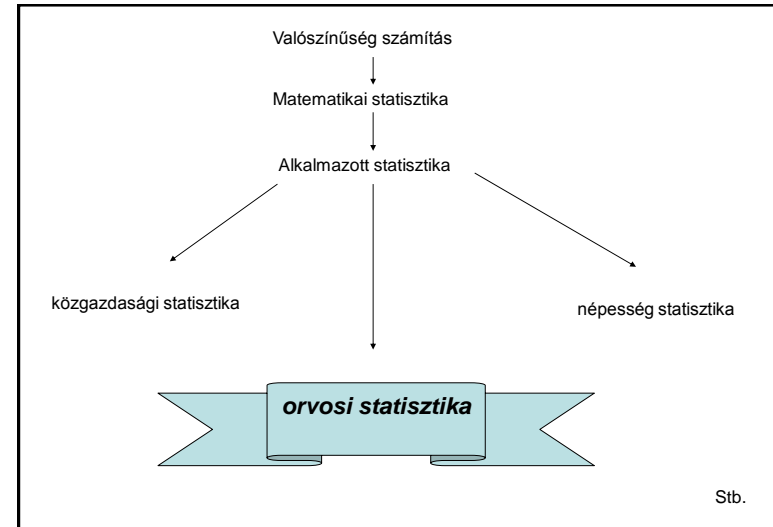
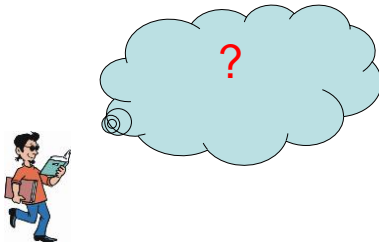
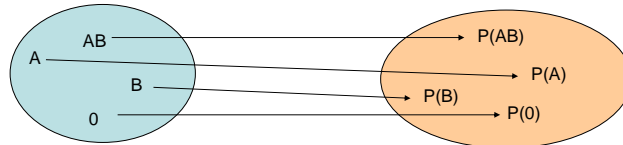


Valószínűesszámitás és a statisztika



Példa: vércsoportok



Elemi esemény: A, B, AB, 0
Teljes eseményhalmaz.
Valószínűségek:
 $P(A)$, $P(B)$, $P(AB)$, $P(0)$
Egymást kizáró események, tehát:

$$P(A) + P(B) + P(AB) + P(0) = 1$$

Összetett esemény pl.:
megtalálható az A antigén:

$$\text{valószínűsége} = P(A) + P(AB)$$

Egy és csak egy antigén található:

$$\text{valószínűsége} = P(A) + P(B)$$

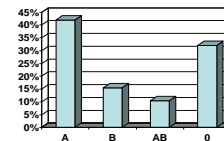
„Be van fejezve a nagy mű, igen.
A gép forog, az alkotó pihen.”

?



Az eloszlás

Vércsoportok eloszlása
Magyarországon



Hogyan juthatunk ilyen információhoz?
Mennyire megbízható?



Elméleti úton (nagyon ritka)
(pl. kocka feldobás:
minden elemi esemény valószínűsége: $1/6$.)

Tapasztalati úton
(kísérletet végzünk.
Kísérlet: mérés,
megfigyelés,
kikérdezés stb.)

Populáció és minta

Ideális, ha minden lehetséges esetet megvizsgálunk.

Populáció (alapsokaság)

Alapsokaság, olyan vizsgálni kívánt egyedek, vagy más tetszőleges elemek véges vagy végtelen összessége, amelyeknek közös megfigyelhető jellemzői vannak. Elméleti összesség is lehet, potenciálisan megfigyelhető elemekkel.

Minta

A populáció relatíve kis méretű kiragadott része valamilyen előírás szerint válogatva.

Mintavétel alapelvei



Következtetés
nagyobb elemszámú minta – kisebb eltérések,
megbízhatóbb eredmény.

- Lehetőleg minél nagyobb elemszám.
(Az ésszerűség határain belül.)
- Véletlen mintavétel.
- Orvosi kiegészítés.
Ha nincs semmi kizáró ok, akkor legyen véletlen.



A hiba forrásai

Mintavételi hiba

Abból adódik, hogy nem az alapsokaságot, hanem csak egy részét (minta) vizsgáljuk.

A statisztika módszereivel elemezhető, számba vehető!



Nem mintavételi hiba

Adatfelvételi hiba pl.:
válaszadási hiba, feldolgozási
hiba stb.

Nőgyógyászat



Extrém példa:
Férfiak és nők
aránya a
társadalomban.

Kérem a következőt!

A becslés



Milyen
magas a fa?

Kb. 7 m magas.



A **becslés** olyan eljárás, amely hiányos, többnyire tapasztalati adatok alapján, egy adott esetre, adott változóhoz egy **becsült** értéket rendel.

A becslés típusai

Pontbecslés

Egyetlen értékkel történő közelítés.



Körözés
...
kb. 175 cm magas
...

Intervallumbecslés

Egy intervallummal (amiben nagy megbízhatósággal megtalálható) történő közelítés.



Körözés
...
170-175 cm magas
...

A jó becslés tulajdonságai

Torzítatlan: A becslés várható értéke minden minta-elemszám esetén éppen a keresett paraméter. (Körülötte ingadoznak)

Hatásos: A becslésnek a paramétertől való közepes négyzetes eltérése minimális. (azaz a szórása kicsi)
Két egyaránt torzítatlan becslés közül az a hatásosabb, amelyre a közepes négyzetes eltérés a kisebb.)

Konzisztens: becsléssorozat, amelyben a becslések torzítatlanok és közepes négyzetes eltérésük a zérushoz közeledik, (sztochasztikusan) konvergál a paraméter valódi értékéhez.
Ingadozása n növekedtével csökken.

Elégséges: Olyan becslés, amely az összes információt tartalmazza a paraméterre, amit a mintából kaphatunk. (Pl. a normális eloszlásra középérték és a szórás elégséges statisztika).

Kategoriális változó

Kísérlet: kiválasztunk egy embert és elvégezzük a vizsgálatot.



Kiválasztunk elegendő számú embert.

n : elemszám.

Minta: a kiválasztott n számú ember a sokaságból.

| vércsoport | gyakoriság |
|------------|------------|
| A | k_A |
| B | k_B |
| AB | k_{AB} |
| 0 | k_0 |

Kimenetel:
A vagy B vagy AB vagy 0.

Egy valószínűség becslése

$P(A)$ az A vércsoport előfordulásának valószínűsége.
Az A vércsoport előfordulásának a várható értéke $nP(A)$.

Az $nP(A)$ becslése a minta alapján: k_A

A $P(A)$ pontbecslése: k_A/n .



Rendben van, de egy másik mintából más érték származik.
Mennyire megbízható ez az érték?

A relatív gyakoriság hibája



Binomiális eloszlás.
várható érték: np
variancia: $np(1-p)$

(no lám! Valószínűségszámítás?)

n elemű minta:
 k elem A vércsoportú, $(n-k)$ nem.

$s_{k/n}$ érték a k/n szórása, vagy
standard hibája.



A k_x érték szórásának becslése:

$$s_k = \sqrt{nP(A)(1-P(A))}$$

A k_x/n érték szórásának becslése:

$$s_{k/n} = \frac{\sqrt{nP(A)(1-P(A))}}{n} = \sqrt{\frac{P(A)(1-P(A))}{n}}$$

$P(A)$ helyett a k_x/n -t használjuk!

Konfidencia intervallum

Ennek segítségével megadhatunk egy intervallumot.
(intervallumbecslés)

$$\left(\frac{k}{n} \pm s_{k/n} \right)$$

68%-os konfidencia
(megbízhatósági) intervallum,
amihez 68%-os **konfidencia szint**
tartozik.



Jelentése:

Ha nagyon sok mintán
megismételjük a megfigyelést,
akkor a konfidencia intervallumok
68%-a tartalmazza a $P(A)$ -t.

Vagyis az intervallumbecslés
megbízhatósága 68%

Folytonos változó

Példa: testmagasság

testmagasság: 172 cm.



Helyes
kijelentés?



Nem!

- Pontos mérés nem lehetséges,
- végtelen pontosságú eszköz kellene.

Az eseménytér végtelen nagy!

Véges elemszámú minta.
Nincs két azonos elem.
(gyakoriság értékek: 1 vagy 0)



Hamis következtetés,
gyakorlatban nem
kivitelezhető.

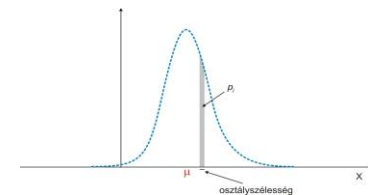
Mintavétel folytonos változó esetében

Helyes kijelentés:

A testmagasság (x):
 $171,5 \leq x < 172,5$ cm



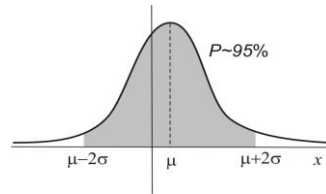
Egy meghatározott érték helyébe, egy
intervallum (osztály) lép.
(Továbbiakban a diszkrét eloszláshoz
hasonlóan használjuk)



p_j – annak a valószínűsége, hogy:
 x az adott osztályba tartozzon.

A μ és a σ

A σ az adatok szóródását jellemzi a μ körül. Az adatoknak kb. 68%-a a μ körüli 2σ széles intervallumban van.



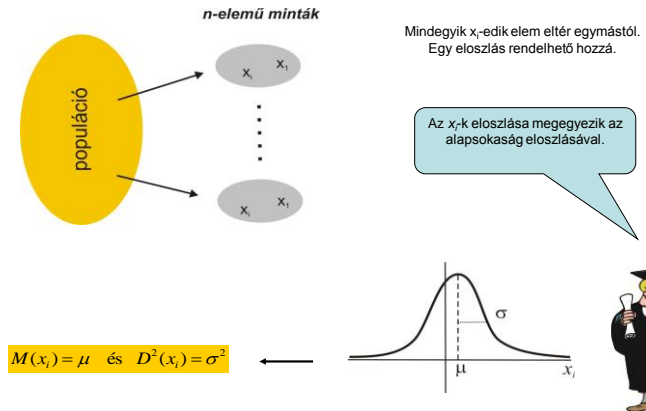
$$(\mu \pm \sigma) \approx 68\%$$

$$(\mu \pm 2\sigma) \approx 95\%$$

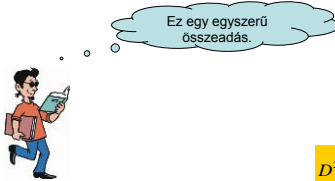
$$(\mu \pm \infty) = ?$$



A minták eloszlása



Az átlag várható értéke és varianciája



$$\bar{x} = \frac{1}{n} \sum_i x_i$$

$$M(\bar{x}) = \frac{1}{n} \sum_i M(x_i) = \frac{1}{n} (n\mu) = \mu$$

$$D^2(\bar{x}) = \frac{1}{n^2} \sum_i D^2(x_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$



Az átlag várható értéke azonos az alapsokaságéval, varianciája annak n -ed része.

A várható érték becslése

A várható érték becslése az átlag.

$$M(x) = \int x f(x) dx \Rightarrow \sum_j p_j x_j$$

p_j -t közelítsük a k_j/n relatív gyakorisággal!

$$\sum_j \frac{k_j}{n} x_j = \frac{1}{n} \sum_j k_j x_j = \frac{1}{n} \sum_i x_i$$

Torzítatlan becslés, mert:

$$M(\bar{x}) = \mu$$



Korrigált tapasztalati szórás

Az eltérés az átlag és a várható érték különbségéből fakad.

$$(\bar{x} - \mu)^2$$

$$\downarrow$$

$$M[(\bar{x} - \mu)^2]$$

$$\downarrow$$

$$\frac{\sigma^2}{n}$$

A minták közötti eltérések varianciája.



$$\sigma^2 = M(s^2) + \frac{\sigma^2}{n}$$

$$\sigma^2 = \frac{n-1}{n} M(s^2)$$

$$s^{*2} = \frac{n}{n-1} s^2$$

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

A továbbiakban s -el jelöljük a korrigált tapasztalati szórást.

A standard hiba

Az átlag varianciája:

$$\frac{\sigma^2}{n}$$

Az átlag szórása:

$$\frac{\sigma}{\sqrt{n}}$$

s a jó becslése a σ -nek.

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Ez tehát az átlag szórása, vagy standard hibája.

De általában a σ sem ismert.



A várható érték konfidencia intervalluma

Hasonlóan a P becsléséhez, a standard hiba ismeretében megadhatjuk a várható érték konfidencia intervallumát.

$$[\bar{x} \pm s_{\bar{x}}]$$

Ez az intervallum kb. 68% megbízhatósággal tartalmazza μ -t.

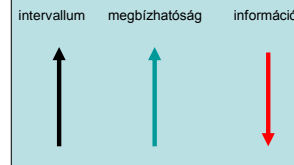


Az intervallum becslés sajátosságai

68%? Nem kevés egy kicsit?

Hát növelhetjük, pl.: a következő esetében kb. 95% a konfidencia szint, de az információ kevesebb.

$$[\bar{x} \pm 2 \cdot s_{\bar{x}}]$$

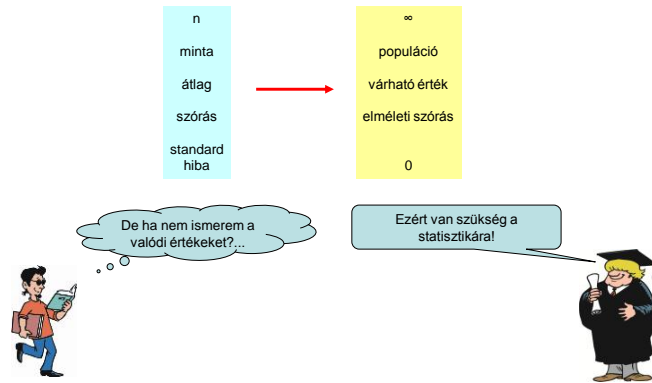


A pontos összefüggés:

$$[\bar{x} \pm t_p \cdot s_{\bar{x}}]$$

ahol t_p : az $(n-1)$ -ed fokú t -eloszlás esetében a p valószínűséghez tartozó érték. (a megbízhatósági szint $(1-p)$)

Kapcsolat a paraméterek között



Normál értékek

