konfidencia intervalluma

- Célváltozó: *Y*
  - Pl.: Allergia {eset, kontroll}
- Magyarázó változó: *X*
  - Pl.: Lakóhely légszennyezettsége
  - {alacsony, normális,magas}

$$\mathrm{Odds}_{X_i^{(s)}} = \frac{P(Y^{(1)} \mid X_i^{(s)})}{P(Y^{(0)} \mid X_i^{(s)})} \qquad \mathrm{OR}_{X_i^{(1,0)}} = \frac{\mathrm{Odds}_{X_i^{(1)}}}{\mathrm{Odds}_{X_i^{(0)}}}$$

- Odds: esély
- OR: odds ratio : esélyhányados

# Hatáserősség konfidencia intervalluma

- A hatáserősség logaritmusa a normális eloszlást közelíti
- $X \sim N(\log(OR), \sigma^2)$
- A konfidenica intervallum normális eloszlást feltételezve számítható:

$$CI_L = \bar{x} - z^* \frac{\sigma}{\sqrt{N}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{N}} = CI_U$$

- Ahol $SE = \sigma/\sqrt{N}$ így számítható

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}}.$$

# Hipotézisvizsgálat

**Hipotézis:** az alapsokaság paramétereire vagy az alapsokaság eloszlására vonatkozó feltevés.

- H0: null - hipotézis
- H1: alternatív hipotézis

- Pl. két minta átlagokra vonatkozóan formailag megfogalmazva
  - H0: $\mu 1 = \mu 2$ vagy $\mu 1 - \mu 2 = 0$
  - H1: $\mu 1 \neq \mu 2$
- **Hipotézisellenőrzés:** az a statisztikai módszer, amelynek segítségével egy véletlen minta alapján eldöntjük, hogy az adott hipotézis *(H0) elfogadható-e vagy sem.*

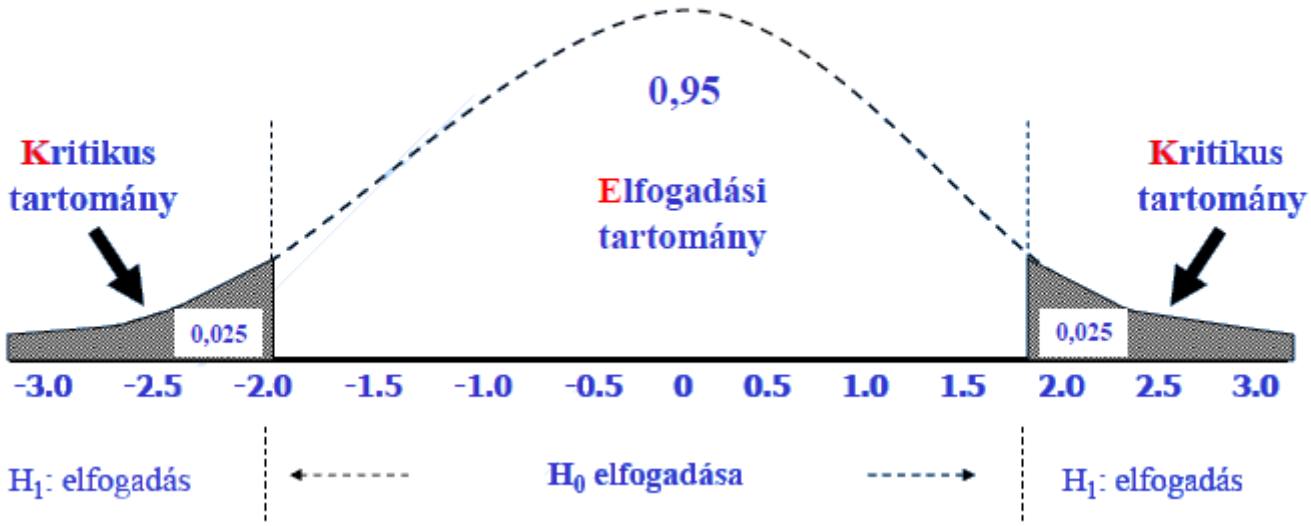
# Statisztikai próba

- Minta alapján dönt a hipotézis elfogadhatóságáról
- Konstruálunk egy próbafüggvényt, amely függ a mintától, és amelynek eloszlása ismert, ha H0 igaz.
- p-érték (empirikus szignifikancia szint): Az a legkisebb szignifikancia szint, amely mellett a vizsgált H0 hipotézist már elutasíthatjuk a H1 hipotézissel szemben.

- Döntés a p értéke alapján:
  - $p < \alpha$ : H0-t elvetjük, elfogadjuk H1-et
  - $p \geq \alpha$ : H0-t elfogadjuk

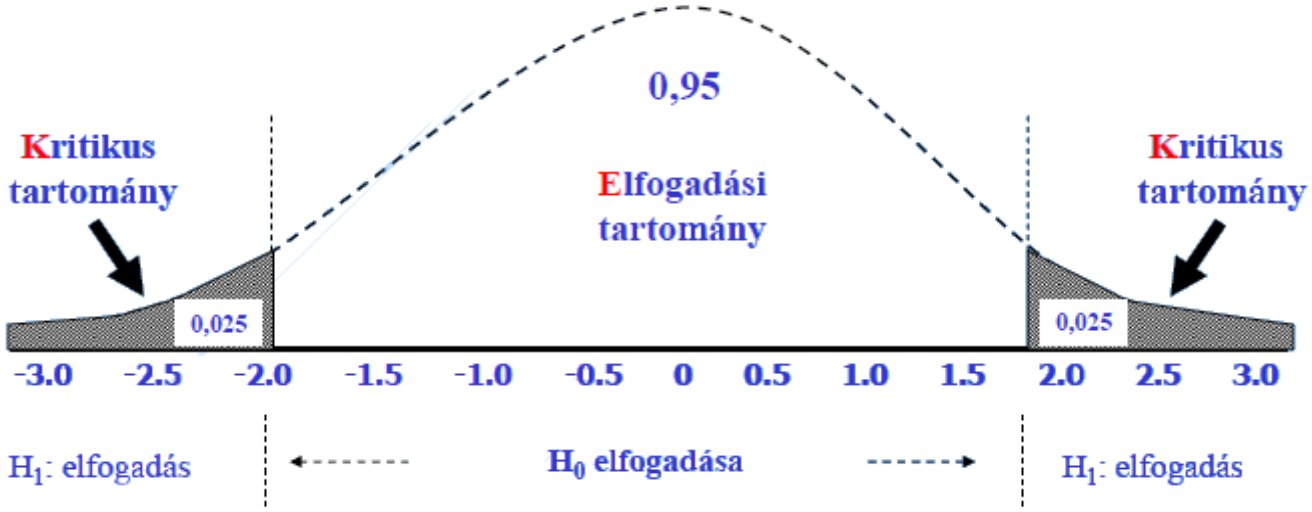
# Statisztikai próba

- Ha a próbafüggvény értéke az elfogadási tartományba esik, akkor a tapasztalati adatok $\alpha$ szignifikancia szinten nem mondanak ellent H0-nak
- Ha a próbafüggvény értéke a kritikus tartományba esik $\alpha$ szignifikancia szinten, akkor elvetjük H0-t



# Statisztikai próba

- Kétoldali próba: két oldalról állít alsó és felső korlátot (a feltételtől való eltérés tényét vizsgáljuk, irányát nem).
- Egyoldali próba: csak az egyik irányban állít korlátot (csak ilyen irányú eltérés lehetséges vagy fontos számunkra).



# Elkövethető hibafajták

- Type I. error ($\alpha$ hiba vagy szignifikancia érték): annak valószínűsége, hogy tévesen elutasítjuk a valós H0 hipotézist.
- Type II. error ($\beta$ hiba): a hibás H0 hipotézis elfogadásának valószínűsége.
- Power: a téves H0 helyes elutasításának valószínűsége. = 1 - β

| Döntés \ Valós helyzet | H0 igaz | H0 hamis |
|---|---|---|
| H0 elfogadása | Helyes döntés ($1-\alpha$) | Másodfajú hiba (Type II. error : $\beta$) |
| H0 elutasítása | Elsőfajú hiba (Type I. error : $\alpha$) | Helyes döntés ($1-\beta$) |

# Döntések értelemzése

- 1 - $\alpha$ : H0 elfogadása, amikor az valóban igaz (H1 elutasítása helyesen) H0: $\mu 1 = \mu 2$ ~~H1: $\mu 1 \neq \mu 2$~~ ✓
- $\alpha$ : H0 elutasítása, amikor az igaz, és H1 elfogadása, holott az nem igaz ~~H0: $\mu 1 = \mu 2$~~ H1: $\mu 1 \neq \mu 2$ ✗
- 1 - $\beta$: H0 elutasítása, amikor az hamis, és H1 elfogadása, amikor az igaz ~~H0: $\mu 1 = \mu 2$~~ H1: $\mu 1 \neq \mu 2$ ✓
- $\beta$ : H0 elfogadása, amikor az hamis, és H1 elutasítása, holott az igaz H0: $\mu 1 = \mu 2$ ~~H1: $\mu 1 \neq \mu 2$~~ ✗

# A hipotézis vizsgálat menete

- A null- és alternatív hipotézis megfogalmazása.
- Próbafüggvény keresése/szerkesztése.
- Előre rögzített szignifikanciaszint mellett az elfogadási és elutasítási tartomány megszerkesztése.
- A próbafüggvény empirikus értékének meghatározása.
- Döntés

# Paraméteres eljárások

- Feltétel: a vizsgált valószínűségi változók eloszlása normális eloszlást követ

- Normalitás vizsgálat
- Egy várható érték összevetése egy adott értékkel
- Két várható érték összevetése
- Varianciák homogenitásának vizsgálata

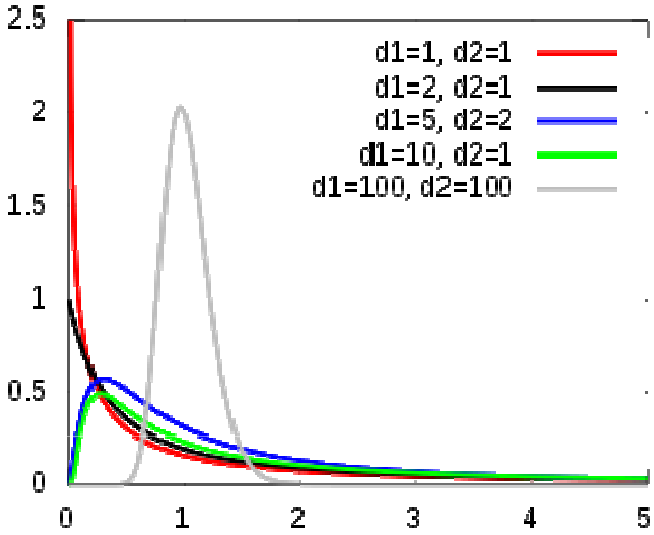
# F-próba

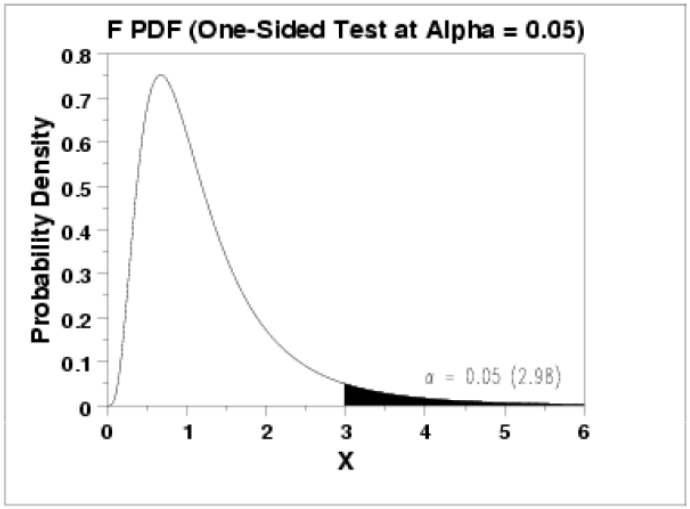
- Két variancia homogenitásának eldöntésére, azaz a két minta azonos varianciájú alapsokaságból származik-e.
- H0: $S_1^2 = S_2^2$ vagyis $S_1^2 - S_2^2 = 0$ (varianciák azonosak)
- H1: $S_1^2 \neq S_2^2$ vagyis $S_1^2 - S_2^2 \neq 0$ (varianciák nem azonosak)

- A két minta elemszámai: *N1 és N2*
- A két szabadsági fok: *df1=n1 -1 és df2=n2 – 1.*
- Az Fkrit, $_{(N1-1,\, N2-1)}$ két szabadsági foktól függ valamint α-tól.

$$F_{df1,df2} = \frac{S_1^2}{S_2^2}$$

# F-próba




- ha F-próbával a két variancia azonos, akkor pl. használhatunk kétmintás t-próbát
- ha a két variancia nem azonos, akkor az ún. d-próbát (Welch próbát) használunk

# Egymintás T-teszt

- Egy a populáció várható értéke megegyezik–e egy feltételezett várható értékkel vagy pedig szignifikánsan eltér attól. Feltétel, hogy a változó legyen normális eloszlású.
- Hipotézisek:
  - Ho: x^- $\mu = 0$ (nem tér el szignifikánsan μ-től)
  - H1: x^- $\mu \neq 0$(szignifikánsan eltér)
- A vizsgálathoz a t – statisztikára van szükség: $t = \frac{\bar{x} - C}{s_{\bar{x}}}$
- Amennyiben a számolt *t* értékünk abszolút értéke kisebb, mint $t_{krit}$, úgy a H0 nullhipotézist α szignifikancia szinten elfogadjuk; ellenkező esetben elvetjük és a H1-t fogadjuk

$$s_{\bar{x}}^2 = \frac{s_x^2}{N}$$

# Párosított T-teszt

- Ugyanazokon a mintákon két mérést végzünk különböző időpontokban: kezelés előtt (t0) és után (t1), így a két N-elemű összetartozó párokból álló mintát kapunk.
- A cél a kezelés hatásosságának vizsgálata, azaz történt e szignifikáns változás $d^i = x_0^i - x_1^i$; $d\hat{} = 1/n \, \Sigma_i \, (d^i)$
  - Ho: $d\hat{}=0$ (nincs szignifikáns eltérés)
  - H1: $d\hat{} \neq 0$ (van szignifikáns eltérés)
- A vizsgálathoz a t – statisztikára van szükség: $t_f = \frac{\bar{d} - 0}{s_{\bar{d}}}$

$$s_{\bar{d}}^2 = \frac{s_d^2}{N} \qquad s_d^2 = \frac{\sum_{i=1}^{N} (d_i - \bar{d})^2}{N - 1}$$

# Kétmintás T-teszt

Két független minta összehasonlítására használjuk.

- Feltételek:
  - a) csoportok függetlensége
  - b) adatok normalitása
  - c) csoportok varianciája legyen azonos (F-próba).

- Hipotézisek:
  - Ho: $x_1\hat{} = x_2\hat{}$ (nem tér el szignifikánsan egymástól)
  - H1: $x_1\hat{} \neq x_2\hat{}$ (szignifikánsan eltér)

- A vizsgálathoz a t – statisztikára van szükség

# Kétmintás T-teszt

A.) Csoportok közötti variancia egyenlő

A populáció torzítatlan becslése (pooled variance):

$$s^2 = \frac{\sum_{i=1}^{N_1}\left(x_i - \bar{x}_1\right)^2 + \sum_{j=1}^{N_2}\left(x_j - \bar{x}_2\right)^2}{N_1 + N_2 - 2}$$

A két átlag eltérésének standard hibája:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}} = s \cdot \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Az számítandó t-statisztika:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_{\bar{x}_1 - \bar{x}_2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}$$

# Kétmintás T-teszt

A.) Csoportok közötti variancia nem egyenlő

A variancia becslése (Cochran-Cox metódus)

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sum_{i=1}^{N_1}\left(x_i-\bar{x}_1\right)^2}{N_1(N_1-1)} - \frac{\sum_{j=1}^{N_2}\left(x_j-\bar{x}_2\right)^2}{N_2(N_2-1)}} = \sqrt{s^2_{\bar{x}_1}+s^2_{\bar{x}_2}}$$

A számítandó t-statisztika és annak korrekciója

$$t = \frac{\bar{x}_1-\bar{x}_2}{s_{\bar{x}_1-\bar{x}_2}}$$

$$t_{0.05} = \frac{t_1\cdot s^2_{\bar{x}_1} + t_2\cdot s^2_{\bar{x}_2}}{s^2_{\bar{x}_1}+s^2_{\bar{x}_2}}$$

# Személyiségi jegyek - adathalmaz

```
https://sapa-project.org/
library(psych)
data(sat.act)
summary(sat.act)
```

SAT: **Scholastic Assessment Test**

ACT: **American College Testing**

- Gender: males = 1, females = 2
- Education: self reported education 1 = high school ... 5 = graduate work
- Age: age
- ACT: ACT composite scores may range from 1 - 36. National norms have a mean of 20.
- SATV: SAT Verbal scores may range from 200 - 800.
- SATQ: SAT Quantitative scores may range from 200 - 800

# Feladatok

- Vizsgálja meg az egyes változók eloszlását!
- Képezzen boxplotokat! Hol találunk outliereket?
  - `boxplot(dataset$variable,…)`
- Vizsgálja meg a teszteredményt leíró ACT eloszlását nemekre bontva!
  - `lst <- split(sat.act$..., sat.act$...)`
  - `lapply(lst, summary)`
- Vizsgálja meg a változók hisztogramját!
  - `hist(…)`
- Vizsgálja meg, hogy az életkor átlagértéke szignifikánsan eltér-e 26 évtől!
  - `t.test(…,mu=…)`

# Feladatok

- Normális eloszlású-e az ACT változó?
- Szignifikánsan különbözik-e egymástól a SATV és SATQ változók eloszlása?
- `t.test(…,…,paired = TRUE)`
- Honnan látható, hogy nem normális eloszlásúak?

- Kimutatható-e függés a nem és a képzettség között ?
- `table(sat.act$gender,sat.act$education)`
- `Chisq.test(…)`

# Feladatok

```
library(psych)
data(epi.bfi)
```

| | |
|---|---|
| epiE | EPI Extraversion |
| epiS | EPI Sociability (a subset of Extraversion items |
| epiImp | EPI Impulsivity (a subset of Extraversion items |
| epilie | EPI Lie scale |
| epiNeur | EPI neuroticism |
| bfagree | Big 5 inventory (from the IPIP) measure of Agreeableness |
| bfcon | Big 5 Conscientiousness |
| bfext | Big 5 Extraversion |
| bfneur | Big 5 Neuroticism |
| bfopen | Big 5 Openness |
| bdi | Beck Depression scale |
| traitanx | Trait Anxiety |
| stateanx | State Anxiety |

# Köszönöm a figyelmet!

- (gabor.hullam-*at-mit*.bme.hu)


Budapest University of Technology and Economics
Department of Measurement and Information Systems